

Exploring Traffic Pattern Variability in Vehicular Federated Learning^{*,**}

Giuliano Fittipaldi^{a,b}, Rodrigo S. Couto^a and Luís H. M. K. Costa^a

^aUniversidade Federal do Rio de Janeiro - GTA/DEL-Poli/PEE-COPPE, , Rio de Janeiro, RJ, Brazil

^bSorbonne Université, CNRS, LIP6, F-75005, Paris, France

ARTICLE INFO

Keywords:

Vehicular Networks
Federated Learning
Mobility
Traffic Patterns

ABSTRACT

The emergence of software-defined vehicles has brought machine learning into the vehicular domain. To support these data-driven applications, techniques to incentivize users to share their vehicle data are crucial. Federated learning trains machine learning models in a distributed manner, leveraging client data without compromising its privacy. Nonetheless, in vehicular networks, the dynamic behavior of nodes affects client availability and the global model's performance. Accordingly, this paper evaluates federated learning (FL) in a realistic vehicular network topology, accounting for real vehicle traffic in two Brazilian urban areas. The network simulation covers 3.7 km² with 1,290 vehicles per hour and road speeds, based on real data. Our paper provides a comprehensive analysis of the impact that different traffic behaviors can yield during the training phase of a federated learning model. We observe that there is a performance decay in urban areas with longer vehicle permanence. Interestingly, longer vehicle participation in FL training leads to a biased final model with reduced generalization. We propose a novel approach to verify vehicle variability over time, by using the Dice-Sørensen coefficient to compare the set of clients participating in different rounds of training. By maintaining the vehicle variability over the rounds we can reduce the effect of the bias on the model, and – with a 47% reduction of the communication overhead – achieve faster learning, higher convergence in the first 15 rounds, and an equivalent final accuracy. Additionally, we extend our analysis by conducting simulations under more extreme traffic scenarios across multiple datasets, using a MobileNetV3. The results confirm that sustaining high vehicle variability – in scenarios with a brief participation of vehicles in the training – yields comparable model performance while saving up to 83.5 GB in communication costs.

1. Introduction


The rise of software-based vehicular technologies culminating in software-defined vehicles (SDVs) has made AI essential for applications such as resource management, maintenance prediction, driver assistance, and autonomous driving. AI's growth potential is significant [27], with major corporations adopting it and governments enacting supportive regulations [6]. Data is needed to develop and enhance AI models, and modern vehicles generate vast amounts of data. However, this also raises privacy concerns [29]. To address privacy, Google proposed the federated learning framework [28].

In federated learning, each client receives the same initial global Neural Network (NN) model from a central server, trains it on local data, and sends only the updated NN weights back to the server, keeping the data private [11]. The server aggregates these weights into a global model, which is later redistributed to the clients for the next round, completing one federated learning training round.

When vehicles act as clients, relying on connectivity, the dynamic nature of vehicular networks poses a challenge due to the constant movement of nodes, which may frequently enter and leave the network [12]. This variability is influenced by factors such as road topology, speed limits, and queue factors. Distributed systems reliant on wireless communications, such as 5G, depend on the availability of nodes within a coverage area. Here the coverage translates

*This invited article is a journal extension of our conference paper by the same authors [13], which focused in analyzing the traffic patterns simulated from two different geographical regions. We extend this analysis by stressing the system in more extreme cases of traffic patterns and evaluating the performance of a state-of-the-art neural network for embedded systems across datasets of varying complexity.

**This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, CNPq grants 408255/2023-4 and 309304/2021-0, FAPERJ grants E-26/204.122/2024 and E-26/204.562/2024, FAPESP grants 23/00673-7 and 23/00811-0, and Development and Research Foundation - Fundep - Rota 2030, and our partners Stellantis and Mobway.

 giuliano.fittipaldi@lip6.fr (G. Fittipaldi); rodrigo@gta.ufrj.br (R.S. Couto); luish@gta.ufrj.br (L. H. M. K. Costa)
ORCID(s):

to a specific region of interest. In vehicular federated learning, this means each training round will vary in the number and identity of participating vehicles, affecting the combined set of local datasets and, consequently, the global model. Therefore, to evaluate the performance of vehicular federated learning systems realistically, it is essential to account for the dynamic availability of vehicles resulted from their high mobility patterns.

This study investigates the role of client availability in federated learning (FL) within large-scale vehicular networks, focusing on how traffic patterns influence model performance. The research aims to analyze the intermittent client participation effect due to vehicles entering and exiting a defined area during training. We evaluate the performance of federated learning, focusing on the final global model's generalization capability. Our analysis is conducted over different datasets of image classification due to its relevance in sensitive role in vehicular applications, ranging from driver assistance systems that provide situational alerts to autonomous driving systems requiring high model accuracy for safety-critical decisions.

Preliminary experiments use real traffic data from Rio de Janeiro and Niterói to simulate vehicle participation during FL training. A convolutional neural network (CNN) is trained on the MNIST dataset [8] to evaluate the impact of vehicle density and variability. The results reveal disparities in model performance between regions, attributed to differences in vehicle stay duration and participation patterns. To explain these disparities, a hypothesis is formulated that states that longer vehicle stays often lead to over-representation of specific clients, potentially introducing bias in the training process towards those favored clients. To evaluate the impact of vehicle stay, we employ the Dice-Sørensen (DS) coefficient [22], a metric that quantifies the overlap between two sets. This coefficient is used to measure the similarity of client data across federated learning rounds, providing insights into how traffic patterns influence the training process.

To mitigate this potential bias, a participation limit is imposed, reducing data overlap while improving model performance, even with a 47% drop in vehicle participation density. Building on these findings, we expand our analysis to more complex datasets—CIFAR-10, CIFAR-100 (coarse and fine), and GTSRB—to examine the broader implications of traffic patterns. To analyze more complex data, we also improve our machine learning model and pipeline by using a state-of-the-art model, MobileNetV3, for its suitability in embedded systems, along with a hyperparameter optimization for the individual clients in federated learning to enhance performance specifically in the federated case.

The results show that while simpler models like CNNs are more sensitive to bias, MobileNetV3 leverages larger datasets more effectively, though extended vehicle stays can still cause minor biases in extreme scenarios. Interestingly, regions with shorter vehicle stays achieve comparable performance to regions with longer vehicle stays while using a fraction of the total vehicles in training, which significantly reduces communication and computational costs. This highlights the potential for optimizing FL in vehicular networks by balancing data volume, client diversity, and model capacity.

In summary, this paper makes the following contributions: (1) We conduct a detailed analysis of the impact of traffic dynamics on federated learning (FL) at scale, addressing a gap in prior research by evaluating a wider range of scenarios, including extreme mobility conditions, and using four datasets of increasing complexity, including GTSRB with 43 traffic sign classes. (2) We leverage MobileNetV3 to better represent federated learning deployments on resource-constrained embedded systems and enhance the FL process through hyperparameter optimization. (3) We propose a novel approach to verify vehicle variability over time using the Dice-Sørensen coefficient to compare the set of clients participating in different rounds of training. (4) We demonstrate that maintaining vehicle variability over training rounds reduces bias, achieving a 47% reduction in communication overhead while enabling faster learning, higher convergence in the first 15 rounds, and equivalent final accuracy.

The paper is organized as follows: Section 2 discusses related work on vehicular federated learning. Section 3 details the methodology, including how we defined the traffic patterns analyzed, specifics regarding the datasets and models used, and the hyperparameter optimization conducted. Section 4.1 evaluates the performance between the patterns of two different regions in a simple task. In Section 4.2, we then evaluate the results on the generated distributions reflecting more extreme cases of traffic patterns, with the aim of understanding the implications for more complex tasks using a state-of-the-art model. Section 5 concludes the paper and discusses future work.

2. Related Work

Compelling work related to the process of vehicular federated learning has been developed, focusing on client availability and reputation, vehicular mobility, and V-2-V communications. Mobility, for instance, is one of the main

aspects to account for when developing vehicular systems. In [23], a communication-efficient federated learning framework is evaluated considering 1,000 devices. However, their approach, although focused on the vehicular context, does not account for mobility. In [39], an aggregation approach is proposed, and federated learning is evaluated in the vehicular context. However, the work accounts for uniform mobility, which does not impact the node availability throughout the training. The paper [42] proposes a mobility-aware caching for federated learning in vehicular networks. In this work the vehicle speed is modeled following a truncated Gaussian distribution, which is used in other vehicular network papers [2, 1, 41]. Nevertheless, these works are deeply focused on modeling vehicles velocity, accounting for simpler road topology, and do not cover the aspects of complex traffic patterns observed in urban areas.

Other core aspects, such as client behaviour, reliability and availability, highly influence the deployment of vehicular federated learning systems. The study in [31] introduces a greedy algorithm that selects clients for training based on their computational resources and wireless link quality. It assumes that all clients are always available and willing to participate, allowing the centralized server to decide which ones to involve. As a result, the work focuses on server-side decision-making rather than the unpredictable nature of client availability—an important factor in vehicular networks. Although the results are promising, they do not address the randomness introduced by client mobility, which is a key aspect we aim to explore.

The work in [40] takes into account limited client participation due to bandwidth and time constraints and proposes a client selection algorithm for federated learning. However, in their setup, clients join or skip training based on their willingness or the state of their local datasets, not focusing on the mobility aspect. The authors also adopt a hybrid strategy that includes using parts of client data in a centralized training setup. Despite yielding the best performance in terms of accuracy, the hybrid approach bypasses the privacy issue, raising the question of comparability with other federated learning solutions.

In [21], client participation is handled through a reputation system that scores each client based on their reliability and data quality. However, similar to [31], the variation in client participation comes from choices made by the server, not from actual client behavior. This means the approach assumes clients are always available, which does not reflect the reality of highly dynamic settings like vehicular networks. While both works consider differences in client participation, they do not fully explore the challenges caused by clients becoming unavailable due to factors outside the server's control, such as unstable connections. Their focus is mainly on choosing clients, not on dealing with unpredictable availability, which can have a substantial impact on real-world model performance.

Additional work can be found regarding client's reputation. In [25], the authors focus on updating client reputations in a privacy-preserving way for vehicular networks. They leverage cloud-fog collaboration and cryptographic masking to protect identity and reputation data, while ensuring accurate updates. Unlike other approaches, it addresses both privacy and correctness without relying on fully trusted entities.

Some other works focus in assessing client reliability in vehicular networks. In [17], the authors propose a trust evaluation scheme for federated learning based on digital twins. For the simulations, clients assume different behavior patterns that reflect if they have a benign or a malign contribution to the model's training. The paper demonstrates that their method is capable of identifying the different behaviors and adapting the system's trust value according to the contribution of the client to the model. However, the work does not focus directly on vehicular mobility or on traffic patterns, nor on their direct impact on the federated learning training.

There is more limited research regarding client availability in federated learning. In [36], a convergence analysis is conducted where client availability follows an arbitrary finite-state Markov chain. This modeling choice captures scenarios where client activities exhibit correlation patterns. Nevertheless, the availability is not related to vehicular networks as it does not consider the impact of traffic patterns. In [35], an analysis over the client availability issue is performed with a proposed solution. In this work, the availability models are defined in 5 categories. Always available clients form the baseline model; scarce availability occurs with a probability of being available 20% of the time; home devices and smartphones have independent availability patterns based on log-normal distributions and sine-modulated functions, respectively; and uneven availability scales inversely with each client's dataset size. In this context, the study does not incorporate client availability models based on mobility, thereby not capturing the dynamic nature of vehicular networks.

Although not the main focus of this work, cybersecurity is also an important concern in federated learning, especially in vehicular networks. Studies such as [30],[26], and [15] address security threats like backdoor attacks, vulnerabilities in V2V communication, and other malicious behaviors. These works highlight potential risks that should be considered when deploying federated learning in dynamic and open environments.

Table 1

Comparison of works regarding federated learning, vehicular scenarios, and client availability.

Works	Vehicular Scenario	Intermittent Client Participation	Mobility	Model for embedded devices
[23]	Yes	No	No	No
[31]	No	Yes	No	No
[40]	No	Yes	No	No
[21]	No	Yes	No	No
[36]	No	Yes	No	No
[35]	No	Yes	No	No
[39]	Yes	Yes	Limited	No
[32]	Yes	Yes	Yes	No
Ours	Yes	Yes	Yes	Yes

In [13], we approach the gap in the literature regarding the analysis of the dynamic behavior of vehicular networks in a federated learning context. We tackle this by analyzing mobility and client availability in two distinct urban regions, evaluating how traffic patterns and vehicle stay impact federated learning performance on the MNIST classification task while using a CNN model. The results show that a potential bias can appear towards vehicles that stay for longer periods in the training regions. By limiting the participation of these vehicles and consequently enhancing the data variability in training, we observed a reduction in 47% of the total vehicles involved in training while sustaining equivalent or better performance in the classification task.

The present work expands substantially on this foundation by providing a deeper and broader evaluation of traffic patterns and their impact on FL. While prior research has largely overlooked how traffic dynamics affect FL at scale, we close this gap by conducting a detailed analysis across a wider range of scenarios, including extreme mobility conditions. Our experiments span four datasets of increasing complexity – including GTSRB with its 43 traffic sign classes – and leverage MobileNetV3 to better represent deployment on resource-constrained embedded systems. We further enhance the federated learning (FL) process through hyperparameter optimization. To evaluate the practical implications of our approach, we also conduct communication cost simulations. Our findings reveal that training in regions characterized by high client variability, significantly improves convergence in the early stages and preserves strong model performance throughout training. Importantly, this is accomplished while engaging fewer vehicles in the training process, resulting in substantial reductions in communication overhead. In some configurations, total communication costs were reduced by as much as 83.5 GB. These results highlight the advantages of strategically selecting training regions with rich client diversity: they offer a favorable balance between learning efficiency, resource usage, and real-world feasibility in federated learning deployments. Table 1 presents an overview of the literature considered, highlighting the presence of a vehicular scenario, client availability evaluation, mobility from the federated clients, and the use of state-of-the-art models for edge devices.

3. Methodology

In this section, we explain the mobility simulation characteristics and how the federated learning is modeled based on this mobility. We also present details regarding the preliminary results and the expanded simulations.

3.1. Federated learning modeling

In the vehicular federated learning simulations, each vehicle is indexed and treated as an independent client with a dataset that is distinct from those of other vehicles. To achieve this, the complete dataset is partitioned by the total number of clients in the simulation, in an iid fashion, and then distributed across those clients. Additionally, vehicles are presumed to have identical hardware resources, resulting in both uniform individual training times and identical local dataset sizes.

An important factor we investigate is the vehicle training frequency, defined as the number of rounds a particular vehicle participates in during the federated learning process. This concept is crucial because each client (vehicle) maintains its own static local dataset throughout the training rounds. Intuitively, the participation by the same clients can lead to an over-representation of specific datasets, which may not accurately reflect the true distribution of data in a broader environment. This lack of diversity in data samples could potentially affect the performance and generalization

of the global model. Regarding the communication between the vehicles and the centralized server, we assume that a vehicle is only disconnected when it leaves the region of interest.

3.2. Preliminary simulations setup

3.2.1. Region selection

To understand the impact of different urban scenarios on vehicular FL, two distinct regions of the metropolitan area of Rio de Janeiro were chosen for evaluation based on several criteria to ensure a fair comparison. Firstly, both regions have an equal area of 3.7 km^2 , assuming the centralized server has the same limited coverage area. Secondly, the regions are situated in central parts of the cities of Rio and Niterói to ensure a high density of vehicles and mobile network coverage. Thirdly, the regions have a similar level of complexity in road topology. This is determined by combining spatial metrics such as edge density, intersection density, and lane kilometer density. This criterion ensures the avoidance of regions like express highways, where node behavior can be much simpler, which is not the focus of this study.

The combined complexity metric of the road topology is defined as:

$$S = w_e \cdot \frac{d_e}{d_{e,\max}} + w_i \cdot \frac{d_i}{d_{i,\max}} + w_l \cdot \frac{d_l}{d_{l,\max}}, \quad (1)$$

where:

- w_e, w_i, w_l are the weights assigned to the edge density, intersections, and lane kilometers respectively.
- d_e is the edge density.
- d_i is the density of intersections.
- d_l is the density of lane kilometers.
- $d_{e,\max}$ is the maximum observed value of edge density between the two regions.
- $d_{i,\max}$ is the maximum observed value of intersection density between the two regions.
- $d_{l,\max}$ is the maximum observed value of lane kilometer density between the two regions.

The values of the weights are defined based on importance in complexity. The edge density is considered the most important factor, as it directly accounts for the quantity of possible end-to-end routes in the region. The second most important factor is the density of intersections, as they also impact the quantity of possible routes. Given a pair of edges defining the start and finish of a route, each added intersection represents a decision point where different paths can be taken. The more intersections, the greater the number of decision points, which increase the probability of possible paths, depending on the road topology.

The last factor considered is the density of lane kilometers. While more lanes can accommodate more traffic, they don't necessarily increase the number of different routes available. Lane kilometers are more related to the capacity and efficiency of the roads rather than the complexity of the network structure. As such, we give less weight to the density of lane kilometers in the combined complexity metric. The weights were defined as follows:

$$[w_e, w_i, w_l] = [0.5, 0.4, 0.1]. \quad (2)$$

Initially, two visually similar regions are chosen for the analysis, in which the 3.7 km^2 area defined contained the geographical boundaries of the neighborhoods. Then the combined complexity is calculated to infer the similarity between the regions. The neighborhoods chosen were Botafogo and Icaraí, located in the state of Rio de Janeiro, Brazil. Table 2 summarizes the main characteristics of the two simulated areas.

3.2.2. Dataset and Model

For the initial analysis we use the MNIST dataset [8] due to its broad use in the literature [23, 5, 24] and substantial size, making it suitable for partitioning in large-scale federated learning experiments. MNIST comprises 60,000 grayscale images of handwritten digits, with size 28×28 pixels and categorized into ten classes (digits 0-9).

A simple convolutional neural network (CNN) design is used. It comprises two convolutional layers with ReLU activation functions followed by a 2×2 max-pooling layer. The output of these convolutional layers is fed into three fully connected layers with ReLU activation functions. The final softmax layer has 10 outputs to classify the images.

Table 2

Main characteristics of the two urban regions simulated.

	Botafogo	Icaraí
Area (km^2)	3.7	3.7
Intersections	347	347
Edges	591	534
Lane length (km)	86.4	80.5
S	0.94	1

3.2.3. Traffic Patterns simulated from regions

The mobility simulation regarding the regions is done using the *SUMO* simulator [9]. The maps of the selected regions were exported from OpenStreetMap [14]. Vehicle flow rates were derived from traffic observed between 9 a.m. and 1 p.m. on major streets in Botafogo (Voluntários da Pátria and São Clemente), with an average of 1,290 vehicles per hour used to define the inflow for both Botafogo and Icaraí regions. All simulations used the Krauss car-following model, with acceleration set to 2.6 m/s^2 , deceleration to 4.5 m/s^2 , and the LC2013 lane-changing model. To maintain fairness, the Traffic Factor parameter was set to 10 for both regions, ensuring a higher probability of vehicle initialization at the region boundaries. Each vehicle in our simulation has a dataset of identical size, independent from the datasets of other vehicles. The dataset is split in a stratified fashion, guaranteeing a uniform class distribution across different vehicles. The vehicles participating in a federated round are those actively traveling within the region during that specific time interval. To integrate the mobility simulation into the federated learning training, we generate a file containing a dictionary with the respective vehicles that are active in the region throughout the whole simulation. This dictionary is queried during the FL training in order to allow or not a client to participate in a given round.

3.2.4. Dice-Sørensen coefficient

The Dice-Sørensen coefficient measures the similarity between two sets. It is widely used in medical image segmentation to quantify spatial similarity in images [45]. We apply this metric in the context of federated learning to assess the similarity between the set of datasets of each round. We can represent the set of datasets for each round as D_{R_j} for that respective round j . Figure 1 helps with the visualization.

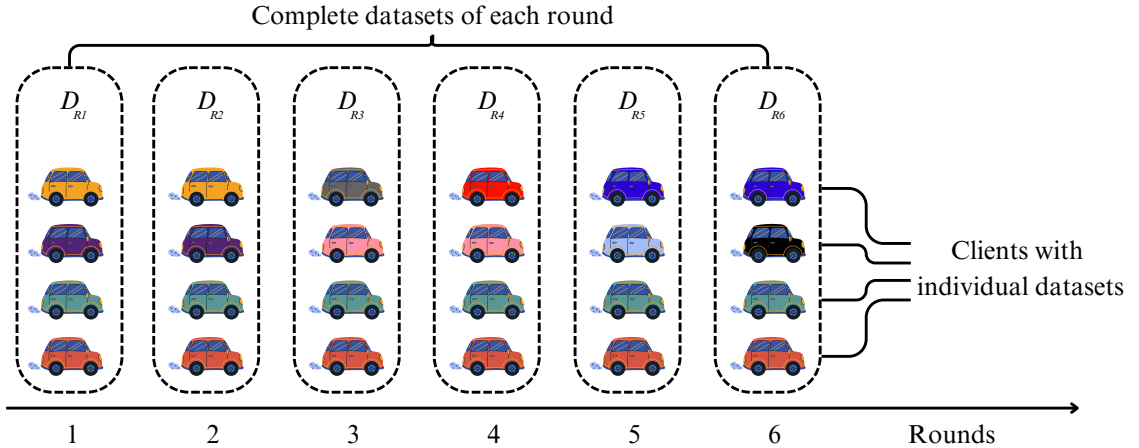


Figure 1: Set of datasets constituting the complete dataset of each round.

The Dice-Sørensen (DS) coefficient can thus effectively measure the variety between the complete datasets across different rounds by quantifying the degree of overlap between them as follows:

$$DS_{(D_{R_j}, D_{R_k})} = \frac{2|D_{R_j} \cap D_{R_k}|}{|D_{R_j}| + |D_{R_k}|}. \quad (3)$$

Table 3

Number of training samples per vehicle in each dataset configuration.

Dataset	N° training samples
Cifar datasets	250
GTSRB	133

In federated learning, particularly in vehicular environments, the set of participating clients—and consequently their datasets—can vary significantly from one training round to the next. This variability affects the composition of the aggregated dataset seen by the global model at each round. The Dice-Sørensen coefficient provides an interpretable way to quantify this change by measuring the similarity between the sets of data contributing to model updates over time. High Dice values between rounds suggest that many of the same clients (and thus similar data) are participating in consecutive rounds, leading to limited variability in the aggregated data. Conversely, lower Dice values indicate that the contributing clients—and their data—are changing more substantially across rounds, which can enhance dataset diversity and mitigate model overfitting to a narrow subset of the population.

This metric is particularly well-suited to our study because we are not interested merely in the number of clients participating, but in how their presence shifts across rounds. By comparing the aggregated datasets $D_{R,j}$ and $D_{R,k}$ of two federated learning rounds, we gain insight into the degree of continuity or novelty in the data presented to the global model. This perspective is important for evaluating the effects of different traffic patterns, where clients (vehicles) may enter and leave the area of interest frequently. A region with high vehicle turnover, for instance, will likely result in lower Dice coefficients between rounds, which we hypothesize can promote better generalization in the model by exposing it to a wider variety of data. Thus, using the Dice-Sørensen coefficient allows us to connect client participation dynamics directly to the statistical properties of the training data and, ultimately, to the performance of the learned model.

3.3. Expanded simulations setup

Building on the analysis of traffic patterns in the regions of Botafogo and Icaraí, we designed a more controlled experiment leveraging synthetic vehicle distributions. This approach allows us to systematically evaluate the impact of varying vehicular participation levels on federated learning performance. To strengthen our assumptions, we utilize four datasets with varying levels of complexity, providing insights into how model performance adapts to different challenges and data distributions. Additionally, we employ a state-of-the-art model specifically designed for embedded systems, ensuring that our simulations align closely with models suitable for real-world vehicular applications. This combination of diverse datasets and a practical, efficient model enriches the relevance and applicability of our findings.

3.3.1. Datasets

For the expanded simulations, four datasets with different data complexity levels are considered. The **CIFAR-10** dataset—which is widely used as a benchmark for image classification tasks [32, 43]—contains 60,000 32×32 pixel images across ten classes, offering greater complexity than MNIST due to its higher resolution and diversity of visual features. Achieving competitive accuracy with the literature on CIFAR-10 often requires advanced architectures like ResNets, unlike the simpler CNNs sufficient for MNIST. To increase classification difficulty, the **CIFAR-100** dataset extends CIFAR-10 by dividing the same number of images into 100 classes with two levels of granularity: *coarse*, which groups the 100 classes into 20 superclasses, and *fine*, which separates all 100 classes individually. This results in fewer images per class and more challenging tasks. Additionally, the **German Traffic Sign Recognition Benchmark** (GTSRB) dataset introduces 39,200 images in 43 traffic sign classes, such as stop and speed limit signs. While GTSRB has more classes than CIFAR-10 and CIFAR-100 coarse, it presents a simpler task, as traffic signs are designed to be easily distinguishable. This progression from CIFAR-10 to CIFAR-100 and GTSRB enables the evaluation of models across varying levels of complexity. Table 3 provides the discriminated number of unique samples that each vehicle carries throughout the simulation.

3.3.2. Embedded systems state-of-the-art model

To extend the analysis, more complex tasks require more advanced neural networks, and thus we focus on the **MobileNetV3**, a state-of-the-art architecture for embedded systems. Designed for resource-constrained environments, MobileNetV3 offers an excellent balance of computational efficiency and performance. While it may not match

Table 4

Model comparison for image classification on ImageNet-1K [7].

Model	Accuracy (%)	GFLOPS	Parameters (M)	Size (MB)
MobileNetV3-Large	75.274	0.22	5.48	21.1
MobileNetV3-Small	67.668	0.06	2.54	9.8
MobileNetV2 [37]	71.878	0.30	3.50	13.6
Resnet 50 [20]	80.858	4.09	25.56	97.8
EfficientNetV2 [38]	85.808	56.08	118.51	454.6
ViT-H/14 [10]	88.552	1016.72	633.47	2416.6

Table 5

Parameters used for MobileNetV3-Large.

Parameters	Value
Weights	IMAGENET1K V1
Criterion	Cross Entropy
Optimizer	SGD
Learning Rate (LR)	0.1
LR Scheduler	StepLR

the accuracy of some larger models, its design prioritizes practicality in embedded applications. Table 4 compares MobileNetV3 to other models like ViT, ResNet-50, and EfficientNetV2, showcasing metrics such as accuracy, computational cost (GFLOPs), parameter count, and model size. For this study, the PyTorch implementation of MobileNetV3-Large, pre-trained on ImageNet-1K, was used. The model implementation details are summarized in Table 5. Specifically, the optimizer was configured with a learning rate of 0.1, momentum of 0.9, and a weight decay of 0.0001. The learning rate scheduler was set with a step size of 25 and a decay factor (gamma) of 0.3. The model parameters were chosen according to preliminary experiments of centralized training over the CIFAR-10 dataset. These parameters yielded the best performance in terms of accuracy and loss, matching the benchmark provided in [18]. With 5.5 million parameters, 0.22 GFLOPs, and a size of 21 MB, the MobileNetV3-Large combines efficiency with strong performance. Its lightweight nature makes it ideal for vehicular systems, where models must operate on embedded devices powered by limited battery resources.

3.4. Hardware setup

The federated learning simulations were conducted using an NVIDIA Tesla V100S-PCIE-32GB GPU environment with CUDA version 12.2. Two GPUs were available, each with 32 GB of memory. To run the computation in parallel, we allocated a fixed amount of computational resources per virtual client: 3 CPU cores and 0.1 GPU (i.e., 10% of a GPU).

4. Simulation and Analysis

Our simulations begin with an initial setup phase in which we generate the synthetic traffic patterns to be analyzed and conduct the hyperparameter optimization for the federated learning framework. Further on, we focus on comparing performance across two regions with distinct traffic patterns, considering a simple machine learning task. Following this, we explore the effects of more extreme traffic scenarios, represented by synthetic distributions, to assess their influence on complex tasks with a state-of-the-art model for embedded systems.

4.0.1. Synthetic Traffic Patterns

To better isolate and evaluate the impact of traffic dynamics on federated learning, we introduce synthetic traffic patterns that simulate different levels of vehicle availability over time. Although these patterns are generated in advance, they are constructed to reflect varying degrees of vehicle permanence in a region, thereby introducing dynamic participation variability into the training process. This design choice allows us to systematically assess how different traffic scenarios affect convergence, model performance, and communication efficiency.

For the synthetic traffic patterns, the maximum number of rounds a vehicle can participate in is the total number of rounds from the whole simulation, 40 rounds. The goal of observing distributions with such a range is to evaluate extreme scenarios, since in our case 40 rounds would sum up to approximately three hours of vehicle permanence in a given region. The distributions generated assign the number of rounds each vehicle participates in the federated training.

Regarding the shape of the distributions, three scenarios were generated. The first scenario simulates a region with light traffic, with most vehicles staying for a brief period of time within that region. The second scenario considers moderate traffic of vehicles, with a traffic pattern that approximates the ones observed in the simulations of Botafogo and Icarai. The third scenario is a heavy-traffic region, with most of the vehicles staying for an extended period of time. The details are described next. -

- **Brief** – As shown in Figure 2(a), this distribution simulates *light traffic conditions*, where vehicles pass quickly through the region and have very limited time to participate in training. The participation frequency follows an *exponentially decaying distribution* (scale = 10), resulting in most vehicles contributing to only one round of federated learning.
- **Moderate** – Illustrated in Figure 2(b), this scenario reflects *moderate traffic flow*, resembling the patterns observed in the *Botafogo and Icarai* mobility simulations. Vehicle participation is modeled with a *Gaussian distribution* centered around 20 rounds (mean = 20, standard deviation = 8), meaning that most vehicles remain in the region long enough to contribute to a moderate number of training rounds.
- **Extended** – Depicted in Figure 2(c), this distribution corresponds to *heavy traffic conditions*, where vehicles remain in the region for extended periods—up to the full simulation time of three hours. Participation frequencies follow a *growing exponential distribution* (scale = 10), leading to high engagement, with most vehicles taking part in many or all training rounds.

The synthetic distributions simulate diverse traffic scenarios, each producing distinct vehicle participation patterns for federated training. Paired with a range of datasets for image classification, this approach enables a comprehensive and controlled analysis of how traffic behaviors influence federated learning outcomes.

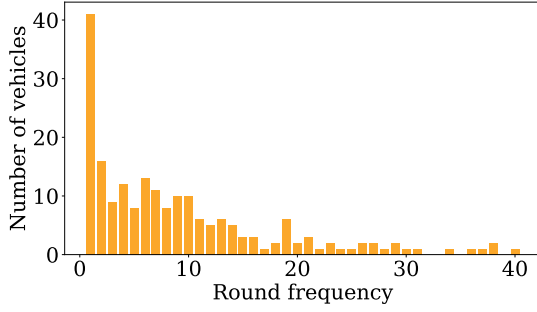
4.0.2. Client Optimization

Hyperparameter optimization (HPO) is one of the foundations of traditional machine learning pipelines [3]. However, when it comes to federated learning, this process is often overlooked. Most research in federated learning does not consider the optimization of hyperparameters in the federated setting, which raises the question of comparability between the centralized model performance, often used as a baseline, and the federated results. Since the optimal hyperparameters depend on specific conditions, such as dataset size and data distribution [3], it is neither logical nor effective to apply those hyperparameters from centralized training to a federated setup. Training conditions are much different for federated clients since they work on a small portion of a dataset and for limited epochs, when considering the simulations in literature.

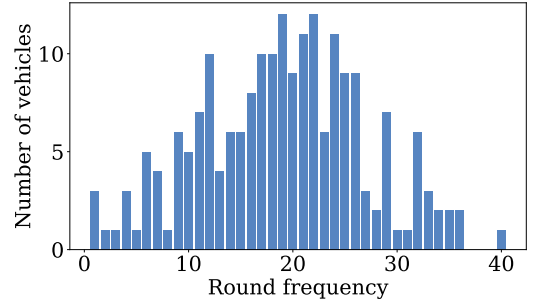
Although HPO is rarely addressed in the context of federated learning, this challenge is taken up by a couple of works. In [34], three optimizers are proposed that adapt client learning rates online and with improved convergence. Similarly, [44] focuses on HPO in federated systems, though its emphasis is on tabular data and non-neural network models, such as decision trees.

4.0.3. Batches

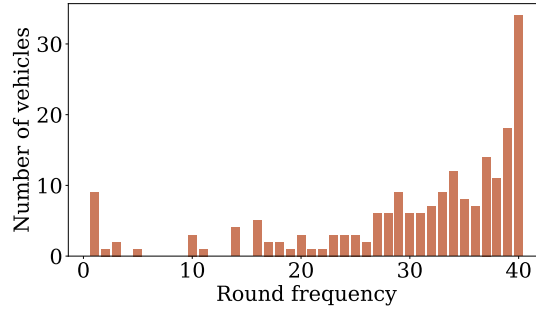
The first hyperparameter optimized in this work is the batch size, which demonstrates the importance of federated learning HPO. Using CIFAR-10 as an example, the optimal batch size was found to be 128 for centralized training, which yields 93.2% accuracy. To clarify, this means that the model processes 128 images per time out of the 50,000 total images on the complete dataset, resulting in 50,000/128 epochs, and thus the weights of the neural network (NN) will be updated 391 times. In the federated learning setup with 200 clients, each client receives 50,000/200 images. Using the 128 batch size would limit each client to just two neural network updates during the whole training, significantly impacting the way in which the model learns the local images. In an attempt to address this, centralized training was performed with a dataset partition of the size of a single federated client to examine the influence of batch size. In this work, CIFAR-10 showed a batch size of 32 as optimal among the sizes tested and had an accuracy increase of 10%



(a) Brief stay.



(b) Moderate stay.



(c) Extended stay.

Figure 2: Vehicle distributions for a maximum of 40 rounds.

Table 6

Batch sizes according to the optimization over each dataset.

Dataset	Batch Size
CIFAR-10	32
CIFAR-100 coarse	32
CIFAR-100 fine	32
GTSRB	16

compared to batch size 16. In CIFAR-100 Coarse and Fine, there was also a similar behavior, with 32 being the best batch size in terms of final accuracy. In GTSRB, however, the best performance was achieved using the batch size of 16.

Finally, the selected batch sizes can be seen in Table 6. These results confirm that the optimal batch size depends on the specific dataset—or data partition in the federated learning case—and directly influences training efficiency in a federated environment.

4.0.4. Local Epochs

The number of local epochs that each client is trained for in federated learning impacts the training dynamics and model performance significantly. While DNNs provide the capacity to model complex data, their large parameter space necessarily demands careful optimization to balance computation cost and model performance [16]. Taking CIFAR-10, for example, experiments made with a batch size of 32 and 30 epochs show that models indeed learn the local datasets effectively and converge at a reduced standard deviation across runs. Although overfitting was observed, this gets mitigated in federated learning since the averaging of weights across clients takes place to ensure that the global model captures the aggregated trends of the data. A similar behavior is observed for the other three datasets,

where increasing the number of classes in the CIFAR-100 coarse and CIFAR-100 fine datasets results in a significant decay in the validation accuracy. However, the training accuracy remains at 100%, demonstrating the model's capacity of learning on local datasets, even if they are not representative of the complete distribution. The primary concern is to avoid underfitting, where the local models fail to learn any local data characteristics and, in turn, affect global performance. In all datasets, 30 epochs seemed to provide a balance between computational efficiency and sufficient training time, taking into consideration realistic vehicular traffic patterns and durations in the simulation. Therefore, the number of local epochs that each vehicle conducts of training before the aggregation phase is 30.

4.1. Preliminary Results

Firstly, we evaluate the node availability behavior in the considered regions. Figure 3 plots the number of vehicles participating in the FL rounds. Both regions start the simulation with a similar number of vehicles. Nevertheless, as the simulation progresses, Botafogo begins to show higher vehicle counts, suggesting that vehicles in this region take longer to complete their routes. This observation is reinforced if we compare the average trip duration in both regions, which is 246 s for Botafogo and 197 s for Icaraí.

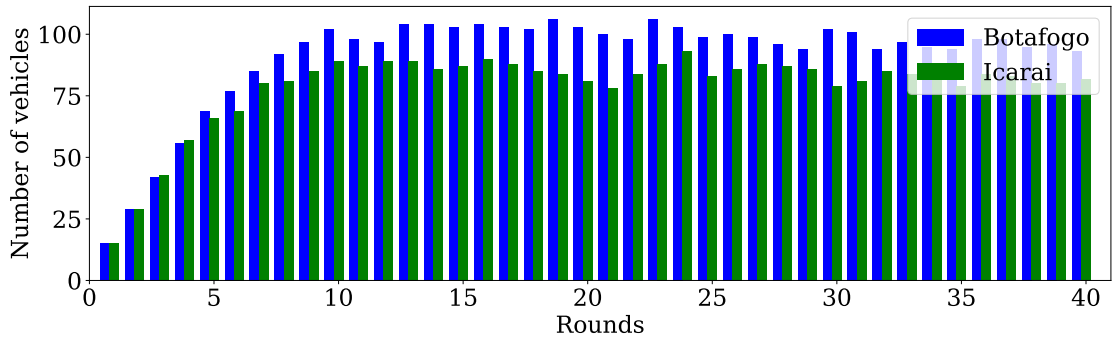


Figure 3: Number of vehicles participating in each training round for Botafogo and Icaraí.

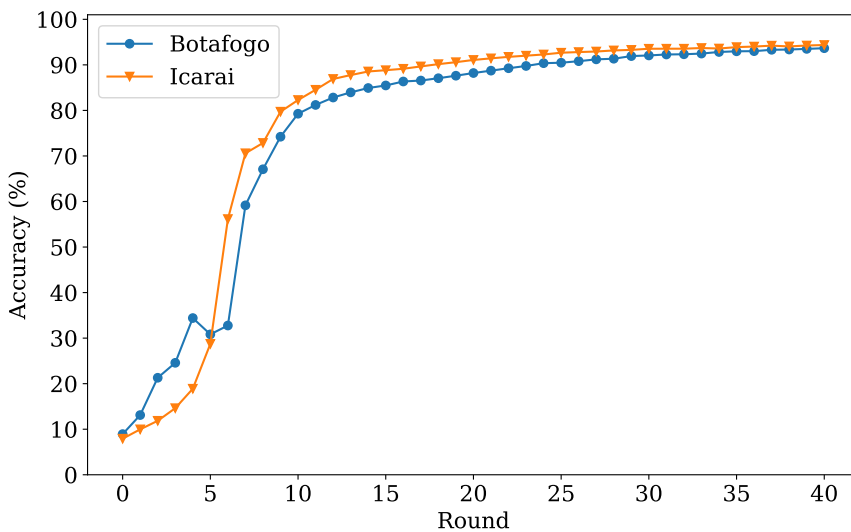


Figure 4: Accuracy obtained at each training round for Botafogo and Icaraí.

From a traditional ML perspective, a larger dataset typically improves model performance [19]. In this sense, the region of Botafogo should intuitively yield a better model performance in federated learning training, since the region shows a higher number of vehicles throughout the simulation and thus more data. Nonetheless, the model accuracy plotted in Figure 4 shows a counterintuitive result. Botafogo exhibits a worse accuracy curve, despite the higher data volume. The difference is more relevant in the first half of the simulation, showing that Icarai provides faster convergence. Based on this result, we formulate the hypothesis: Each client contributes equally to compose a representative dataset (*in sample*) when compared with the real data distribution (*out of sample*). A region that observes a larger average vehicle stay can produce bias towards the contribution of persistent vehicles, thereby reducing the model's performance.

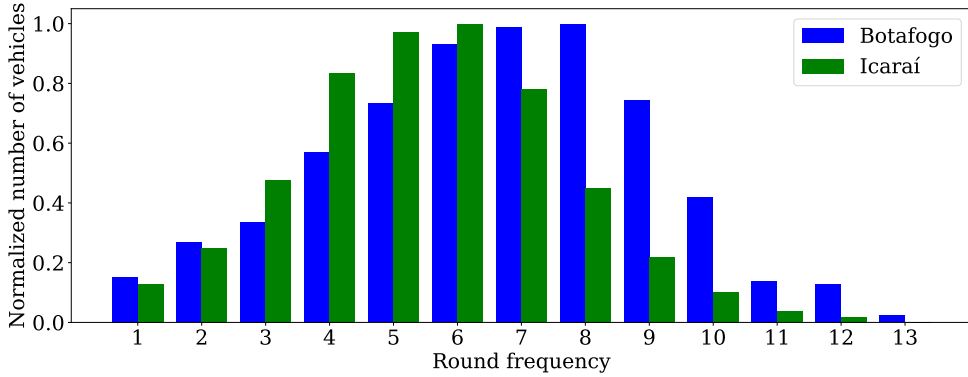


Figure 5: Frequency of vehicle participation per training round.

In addition to the average trip duration and the volume of vehicles in each round, we can directly infer the frequency of vehicle participation throughout the simulation in Figure 5. The frequency of vehicle participation is the number of FL rounds that a vehicle has taken part in. We normalize this frequency by the maximum frequency among all the vehicles in each region.

The green curve, representing the Icarai region, is centered towards 5-6 rounds, which means that most of the vehicles contributed over 5-6 rounds of federated learning training. The green curve also shows a significant decay for values above 8 rounds. The blue curve, representing the region of Botafogo, is centered towards the values of 7-8 rounds, which already translates into more frequent vehicles. In addition to that, the blue curve also presents a significant spread towards higher values, assuming more than double the vehicles that participate in 9 and 10 rounds, when compared with Icarai.

To validate our hypothesis, we then calculate the Dice-Sørensen coefficient between the datasets D_{R_i} , inferring the degree of variability in the datasets throughout the n federated rounds of simulation. The coefficient can be calculated using Equation 3. This analysis results in a DS matrix for each region, represented in Figures 6(a) and 6(b). In this matrix, a higher coefficient value indicates a greater similarity between the datasets of different rounds. This explains why the main diagonal shows the maximum values, as it compares each round's total dataset with itself. Interpreting the matrix, the region of Botafogo has higher values of DS coefficient as well as a higher dispersion from the main diagonal. The Icarai matrix shows both a lower dispersion from the main diagonal as well as lower DS coefficients. Initially, when observing Figures 6(a) and 6(b) the difference is subtle. Nevertheless, in the following more controlled experiments, this subtle difference becomes more evident, showing that a higher rate of vehicles permanence results in a higher similarity between rounds' datasets. This can degrade the model, potentially due to a bias towards the most frequent vehicles.

To better investigate whether the extended presence of vehicles in training rounds impacts the model performance, we perform a controlled experiment for the Botafogo region. A maximum participation threshold is set sequentially across rounds. This implies that if a vehicle has participated in training up to the maximum threshold, its involvement in subsequent rounds is disregarded, and its neural network weights are not aggregated into the global model. We evaluate two values of maximum participation rate and compare the results in Figure 7. The plot demonstrates that lowering the participation threshold improves performance, with the limit set at 3 rounds outperforming the thresholds

of 4 rounds and the original vehicle pooling strategy. The curves constrained by the round participation limit exhibit improved convergence towards the same final accuracy.

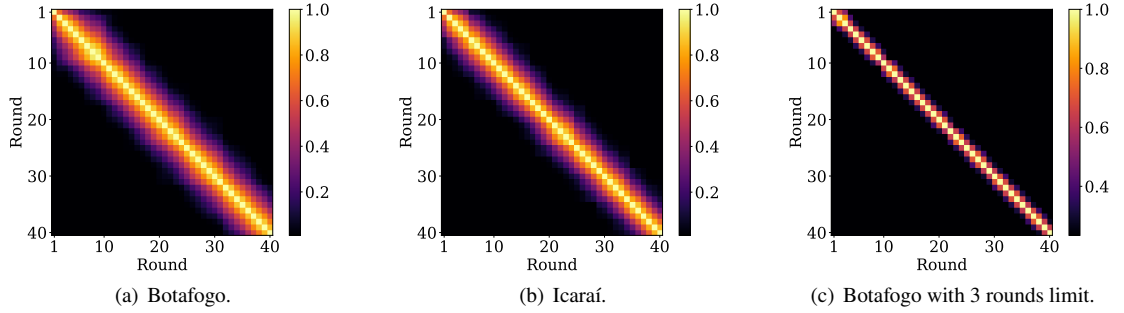


Figure 6: DS matrix for the different regions.

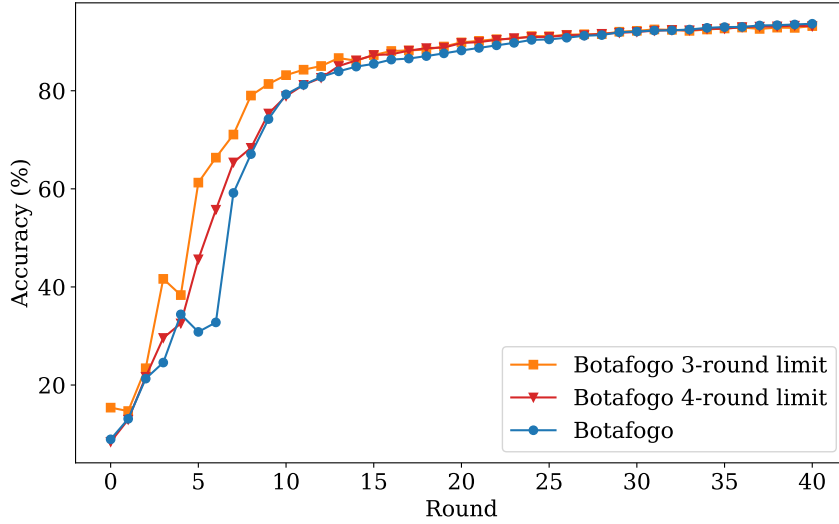


Figure 7: Performance comparison with maximum participation threshold.

We can then compute the DS coefficient matrix for the 3-limit curve in Figure 6(c) and compare the degree of similarity with the original Botafogo matrix. A clear difference can be observed between Figures 6(c) and 6(a). For the 3-limit matrix, we see a drastic reduction in the main diagonal dispersion, as the similarity of datasets between rounds is reduced. This means that the model receives less influence from the same vehicles across the rounds, and consequently the weight vectors are less biased towards those vehicles datasets. This shows that vehicles that stay longer in a region, participating in an excessive number of FL rounds, negatively impact the model's performance.

Figure 9 shows the vehicle distribution with the imposed limit. Despite achieving better performance compared to the original Botafogo setup, implementing a 3-round limit also results in a 47% reduction in the average number of vehicles required for training, as observed in the distribution of vehicles in Figure 9.

Reducing the number of vehicles in training not only implies less clients to motivate to participate in training, but also reduces the communication costs associated. The cost is calculated as the size of one parameter, which is 4 bytes since our model uses floating-point 32 precision, times the number of model parameters (2.5M), times 2 to account for the downlink and uplink communication. In Figure 8 we present the communication cost associated with training in each case. The cost is illustrated in the x -axis, in Gigabytes, for the original Botafogo setup and the respective round

reductions, under various predictions accuracies. Specifically, the 3-round limit results in a cost reduction of 54,7%, 54,9% and 55,7%, according to each accuracy threshold.

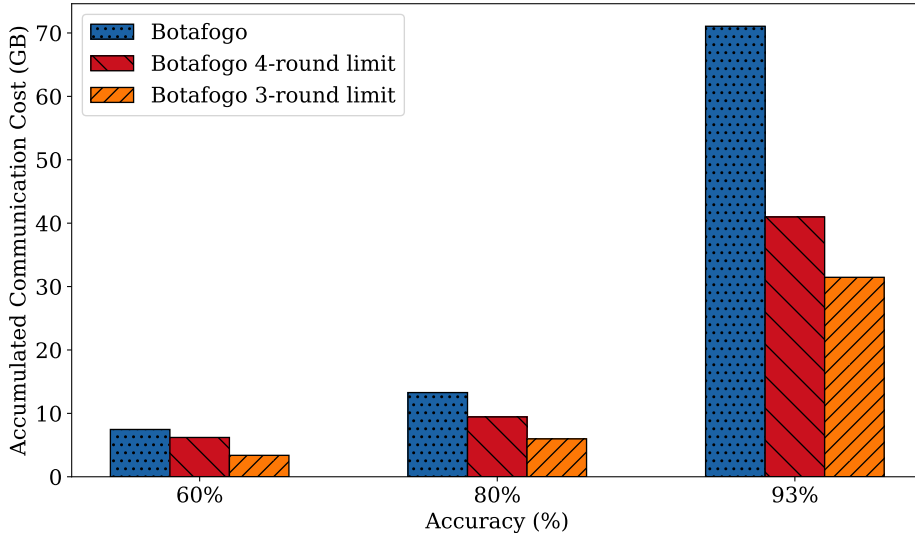


Figure 8: Accumulated communication cost for different accuracies.

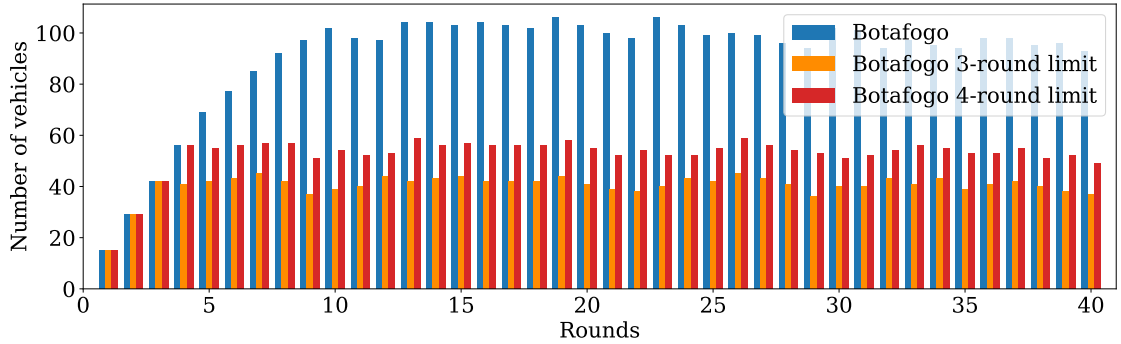


Figure 9: Number of vehicles participating in each training round for Botafogo with the original distribution and the three and four rounds limit distributions.

The performance increase over the original Botafogo setup, combined with the reduction of the average dataset size across the rounds, validates our hypothesis that a higher rate of vehicles permanence can produce bias towards the frequent vehicles' contribution, thereby reducing model performance. This bias can be explained by the federated learning aggregation process. The traditional federated learning aggregation strategy, *FedAvg* [28], generates a global model in every round by applying an average between the neural networks weights. When vehicles start participating in a large amount of rounds, the average weight vectors tend to point towards the weights of those frequent vehicles, creating the bias. The impact on performance is explained by classical concepts in machine learning [4], which assert that a model biased towards the training data tends to have a reduced generalization capacity. This bias can lead to a loss in performance proportional to the difference between the error observed in the training sample and the error expected on new, unseen data.

4.2. Diverse Traffic Patterns Analysis with a state-of-the-art neural network model

Expanding on the evaluation of traffic patterns in the Botafogo and Icarai regions, we shift our focus to include a broader set of patterns applied to more diverse datasets. This section focuses on training a state-of-the-art model

optimized for embedded devices, aiming to evaluate the effects of extreme vehicle distribution scenarios. Our goal is to understand the potential outcomes of these scenarios while leveraging a high-performance neural network to handle datasets of varying complexity.

In the new experiments, the variation in participation frequencies across the Brief, Moderate, and Extended scenarios leads to significant differences in the number of vehicles involved in training during each round. The number of vehicles for these experiments is shown in Figure 10. A clear reduction in vehicle counts is observed when comparing the Brief scenario, represented by the yellow bars, to the Moderate and Extended scenarios, represented by the blue and brown bars, respectively. This outcome aligns with the inverse relationship between participation frequency and the occurrence of overlapping vehicles across rounds, where lower participation frequency naturally results in fewer vehicles per round.

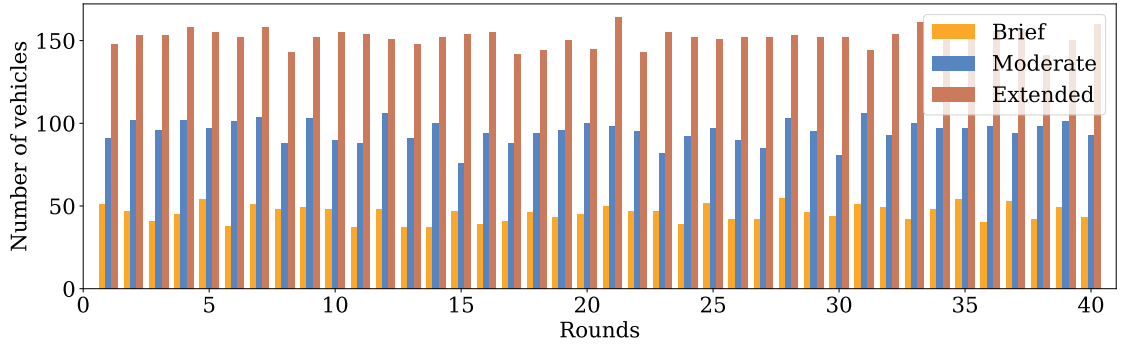


Figure 10: Number of vehicles participating in each training round for different scenarios with maximum individual participation of 40 rounds.

In Figure 13 we can analyze aspects in terms of complexity of the datasets and influence from the different traffic patterns when subjected to a state-of-the-art model. The simulations validate our previous assumption that the CIFAR datasets are increasingly more complex when going from CIFAR-10 up to CIFAR-100 fine. While for the CIFAR-10 dataset, the Extended distribution yields the best accuracy of 85.02%, for the CIFAR-100 fine, the maximum accuracy reached is 52.8%. It is interesting to notice that with varying vehicle participation, not only the global model performance decays in its absolute value, but also the training noise is increased when enhancing the complexity of the task, which is clear when comparing the curves 13(a) and 13(c). Across all 4 datasets, the deep neural network model is able to handle in a comparable manner all three synthetic traffic patterns generated. This result is interesting if we recall Figure 10, which shows that the Brief distribution observes approximately 60% less vehicles training when compared to the Extended distribution. In other words, if we maintain a higher vehicle variability while using a state-of-the-art model, we are capable of reducing the number of vehicles involved in training while achieving comparable performance.

Some other observations can be made regarding the specifics of each dataset experiment. For the CIFAR-100 coarse dataset, illustrated in Figure 13(b), we observe an overfitting behavior as training progresses across all distributions. The maximum performance for all three distributions is observed at round 7. Beyond that, not only is the accuracy observed to decrease, but the noisy behavior also seems to become significantly more pronounced. The CIFAR-100 fine dataset in Figure 13(c) exhibits the highest level of training noise among all experiments, which is expected given the increased complexity of classifying a larger number of classes. Assuming that every vehicle has an identical storage capacity across all experiments and datasets, this implies that each client retains fewer samples per class. As a result, the local datasets become more narrowly focused, representing specific partitions of the overall data distribution. Consequently, each local training phase is more biased towards that specific partition of the data distribution, causing more abrupt changes in the neural network weight at each training step.

Although the GTSRB dataset includes many classes, with vehicles holding only five samples per class, its complexity is lower than that of CIFAR-10, which has just 10 classes. This is because traffic signs are intentionally designed to be easily distinguishable, featuring distinctive visual traits like contrasting colors or shapes for quick recognition. As a result, the model performs exceptionally well across all distributions on the GTSRB dataset, as

Experiment	CIFAR-10	CIFAR-100 Coarse	CIFAR-100 Fine	GTSRB
Brief	84.74%	64.52%	52.80%	96.01%
Moderate	84.46%	64.46%	52.64%	95.98%
Extended	85.02%	64.24%	50.48%	95.66%
Centralized	93.40%	81.00%	73.80%	95.20%

Table 7

Top accuracy achieved for different synthetic distributions and the centralized setup across datasets.

shown in Figure 13(d). The classification simplicity of the GTSRB task is further highlighted by the federated setup achieving a peak accuracy of 96.32%, surpassing the centralized setup's accuracy of 95.2%. This outcome is surprising, as centralized performance typically serves as the upper bound for federated learning accuracy.

Figure 11 shows the total communication costs for the model to reach its top accuracy on each dataset. Across all datasets, the **Brief** distribution consistently outperforms the other two. Interestingly, training on the *Cifar-100 Coarse* dataset incurs the **lowest** communication cost, while *GTSRB* results in the **highest**, which appears counter-intuitive. This is because, although GTSRB is an easier task for the model and *Cifar-100 Coarse* one of the hardest, the model quickly reached peak performance on *Cifar-100 Coarse* in early rounds and then experienced a decline due to overfitting, as discussed earlier. In contrast, performance on GTSRB continued to improve steadily with more training rounds.

To better understand this discrepancy, Figure 12 presents the wasted communication cost—that is, the portion of training that did not contribute to improved model performance. This figure reveals that GTSRB had the lowest amount of wasted communication, while *Cifar-100 Coarse* had the highest, with nearly 100 GB of communication cost yielding no performance gain. Additionally, this graph emphasizes the benefits of deploying a federated learning model in a region with a Brief traffic pattern, which can save up to 83.54 GB in communication costs compared to an Extended distribution.

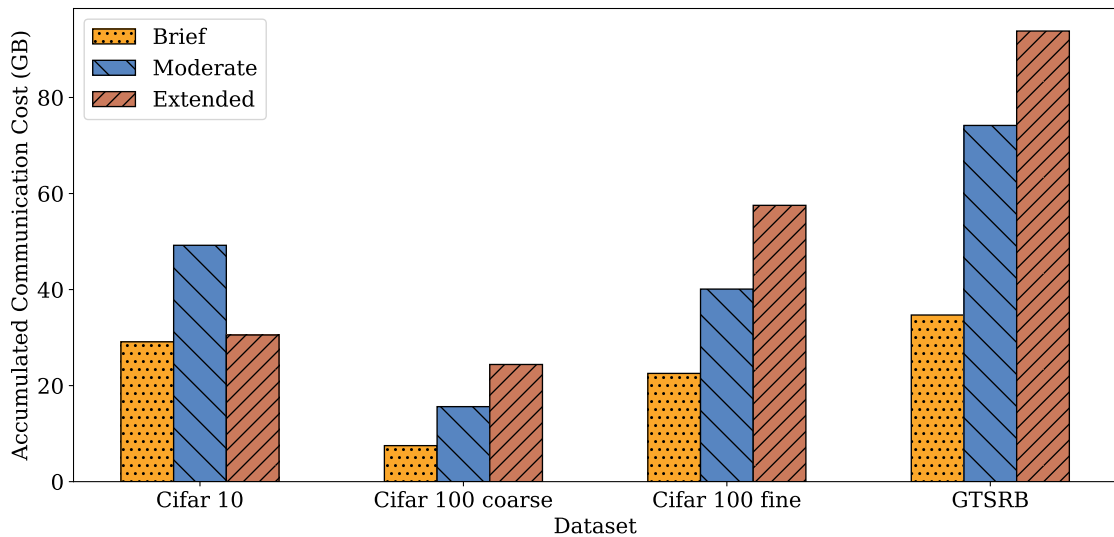


Figure 11: Accumulated communication cost for each distribution across datasets.

For benchmarking purposes, Table 7 depicts the top accuracy of each experiment, as well as the centralized performance, all using the MobileNetV3.

4.2.1. Assessing bias

An imbalance in vehicle participation can introduce bias towards frequently participating vehicles. While preliminary results with a simple CNN revealed such bias, experiments with a state-of-the-art model showed differing outcomes. This section aims to determine whether bias persists in federated learning with varying client frequencies while using a state-of-the-art model for embedded systems. To assess this, we evaluate the global model's performance

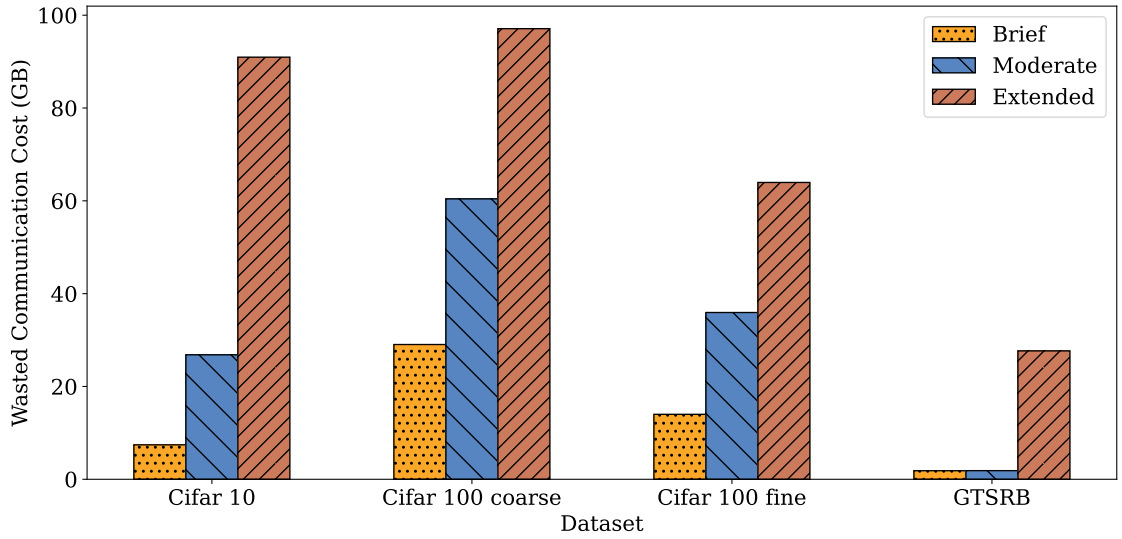


Figure 12: Wasted communication cost for each distribution across datasets.

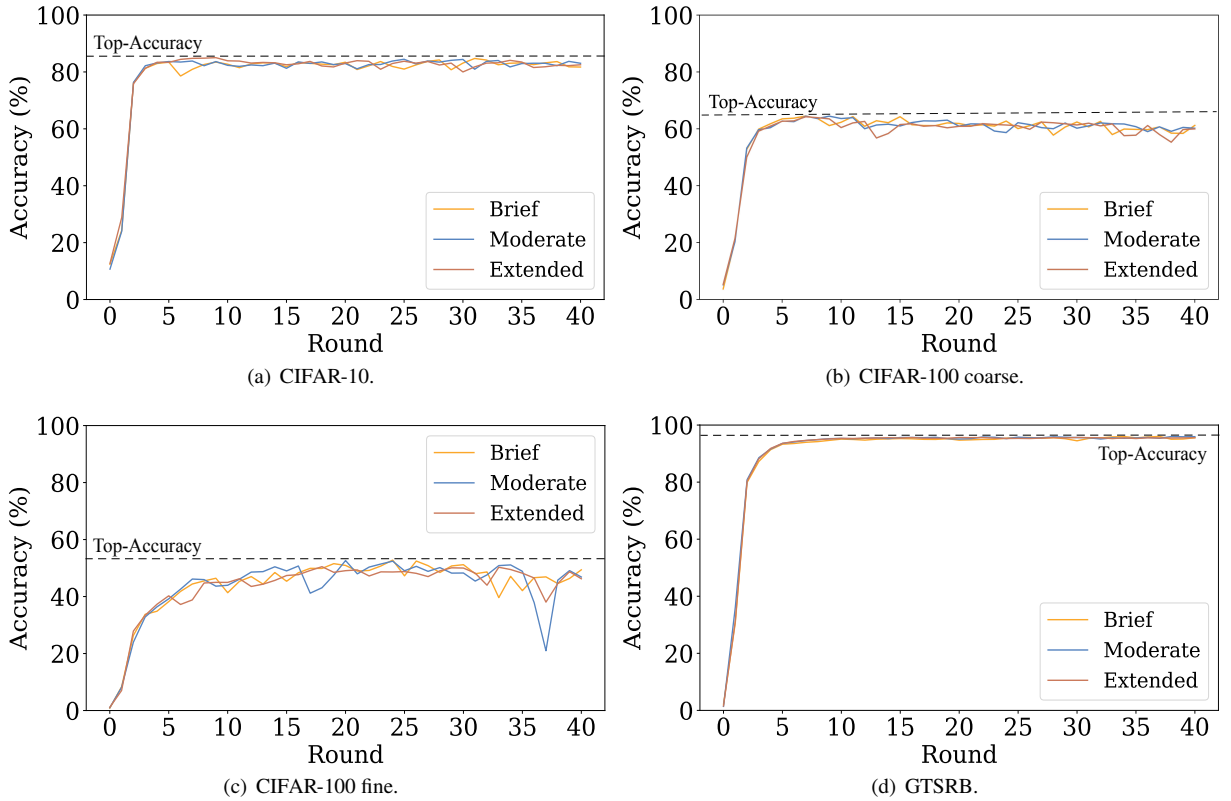
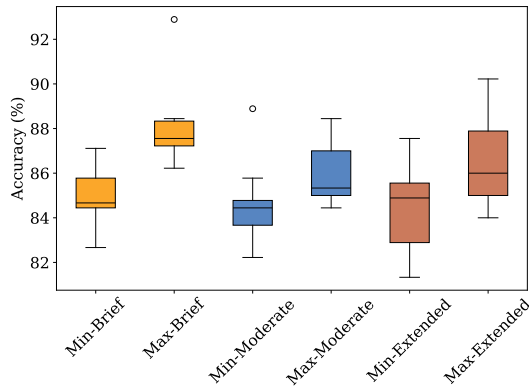
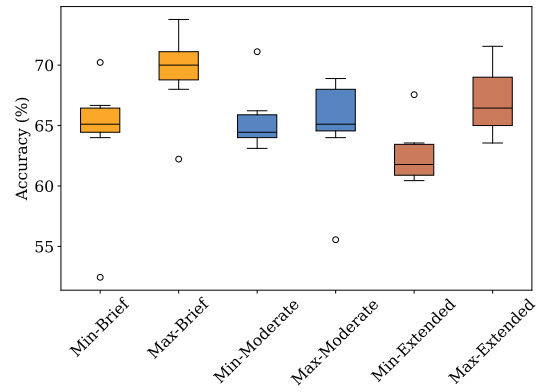


Figure 13: Accuracy comparison for different datasets and distributions with a maximum participation of 40 rounds.

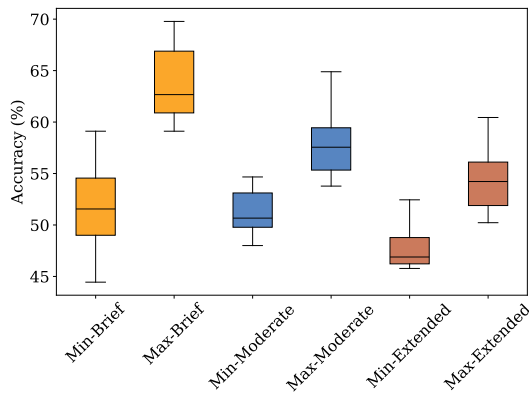
on training data from two groups: vehicles with the least and most participation rounds during training. By comparing accuracy between these groups, we aim to determine whether clients that participate in the least amount of rounds experience lower performance due to the global model being less exposed to their local datasets.



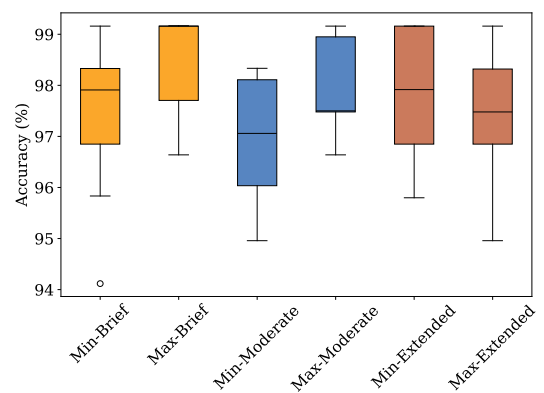
(a) CIFAR-10.



(b) CIFAR-100 coarse.



(c) CIFAR-100 fine.



(d) GTSRB.

Figure 14: Accuracy of the global model on the local datasets from clients that participated in the minimum and maximum number of federated rounds.

The figures present, for each dataset, boxplots depicting the individual accuracy values for vehicles in each participation group. "Min-Brief" corresponds to the ten vehicles with the least participation in training under the Brief distribution, while "Max-Brief" represents the ten vehicles with the highest participation in the same simulation. This naming convention is consistently applied across the Moderate and Extended distributions.

In datasets other than GTSRB, a discernible trend indicates higher accuracy for vehicles with maximum participation. This is evident in Figure 14(a) for the CIFAR-10 dataset, where differences in median, first, and third percentile values between the "Min" and "Max" cases are observed across all distributions. Similar patterns are seen in Figure 14(b) for the CIFAR-100 coarse dataset and Figure 14(c) for the CIFAR-100 fine dataset. To support this analysis, Table 8 presents the median accuracy values from the box plots, while Table 9 provides accuracy at the Q1 (25th percentile) and Q3 (75th percentile), with the highest values emphasized in bold.

The presence of this trend indicates that the global model is more adapted, or in other words, biased to the local datasets of the vehicles that are more frequent in training. The most pronounced bias is observed in the CIFAR-100 fine dataset under the Brief scenario, with a gap of approximately 11.1% in median accuracy, as detailed in Table 8. This finding aligns with our earlier assumption that more complex datasets amplify the impact of each vehicle's contribution, thereby intensifying the bias caused by imbalanced vehicle participation. For the GTSRB experiment, we have seen that the model is capable of handling the data complexity as if the training was conducted in a centralized fashion. Because of that, differences in client participation for the federated learning setup do not influence the performance overall.

Table 8

Median for the accuracy over the training set of the minimum and maximum participating vehicles across all datasets.

Datasets	Distributions	Median (min)	Median (max)
CIFAR-10	Brief	84.6%	87.5%
CIFAR-10	Moderated	84.4%	85.3%
CIFAR-10	Extended	84.8%	86%
CIFAR-100 coarse	Brief	65.1%	70%
CIFAR-100 coarse	Moderated	64.4%	65.1%
CIFAR-100 coarse	Extended	61.7%	66.4%
CIFAR-100 fine	Brief	51.5%	62.6%
CIFAR-100 fine	Moderated	50.6%	57.5%
CIFAR-100 fine	Extended	46.8%	54.1%
GTSRB	Brief	97.9%	99.1%
GTSRB	Moderated	97.0%	97.5%
GTSRB	Extended	97.9%	97.5%

Table 9

Q1 (25th percentile) and Q3 (75th percentile) for the accuracy over the training set of the minimum and maximum participating vehicles across all datasets.

Datasets	Distributions	Q1 (min)	Q1 (max)	Q3 (min)	Q3 (max)
CIFAR-10	Brief	84.4%	87.2%	85.7%	88.3%
CIFAR-10	Moderated	83.6%	85.0%	84.7%	87%
CIFAR-10	Extended	82.8%	85%	85.5%	87.8%
CIFAR-100 coarse	Brief	64.4%	68.7%	66.4%	71.1%
CIFAR-100 coarse	Moderated	64.0%	64.5%	65.8%	68%
CIFAR-100 coarse	Extended	60.8%	65.0%	63.4%	69%
CIFAR-100 fine	Brief	49.0%	60.9%	54.5%	66.8%
CIFAR-100 fine	Moderated	49.7%	55.3%	53.1%	59.4%
CIFAR-100 fine	Extended	46.2%	51.8%	48.7%	56.1%
GTSRB	Brief	96.8%	97.7%	98.3%	99.1%
GTSRB	Moderated	96.0%	97.5%	98.1%	98.9%
GTSRB	Extended	98.8%	96.8%	99.1%	98.3%

4.3. Impact over Non-IID Data Distributions

To further explore the effects of traffic patterns on federated learning, we extend our analysis to scenarios where the data distribution across clients is non-identically and independently distributed (non-iid). This setting more closely reflects real-world deployments, where each vehicle is likely to collect data from its unique driving context, potentially leading to both class imbalance and unequal dataset sizes.

We simulate non-iid data by partitioning the dataset using a *Dirichlet distribution* with two different concentration parameters: $\alpha = 5.0$ and $\alpha = 0.5$. The $\alpha = 5.0$ configuration introduces mild heterogeneity, while $\alpha = 0.5$ results in a highly skewed distribution, where clients receive samples from only a few classes, and the number of samples per client varies significantly. For comparison, we also include the iid case used in the previous sections.

Figure 15 shows the data distribution among clients for the three configurations: iid, Dirichlet $\alpha = 5.0$, and Dirichlet $\alpha = 0.5$. These visualizations help illustrate the increasing levels of heterogeneity in both class distribution and dataset size. To isolate the effects of data heterogeneity, we conduct these experiments exclusively on the CIFAR-10 dataset. CIFAR-10 serves as a standard benchmark for image classification tasks and offers a balanced level of complexity suitable for evaluating federated learning behavior. In contrast, CIFAR-100 (coarse and fine) shares the same image set but rearranged into more granular class structures, which would introduce confounding factors related to task difficulty. Additionally, the GTSRB dataset was found to be too simple for the chosen model, yielding limited insight in the context of non-iid evaluation. To further focus our analysis, we consider only the Brief and Extended traffic distributions—representing the two extremes in client participation frequency. Since our earlier results showed

minimal accuracy differences across traffic scenarios under iid conditions, this selection allows us to assess whether the combination of non-iid data and varying client participation introduces new performance variations during training.

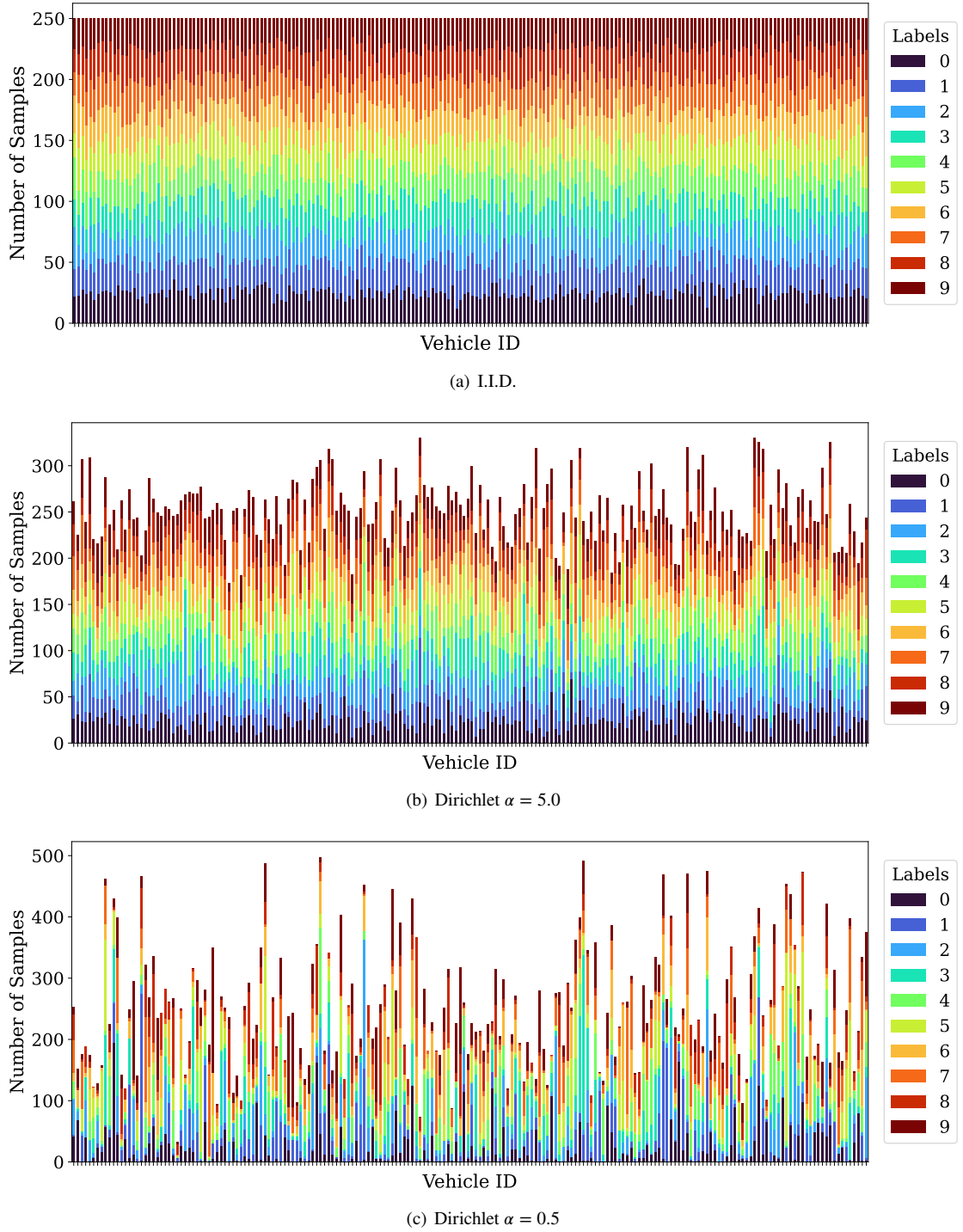


Figure 15: Client data distributions under different partitioning schemes.

Due to the additional challenge introduced by heterogeneous data and client availability, we also explore a more robust aggregation strategy proposed by [33], **FedOpt**. Unlike FedAvg, which applies simple averaging, FedOpt

introduces an adaptive optimization layer on the server that can better handle discrepancies in local updates—making it a more suitable candidate for FL scenarios involving highly variable client behavior and data distributions. In our experiments, we specifically implement the **FedYogi** variant of FedOpt, which builds on the Yogi optimizer to adaptively adjust learning rates and stabilize training under non-iid conditions.

Figures 16 and 17 compare the training performance of FedAvg and FedYogi, respectively, across different traffic distributions (Brief, Moderate, and Extended) and Dirichlet α values. Each curve represents a unique combination of traffic scenario and data heterogeneity level.

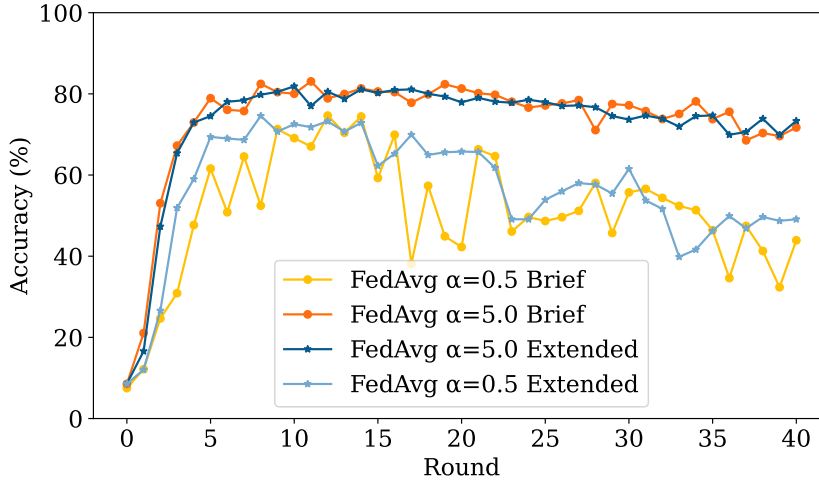


Figure 16: Performance of FedAvg across non-iid data settings and traffic distributions.

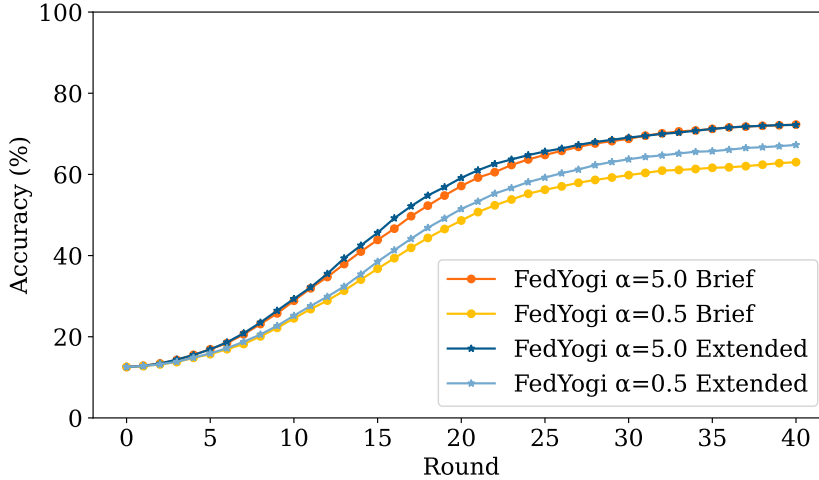


Figure 17: Performance of FedYogi across non-iid data settings and traffic distributions.

FedAvg Results. Figure 16 presents the performance of FedAvg under different traffic patterns and levels of data heterogeneity. With a Dirichlet distribution parameter of $\alpha = 5.0$, both the **Brief** and **Extended** scenarios achieve comparable top-1 accuracies—83.07% and 81.07%, respectively. Although these values are relatively close to each other, they remain consistently lower than those observed under iid conditions (84.74% for Brief and 85.02% for Extended). Moreover, both configurations exhibit signs of overfitting in the later training rounds (particularly between rounds 20 and 40), as accuracy declines despite ongoing training. This overfitting behavior contributes to increased **wasted communication cost**, as additional communication and computation yield negative returns in model performance.

When the degree of heterogeneity is increased—using a Dirichlet distribution with $\alpha = 0.5$ —the impact on model performance becomes more pronounced. In this highly skewed setting, the **Brief** distribution suffers the most, reaching a reduced top-1 accuracy of 74.64% and exhibiting considerable instability in the training process, with high variance and noise in the accuracy curve. The overfitting effect observed earlier becomes even more severe under this configuration. In contrast, the **Extended** distribution demonstrates greater robustness, achieving a higher top-1 accuracy of 77.64% and maintaining more stable convergence during the initial rounds. Nevertheless, even this more stable scenario suffers from overfitting, with performance degrading in the final rounds of training.

Taken together, these results underscore an important trade-off in the design of federated learning deployments. On the one hand, traffic scenarios with higher client variability—such as the Brief distribution—can improve communication efficiency by increasing data diversity across rounds, particularly when the underlying data distribution is close to iid or only moderately non-iid. On the other hand, when client data becomes highly skewed, as in the $\alpha = 0.5$ case, this variability introduces instability that can hinder model generalization and lead to poorer performance. In such cases, traffic patterns with more persistent client participation—such as the Extended scenario—can help stabilize training and mitigate the negative effects of extreme data heterogeneity. These findings suggest that selecting a deployment region for federated learning should consider both the expected client data distribution and the local traffic behavior, as their interaction plays a critical role in overall system performance.

FedYogi Results. Figure 17 shows the performance of FedYogi, a variant of FedOpt designed to handle client and data heterogeneity through adaptive learning rate adjustment. Compared to FedAvg, FedYogi exhibits a distinct training dynamic: although convergence is noticeably slower, the learning process is highly stable, with smooth and consistent accuracy gains throughout all 40 training rounds. Notably, there is no sign of overfitting, even in the later stages of training—contrasting with the behavior observed in FedAvg under similar conditions.

For the **Brief** traffic distribution, the $\alpha = 5.0$ setting achieved a top-1 accuracy of 72.28%, outperforming the more heterogeneous $\alpha = 0.5$ configuration, which reached 63.02%. Similarly, in the **Extended** traffic scenario, the $\alpha = 5.0$ configuration yielded a top-1 accuracy of 72.19%, closely matching the Brief case. Interestingly, while the Extended scenario starts off slightly stronger than Brief—showing better accuracy around round 20—the final accuracy values for both traffic patterns converge to nearly the same level.

Under the more extreme heterogeneity of $\alpha = 0.5$, however, the Extended distribution shows a superior performance, reaching a top-1 accuracy of 67.27% compared to 63.02% in the Brief case. This result suggests that longer client participation, as enabled by the Extended distribution, can help offset the instability caused by highly skewed data partitions. The extended presence allows for more consistent updates from individual clients, which may better align local model trajectories with the global objective over time.

These findings highlight a trade-off: while FedYogi improves stability across all configurations, the benefits of extended client participation become more apparent as data heterogeneity increases. The slower convergence rate observed in FedYogi is characteristic of adaptive optimizers like Yogi, which apply conservative updates early on to stabilize training under noisy and heterogeneous gradients. This behavior, while delaying peak accuracy, enhances reliability and prevents premature overfitting—particularly valuable in federated settings with dynamic client participation.

Overall, these results suggest that in scenarios with high variability and non-iid data, FedYogi offers a more robust alternative to FedAvg by favoring long-term stability over short-term gains, and that traffic patterns with longer client presence may help further mitigate the effects of extreme heterogeneity.

5. Conclusion and Future Work

In this work, we analyzed the vehicular federated learning behavior when subjected to different mobility traffic patterns. First, we have simulated two realistic federated learning scenarios considering the mobility patterns of two distinct regions. The regions selection was conducted based on a proposed criteria that compares traffic complexity between regions. We then evaluated the impact of vehicle permanence in the regions regarding the similarity between the total datasets across the rounds, using the *Dice-Sørensen* (DS) coefficient. Surprisingly, lower performance was observed for the scenario with a higher vehicular count. By analyzing this scenario's data variability across training, we observed a higher dispersion on the main diagonal of the DS matrix. We have then formulated the hypothesis that this degradation was due to the presence of bias in the model towards the vehicles with an excessive prevalence in training rounds, which affected the model's generalization capabilities.

The hypothesis was then tested by setting a limit to the maximum rounds a vehicle could participate in the federated learning. That drastically reduced the dispersion in the DS matrix and increased accuracy performance, while observing a significant 47% reduction in the number of participating vehicles across the rounds, confirming our hypothesis.

To expand our findings, we then conducted simulations over more distinct cases of traffic patterns while considering a state-of-the-art model for embedded systems and four image classification datasets with different levels of complexity. For the expanded simulations, bias was observed across all CIFAR datasets and distributions, which validates that frequent vehicles do generate bias in the global model towards their own local data. However, a state-of-the-art model is capable of countering this issue in terms of accuracy since its architecture leverages better the available data. The GTSRB dataset showed minimal performance impact in all scenarios, with federated learning often outperforming centralized training. This high performance is expected, as GTSRB classes (traffic signs) are designed to be easily distinguishable, simplifying the model's feature extraction and class separation.

An important consideration when comparing federated learning methods is how to assess fairness in performance evaluations. Federated learning introduces additional hyperparameters compared to traditional machine learning, such as batch size and the number of local epochs per round, which can significantly impact performance. However, many studies do not optimize these parameters, which limits the analysis and comparability of results, and even in some cases, these parameter values are not specified, making it difficult to reproduce results. Therefore, it is important for federated learning studies to clearly specify the number of local epochs, client hyperparameters, ensuring a fair optimization of the federated learning setup in comparison with other works or even with the centralized approach.

Traffic patterns play a significant role in the deployment of vehicular federated learning solutions, yet research in this area remains limited. This work offers initial analyses, debating the impact of traffic patterns for classification tasks and emphasizing the importance of exploring this topic in a thorough way.

Future studies should explore a broader range of machine learning tasks, as bias behavior may vary depending on data distribution and the specific task. Real-world vehicular applications, such as object detection and image segmentation, should be considered in the analysis of federated learning, while incorporating traffic patterns. This extension would provide valuable insights regarding the impact of the mobility over the system while reducing the gap between research and deploy. However, several challenges would need to be addressed. First, the dataset size is crucial, as traffic pattern analysis often requires a large number of clients, necessitating datasets with enough data per client for effective learning. Additionally, the context in vehicular networks, in terms of the time of day and the geographical location, should be considered for the deploy of federated learning schemes, since different contexts might require different models. For example, the behavior of vehicles might change during the late night period, when compared to rush hours. The same can be expected when observing traffic between two different cities, in different countries. The idea is that federated learning could offer the ability to train global models tailored for these specific scenarios, similarly to a transfer learning approach. However, the definition of context in vehicular networks remains an area to explore. How to precisely define if a given geographical region should benefit from the same global model or if the region should be segmented into smaller regions, each with its specific model, it is still an open challenge. Lastly, for effective vehicular federated learning deployment, we need a better understanding of the mobility of the nodes across the city. Tracking the participating clients and understanding their dynamic availability patterns, rather than relying on simulators or synthetic distributions, offers a more realistic analysis for developing not only vehicular federated learning solutions but also any user-based mobile system.

CRediT authorship contribution statement

Giuliano Fittipaldi: Conceptualization of this study, Methodology, Simulations development, Data analysis, Writing. **Rodrigo S. Couto:** Conceptualization of this study, Methodology, Writing, Work supervision, Funding acquisition. **Luís H. M. K. Costa:** Conceptualization of this study, Methodology, Writing, Work supervision, Funding acquisition.

References

- [1] Abuelenin, S.M., Abul-Magd, A.Y., 2014. Empirical study of traffic velocity distribution and its effect on vanets connectivity, in: 2014 International Conference on Connected Vehicles and Expo (ICCVE), pp. 391–395. doi:10.1109/ICCVE.2014.7297577.
- [2] AlNagar, Y., Hosny, S., El-Sherif, A.A., 2019. Towards mobility-aware proactive caching for vehicular ad hoc networks, in: 2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW), pp. 1–6. doi:10.1109/WCNCW.2019.8902903.

- [3] Bardenet, R., Lindauer, M., Kégl, B., 2019. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research* 20, 1–32. Submitted 7/18; Revised 2/19; Published 3/19.
- [4] Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*. Springer.
- [5] Cao, X., Jia, J., Gong, N., 2021. Provably secure federated learning against malicious clients. *ArXiv abs/2102.01854*. doi:10.1609/aaai.v35i8.16849.
- [6] Council of the European Union, 2024. Artificial intelligence act, text of the provisional agreement, 2 february 2024. URL: <https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf>. accessed: January 12, 2025.
- [7] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.
- [8] Deng, L., 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* 29, 141–142.
- [9] (DLR), G.A.C., 2024. SUMO: Simulation of Urban MObility. DLR Institute of Transportation Systems. URL: <https://www.eclipse.org/sumo/>. accessed: 2025-01-01.
- [10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR abs/2010.11929*. URL: <https://arxiv.org/abs/2010.11929>, arXiv:2010.11929.
- [11] Elbir, A.M., Soner, B., Çöleri, S., Gündüz, D., Bennis, M., 2022. Federated learning in vehicular networks, in: 2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), IEEE. pp. 72–77.
- [12] Fiore, M., Härris, J., 2008. The networking shape of vehicular mobility , 261–272doi:10.1145/1374618.1374654.
- [13] Fittipaldi, G., Couto, R.S., Henrique, M.C.L., 2024. On the impact of the traffic pattern on vehicular federated learning, in: 2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), IEEE. pp. 365–370.
- [14] Foundation, O., 2024. OpenStreetMap. OpenStreetMap Foundation. URL: <https://www.openstreetmap.org/>. accessed: 2024-01-01.
- [15] Gao, J., Zhang, B., Guo, X., Baker, T., Li, M., Liu, Z., 2022. Secure partial aggregation: Making federated learning more robust for industry 4.0 applications. *IEEE Transactions on Industrial Informatics* 18, 6340–6348. doi:10.1109/TII.2022.3145837.
- [16] Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- [17] Guo, J., Liu, Z., Tian, S., Huang, F., Li, J., Li, X., Igorevich, K.K., Ma, J., 2023. Tfl-dt: A trust evaluation scheme for federated learning in digital twin for mobile networks. *IEEE Journal on Selected Areas in Communications* 41, 3548–3560. doi:10.1109/JSAC.2023.3310094.
- [18] Haase, D., Amthor, M., 2020. Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets. URL: <https://arxiv.org/abs/2003.13549>, arXiv:2003.13549.
- [19] Halevy, A., Norvig, P., Pereira, F., 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 8–12. doi:10.1109/MIS.2009.36.
- [20] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* , 770–778.
- [21] Kang, J., Xiong, Z., Niyato, D., Xie, S., Zhang, J., 2019. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE Internet of Things Journal* 6, 10700–10714. doi:10.1109/JIOT.2019.2940820.
- [22] Kosman, E., Leonard, K., 2005. Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Molecular Ecology* 14, 415–424. doi:10.1111/j.1365-294X.2005.02416.x.
- [23] Liu, S., Yu, J., Deng, X., Wan, S., 2021. Fedcpf: An efficient-communication federated learning approach for vehicular edge computing in 6g communication networks. *IEEE Transactions on Intelligent Transportation Systems* 23, 1616–1629.
- [24] Liu, W., Chen, L., Chen, Y., Zhang, W., 2019. Accelerating federated learning via momentum gradient descent. *arXiv:1910.03197*.
- [25] Liu, Z., Wan, L., Guo, J., Huang, F., Feng, X., Wang, L., Ma, J., 2025. Ppru: A privacy-preserving reputation updating scheme for cloud-assisted vehicular networks. *IEEE Transactions on Vehicular Technology* 74, 1877–1892. doi:10.1109/TVT.2023.3340723.
- [26] Lu, Y., Huang, X., Zhang, K., Maharjan, S., Zhang, Y., 2020. Blockchain empowered asynchronous federated learning for secure data sharing in internet of vehicles. *IEEE Transactions on Vehicular Technology* 69, 4298–4311. doi:10.1109/TVT.2020.2973651.
- [27] Makridakis, S., 2017. The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures* 90, 46–60.
- [28] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data, in: *Artificial intelligence and statistics*, PMLR. pp. 1273–1282.
- [29] Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., Guo, S., 2016. Protection of big data privacy. *IEEE access* 4, 1821–1834.
- [30] Miao, Y., Xie, R., Li, X., Liu, Z., Choo, K.K.R., Deng, R.H., 2024. Efficient and secure federated learning against backdoor attacks. *IEEE Transactions on Dependable and Secure Computing* 21, 4619–4636. doi:10.1109/TDSC.2024.3354736.
- [31] Nishio, T., Yonetani, R., 2019. Client selection for federated learning with heterogeneous resources in mobile edge, in: *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, IEEE. URL: <http://dx.doi.org/10.1109/ICC.2019.8761315>, doi:10.1109/icc.2019.8761315.
- [32] Pervej, M.F., Jin, R., Dai, H., 2023. Resource constrained vehicular edge federated learning with highly mobile connected vehicles. URL: <https://arxiv.org/abs/2210.15496>, arXiv:2210.15496.
- [33] Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., McMahan, H.B., 2021. Adaptive federated optimization. URL: <https://arxiv.org/abs/2003.00295>, arXiv:2003.00295.
- [34] Reddi, S.J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., McMahan, H.B., 2020. Adaptive federated optimization. *CoRR abs/2003.00295*. URL: <https://arxiv.org/abs/2003.00295>, arXiv:2003.00295.
- [35] Ribero, M., Vikalo, H., de Veciana, G., 2023. Federated learning under intermittent client availability and time-varying communication constraints. *IEEE Journal of Selected Topics in Signal Processing* 17, 98–111. doi:10.1109/JSTSP.2022.3224590.
- [36] Rodio, A., Faticanti, F., Marfoq, O., Neglia, G., Leonardi, E., 2023. Federated learning under heterogeneous and correlated client availability, in: *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, pp. 1–10. doi:10.1109/INFOCOM53939.2023.10228876.

- [37] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2019. Mobilenetv2: Inverted residuals and linear bottlenecks. URL: <https://arxiv.org/abs/1801.04381>, arXiv:1801.04381.
- [38] Tan, M., Le, Q.V., 2021. Efficientnetv2: Smaller models and faster training. URL: <https://arxiv.org/abs/2104.00298>, arXiv:2104.00298.
- [39] Ye, D., Yu, R., Pan, M., Han, Z., 2020. Federated learning in vehicular edge computing: A selective model aggregation approach. IEEE Access 8, 23920–23935.
- [40] Yoshida, N., Nishio, T., Morikura, M., Yamamoto, K., Yonetani, R., 2020. Hybrid-fl for wireless networks: Cooperative learning mechanism using non-iid data. URL: <https://arxiv.org/abs/1905.07210>, arXiv:1905.07210.
- [41] Yousefi, S., Altman, E., El-Azouzi, R., Fathy, M., 2008. Analytical model for connectivity in vehicular ad hoc networks. IEEE Transactions on Vehicular Technology 57, 3341–3356. doi:10.1109/TVT.2008.2002957.
- [42] Yu, Z., Hu, J., Min, G., Zhao, Z., Miao, W., Hossain, M.S., 2021. Mobility-aware proactive edge caching for connected vehicles using federated learning. IEEE Transactions on Intelligent Transportation Systems 22, 5341–5351. doi:10.1109/TITS.2020.3017474.
- [43] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V., 2018. Federated learning with non-iid data. ArXiv abs/1806.00582. doi:10.48550/arXiv.1806.00582.
- [44] Zhou, Y., Ram, P., Salonidis, T., Baracaldo, N., Samulowitz, H., Ludwig, H., 2021. Flora: Single-shot hyper-parameter optimization for federated learning. CoRR abs/2112.08524. URL: <https://arxiv.org/abs/2112.08524>, arXiv:2112.08524.
- [45] Zou, K., Warfield, S., Bharatha, A., Tempany, C., Kaus, M., Haker, S., Wells, W., Jolesz, F., Kikinis, R., 2004. Statistical validation of image segmentation quality based on a spatial overlap index. Academic radiology 11 2, 178–89. doi:10.1016/S1076-6332(03)00671-8.



Giuliano Prestes Fittipaldi received his cum laude Eng. degree in electronics and computer engineering and his M.Sc. in electrical engineering from the Universidade Federal do Rio de Janeiro (UFRJ), Brazil, in 2024. He is currently a PhD candidate at LIP6-Sorbonne Université (Paris 6), in the Networks and Performance Analysis team. His major research interests include vehicular networks, mobile networks, and machine learning.



Rodrigo de Souza Couto received a cum laude B.Sc. degree in electronics and computing engineering from Universidade Federal do Rio de Janeiro, Rio de Janeiro, in 2011, and the D.Sc. degree in electrical engineering also from Universidade Federal do Rio de Janeiro in 2015. He has been an Associate Professor at Universidade Federal do Rio de Janeiro, Rio de Janeiro, since 2018. His primary research interests include internet of things, cloud/edge computing, network optimization, and deep learning. He is an associate editor of the Journal of the Brazilian Computer Society (JBACS).



Luís Henrique Maciel Kosmalski Costa received his Eng. and M.Sc. degrees in electrical engineering from Universidade Federal do Rio de Janeiro (UFRJ), Brazil, and the Dr. degree from Université Pierre et Marie Curie (Paris 6), Paris, France, in 2001. Since August 2004, he has been an associate professor with Poli/COPPE/UFRJ. His major research interests are in the areas of routing, wireless, and vehicular networks.