Journal of Industrial Information Integration 50 (2026) 101052

Contents lists available at ScienceDirect

Journal of Industrial Information Integration

journal homepage: www.elsevier.com/locate/jii

Full length article

A thorough assessment of the non-IID data impact in federated learning

Daniel M. Jimenez-Gutierrez *, Mehrdad Hassanzadeh, Aris Anagnostopoulos, Ioannis Chatzigiannakis, Andrea Vitaletti

Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy

ARTICLE INFO

Keywords:
Federated learning
Machine learning
Non-IID data
Data heterogeneity quantification
Spatiotemporal skew

ABSTRACT

Federated learning (FL) allows collaborative machine learning (ML) model training among decentralized clients' information, ensuring data privacy. The decentralized nature of FL deals with non-independent and identically distributed (non-IID) data. This open problem has notable consequences, such as decreased model performance and longer convergence times. Despite its importance, experimental studies systematically addressing all types of data heterogeneity (a.k.a. non-IIDness) remain scarce. This paper aims to fill this gap by assessing and quantifying the non-IID effect through an empirical analysis. We use the Hellinger Distance (HD) to measure differences in distribution among clients. Our study benchmarks five state-of-the-art strategies for handling non-IID data, including label, feature, quantity, and spatiotemporal skews, under realistic and controlled conditions. This is the first comprehensive analysis of the spatiotemporal skew effect in FL. Our findings highlight the significant impact of label and spatiotemporal skew non-IID types on FL model performance, with notable performance drops occurring at specific HD thresholds. The FL performance is also heavily affected, mainly when the non-IIDness is extreme. Thus, we provide recommendations for FL research to tackle data heterogeneity effectively. Our work represents the most extensive examination of non-IIDness in FL, offering a robust foundation for future research.

1. Introduction

In today's digital age, the interaction of machine learning (ML) and healthcare or financial data holds immense promise for improving disease diagnosis [1] and combating financial crimes [2]. These advancements have traditionally relied on centralized learning (CL), such as Machine Learning as a Service (MLaaS) platforms — including AWS, Azure, and Google Cloud [3] — where raw data is aggregated on a central server for model training. However, it raises critical questions about data privacy when dealing with sensitive information from hospitals or banks. In this context, trusted research environments emerge as a mechanism for balancing ML research and protecting individual privacy [4,5].

Federated learning (FL) [6] has emerged as a transformative approach for training ML models across decentralized data sources, preserving data privacy and security. This paradigm is particularly beneficial in cross-silo settings, where entities such as hospitals, banks, and other organizations collaborate without sharing sensitive data. However, a significant challenge inherent in FL is the variation in data distributions across clients, referred to as non-IID (Non-Independent and Identically Distributed) data. This non-IID data (i.e., non-IIDness, data heterogeneity) hinders model performance and convergence during training [7,8]. Such non-IID data is classified into four categories: label, feature, quantity, and spatiotemporal skew [9].

Spatiotemporal skew presents unique challenges that are particularly critical yet underexplored in FL research. This skew occurs when data distributions vary across both geographical locations (spatial) and periods (temporal) [10]. Such skew fundamentally differs from the label, feature, or quantity skew by introducing dynamic variations that standard FL aggregation algorithms often fail to address [11]. Understanding spatiotemporal skew is crucial because it directly impacts the model's ability to generalize across diverse real-world environments while maintaining temporal relevance, making it a key frontier for robust FL systems.

Furthermore, diagnosing and quantifying the level of non-IID data in FL is a significant challenge, as emphasized by Pei et al. [12] and Li et al. [13], who identify critical research directions in this domain. Numerous studies have introduced metrics to quantify the level of non-IID data in FL [14–17], with the Hellinger Distance (HD) [18] emerging as one of the most reliable options. HD = 0.0 corresponds to fully IID data, while higher values (e.g., 0.25, 0.5, 0.75, 0.9) represent increasing

* Corresponding author.
E-mail addresses: jimenezgutierrez@diag.uniroma1.it (D.M. Jimenez-Gutierrez), hassanzadeh.1961575@studenti.uniroma1.it (M. Hassanzadeh), aris@diag.uniroma1.it (A. Anagnostopoulos), ichatz@diag.uniroma1.it (I. Chatzigiannakis), vitaletti@diag.uniroma1.it (A. Vitaletti).

https://doi.org/10.1016/j.jii.2025.101052
Received 16 June 2025; Received in revised form 7 November 2025; Accepted 23 December 2025
Available online 27 December 2025

degrees of non-IID data, with HD = 0.9 approaching the most non-IID scenario considered in our study. HD offers a fine-grained measurement of distribution differences, achieving values close to 1 under extreme non-IID conditions, unlike the Jensen–Shannon Distance (JSD), which tends to plateau at lower levels. Furthermore, HD is versatile and applicable across various types of skews.

Motivation. Recent advances in FL research have significantly advanced our understanding of non-IID data challenges, with notable progress in addressing isolated aspects of heterogeneity such as label skew [19–21]. Based on this foundation, there is now a timely opportunity to unify these insights through comprehensive empirical benchmarks that span the full spectrum of non-IID skews. While pioneering theoretical frameworks [12,22] have established critical mitigation principles, translating these into practice requires systematic quantification of how diverse non-IID data types — from feature skew to spatiotemporal drift — affect real-world FL performance metrics. Closing this knowledge gap through rigorous experimental analysis will empower the community to develop FL systems that are theoretically sound and empirically robust across application domains.

Our study addresses this gap by using the HD to quantify differences in client data distributions, enabling a rigorous empirical analysis of non-IID effects across multiple dimensions, including label, feature, quantity, and spatiotemporal skews. Throughout the assessment of spatiotemporal skew, we capture the impact of dynamic data shifts over time and space, which are particularly relevant in real-world applications such as banking credit risk [23] and personalized healthcare [24]. This approach ensures that our conclusions are robust and generalizable across diverse scenarios.

Contribution. The subsequent points encapsulate the contributions of our study:

1. We benchmark five of the most employed state-of-the-art aggregation and client selection of FL algorithms to tackle non-IID data distributions among clients under realistic, controlled, and quantifiable methods for synthetic data partitioning and all non-IID types (label, feature, quantity, and spatiotemporal skews). Previous empirical works have focused on label skew [19–21] or in combinations of label, feature, and quantity skew [25] (see Section 2.1 for more details). Thus, this is the first study empirically analyzing how the spatiotemporal skew affects the performance of FL models.
2. We motivate using HD to quantify the differences among data distributions, standardizing the guidelines for systematic studies of non-IID data in FL. This is the first work to demonstrate that the effect of the non-IID data is not the same under all levels of heterogeneity. We use HD to quantify differences in distribution as it provides more granular information, and we leave as future work the exploration of other measures; see our section on design insights and opportunities.
3. We provide a reference to researchers about which methods are robust to which kind of non-IID data on highly benchmarked datasets.
4. We give highlights and relevant recommendations for FL researchers based on quantifying the level of non-IID data.

To the best of our knowledge, this is the most comprehensive and complete empirical study of non-IID data and its effects on FL models.

Positioning within industrial information integration. This study contributes to the Journal of Industrial Information Integration's focus on industrial non-IID data and privacy-preserving analytics by thoroughly evaluating non-IID data effects in FL. FL is increasingly adopted in industrial domains such as healthcare, intrusion detection in IoT, and digital twins for industrial IoT, where data is distributed across multiple silos and devices with inherent heterogeneity. Prior works in this journal have addressed related challenges, including emotion recognition based on electroencephalography (EEG) as a crucial research area in the Internet of Medical Things (IoMT) [26], federated ensemble model for intrusion detection in distributed IoT networks for enhancing cybersecurity [27], and adaptive optimization for FL enabled digital twins in industrial IoT [28].

Our work advances these efforts by systematically quantifying the impact of diverse non-IID data types on FL model performance using the HD metric. Furthermore, we benchmark state-of-the-art aggregation and client selection algorithms, offering practical guidance for deploying FL in industrial scenarios characterized by complex data distributions. This positioning situates our research within the ongoing scholarly conversation on industrial information integration and FL, underscoring its significance for robust, privacy-aware industrial analytics.

Ethical considerations in FL:. FL inherently aligns with ethical principles related to data minimization and user privacy, as it allows individual clients to retain their raw data locally. However, despite these benefits, FL is not immune to ethical concerns. Potential privacy risks remain due to model inversion or gradient leakage attacks [29], and there is a need for transparency and informed consent when deploying FL in real-world applications [30]. Future work must integrate privacy-preserving mechanisms (e.g., differential privacy [31], secure aggregation [32]) and conduct rigorous audits to ensure ethical compliance in decentralized learning scenarios [33].

2. Related work

In this section, we present recent studies that empirically evaluate the behavior of non-IID data in FL. Additionally, we compare our work with relevant surveys on non-IID data in FL.

2.1. Empirical studies

Studies that analyze and benchmark the performance of methods to tackle the non-IID data effect on the FL models under controlled and systematic scenarios are scarce. Nevertheless, in this section, we introduce those works that, to some extent, provide empirical analysis about the repercussions of non-IID data.

A study by Vahidian et al. [19] challenges conventional thinking regarding non-IID data in FL. They posit that dissimilar data among participants is not always detrimental and can be advantageous, and we found similar results. Their argument centers on two main points: firstly, that differences in labels (label skew) are not the sole determinant of non-IID data, and secondly, that a more effective measure of heterogeneity is the angle between the data subspaces of participating clients. Complementary, *we encompass a broader spectrum of non-IID data types and include images and tabular data*.

Wong et al. [20] conduct extensive experiments on a large network of IoT and edge devices to present FL real-world characteristics, including learning performance and operation (computation and communication) costs. Moreover, they mainly concentrate on heterogeneous scenarios, the most challenging issue of FL. While they thoroughly analyze the impact of non-IID data, the focus is primarily on image datasets, and they do not explore comparisons of aggregation algorithms to address highly heterogeneous scenarios. In contrast, *our work expands on this by incorporating diverse datasets and benchmarking state-of-the-art methodologies to tackle non-IID data in FL effectively*, offering a broader and more practical perspective.

In their study, Mora et al. [21] examine existing solutions in the literature to mitigate the challenges posed by non-IID data. On the one hand, they emphasize the underlying rationale behind these alternative strategies and discuss their potential limitations. On the other hand, they identify the most promising approaches based on empirical results and critical defining characteristics, such as any assumptions made by each strategy. In addition, they focused on label skew and considered

Table 1
Comparison against surveys (resources) for non-IID data in FL (✓: Included, −: Partially included, ✗: Not included).

Resource	Publication year	Label skew	Feature skew	Quantity skew	Spatiotemporal skew	Non-IID data Quantification	Empirical highlights
[22]	2024	✓	✗	✗	✗	✗	✗
[12]	2024	✓	✓	✓	✗	−	✗
[34]	2022	✓	✓	✓	✗	✗	✗
[10]	2022	−	✗	✗	✓	✗	✗
[9]	2021	✓	✓	✓	−	✗	✗
Ours	2025	✓	✓	✓	✓	✓	✓

one dataset in their experiments. In our paper, *we alternatively analyze broader datasets and data skewness types*, identifying limitations and potential approaches to overcome them.

Li et al. [25] conduct a comprehensive experimental evaluation of FL aggregation algorithms under non-IID data settings. They systematically analyze the strengths and limitations of state-of-the-art FL aggregation algorithms while introducing diverse data partitioning methods to simulate various non-IID scenarios. Their work highlights the challenges posed by non-IID data, such as accuracy degradation and training instability, and provides empirical insights into the performance of each algorithm across different settings. In our paper, *we build upon their work by analyzing spatiotemporal skew and introducing metrics to quantify its effects*, offering a broader perspective on non-IID data and its impact on model performance.

2.2. Surveys

Some surveys that explore the effects of non-IID data in the model performance have been proposed, and they are depicted in Table 1. Earlier works, such as those from 2024, focus on label skew, providing valuable insights into this dimension. Resources from 2022 partially mention spatiotemporal skew, contributing to a deeper understanding of these aspects. The 2021 study offers an essential foundation by addressing label skew, paving the way for more comprehensive analyses in subsequent years.

Compared to the previous surveys summarized in Table 1, our work builds upon and extends their contributions by offering a more comprehensive approach. *We include empirical evaluations of the effects of non-IID data, provide quantification of the non-IID data level, and conduct extensive experiments to assess the impact of spatiotemporal skew thoroughly*. These additions enhance our study's depth and practical relevance, complementing earlier works.

3. Background

This section provides an overview of FL, its fundamental training process, and the challenges posed by non-IID data. We categorize different types of data skew that impact FL performance and introduce methods for quantifying the level of non-IID data. Additionally, we review state-of-the-art aggregation and client selection strategies designed to address these challenges, such as FedAvg [6], FedProx [13], Random size-proportional selection (Rand) [35], Power-Of-Choice (POC) [35], and Model Contrastive Learning (MOON) [36].

3.1. Basics of FL

FL [6] suits siloed data (a.k.a. clients, local nodes, parties, participants) where multiple organizations or institutions hold their datasets. This decentralized approach enables collaborative model training without sharing the raw data, contributing to data privacy and ownership for each participating organization [37].

Fig. 1 presents an overview of the cross-silo FL training process, where participating clients train local models (M_i) on their datasets

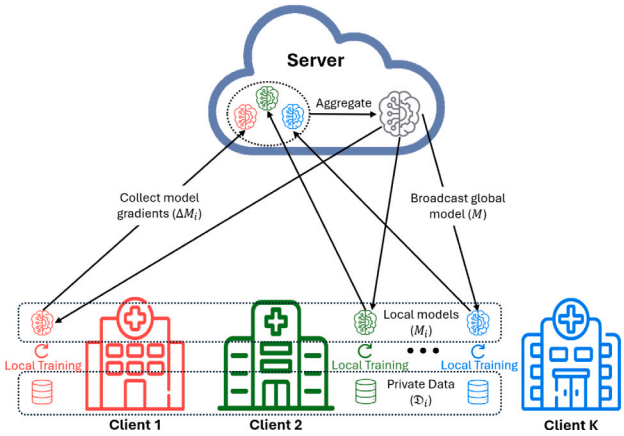

Fig. 1. FL training process overview.

(D_i), all based on a pre-distributed global model (M) [38,39]. Instead of sharing raw data, clients exchange model updates (M_i) without sensitive information, aggregating centrally to enhance the global model. By utilizing a central server for coordination, clients transmit their updates, aggregated to improve the global model. This iterative process allows collaboration without compromising data privacy, as each client receives the updated global model without sharing raw data, ensuring data privacy while facilitating collaboration.

3.2. Data skew types.

A *centralized dataset*[1] $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, is a collection of n tuples where $\mathbf{x}_i = [(x_i)_1, \ldots, (x_i)_m]$ is the feature representation of the ith element (sample) in the dataset, and $y_i \in \{1, \ldots, \ell\}$ is the (true) label of the ith element.

In the FL setting, the dataset $\mathcal{D}$ is distributed over K clients. We let $\mathcal{D}_i$ be the set of elements of the ith client. That is:

$$\mathcal{D} = \cup_{i=1}^{K} \mathcal{D}_i \qquad \text{and for } i \neq j: \qquad \mathcal{D}_i \cap \mathcal{D}_j = \emptyset.$$

Defining the type of non-IID data in FL is relevant since it can drastically influence the performance of the models. We follow the settings of previous work [9,40]. For a supervised learning task on client i (local node i), we assume that each data sample $(x, y) \in \mathcal{D}_i$, where x is the input attributes or features, and y is the label, following a local distribution $P_i(\mathbf{x}, y)$. Let us define:

$$P_i^Y(y) = \sum_{\substack{(\mathbf{x},z)\in\mathcal{D}_i \\ z=y}} P_i(\mathbf{x}, z) \quad \text{and} \quad P_i^{X_\ell}(x) = \sum_{\substack{(\mathbf{x},y)\in\mathcal{D}_i \\ x_\ell=x}} P_i(\mathbf{x}, y) \tag{1}$$

[1] Notice that this definition of a centralized dataset includes tabular data, images, medical data, and graph data, and any dataset expressible as a collection of arrays.

with $P_i^Y(y)$, the ith client labels' distribution and $P_i^{X_\ell}(x)$ the distribution over the ℓth input feature of the ith client. Then, the classification for non-IID data (i.e., data skew types) is as follows:

- Regarding the concept of *identically distributed*:
  1. **Label skew**: Means that the label distribution $P_i^Y(y)$ of different clients is different.
  2. **Feature skew**: Occurs when the distribution of the features $P_i^{X_\ell}(x)$ varies from client to client.
  3. **Quantity skew**: Refers to the significant difference in the number of examples of different client data $P_i(\mathbf{x}, y)$.
- Regarding the concept of *independent*:
  4. **Spatiotemporal skew**: Also known as spatial–temporal skew under federated continual learning (FCL) [41–43]. It refers to the inner correlation of data in the time (or space) domain. In other words, the distribution $P_i(\mathbf{x}, y)$ is not stationary but depends on time or space.

3.3. Quantifying the degree of non-IID data.

Regarding selecting valuable scenarios to demonstrate the effects of non-IID data in FL, the current literature often relies on ad-hoc partitions [44–46]. Therefore, in this work, we use a *metric that systematically evaluates the level of non-IID data to select scenarios for measuring the effect of non-IID data*. We opted for the Hellinger Distance (HD), a metric widely used to gauge the separation between two probability distributions calculated as in Eq. (2) [18].

$$\mathrm{HD}(P_1^Y(y), P_2^Y(y)) = \frac{1}{\sqrt{2}}\sqrt{\sum_{y\in Y}\left(\sqrt{P_1^Y(y)} - \sqrt{P_2^Y(y)}\right)^2} \tag{2}$$

The HD provides a fine-grained and sensitive measurement of distributional differences, reaching values close to 1 under extreme non-IID conditions, unlike the JSD, which often saturates and fails to capture high skew levels. In contrast to the Earth Mover's Distance (EMD) [14], whose values depend heavily on the choice and scale of the ground distance, HD offers normalized and consistent comparisons across tasks and datasets, making it particularly suitable for FL scenarios. The ablation study in Jimenez et al. [14] empirically demonstrated that HD can capture subtle variations between clients' data distributions even under moderate non-IID data. Through its continuous quantification of distributional divergence, HD enables a precise assessment of how the level of non-IID data evolves and impacts model performance, offering a detailed understanding of how diverse forms of heterogeneity, across labels, features, quantities, and spatiotemporal aspects, affect convergence and generalization.

3.4. Aggregation and client selection algorithms

In an FL process, the server aggregates the weights obtained from each client and communicate them back to each participant.

In this section, we explain the five state-of-the-art aggregation and client-selection algorithms assessed in our experiments.

FedAvg: It is a fundamental algorithm in FL [6] designed to train ML models across a network of decentralized devices while preserving data privacy. In FedAvg, each client computes model updates and sends them to a central server using local data. The server averages these updates to calculate a global model update and then sends it back to the clients. This process iterates until it converges (the model's performance gets stable). FedAvg suffers from three key limitations: (1) degraded performance under non-IID data distributions across clients, (2) inefficient communication rounds caused by straggling devices or disproportionate local dataset sizes, and (3) a uniform aggregation approach that fails to account for variability in client data quality or device reliability, potentially biasing the global model.

FedProx: It is a framework designed to address non-IID data in FL [13], offering a generalized and reparametrized version of FedAvg. This approach incorporates a regularization term (μ) to minimize the difference between local and global weights. Finally, the framework aggregates local model updates from all devices to obtain an updated global model. Using this proximal term, FedProx aims to improve convergence and performance in heterogeneous federated learning environments. FedProx ensures convergence even with non-IID data while requiring only minor adjustments to implementation. However, it has notable drawbacks: (1) its effectiveness depends significantly on careful hyperparameter selection, and incorrect settings may slow convergence or raise communication overhead; (2) in real-world applications, its advantages weaken under extreme levels of non-IID data, particularly when client datasets contain completely distinct classes. [25].

Rand: Rand is a baseline client selection strategy introduced to handle non-IID data that is not biased toward clients with higher local losses. Most current analysis frameworks consider a scheme that selects the training set of clients $S(t)$ by sampling m clients randomly (with replacement) such that client k gets selected with probability p_k, the fraction of data at that client [47]. Rand provides unbiased client selection but suffers from key limitations: (1) it fails to prioritize clients with informative updates, hindering convergence speed; and (2) in highly non-IID settings, it may underrepresent rare data distributions, reducing model generalization.

POC: The POC algorithm performs well under a non-IID distribution. It is inspired by the power of d choices load balancing strategy, which queueing systems commonly use [35,47]. The central server first samples a candidate set of d clients, where d is between m (the number of clients to be selected) and K. These candidates are chosen based on their data fraction (p_k). The server then sends the current global model to these candidates, who compute and return their local losses. Finally, the server selects m clients with the highest losses to participate in the next training round. This approach aims to balance the workload and prioritize clients with more informative updates, improving the efficiency of the FL process. While this approach enhances training efficiency and manages non-IID data effectively, it also has limitations: (1) Selection bias can marginalize underrepresented clients, thereby weakening the model's generalization ability. (2) The method introduces higher complexity and greater communication costs in the client selection process.

MOON: It is a simple and effective FL framework designed to tackle non-IID data. It uses the similarity between model representations to correct the local training of individual parties (i.e., conducting contrastive learning at the model level). The network proposed in MOON has three components: a base encoder, a projection head, and an output layer. The base encoder extracts representation vectors from inputs. Le et al. [36] introduce an additional projection head to map the representation to a space with a fixed dimension. Last, the output layer produces predicted values for each class. For ease of presentation, with model weight w, they use $F_w(\cdot)$ to denote the whole network and $R_w(\cdot)$ to denote the network before the output layer (i.e., $R_w(X_\ell)$ is the mapped representation vector of input X_ℓ). Like the previous methods, this approach is effective but has drawbacks: (1) higher computational costs due to generating and comparing augmented data views, and (2) restricted use for non-visual data (e.g., text or time-series), where creating meaningful augmentations is difficult because of text context-dependence and time-series structural limitations. [48]

3.5. Models used in FL

In FL, the choice of model architecture plays a critical role in determining both performance and communication efficiency across

Table 2
Characteristics of the datasets.

Dataset	Type	#training examples	#test examples	#features	#classes	Classes distribution
CIFAR10	Images	50,000	10,000	3072	10	Balanced
FMNIST	Images	60,000	10,000	784	10	Balanced
CIFAR100	Images	50,000	10,000	3072	100	Balanced
Physionet	Tabular	39,895	2095	120	27	Balanced
Covtype	Tabular	522,910	58,102	54	7	Unbalanced
Serengeti	Tabular	257,927	28,659	64	13	Unbalanced
5G NTF	Tabular	74,838	13,207	7	12	Unbalanced
MHEALTH	Tabular	851,021	364,724	14	13	Unbalanced

distributed clients. Depending on the nature of the data and the target task, different types of models may be employed to balance expressiveness, computational cost, and generalizability [49]. Below, we outline several common model types used in FL, highlighting their core characteristics and suitability for decentralized training environments.

- *Deep Neural Networks (DNNs)*: DNNs are feedforward networks with multiple hidden layers, capable of learning complex patterns through hierarchical feature extraction [50]. They serve as foundational models in FL due to their flexibility.
- *Convolutional Neural Networks (CNNs)*: CNNs specialize in processing grid-like data (e.g., images) using convolutional layers for local feature detection, pooling for dimensionality reduction, and fully connected layers for classification [51]. Their parameter-sharing property makes them efficient for FL tasks.
- *Transfer Learning Models*: Pre-trained architectures like ResNet9 [52], EfficientNetB0 [53], and MobileNetV2 [54] leverage transfer learning by adapting learned features from large datasets (e.g., ImageNet) to new tasks with limited data [55]. In FL, such models reduce communication overhead and improve convergence by starting from robust initial weights.

4. Experimentation setup

Datasets. This work considers eight widely employed real datasets to train the centralized learning (CL) and FL models. Four of these, i.e., CIFAR10 [56], FMNIST [57], Physionet 2020 [58], and Covtype [59] serve to simulate label, feature, and quantity skew. To further examine label skew in scenarios with a considerably larger number of classes, we also included CIFAR100 [60]. The remaining three datasets, i.e., 5G Network Traffic flows [61], MHEALTH [62], and Snapshot Serengeti [63], are used to simulate spatiotemporal skew. Table 2 provides an overview of the main characteristics of each dataset.

Models. We adopt a well-studied CNN broadly applied in computer vision [64] for the CIFAR10 and FMNIST datasets. It includes one input layer and three convolutional blocks, where the first two blocks each have a convolutional layer followed by a max pooling layer, and the final block contains a convolutional layer and a flattened layer. The initial convolutional layer has 32 filters, whereas the subsequent two layers each have 64 filters with a 3×3 filter size and ReLu as the activation function. In the dense section of the network, there is one dense layer with 64 neurons using ReLU as the activation function. We utilized a ResNet9 for the CIFAR100 dataset. Additionally, we employed in our tests the transfer learning models EfficientNetB0 and MobileNetV2 since they produce higher classification power results for the datasets studied.

For the tabular datasets, we use a DNN, selected because it is widely employed in classification tasks with tabular data [65]. It comprises one input layer, three hidden layers, and one output layer [66]. The input layer uses as many units as the number of features in the training set. The three layers contain 500 hidden units each, and the last layer is formed by considering the neurons equal to the number of classes to predict. Additionally, the hidden layers used the ReLu activation function, and the output layer used a SoftMax function. We use Adam as our optimizer with a learning rate of 0.001 for $K = 30$ clients and a batch size of 64. In our simulations, models were trained for 40 communication rounds and 10 local epochs, except for those using the MOON aggregation algorithm and those involving the CIFAR100 dataset, which were trained for 100 communication rounds to ensure convergence in those settings. To ensure reproducibility and statistical validity, we executed all experiments across the datasets using five and ten distinct data partitions generated from fixed random seeds.

4.1. Hyperparameters tuning

For a fair comparison, we base our hyperparameter grids on the best-performing hyperparameters presented in the original papers as follows:

- **FedAvg**: We do not set any specific tuning process for this algorithm [6].
- **Rand**: The fraction of clients considered in each communication round gets fine-tuned from $\{0.3, 0.5, 0.7\}$ [47].
- **FedProx**: The μ parameter gets fine-tuned from {0, 0.001, 0.01, 0.1, 1, 10, 100} [13].
- **POC**: The parameter C is equal to 0.5. The parameter d gets fine-tuned from $\{15, 18, 19, 21\}$ [47].
- **MOON**: The μ is tuned from the grid of $\{0.1, 1, 5, 10\}$, and we find the best μ of 0.1, and we set the value of *temperature* to 0.5 [36].

4.2. Hardware specification

We used an Ubuntu 22.04.4 LTS machine with 200 GB of disk, Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50 GHz processor, 16 processors, 125 GB of RAM, and Python 3.10.12 to run the experiments. The FL models were trained using the Flower [67] platform.

4.3. Performance metrics

This subsection describes the performance and convergence metrics considered in our experiments and a justification for their use.

Accuracy [19–21]. It refers to the proportion of correctly classified instances compared to the total data size. Higher values of accuracy indicate a better model performance. It can be calculated as follows:

$$Acc = \frac{\sum_{k=1}^{K} C_k}{\sum_{k=1}^{K} n_k} \tag{3}$$

where C_k indicates the number of correctly classified samples on client k and n_k is the number of data samples on client k. We performed five and ten independent trials using different random seeds for the experiments. To ensure robust and reliable results, we report the mean accuracy and the standard deviation across these trials, providing a comprehensive view of the model's performance variability.

Curvature [68,69]. We incorporate curvature as a metric to identify points along the accuracy curve with respect to HD where model performance changes considerably. Given a parametric curve $\alpha(t) = (x(t), y(t))$, where $x(t)$ denotes the level of non-IID data quantified by HD and $y(t)$ the corresponding model accuracy, the curvature $\kappa(t)$ at that point is defined as:

$$\kappa(t) = \frac{x'(t)y''(t) - x''(t)y'(t)}{\left(x'(t)^2 + y'(t)^2\right)^{3/2}} \tag{4}$$

where $x'(\cdot)$ denotes the first-order derivative and $x''(\cdot)$ is the second-order derivative.

Number of times detected as critical point (#Detected as critical point). To identify critical points related to the effects of non-IID data, we count how many times each HD value is detected as critical based on curvature. Specifically, points where the curvature satisfies $\kappa \geq$

1 are considered indicators of sharp performance degradation. This count-based metric highlights HD values where degradation occurs consistently across the different models built by varying the label skew of the clients' data distributions, reflecting the robustness and consistency of a critical point of change across different settings.

Average curvature. This metric captures the overall sharpness of performance change at each HD value by averaging curvature values across different models built under the same non-IID data. By averaging out the curvature values across models, it offers a more robust and reliable estimate of how critical each point truly is. A higher average curvature indicates a sharper and more consistent performance drop, pointing to greater sensitivity or instability at that level of non-IID data. This metric aims to identify the points where performance degradation is not only present but also most pronounced, helping us more effectively evaluate and compare the flagged HD values (0.5 and 0.75) as critical points.

Number of times that performed the best [25]. This metric quantifies how frequently an aggregation algorithm performs better than other approaches across multiple experimental trials. A higher count indicates greater consistency and robustness. This metric helps identify which methods consistently excel across different data partitions, which is particularly valuable in FL, where non-IID data distributions can lead to high-performance variance between trials.

Rounds-to-accuracy (RTA) [70]. This metric measures the minimal number of communication rounds needed for the global model to achieve at least 90% of the maximum accuracy among the aggregation algorithms. It reflects the efficiency of the FL process since lower values indicate a faster model's convergence.

5. Label skew results

This section examines how label skew in the client data affects the models' performance. Consider that the Covtype is an unbalanced dataset regarding the labels, and CIFAR10, FMNIST, CIFAR100, and Physionet are balanced datasets. The accuracy of the models on each CL is our baseline for comparing the accuracy of the models generated in FL.

The results in this section are presented as a step toward *completeness* with respect to three properties: (1) reproducibility across datasets, (2) consistency across aggregation algorithms, and (3) visibility across the non-IID spectrum. A consolidated statement of the criteria and a compact evidence map tying each highlight to its supporting results appear in Section 5.4.

5.1. Synthetic partitioning method

Using the FedArtML tool [14], we employed the Dirichlet distribution (DD) to partition data among clients based on label distribution. The DD generates random numbers summing to one, controlled by the parameter α. Higher α values (e.g., 1000) create similar local distributions, while lower values increase the chance of clients having examples from a single, randomly chosen class [44]. The selected values {1000, 6, 1, 0.3, 0.03} allow us to examine the impact of varying degrees of non-IID data on FL performance. Note that the Dirichlet distribution is the multivariate analogue of the Beta distribution; in particular, $\text{Unif}(0,1) = \text{Beta}(1,1)$. Therefore, the partition of the datasets using DD is a skewed split of the data distribution [71].

We quantify the degree of non-IID data across clients using the HD. Specifically, we partition the data using the DD's α values $\{1000, 6, 1, 0.3, 0.03\}$ to achieve distinct HD levels $\{0.0, 0.25, 0.5, 0.75, 0.9\}$ to cover a representative spectrum of non-IID data ranging from fully IID to highly non-IID partitions. For instance:

- A DD concentration parameter $\alpha = 1000$ yields IID data (HD $\approx$ 0.0), as labels are uniformly distributed.

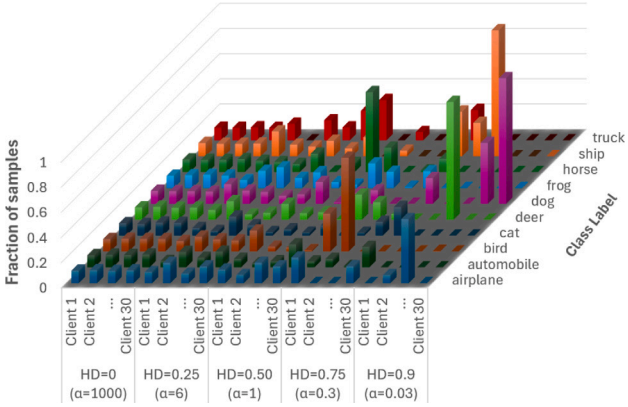

Fig. 2. Distribution of CIFAR10 among 30 clients for different levels of non-IID data. The x-axis shows distinct α values used to partition the data and the resulting HD for clients from 1 to 30. The y-axis shows the participation of each class depicted on the z-axis.

- Conversely, $\alpha = 0.03$ produces highly skewed partitions (HD $\approx$ 0.9).

Thus, each α value maps to a unique HD based on the label distribution, enabling controlled experimentation across non-IID scenarios. Fig. 2 exemplifies the partition distribution for label skew using thirty clients. All ten classes get evenly distributed among every client in the IID scenario ($\alpha = 1000, \text{HD} = 0.0$). As we decrease the α parameter in the DD, the distribution of classes among clients becomes more diverse. In the extreme case of $\alpha = 0.03, \text{HD} = 0.9$, certain classes are absent in some clients.

5.2. Classification power

In this subsection, we focus on the findings from the simulations to compare different aggregation algorithms and datasets regarding their classification power (a.k.a. accuracy).

Highlight 1: The drop in the model's performance for label skew appears in a double threshold. A notable performance decline is immediately evident when the HD exceeds 0.5 and 0.75.

Previous works claim that non-IID data affects the performance of FL models [22,34,72]. Nevertheless, for the first time, we showcase that the effect of non-IID data is not the same under all levels of heterogeneity. Fig. 3 depicts the accuracy change by varying the level of non-IID data distributions among the clients measured by HD concerning the baseline model created in the centralized setting. As the non-IID data partitions increase, the model's accuracy decreases. When the HD between data distributions of the clients exceeds 0.75, the drop becomes more drastic compared to previous levels.

To more precisely characterize the inflection points in model performance, we compute the curvature of the accuracy curves across multiple datasets. As shown in Table 4, curvature values highlight sharper changes at HD = 0.75 and beyond. Notably, models also start to experience a sharper decline in accuracy after HD = 0.5, indicating the beginning of instability. The aggregated metric "#Detected as critical point" shows that HD = 0.75 is identified as a critical point in all five datasets, and the average curvature at this point (3.7) is the highest across all HD values.

One possible explanation for this double-threshold effect is that when HD surpasses 0.5, the model starts experiencing a noticeable decline due to increasing divergence in local distributions, leading to a degradation in the global model's generalization. However, beyond HD = 0.75, the level of heterogeneity may reach a critical point where

Table 3
Mean and standard deviation Accuracy for each dataset for CL, FedAvg, Rand, FedProx, Power-Of-Choice, and MOON, considering different levels of non-IID data as measured by HD for $K = 30$. Each model has undergone ten different trials (random seeds).

Category	Dataset	HD	CL	FedAvg	Rand	FedProx	POC	MOON
Label distribution skew	CIFAR10	0	70.50% ± 0.60%	66.12% ± 0.73%	66.16% ± 0.74%	**66.35% ± 0.72%**	66.26% ± 0.70%	64.45% ± 1.05%
		0.25		**65.91% ± 0.49%**	65.49% ± 0.52%	65.86% ± 0.60%	65.61% ± 0.67%	63.40% ± 0.74%
		0.5		63.41% ± 0.95%	62.93% ± 1.55%	**63.56% ± 0.83%**	62.86% ± 1.71%	60.95% ± 1.33%
		0.75		**58.85% ± 1.06%**	56.80% ± 2.24%	58.80% ± 1.04%	55.84% ± 1.48%	55.27% ± 0.51%
		0.9		43.22% ± 2.24%	40.95% ± 2.43%	**44.33% ± 2.83%**	39.04% ± 2.99%	38.84% ± 2.31%
	FMNIST	0	90.90% ± 0.20%	90.68% ± 0.18%	90.63% ± 0.18%	**90.69% ± 0.15%**	90.62% ± 0.17%	88.70% ± 0.27%
		0.25		90.44% ± 0.14%	**90.52% ± 0.17%**	90.51% ± 0.17%	90.51% ± 0.18%	88.17% ± 0.22%
		0.5		89.92% ± 0.11%	89.84% ± 0.31%	**89.96% ± 0.20%**	89.74% ± 0.35%	87.37% ± 0.22%
		0.75		88.15% ± 0.54%	87.51% ± 0.81%	**88.17% ± 0.47%**	87.23% ± 0.78%	84.70% ± 0.84%
		0.9		80.37% ± 3.78%	79.08% ± 4.67%	**81.10% ± 2.21%**	77.83% ± 3.28%	70.79% ± 5.73%
	CIFAR100	0	67.47% ± 0.46%	62.88% ± 0.28%	62.72% ± 0.35%	**63.05% ± 0.12%**	62.80% ± 0.38%	56.51% ± 0.30%
		0.25		**62.66% ± 0.35%**	62.45% ± 0.25%	62.55% ± 0.32%	62.30% ± 0.21%	56.32% ± 0.49%
		0.5		61.72% ± 0.60%	61.49% ± 0.39%	**61.74% ± 0.26%**	61.23% ± 0.17%	56.47% ± 0.19%
		0.75		59.47% ± 0.37%	58.46% ± 0.73%	**59.72% ± 0.51%**	59.10% ± 0.25%	56.24% ± 0.42%
		0.9		54.38% ± 0.69%	52.87% ± 1.17%	**54.80% ± 0.63%**	52.85% ± 0.99%	51.45% ± 1.18%
	Physionet	0	63.74% ± 1.24%	57.97% ± 0.49%	57.48% ± 0.40%	58.16% ± 0.62%	57.86% ± 0.76%	**61.80% ± 0.62%**
		0.25		57.65% ± 0.47%	57.42% ± 0.54%	57.79% ± 0.55%	57.48% ± 0.53%	**60.94% ± 0.60%**
		0.5		55.69% ± 0.93%	55.26% ± 1.05%	56.29% ± 1.00%	55.24% ± 1.48%	**58.76% ± 0.71%**
		0.75		50.88% ± 1.18%	50.19% ± 2.07%	51.47% ± 1.20%	49.51% ± 2.30%	**53.51% ± 1.37%**
		0.9		41.35% ± 2.70%	39.68% ± 3.01%	41.95% ± 2.07%	38.81% ± 3.68%	**42.49% ± 3.00%**
	Covtype	0	95.60% ± 0.10%	94.95% ± 0.06%	94.89% ± 0.08%	94.96% ± 0.09%	94.84% ± 0.10%	**95.64% ± 0.06%**
		0.25		92.05% ± 1.14%	91.63% ± 1.52%	92.10% ± 1.28%	**93.89% ± 0.73%**	93.36% ± 1.18%
		0.5		84.92% ± 3.71%	83.10% ± 2.79%	85.54% ± 3.39%	**88.21% ± 2.17%**	84.61% ± 3.44%
		0.75		**77.55% ± 3.63%**	76.25% ± 4.10%	77.55% ± 3.64%	74.68% ± 5.31%	57.79% ± 12.66%
		0.9		59.10% ± 8.70%	59.02% ± 9.19%	**59.46% ± 8.74%**	57.29% ± 10.72%	50.51% ± 5.57%
Number of times that performed the best				4	1	12	2	6

Table 4
Curvature values derived from the CNN model's accuracy trends at various levels of non-IID data, highlighting points of sharp change in performance under increasing label skew across five datasets.

Dataset	HD = 0.25	HD = 0.50	HD = 0.60	HD = 0.65	HD = 0.70	HD = 0.75	HD = 0.80	HD = 0.85
CIFAR10	0.2	**0.3**	1.1	0.5	1.6	**3.9**	4.4	2.8
Covtype	0.3	**0.4**	0.2	0.3	0.2	**3.5**	6.6	2.9
FMNIST	0.0	**0.0**	0.5	0.6	0.9	**1.3**	1	3.3
Physionet	0.1	**0.2**	0.5	2.2	1.3	**1**	3.3	3.2
CIFAR100	0.0	**0.2**	0.1	0.1	1.2	**1.5**	2.5	1.8
#Detected as critical point	0	**0**	1	1	3	**5**	5	5
Average curvature	1.2	**0.2**	0.5	0.75	1	**3.7**	3.57	2.8

Table 5
Curvature values derived from the accuracy trends of CNN, EfficientNetB0, and MobileNetV2 models on the CIFAR-10 dataset at various levels of non-IID data, indicating points of sharp performance change.

Model	HD = 0.25	HD = 0.50	HD = 0.60	HD = 0.65	HD = 0.70	HD = 0.75	HD = 0.80	HD = 0.85
CNN	0.2	**0.3**	1.1	0.5	1.6	**3.9**	4.4	2.8
EfficientNetB0	0.2	**0.5**	2.3	3.1	4.7	**3.7**	2.1	0.7
MobileNetV2	0.2	**0.3**	2.0	2.9	5.5	**4.8**	1.1	0.3
#Detected as critical point	0	**0**	3	2	3	**3**	3	1
Average curvature	0.2	**0.4**	1.8	2.2	3.9	**4.1**	2.5	1.3

client models become overly specialized to their local data, significantly reducing the effectiveness of global aggregation. This sharp accuracy drop suggests that at extreme levels of non-IID data, FedAvg struggles to find a well-generalized solution, potentially due to conflicting optimization directions from highly dissimilar client updates.

Highlight 2: Transfer learning models exhibit greater sensitivity to variations in clients' label distributions, with performance degrading more rapidly and sharply as non-IID data increases.

Fig. 4 illustrates the accuracy of models created using three different architectures: the CNN discussed earlier, EfficientNetB0, and MobileNetV2, both of which utilize transfer learning. As HD increases, all models experience performance degradation. Transfer learning models exhibit a more pronounced decline under high non-IID data compared to the CNN which stems from their reliance on frozen feature extractors, limiting adaptation to heterogeneous data. As HD increases, local updates to the final layers create misaligned feature representations, reducing the effectiveness of global aggregation. In contrast, the CNN trained from scratch better adapts to decentralized data, making it more robust in extreme non-IID settings.

Table 5 presents the curvature analysis for the same models. The curvature values for the transfer learning models not only confirm the sharper performance drops but also reveal that these declines begin earlier along the HD spectrum. The peaks in curvature, which quantify the steepness of accuracy degradation, occur sooner and with higher intensity compared to those in Table 4. This suggests that transfer learning models are more sensitive to label distributional shifts over the clients' data. Still, the presence of distinct inflection points after HD = 0.5 and at more prominent at HD = 0.75 further supports the existence of a double threshold pattern in performance degradation under non-IID conditions.

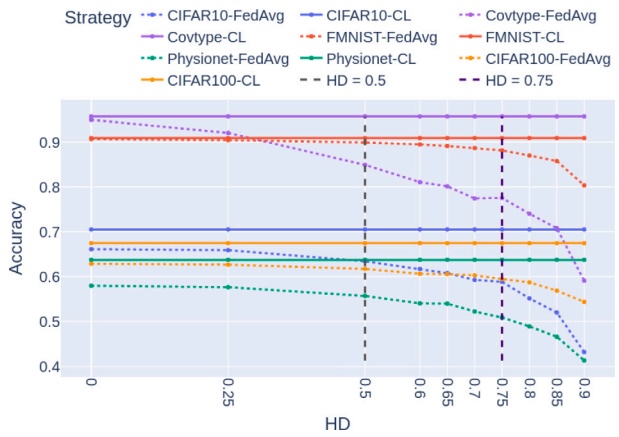

Fig. 3. Changes in the models' accuracy considering different levels of non-IID data measured by HD for $K = 30$.

Fig. 4. Changes in the models' accuracy considering different levels of non-IID data measured by HD for $K = 30$ using CNN, EfficientNetB0, and MobileNetV2 on the CIFAR10 dataset.

Highlight 3: Aggregation algorithms such as Rand, POC, and MOON are particularly vulnerable to performance degradation under conditions of high non-IID data.

Consider the case when the HD is 0.9 (i.e., high non-IID data) for all the datasets presented in Table 3. When comparing the accuracy of Rand and POC versus their corresponding CL performance, those algorithms tend to have a higher drop in performance than the other aggregation algorithms. In FL scenarios with high non-IID data, the distribution of data labels across clients is highly uneven. It means that specific clients may have more or different data types than others. Under such a scenario, Rand and POC may inadvertently select clients with skewed or unrepresentative data distributions, leading to poor generalization performance when aggregating their updates.

MOON has the sharpest decrease in performance when the paradigm switches from CL to FL. MOON's contrastive loss, designed to align local and global representations, becomes ineffective when client distributions are too divergent, as past representations no longer serve as meaningful anchors. This misalignment exacerbates performance degradation, making these methods less suited for extreme non-IID scenarios.

Highlight 4: Unbalanced class datasets experience a sharper drop in performance when moving from IID to highly non-IID settings compared to balanced datasets.

Table 6
The performance's decrease range of the model for the four datasets, moving from the lowest (HD = 0) to the highest (HD = 0.9) levels of non-IID data and considering FedAvg.

Dataset	(HD = 0) - (HD = 0.9)
CIFAR10	22.9%
FMNIST	10.31%
CIFAR100	8.5%
Physionet	16.62%
Covtype	35.85%

Consider the IID case (i.e., HD = 0) and the most extreme non-IID case (i.e., HD = 0.9) for each dataset as depicted in Table 6. Notice that the range of decrease in the reported performance (accuracy of HD = 0 - accuracy of HD = 0.9) is higher for the unbalanced dataset (Covtype) than for balanced datasets (CIFAR10, FMNIST, CIFAR100, and Physionet).

The sharper performance drop in unbalanced datasets under high non-IID data derives from the compounding effects of class imbalance and non-IID data. In such cases, certain classes may be overrepresented in specific clients while being nearly absent in others, leading to biased local models. These biased updates fail to capture the overall class distribution when aggregated, resulting in poor generalization. In contrast, balanced datasets distribute class information more evenly across clients, mitigating this effect and leading to a less severe performance decline.

5.3. Convergence

In this subsection, we focus on the findings obtained using the CIFAR10 dataset to compare different aggregation algorithms regarding their learning process and the smoothness of training.

Highlight 5: The higher the non-IID data level of labels, the more rounds are required to achieve convergence [25].

Table 7 examines the algorithms' convergence from a different perspective. We investigate the performance of each aggregation approach across a certain level of non-IID scenario, independent of the others. For each specific non-IID situation described by HD, we determine how many rounds each aggregation algorithm requires to achieve 90% of its maximum accuracy. Therefore, regardless of the aggregation algorithm, as the non-IID data partitioning over the clients increases, more communication rounds are necessary to reach a convergence point (where the accuracy gets stable). Such behavior aligns with the findings of Li et al. [25].

We observe the mentioned behavior because the data distributions among clients are sufficiently dissimilar, so the models built for each client are only optimal for their data, which diverges from the optimal case. As the training progresses, the weights delivered by the server to the clients improve because they get optimized by considering all of the data across all clients.

Furthermore, FedAvg, Rand, FedProx, and POC exhibit similar behavior in achieving a convergence point, and they do so after roughly the same number of communication rounds. On the other hand, MOON, as expected, requires more rounds to reach the same convergence state.

5.4. Completeness of the label–skew highlights

Establishing the *completeness* of our highlights is essential because these claims are intended to guide practitioners and researchers beyond the specific runs reported here. We consider the label–skew highlights *complete* if each stated pattern is (i) reproducible across multiple datasets — Covtype (unbalanced), CIFAR-10, FMNIST, CIFAR-100, and Physionet — (ii) consistent across aggregation algorithms

Table 7
RTA for FedAvg, Rand, FedProx, POC, and MOON reached in different levels of non-IID cases as determined by HD over CIFAR10 for K = 30.

Category	Dataset	Aggregation algorithm	HD = 0	HD = 0.25	HD = 0.5	HD = 0.75	HD = 0.9
Label distribution skew	CIFAR10	FedAvg	5	5	6	9	15
		Rand	5	5	6	10	14
		FedProx	5	5	6	9	15
		POC	5	5	6	8	15
		MOON	8	8	10	12	20

(FedAvg, Rand, FedProx, POC, MOON), (iii) visible across a representative non-IID grid defined as HD $\in \{0, 0.25, 0.5, 0.75, 0.9\}$ generated via our Dirichlet partitioning, (iv) robust to random seeds (reporting mean±standard deviation), and (v) corroborated by a model-agnostic change-point proxy (curvature κ) rather than accuracy alone.

Demonstrating completeness strengthens both internal validity (the findings are not contingent on a narrow experimental choice) and external validity (the findings are likely to generalize), improves reproducibility, and clarifies the scope and boundary conditions under which our conclusions hold. In turn, this makes the highlights actionable: they can reliably inform system design, algorithm selection, and risk assessment when deploying FL under label skew.

Evidence map. The following list is a *reading guide* that links each stated highlight to the exact figures/tables that substantiate it.

- *Double threshold at HD ≈ 0.5 and 0.75.* The inflection is visible in the per-dataset accuracy trends (Fig. 3 and Table 3) and is quantified by the curvature analysis (Table 5), where the "#Detected as critical point" metric peaks at HD = 0.75 and non-zero detections begin after HD = 0.5 (see also cross-architecture confirmation in Table 5 and Fig. 4).
- *Cross-architecture confirmation.* The same HD≈0.5 and 0.75 pattern appears with alternative model families (e.g., CNN vs. transfer learning), indicating the effect is not tied to a single architecture (see Fig. 4, Table 4).
- *Algorithm sensitivity at high label skew.* Under extreme skew (e.g., HD = 0.9) across datasets in Table 3, Rand and POC tend to suffer larger drops relative to their CL counterparts, and MOON shows the sharpest CL to FL decrease, indicating these approaches are particularly vulnerable in high non-IID settings (discussion following Table 3).
- *Unbalanced datasets are more brittle under high skew.* Comparing IID (HD = 0) to extreme skew (HD = 0.9), the drop is larger for Covtype (unbalanced) than for balanced datasets (Table 6).
- *Convergence slows as label skew increases.* RTA increases monotonically with HD for all five algorithms on CIFAR-10 (Table 7). The same trend holds across the other datasets; for brevity, we report CIFAR-10 as a representative case.

Robustness notes. (A) *Across seeds:* all accuracy tables report mean±standard deviation over repeated trials; the curvature-based critical-point counts (#Detected, with criterion curvature ≥ 1) aggregate over runs and datasets, indicating stable change-points rather than single-seed artifacts. (B) *Across models/algorithms:* the double-threshold pattern appears for CNN and transfer-learning models (Table 4), and the accuracy drop with HD is visible across all five aggregation algorithms (Table 3). (C) *Effect sizes at the thresholds:* the accuracy deltas at HD = 0.5 and 0.75 versus IID are consistently larger than deltas between adjacent lower HD levels; see Table 3 (per-dataset rows) for exact values.

Boundary conditions (scope of validity). Our thresholds are empirical and refer to the non-IID range induced by the Dirichlet partitioning with $\alpha \in \{1000, 6, 1, 0.3, 0.03\}$ mapping to HD $\in \{0, 0.25, 0.5, 0.75, 0.9\}$, the datasets and architectures reported, and the training regimen detailed in Section 4. Within this scope, curvature and accuracy jointly indicate HD ≈ 0.5 (onset) and HD = 0.75 (critical) as consistent change points.

Takeaway. The evidence satisfies our completeness criteria: (i) it reproduces across balanced and unbalanced datasets; (ii) it is consistent across FedAvg, Rand, FedProx, POC, and MOON; (iii) it appears across the full non-IID range HD $\in \{0, 0.25, 0.5, 0.75, 0.9\}$; (iv) it is robust to random seeds (mean±standard deviation reported); and (v) it is corroborated by a curvature-based change-point analysis rather than accuracy alone.

6. Feature skew results

This section examines how feature skew in the client data affects the models' performance.

6.1. Synthetic partitioning method

For simulating feature skew, we employed two diverse methods from FedArtML [14] to test their properties:

Gaussian noise method: This approach introduces diverse Gaussian noise levels to each client's local dataset to achieve diverse feature distributions. Specifically, for each client i, noise levels $\hat{x}$ are added according to the user-defined noise level σ, with $\hat{x} \sim \text{Gau}\left(\sigma \cdot \frac{i}{K}\right)$, where $\hat{x}$ represents the resultant features after applying the noise level to the original features. Here, $\text{Gau}\left(\sigma \cdot \frac{i}{K}\right)$ denotes a Gaussian distribution with a mean of 0 and a variance of $\sigma \cdot \frac{i}{K}$, and K represents the total number of clients.

Hist-Dirichlet-based method: The process starts by characterizing the attributes of each client using other average values and then subjecting them to a binning procedure. Subsequently, it establishes the participation of each feature category within each client using the DD with a specified α. Unlike the Gaussian Noise approach, this method distributes the data among the clients without modifying the features. We measure the non-IID data with the HD among the features across clients (FHD) within the range $\{0, 0.25, 0.5, 0.75, 0.9\}$.

6.2. Classification power

In this subsection, we focus on the findings from the simulations to compare different aggregation techniques and datasets regarding their classification power (a.k.a. accuracy) in the presence of feature skew over the clients' data. Table 8 summarizes the models' performance derived from different aggregation algorithms under varying degrees of non-IID feature distributions, as indicated by FHD.

Highlight 6: The performance of FL-generated models is lower than that of CL-generated models.

Transitioning the training methodology from CL to FL depicts a decline in the model's performance, notably more pronounced in image datasets (CIFAR10) compared to tabular datasets (Covtype). This outcome is predictable since the model gets trained without access to the complete dataset, and each client optimizes the weights based on its available data. The more significant decrease observed in image datasets stems from the heightened complexity inherent in classification tasks compared to tabular datasets.

Table 8
Mean and standard deviation Accuracy for each dataset for CL, FedAvg, Rand, FedProx, Power-Of-Choice, and MOON, considering different levels of feature skewness measured by FHD for K = 30. Each model has undergone ten different trials (random seeds).

Category	Method	Dataset	FHD	CL	FedAvg	Rand	FedProx	POC	MOON
Feature distribution skew	Gaussian Noise	CIFAR10	0	70.50% ± 0.60%	66.12% ± 0.70%	66.16% ± 0.74%	**66.35% ± 0.72%**	66.26% ± 0.70%	64.45% ± 1.05%
			0.35	70.81% ± 0.45%	66.18% ± 0.57%	66.04% ± 0.44%	**66.29% ± 0.63%**	66.24% ± 0.50%	64.23% ± 0.53%
			0.75	70.44% ± 0.29%	66.17% ± 0.44%	66.31% ± 0.61%	**66.36% ± 0.56%**	66.27% ± 0.56%	64.58% ± 0.55%
			0.9	69.71% ± 0.03%	65.81% ± 0.82%	65.89% ± 0.72%	65.85% ± 0.68%	**65.90% ± 0.77%**	63.52% ± 0.66%
		FMNIST	0	90.90% ± 0.20%	90.68% ± 0.18%	90.57% ± 0.14%	**90.69% ± 0.15%**	90.62% ± 0.17%	88.70% ± 0.27%
			0.35	90.91% ± 0.26%	90.69% ± 0.19%	**90.97% ± 0.15%**	90.71% ± 0.17%	90.97% ± 0.17%	88.67% ± 0.28%
			0.75	90.96% ± 0.11%	90.54% ± 0.09%	90.53% ± 0.11%	**90.57% ± 0.23%**	90.51% ± 0.19%	88.64% ± 0.19%
			0.9	90.76% ± 0.13%	90.59% ± 0.09%	**90.71% ± 0.15%**	90.70% ± 0.29%	90.51% ± 0.11%	88.55% ± 0.21%
		Physionet	0	63.74% ± 1.24%	57.97% ± 0.49%	57.48% ± 0.40%	58.16% ± 0.62%	57.86% ± 0.76%	**61.80% ± 0.62%**
			0.35	63.30% ± 1.31%	57.89% ± 0.39%	57.92% ± 0.64%	58.08% ± 0.61%	57.47% ± 1.13%	**61.22% ± 0.72%**
			0.75	60.13% ± 1.11%	52.81% ± 0.83%	52.84% ± 0.74%	52.28% ± 0.38%	51.39% ± 0.25%	**56.10% ± 0.90%**
			0.9	28.97% ± 3.22%	29.43% ± 1.49%	29.88% ± 1.27%	29.43% ± 1.15%	25.25% ± 1.97%	**32.06% ± 1.30%**
		Covtype	0	95.60% ± 0.10%	94.95% ± 0.06%	94.89% ± 0.08%	94.96% ± 0.09%	94.84% ± 0.10%	**95.53% ± 0.04%**
			0.35	95.68% ± 0.10%	94.94% ± 0.06%	94.90% ± 0.04%	94.90% ± 0.08%	94.88% ± 0.08%	**95.65% ± 0.06%**
			0.75	95.53% ± 0.04%	94.79% ± 0.08%	94.62% ± 0.04%	94.74% ± 0.07%	94.68% ± 0.13%	**95.11% ± 0.07%**
			0.9	68.53% ± 1.49%	50.01% ± 0.35%	49.81% ± 0.50%	50.03% ± 0.42%	**50.10% ± 1.67%**	49.20% ± 0.11%
	Hist-Dirichlet	CIFAR10	0	70.50% ± 0.60%	**66.42% ± 0.34%**	66.35% ± 0.35%	66.21% ± 0.59%	66.40% ± 0.70%	64.79% ± 0.32%
			0.25		66.09% ± 0.49%	66.13% ± 0.48%	65.82% ± 0.46%	**66.15% ± 0.53%**	65.12% ± 0.85%
			0.5		66.30% ± 0.67%	66.04% ± 0.66%	66.22% ± 0.40%	**66.52% ± 0.56%**	64.52% ± 1.08%
			0.75		66.23% ± 0.33%	66.04% ± 0.67%	66.17% ± 0.55%	**66.34% ± 0.51%**	64.99% ± 0.31%
			0.9		**66.25% ± 0.47%**	65.25% ± 0.51%	65.25% ± 0.90%	66.17% ± 0.47%	64.12% ± 0.56%
		FMNIST	0	90.90% ± 0.20%	**90.72% ± 0.20%**	90.68% ± 0.12%	90.68% ± 0.15%	90.59% ± 0.22%	88.75% ± 0.13%
			0.25		90.62% ± 0.10%	90.66% ± 0.27%	**90.70% ± 0.15%**	90.62% ± 0.18%	88.72% ± 0.32%
			0.5		**90.78% ± 0.12%**	90.66% ± 0.09%	90.63% ± 0.17%	90.78% ± 0.23%	88.43% ± 0.30%
			0.75		90.53% ± 0.24%	90.67% ± 0.18%	**90.73% ± 0.17%**	90.56% ± 0.12%	88.26% ± 0.30%
			0.9		89.77% ± 0.32%	89.73% ± 0.35%	89.66% ± 0.24%	**89.81% ± 0.29%**	87.38% ± 0.16%
		Physionet	0	63.74% ± 1.24%	57.73% ± 0.72%	57.76% ± 0.51%	58.02% ± 0.76%	57.52% ± 0.68%	**61.13% ± 0.42%**
			0.25		57.67% ± 0.61%	57.62% ± 0.69%	58.16% ± 0.63%	57.05% ± 0.37%	**61.91% ± 0.39%**
			0.5		57.92% ± 0.40%	57.50% ± 0.68%	57.86% ± 0.33%	57.35% ± 0.47%	**61.20% ± 0.83%**
			0.75		57.27% ± 0.64%	57.18% ± 0.97%	57.34% ± 0.88%	56.99% ± 0.86%	**62.11% ± 0.44%**
			0.9		56.47% ± 0.80%	55.49% ± 1.34%	56.49% ± 0.60%	55.80% ± 1.24%	**59.82% ± 1.02%**
		Covtype	0	95.60% ± 0.10%	94.95% ± 0.03%	94.81% ± 0.02%	95.00% ± 0.03%	**98.84% ± 0.09%**	95.62% ± 0.11%
			0.25		94.95% ± 0.05%	94.83% ± 0.09%	94.97% ± 0.09%	94.77% ± 0.03%	**95.63% ± 0.05%**
			0.5		94.90% ± 0.02%	94.88% ± 0.11%	94.90% ± 0.07%	94.88% ± 0.02%	**95.65% ± 0.02%**
			0.75		94.80% ± 0.05%	94.67% ± 0.10%	94.78% ± 0.08%	94.74% ± 0.10%	**95.57% ± 0.07%**
			0.9		93.30% ± 0.42%	93.11% ± 0.52%	93.38% ± 0.31%	93.47% ± 0.33%	**94.51% ± 0.07%**
Number of times that performed the best					4	2	7	7	16

Highlight 7: The model's performance in image datasets remains unaffected by increasing the feature non-IID data [25].

Consider only the image datasets (CIFAR10, FMNIST) and the models generated in FL. Regardless of the aggregation algorithm employed, the performance of the final model remains stable across different levels of feature non-IID data, consistently converging to specific values for each aggregation algorithm. This behavior is consistent with the observations reported by Li et al. [25].

Such a pattern arises from the robustness of convolutional layers, which extract spatial features while suppressing minor pixel variations. In the Gaussian-noise method, small perturbations do not significantly alter key patterns, as convolutional filters average out noise. Deeper layers further aggregate features, preserving essential information and minimizing the impact on performance.

Highlight 8: For tabular datasets, using Gaussian noise levels that exceed FHD=0.9 results in a notable performance decline in the model, emphasizing the acute dissimilarity among samples.

Having tabular datasets (Covtype, Physionet) and using the Hist-Dirichlet approach shows that increasing the degree of feature non-IID data does not impact the performance. However, when dealing with Gaussian noise, if we surpass FHD = 0.9, there is a noticeable decline in performance.

This decline occurs because the data becomes highly dissimilar and noisy, making it difficult for the model to extract meaningful patterns. Even in CL, where data is typically more stable, excessive noise disrupts feature learning, reducing the model's generalization ability and leading to performance degradation.

Highlight 9: In scenarios where the features are non-IID partitioned across clients, MOON performs better than all other aggregation algorithms for tabular datasets.

Table 8 validates that no particular algorithm outperforms others in image datasets, as they yield comparable final performance metrics. MOON emerges as the top-performing algorithm in tabular datasets, surpassing all other algorithms. Its performance is nearly equivalent to that of models trained in CL.

This occurs since MOON can immediately start learning meaningful contrasts between differences in the label using the provided features of the tabular dataset. On the other hand, for images, the model first needs to learn to extract meaningful features from raw pixels before it can start contrasting different object classes effectively. For example, consider the Physionet dataset, which contains features such as age, sex, heart rate, and P–R interval. In that case, each feature has a clear medical interpretation, and the model can directly use the mentioned feature values without learning initial representations. Conversely, for CIFAR10, the model must learn to extract meaningful features from raw pixels before contrastive learning can be effective.

Table 9
RTA for FedAvg, Rand, FedProx, POC, and MOON reached in different levels of feature non-IID cases as determined by FHD over CIFAR10 and Covtype using Hist Dirichlet for K = 30.

Category	Method	Dataset	Aggregation algorithm	FHD = 0	FHD = 0.25	FHD = 0.50	FHD = 0.75	FHD = 0.9
Feature distribution skew	Hist Dirichlet	CIFAR10	FedAvg	5	5	5	5	4
			Rand	5	5	5	5	4
			FedProx	5	5	5	5	4
			POC	5	5	5	5	4
			MOON	8	7	7	8	7
		Covtype	FedAvg	3	3	3	3	4
			Rand	3	3	3	3	4
			FedProx	3	3	3	3	4
			POC	3	3	3	3	4
			MOON	4	4	4	4	5

6.3. Convergence

In this subsection, we concentrate on the results of the simulations, aiming to contrast various aggregation algorithms and datasets in terms of their convergence.

Highlight 10: Feature skew does not alter the model convergence point.

Table 9 presents an alternative viewpoint on the previous highlight. It outlines the iterations needed for FedAvg, Rand, FedProx, POC, and MOON to achieve 90% of the maximum accuracy across various degrees of feature non-IID conditions, as characterized by FHD, using the CIFAR10 dataset. Increasing the non-IID data of features within the data has minimal impact on the model's ability to converge to its optimal performance.

The minimal impact of feature skew on convergence suggests that while feature distributions differ across clients, the underlying task remains learnable. Unlike label skew, which directly affects class representation in local updates, feature skew primarily alters input variations without disrupting the overall decision boundary. As a result, the global model can still generalize effectively across clients, leading to similar convergence behavior regardless of the degree of feature non-IID data.

7. Quantity skew results

This section examines how the quantity skew in the client data affects the models' performance.

7.1. Synthetic partitioning method

We use the MinSize-Dirichlet method included in the FedArtML [14] tool, which specifies the DD's α value and generates the desired participation proportions for each client. Subsequently, a minimum required size, referred to as "the minimum number of examples", is established for each client. Thus, the minimum proportion size, denoted as $MinSize$, is calculated as $MinSize = \frac{MinRequiredSize}{n}$, where n represents the total number of examples in the centralized dataset. If the designated proportions fall below $MinSize$, it substitutes them with $MinSize$. Finally, the proportions are normalized to fall from 0 to 1.

We assess the level of non-IID data using the HD for quantity skew (QHD) within the range $\{0, 0.10, 0.17\}$. This small range arises because the finite size of the dataset constrains quantity skew. Unlike other skews, the proportions derived from the quantity distribution cannot exhibit extreme divergence, as the total number of samples restricts how unevenly clients can receive data.

7.2. Classification power:

In the following paragraphs, we concentrate on the simulation results to evaluate various aggregation algorithms and datasets regarding their classification accuracy, particularly considering the impact of quantity skew on the clients' data.

Highlight 11: The quantity skew in the client's data does not affect the performance of the final model [25].

Table 10 depicts the performance of each aggregation algorithm and dataset, using various levels of non-IID for quantity skew. Examining this table and considering each aggregation algorithm separately, it is evident that the performance of the final models remains consistent across various levels of quantity skewness in the clients' records. This phenomenon occurs regardless of the chosen aggregation algorithm, confirming that quantity skewness does not affect model performance. This behavior pattern agrees with the results documented by Li et al. [25].

The invariance of model performance to quantity skew suggests that FL aggregation algorithms effectively balance updates regardless of varying client sample sizes. Since clients contribute proportionally to the global model, those with fewer samples still provide functional gradients without disproportionately influencing training. Additionally, standard optimization techniques, such as weighted averaging, mitigate potential biases from data imbalance, ensuring stable performance across different levels of quantity skewness.

7.3. Convergence

In this subsection, we focus on the findings obtained when different aggregation algorithms are considered regarding their learning process and the smoothness of training.

Highlight 12: All aggregation algorithms converge after the same number of communication rounds in the presence of quantity skew.

Table 11 shows the RTA for the aggregation algorithms on various levels of quantity non-IID cases for the analyzed datasets. Looking at this table, it is clear that regardless of the degree of non-IID data in the number of records from each label across clients, all aggregation algorithms converge after a consistent number of communication rounds.

The consistent convergence across aggregation algorithms stems from the redundancy in client datasets, where each client's data mirrors the overall distribution. This redundancy allows the global model to learn similar patterns from any subset of clients, ensuring that convergence remains stable regardless of quantity skew.

Table 10
Mean and standard deviation Accuracy for each dataset for CL, FedAvg, Rand, FedProx, Power-Of-Choice, and MOON, considering different levels of non-IID partitioning of the record quantities measured by QHD for $K = 30$. Each model has undergone five different trials.

Category	Method	Dataset	CL	QHD	FedAvg	Rand	FedProx	POC	MOON
Quantity distribution skew	Min-size Dirichlet	CIFAR10	70.50% ± 0.6%	0	65.77% ± 0.57%	65.92% ± 0.72%	**66.45% ± 0.61%**	66.04% ± 0.35%	64.01% ± 0.46%
				0.10	66.69% ± 0.72%	66.29% ± 0.67%	66.71% ± 0.76%	**68.82% ± 0.89%**	63.24% ± 0.37%
				0.17	66.04% ± 0.65%	67.91% ± 0.46%	68.07% ± 0.42%	**68.44% ± 0.51%**	63.49% ± 1.25%
		FMNIST	90.90% ± 0.02%	0	**90.72% ± 0.23%**	90.69% ± 0.25%	90.64% ± 0.15%	90.67% ± 0.11%	88.49% ± 0.24%
				0.10	90.37% ± 0.16%	90.39% ± 0.22%	90.30% ± 0.45%	**90.78% ± 0.13%**	88.04% ± 0.27%
				0.17	90.39% ± 0.32%	90.43% ± 0.19%	90.44% ± 0.27%	**90.65% ± 0.31%**	88.10% ± 0.26%
		Physionet	63.74% ± 1.24%	0	58.23% ± 0.68%	57.13% ± 0.56%	58.04% ± 0.29%	57.08% ± 0.53%	**61.29% ± 0.81%**
				0.10	59.48% ± 2.09%	59.06% ± 1.71%	59.42% ± 1.23%	**61.29% ± 0.50%**	58.67% ± 1.97%
				0.17	64.22% ± 1.27%	64.40% ± 0.55%	**64.92% ± 0.71%**	64.82% ± 0.85%	63.86% ± 0.40%
		Covtype	95.60% ± 0.1%	0	94.97% ± 0.04%	94.87% ± 0.06%	94.96% ± 0.07%	94.83% ± 0.05%	**95.15% ± 0.09%**
				0.10	95.67% ± 0.10%	95.65% ± 0.07%	95.70% ± 0.10%	**95.79% ± 0.10%**	95.13% ± 0.12%
				0.17	90.65% ± 1.57%	94.77% ± 0.43%	**95.27% ± 0.20%**	94.94% ± 0.44%	94.23% ± 0.47%
Number of times that performed the best					1	0	3	6	2

Table 11
RTA for FedAvg, Rand, FedProx, POC, and MOON reached in different levels of quantity non-IID cases as determined by QHD over CIFAR10 and Covtype using Min-size Dirichlet method for $K = 30$.

Category	Method	Dataset	Aggregation algorithm	QHD = 0	QHD = 0.10	QHD = 0.17
Quantity distribution skew	Min-size Dirichlet	CIFAR10	FedAvg	5	3	2
			Rand	5	3	1
			FedProx	5	3	2
			POC	5	3	2
			MOON	6	3	2
		Covtype	FedAvg	3	2	1
			Rand	3	2	1
			FedProx	3	2	1
			POC	3	2	1
			MOON	4	2	1

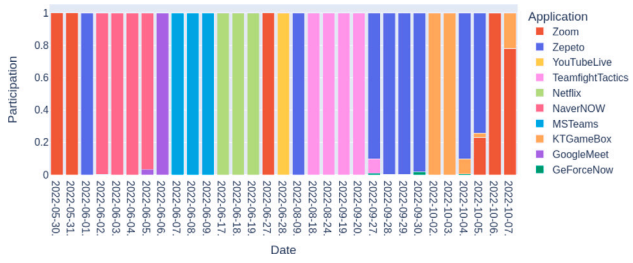

Fig. 5. Distribution of 5GNTF applications (label) along date (spatiotemporal variable expressed in YYYY-MM-DD).

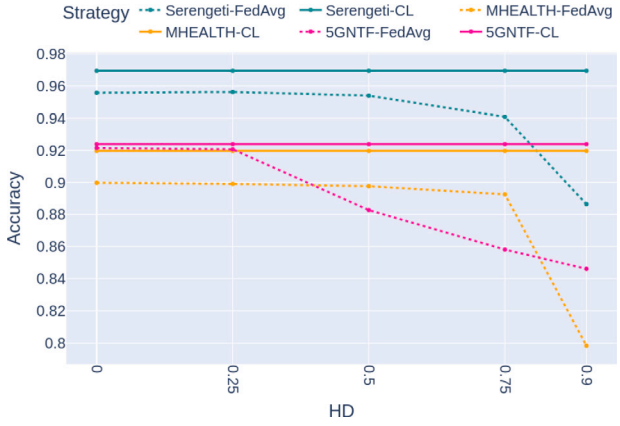

Fig. 6. Changes in the models' accuracy considering different levels of non-IID data measured by STHD for $K = 30$.

8. Spatiotemporal skew results

This section examines how varying levels of data disparity among clients, based on time and location, impact model performance.

8.1. Synthetic partitioning method

The primary constraint in this partition process is that the dataset must contain a categorical variable of space (i.e., places, cities, latitude, longitude, etc.) or time (i.e., hours, months, years, etc.) to use as the partitioning variable. For instance, Fig. 5 depicts the distribution of labels along the date of the 5GNTF dataset. In this case, the space variable employed to create the federated data is the flow's *date* expressed in year-month-day format (categorical).

We use the St-Dirichlet method from FedArtML [14], which employs the DD to segment the data based on spatial (SP skew) or temporal (TMP skew) categories to distribute the data among federated clients. We assess the level of non-IID data using the HD for spatiotemporal skew (STHD) within the range {0, 0.25, 0.5, 0.75, 0.9}.

8.2. Classification power

In this subsection, we focus on the simulation results to evaluate various aggregation algorithms and datasets regarding classification accuracy. We pay particular attention to the impact of different levels of spatiotemporal skewness among the clients' data.

Table 12
Mean and standard deviation Accuracy for each dataset for CL, FedAvg, Rand, FedProx, Power-Of-Choice, and MOON, considering different levels of non-IID partitioning of the records based on their space (SP) and time (TMP) measured by STHD for $K = 30$. Each model has undergone five trials.

Category	Type	Method	Dataset	CL	STHD	FedAvg	Rand	FedProx	POC	MOON
SPT distribution skew	SP skew	St-Dirichlet	Serengeti	96.95% ± 0.11%	0	95.58% ± 0.08%	95.56% ± 0.09%	95.62% ± 0.14	95.48% ± 0.09	**96.69% ± 0.06**
					0.25	95.63% ± 0.05%	95.50% ± 0.10%	95.64% ± 0.03	95.51% ± 0.11	**96.76% ± 0.08**
					0.5	95.40% ± 0.08%	95.32% ± 0.07%	95.36% ± 0.06	95.32% ± 0.06	**96.51% ± 0.04**
					0.75	94.09% ± 0.25%	93.91% ± 0.36%	94.15% ± 0.21	94.06% ± 0.09	**95.80% ± 0.16**
					0.9	88.65% ± 0.18%	87.68% ± 0.58%	88.49% ± 0.13	88.20% ± 0.39	**91.74% ± 0.20**
			MHEALTH	91.97% ± 0.09	0	89.98% ± 0.04%	90.33% ± 0.03	**90.85% ± 0.04**	90.33% ± 0.05	90.56% ± 0.02
					0.25	89.91% ± 0.05	90.34% ± 0.04	**90.85% ± 0.01**	90.33% ± 0.05	90.57% ± 0.02
					0.5	89.77% ± 0.10	90.28% ± 0.04	**90.85% ± 0.02**	90.32% ± 0.04	90.48% ± 0.04
					0.75	89.26% ± 0.31	89.15% ± 0.24	**89.87% ± 0.31**	88.79% ± 0.48	89.69% ± 0.18
					0.90	79.84% ± 0.57	82.34% ± 1.36	82.48% ± 0.78	81.33% ± 0.74	**82.95% ± 1.04**
	TMP skew		5GNTF	92.39% ± 0.10%	0	92.15% ± 0.04%	92.15% ± 0.02%	92.14% ± 0.02%	92.15% ± 0.04%	**92.23% ± 0.02%**
					0.25	92.07% ± 0.12%	92.05% ± 0.15%	92.07% ± 0.18%	92.15% ± 0.05%	**92.28% ± 0.04%**
					0.5	88.28% ± 1.87%	87.92% ± 2.00%	89.19% ± 2.09%	**92.08% ± 0.08%**	89.24% ± 1.80%
					0.75	85.82% ± 0.06%	85.64% ± 0.08%	85.83% ± 0.08%	**91.77% ± 0.32%**	86.51% ± 1.79%
					0.9	84.62% ± 0.36%	84.50% ± 0.20%	84.55% ± 0.37%	**89.23% ± 2.13%**	83.94% ± 0.02%
Number of times that performed the best						0	0	4	3	8

Table 13
HD among clients' label distributions at varying levels of non-IID partitioning by time and space.

Dataset	STHD = 0	STHD = 0.25	STHD = 0.50	STHD = 0.75	STHD = 0.9
Serengeti	0.01	0.09	0.22	0.36	0.53
MHEALTH	0.01	0.01	0.01	0.03	0.07
5GNTF	0.03	0.20	0.29	0.30	0.49

Highlight 13: The higher the non-IID data level in time and space, the worse the model's performance.

Table 12 depicts the performance for each dataset and aggregation algorithms, using multiple levels of non-IID spatiotemporal skew. It demonstrates that, irrespective of the aggregation algorithm employed, model performance deteriorates when data distribution among clients varies concerning time or space. Fig. 6 illustrates the comparison between FedAvg and the centralized model across varying STHD levels, showing the same pattern: model accuracy deteriorates as spatiotemporal non-IID data among clients grows. However, the magnitude of this deterioration differs across datasets. The performance in all datasets drops noticeably once the STHD surpasses 0.75. This phenomenon occurs because increasing the differences among clients' data based on time and location also raises non-IID data in the clients' label distributions. This behavior is evident in Table 13, which displays the HD in label distributions at varying levels of STHD among clients. We also concluded in the label skew study section that higher levels of non-IID data among clients' data distributions negatively impact the performance of the final model.

8.3. Convergence

In this subsection, we showcase the results related to the models' convergence when there are variations in the data concerning time and location, considering the results obtained on the Serengeti, MHEALTH, and 5GNTF datasets.

Highlight 14: Spatial non-IID data shows no consistent impact on convergence rounds; effects vary by dataset dynamics and task difficulty.

According to Table 14, the same level of non-IID data can have a radically different effect depending on the underlying data:

- **Serengeti**: As STHD increases from 0 to 0.90, RTA nearly doubles across all aggregation algorithms (e.g., FedAvg: 6 → 14; POC: 7 → 15). This suggests that data from different sites becomes more heterogeneous, requiring more training rounds for the models to converge.
- **5GNTF**: All aggregation algorithms converge in just one round, despite the STHD. This occurs when the classes are very easy to distinguish and there is a clear gap between the records of one class and those of the others. In such cases, the task becomes trivial, and the time factor has minimal impact on when convergence is reached.
- **MHEALTH**: Across all aggregation algorithms, RTA stays constant—except for MOON, which jumps sharply from 6 to 14 rounds as spatiotemporal non-IID data goes from 0 to 0.90. Since the data come from body-worn sensors and are split by individual subjects, client-specific covariate shifts emerge that particularly undermine representation-based approaches like MOON.

For the MOON algorithm, the gap in RTA between STHD = 0 and STHD = 0.9 varies by dataset — 0 rounds for 5GNTF, 8 for MHEALTH, and 11 for Serengeti — while other aggregation algorithms show no such change. This indicates there is not a one-to-one relationship between STHD and convergence speed; instead, factors like task complexity, temporal patterns, and feature diversity shape the outcome, supporting our earlier point that spatial non-IID data impacts convergence inconsistently, depending on dataset dynamics and problem difficulty.

9. General results

In this section, we provide highlights summarizing the overall results obtained from our experiments, combining the behavior shown before for label, feature, quantity, and spatiotemporal skews.

Highlight 15: Label skew [19,25] and spatiotemporal skew significantly impact the model's performance.

Our experimental analysis reveals that not all forms of non-IID data equally degrade FL performance. Label skew and spatiotemporal skew exhibit the most severe impact. Label skew reduces model accuracy by 10%–40% compared to the CL baseline. Feature and quantity skews show less significant effects (1%–5% accuracy drops). The previous aligns with prior findings [19,25] that label skew disproportionately harms aggregation, as local models overfit to dominant classes.

Table 14
RTA reached for different levels of spatiotemporal non-IID cases as determined by STHD for K = 30.

Dataset	Aggregation algorithm	STHD = 0	STHD = 0.25	STHD = 0.50	STHD = 0.75	STHD = 0.9
Serengeti	FedAvg	6	7	7	10	14
	Rand	7	7	8	10	14
	FedProx	7	7	7	10	14
	POC	7	7	7	10	15
	MOON	8	8	9	12	19
5GNTF	FedAvg	1	1	1	1	1
	Rand	1	1	1	1	1
	FedProx	1	1	1	1	1
	POC	1	1	1	1	1
	MOON	1	1	1	1	1
MHEALTH	FedAvg	2	2	2	3	2
	Rand	2	2	2	3	2
	FedProx	2	2	2	3	2
	POC	2	2	2	4	2
	MOON	6	7	7	13	14

Table 15
The number of cases in which each specific algorithm achieved the best performance for each study.

Study	#Cases	FedAvg	Rand	FedProx	POC	MOON
Label Skew	25	4	1	12	2	6
Feature Skew	36	4	2	7	7	16
Quantity Skew	12	1	0	3	6	2
Spatio Temporal Skew	15	0	0	4	3	8
Total best performance		9	3	26	18	32

Spatiotemporal skew introduces contextual drift (e.g., sensor data varying across locations/times), corrupting the feature space. Similarly to label skew, this cannot be fixed through simple aggregation — our tests show FedAvg suffers 10%–12% higher accuracy loss. The global model fails to perform effectively across all contexts because it averages away crucial environmental patterns unique to specific locations or times.

Highlight 16: FedProx performs better under label skew, POC excels in handling quantity skew, and MOON exhibits greater robustness to feature skew, whereas FedAvg and Rand tend to struggle under high non-IID data levels.

Table 15 summarizes the cases in which each specific algorithm exhibited the best performance compared to other aggregation algorithms across four skewness types considered in our study. In most cases, the FedProx, POC, and MOON aggregation algorithms achieved the best performance, outperforming the simpler FedAvg and Rand algorithms.

Such a superior performance can be attributed to the specific mechanisms to tackle the non-IID data of each aggregation algorithm. FedProx stabilizes training by regulating the influence of the global model on local clients, POC enhances personalization through loss-based selection, and MOON leverages contrastive learning to improve feature representation. These mechanisms enable better adaptation to diverse client distributions, leading to consistently stronger performance across different skewness types.

Although FedProx, POC, and MOON generally outperform FedAvg and Rand, the performance gains are often marginal. This indicates that while these methods offer improvements in handling non-IID data, they do not fully resolve the challenges posed by non-IID data. The relatively small advantage suggests the need for more effective aggregation algorithms to better adapt to diverse client distributions and enhance model performance in FL scenarios.

Table 16
The number of cases in which each aggregation algorithm achieved the best performance for each type of dataset.

Dataset type	#Cases	FedAvg	Rand	FedProx	POC	MOON
Image	39	8	3	19	9	0
Tabular	49	1	0	7	9	32
Total best performance		9	3	26	18	32

Highlight 17: FedProx is more effective on image datasets, while MOON performs better with tabular datasets.

Table 16 examines the best-performing aggregation algorithms from the perspective of the dataset type used for training. It shows that FedProx outperforms all other algorithms on image datasets in nineteen out of thirty-nine cases. In comparison, MOON generally surpasses other algorithms on tabular datasets in thirty-two out of forty-nine cases.

The effectiveness of FedProx on image datasets and MOON on tabular datasets can be attributed to their distinct optimization strategies. FedProx mitigates client drift by stabilizing updates, which is particularly beneficial for complex, high-dimensional image data. In contrast, MOON's contrastive learning framework enhances feature representation, making it more suited for tabular data, where feature relationships play a critical role. These differences explain their varying performance across dataset types.

10. Design insights and opportunities

We provide some design insights and opportunities, intending to help researchers direct their efforts toward solving the effects of non-IID data.

Quantifying the level of non-IID data. . Several works claim that the non-IID data affects the performance of FL models [22,34,72]. Nevertheless, for the first time, we demonstrate that the effect of the non-IID data is not the same under all the levels of heterogeneity (see Fig. 3).

Therefore, it is vital to quantify the non-IID data level in FL. This work uses the HD metric to measure the level of non-IID data. However, we encourage researchers to test different metrics, such as JSD [73], EMD [74], and Total Variation distance [75], among others.

More effective methods to tackle high non-IID data. This work showcases how the state-of-the-art methods to tackle non-IID data (Rand, POC, FedProx, MOON) perform against FedAvg. The conclusion is that no algorithm works better than FedAvg in all the scenarios. Moreover,

the better methods do not greatly improve FedAvg under high non-IID scenarios, as their gain is at most two percentage points. FedAvg remains competitive due to its computational efficiency and adequacy for moderately non-IID data, where its simplicity outperforms complex methods like FedProx or MOON that incur tuning overheads. However, in highly non-IID scenarios, adaptive approaches are essential, revealing a core trade-off between simplicity and adaptability. Thus, optimal algorithm selection depends on the specific non-IID data and system constraints in a given FL deployment.

This phenomenon has also been studied and claimed in the scarce work of empirical analysis of non-IID data and methodologies [19,21,25,76]. Therefore, creating methods to alleviate the effect of high levels of non-IID data appropriately is needed to evolve and preserve FL. This aligns with the open problems reported by Kairouz et al. [77].

Focusing on highly unbalanced data. In our experiments, we claim that the more significant decrease in performance comparing CL and FL occurs in unbalanced datasets since it relates to the challenge of learning from highly skewed and less representative data (see Table 3).

Thus, it is relevant to create solutions to tackle non-IID data by considering the degree of unbalancedness that the labels might have across the clients.

Studying spatiotemporal skew. At the time of writing this work, no analyses or empirical studies about the effect of the spatiotemporal skew on the performance of FL models exist. Thus, for the first time, we produce experiments to understand how different spatiotemporal non-IID data levels affect an FL model's prediction power. The results show (see Table 12) that high levels of space or time skews decrease the performance of the models, more specifically when the HD is higher than 0.75 (severe degree of non-IID data).

Thus, researchers may benchmark techniques to deal with space and time skew in FL [78–80] to determine the behavior under high non-IID data levels.

Methods to compare mixed non-IID data types. Current tools and methods for synthetic partitioning centralized data into federated data [14,46,81–83] focus on simulating one type of non-IID data (label, feature, quantity, spatiotemporal skewness). Nevertheless, a more realistic scenario would be combining two or more types of non-IID data to evaluate the extent to which such mixes can alter the performance of FL models. Therefore, for research in FL purposes, it would be interesting to create methods to partition centralized data into federated clients that permit the control of non-IID data level for two or more data skews simultaneously.

11. Conclusions

This study provides a comprehensive empirical analysis of the non-IID effect in FL. Under controlled conditions, we benchmarked five state-of-the-art strategies for addressing non-IID data distributions, including label, feature, quantity, and spatiotemporal skew, with a particular focus on the relatively unexplored spatiotemporal dimension. We aim to standardize the methodology for studying non-IID data in FL by using HD to quantify data distribution differences. Our findings reveal the significant impact of labels and spatiotemporal skews of non-IID types on FL model performance. We also demonstrate that the model's performance drop appears at a double threshold. When HD is higher than 0.5 and 0.75, higher damage and a steeper decrease in performance slope occur. Moreover, our results suggest that the FL performance is heavily affected, mainly when the degree of non-IID data is extreme. Thus, we offer valuable recommendations for researchers to address non-IID data. This work represents the most thorough examination of non-IID data in FL to date, providing a robust foundation for future research in FL.

CRediT authorship contribution statement

Daniel M. Jimenez-Gutierrez: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **Mehrdad Hassanzadeh:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Conceptualization, Formal analysis. **Aris Anagnostopoulos:** Writing – review & editing, Supervision, Funding acquisition. **Ioannis Chatzigiannakis:** Writing – review & editing, Supervision, Funding acquisition. **Andrea Vitaletti:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Daniel M. Jimenez-Gutierrez was partially supported by PNRR351 TECHNOPOLE – NEXT GEN EU Roma Technopole – Digital Transition, FP2 – Energy transition and digital transition in urban regeneration and construction. Aris Anagnostopoulos was supported by the the PNRR MUR project PE0000013-FAIR, the PNRR MUR project IR0000013-SoBigData.it, and the MUR PRIN project 2022EKNE5K "Learning in Markets and Society." Ioannis Chatzigiannakis was supported by PE07-SERICS (Security and Rights in the Cyberspace) – European Union Next-Generation-EU-PE0000014 (Piano Nazionale di Ripresa e Resilienza – PNRR). Andrea Vitaletti was supported by the project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

Data availability

Data select is publicly available.

References

[1] Anichur Rahman, Tanoy Debnath, Dipanjali Kundu, Md Saikat Islam Khan, Airin Afroj Aishi, Sadia Sazzad, Mohammad Sayduzzaman, Shahab S. Band, Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities, AIMS Public Health 11 (1) (2024) 58–109.

[2] Debidutta Pattnaik, Sougata Ray, Raghu Raman, Applications of artificial intelligence and machine learning in the financial services industry: A bibliometric review, Heliyon (2024).

[3] A.I. Neptune, Best machine learning as a service platforms (MLaaS) that you want to check as a data scientist, 2023, https://neptune.ai/blog/best-machine-learning-as-a-service-platforms-mlaas. (Accessed 14 July 2025).

[4] Mackenzie Graham, Richard Milne, Paige Fitzsimmons, Mark Sheehan, Trust and the goldacre review: why trusted research environments are not about trust, J. Med. Ethics 49 (10) (2023) 670–673.

[5] Peng Zhang, Maged N. Kamel Boulos, Privacy-by-design environments for large-scale health research and federated learning from data, Int. J. Environ. Res. Public Health 19 (19) (2022) 11876.

[6] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.

[7] Yanmeng Wang, Qingjiang Shi, Tsung-Hui Chang, Why batch normalization damage federated learning on non-iid data? IEEE Trans. Neural Netw. Learn. Syst. (2023).

[8] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, Chen Chen, Local learning matters: Rethinking data heterogeneity in federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8397–8406.

[9] Hangyu Zhu, Jinjin Xu, Shiqing Liu, Yaochu Jin, Federated learning on non-IID data: A survey, Neurocomputing 465 (2021) 371–390.

[10] Marcos F. Criado, Fernando E. Casado, Roberto Iglesias, Carlos V. Regueiro, Senén Barro, Non-iid data and continual learning processes in federated learning: A long road ahead, Inf. Fusion 88 (2022) 263–280.

[11] Niklas Babendererde, Moritz Fuchs, Camila Gonzalez, Yuri Tolkach, Anirban Mukhopadhyay, Jointly exploring client drift and catastrophic forgetting in dynamic learning, Sci. Rep. 15 (1) (2025) 5857.

[12] Jiaming Pei, Wenxuan Liu, Jinhai Li, Lukun Wang, Chao Liu, A review of federated learning methods in heterogeneous scenarios, IEEE Trans. Consum. Electron. (2024).

[13] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, Virginia Smith, Federated optimization in heterogeneous networks, Proc. Mach. Learn. Syst. 2 (2020) 429–450.

[14] G. Daniel Mauricio Jimenez, Aris Anagnostopoulos, Ioannis Chatzigiannakis, Andrea Vitaletti, FedArtML: A tool to facilitate the generation of non-IID datasets in a controlled way to support federated learning research, IEEE Access (2024).

[15] Ahmed Elhussein, Gamze Gursoy, A universal metric of dataset similarity for cross-silo federated learning, 2024, arXiv preprint arXiv:2404.18773.

[16] Evgenija S. Novikova, Yang Chen, Aleksej V. Meleshko, Evaluation of data heterogeneity in FL environment, in: 2024 XXVII International Conference on Soft Computing and Measurements, SCM, IEEE, 2024, pp. 344–347.

[17] Yuanli Wang, Joel Wolfrath, Nikhil Sreekumar, Dhruv Kumar, Abhishek Chandra, Accelerated training via device similarity in federated learning, in: Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking, 2021, pp. 31–36.

[18] Roma Goussakov, Hellinger Distance-based Similarity Measures for Recommender Systems (Ph.D. thesis), Umea University, 2020.

[19] Saeed Vahidian, Mahdi Morafah, Mubarak Shah, Bill Lin, Rethinking data heterogeneity in federated learning: Introducing a new notion and standard benchmarks, IEEE Trans. Artif. Intell. (2023).

[20] Kok-Seng Wong, Manh Nguyen-Duc, Khiem Le-Huy, Long Ho-Tuan, Cuong Do-Danh, Danh Le-Phuoc, An empirical study of federated learning on iot-edge devices: Resource allocation and heterogeneity, 2023, arXiv preprint arXiv:2305.19831.

[21] Alessio Mora, Davide Fantini, Paolo Bellavista, Federated learning algorithms with heterogeneous data distributions: An empirical evaluation, in: 2022 IEEE/ACM 7th Symposium on Edge Computing, SEC, IEEE, 2022, pp. 336–341.

[22] Zili Lu, Heng Pan, Yueyue Dai, Xueming Si, Yan Zhang, Federated learning with non-iid data: A survey, IEEE Internet Things J. (2024).

[23] Shuyao Zhang, Jordan Tay, Pedro Baiz, The effects of data imbalance under a federated learning approach for credit risk forecasting, 2024, arXiv preprint arXiv:2401.07234.

[24] Shunxin Guo, Hongsong Wang, Shuxia Lin, Xu Yang, Xin Geng, STHFL: Spatio-temporal heterogeneous federated learning, 2025, arXiv preprint arXiv:2501.05775.

[25] Qinbin Li, Yiqun Diao, Quan Chen, Bingsheng He, Federated learning on non-iid data silos: An experimental study, in: 2022 IEEE 38th International Conference on Data Engineering, ICDE, IEEE, 2022, pp. 965–978.

[26] Weiwei Jiang, Yang Zhang, Haoyu Han, Xiaozhu Liu, Jeonghwan Gwak, Weixi Gu, Achyut Shankar, Carsten Maple, Fuzzy ensemble-based federated learning for EEG-based emotion recognition in internet of medical things, J. Ind. Inf. Integr. (2025) 100789.

[27] Ayushi Chahal, Preeti Gulia, Nasib Singh Gill, Deepti Rani, Design of a federated ensemble model for intrusion detection in distributed IIoT networks for enhancing cybersecurity, J. Ind. Inf. Integr. (2025) 100800.

[28] Wei Yang, Yuan Yang, Wei Xiang, Lei Yuan, Kan Yu, Álvaro Hernández Alonso, Jesús Ureña Ureña, Zhibo Pang, Adaptive optimization federated learning enabled digital twins in industrial IoT, J. Ind. Inf. Integr. 41 (2024) 100645.

[29] Ali Hatamizadeh, Hongxu Yin, Pavlo Molchanov, Andriy Myronenko, Wenqi Li, Prerna Dogra, Andrew Feng, Mona G. Flores, Jan Kautz, Daguang Xu, et al., Do gradient inversion attacks make federated learning unsafe? IEEE Trans. Med. Imaging 42 (7) (2023) 2044–2056.

[30] Pavlos Papadopoulos, Will Abramson, Adam J. Hall, Nikolaos Pitropakis, William J. Buchanan, Privacy and trust redefined in federated machine learning, Mach. Learn. Knowl. Extr. 3 (2) (2021) 333–356.

[31] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q.S. Quek, H. Vincent Poor, Federated learning with differential privacy: Algorithms and performance analysis, IEEE Trans. Inf. Forensics Secur. 15 (2020) 3454–3469.

[32] Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Helen Möllering, Thien Duc Nguyen, Phillip Rieger, Ahmad-Reza Sadeghi, Thomas Schneider, Hossein Yalame, et al., SAFELearn: Secure aggregation for private federated learning, in: 2021 IEEE Security and Privacy Workshops, SPW, IEEE, 2021, pp. 56–62.

[33] Liangqi Yuan, Ziran Wang, Christopher G. Brinton, Digital ethics in federated learning, IEEE Internet Comput. 28 (5) (2024) 66–74.

[34] Xiaodong Ma, Jia Zhu, Zhihao Lin, Shanxuan Chen, Yangjie Qin, A state-of-the-art survey on solving non-IID data in federated learning, Future Gener. Comput. Syst. 135 (2022) 244–258.

[35] Yae Jee Cho, Jianyu Wang, Gauri Joshi, Towards understanding biased client selection in federated learning, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 10351–10375.

[36] Qinbin Li, Bingsheng He, Dawn Song, Model-contrastive federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10713–10722.

[37] Haya Elayan, Moayad Aloqaily, Mohsen Guizani, Deep federated learning for IoT-based decentralized healthcare systems, in: 2021 International Wireless Communications and Mobile Computing, IWCMC, IEEE, 2021, pp. 105–109.

[38] Sadman Sakib, Mostafa M. Fouda, Zubair Md Fadlullah, Khalid Abualsaud, Elias Yaacoub, Mohsen Guizani, Asynchronous federated learning-based ECG analysis for arrhythmia detection, in: 2021 IEEE International Mediterranean Conference on Communications and Networking, MeditCom, IEEE, 2021, pp. 277–282.

[39] Mufeng Zhang, Yining Wang, Tao Luo, Federated learning for arrhythmia detection of non-IID ECG, in: 2020 IEEE 6th International Conference on Computer and Communications, ICCC, IEEE, 2020, pp. 1176–1180.

[40] Xiaodong Ma, Jia Zhu, Zhihao Lin, Shanxuan Chen, Yangjie Qin, A state-of-the-art survey on solving non-IID data in federated learning, Future Gener. Comput. Syst. 135 (2022) 244–258.

[41] Hao Yu, Xin Yang, Xin Gao, Yihui Feng, Hao Wang, Yan Kang, Tianrui Li, Overcoming spatial-temporal catastrophic forgetting for federated class-incremental learning, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 5280–5288.

[42] Xin Yang, Hao Yu, Xin Gao, Hao Wang, Junbo Zhang, Tianrui Li, Federated continual learning via knowledge fusion: A survey, IEEE Trans. Knowl. Data Eng. 36 (8) (2024) 3832–3850.

[43] Hao Yu, Xin Yang, Le Zhang, Hanlin Gu, Tianrui Li, Lixin Fan, Qiang Yang, Addressing spatial-temporal data heterogeneity in federated continual learning via tail anchor, 2024, arXiv preprint arXiv:2412.18355.

[44] Tao Lin, Lingjing Kong, Sebastian U. Stich, Martin Jaggi, Ensemble distillation for robust model fusion in federated learning, Adv. Neural Inf. Process. Syst. 33 (2020) 2351–2363.

[45] Qinbin Li, Yiqun Diao, Quan Chen, Bingsheng He, Federated learning on non-iid data silos: An experimental study, in: 2022 IEEE 38th International Conference on Data Engineering, ICDE, IEEE, IEEE, 2022, pp. 965–978.

[46] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, Phillip Gibbons, The non-iid data quagmire of decentralized machine learning, in: International Conference on Machine Learning, PMLR, unknown, 2020, pp. 4387–4398.

[47] Yae Jee Cho, Jianyu Wang, Gauri Joshi, Client selection in federated learning: Convergence analysis and power-of-choice selection strategies, 2020, arXiv preprint arXiv:2010.01243.

[48] Sheng Chen, Jiancheng Peng, Andi Tong, Cong Wu, PFL-NON-IID framework: Evaluating MOON algorithm on handling non-IID data distributions, in: 2023 International Conference on Image, Algorithms and Artificial Intelligence, ICIAAI 2023, Atlantis Press, 2023, pp. 224–235.

[49] Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, Xing Xie, Communication-efficient federated learning via knowledge distillation, Nat. Commun. 13 (1) (2022) 2032.

[50] Besher Alhalabi, Shadi Basurra, Mohamed Medhat Gaber, Fednets: Federated learning on edge devices using ensembles of pruned deep neural networks, IEEE Access 11 (2023) 30726–30738.

[51] Chaoyang He, Murali Annavaram, Salman Avestimehr, Group knowledge transfer: Federated learning of large cnns at the edge, Adv. Neural Inf. Process. Syst. 33 (2020) 14068–14080.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[53] Mingxing Tan, Quoc Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[54] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[55] Hui-Po Wang, Sebastian Stich, Yang He, Mario Fritz, Progfed: effective, communication, and computation efficient federated learning by progressive training, in: International Conference on Machine Learning, PMLR, 2022, pp. 23034–23054.

[56] Alex Krizhevsky, Vinod Nair, Geoffrey Hinton, CIFAR-10 (Canadian institute for advanced research), 2009, Unknown. URL http://www.cs.toronto.edu/~kriz/cifar.html.

[57] Han Xiao, Kashif Rasul, Roland Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017, arXiv preprint arXiv:1708.07747.

[58] Daniel Mauricio Jimenez Gutierrez, Hafiz Muuhammad Hassan, Lorella Landi, Andrea Vitaletti, Ioannis Chatzigiannakis, Application of federated learning techniques for arrhythmia classification using 12-lead ecg signals, 2022, arXiv preprint arXiv:2208.10993.

[59] Jock Blackard, Covertype, 1998, http://dx.doi.org/10.24432/C50K5N, UCI Machine Learning Repository.

[60] Alex Krizhevsky, Geoffrey Hinton, et al., Learning Multiple Layers of Features from Tiny Images, Toronto, ON, Canada, 2009.

[61] Yong-Hoon Choi, Daegyeom Kim, Myeongjin Ko, Kyung-yul Cheon, Seungkeun Park, Yunbae Kim, Hyungoo Yoon, ML-based 5G traffic generation for practical simulations using open datasets, IEEE Commun. Mag. 61 (9) (2023) 130–136.

[62] Oresti Banos, Rafael Garcia, Alejandro Saez, MHEALTH, 2014, http://dx.doi.org/10.24432/C5TW22, UCI Machine Learning Repository.

[63] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, Craig Packer, Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African Savanna, Sci. Data 2 (1) (2015) 1–14.

[64] Rahul Chauhan, Kamal Kumar Ghanshala, R.C. Joshi, Convolutional neural network (CNN) for image detection and recognition, in: 2018 First International Conference on Secure Cyber Computing and Communication, ICSCCC, IEEE, IEEE, 2018, pp. 278–282.

[65] Maryam Saeed, Olev Märtens, Benoit Larras, Antoine Frappé, Deepu John, Barry Cardiff, ECG classification with event-driven sampling, IEEE Access (2024).

[66] Fatma Murat, Ozal Yildirim, Muhammed Talo, Ulas Baran Baloglu, Yakup Demir, U. Rajendra Acharya, Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review, Comput. Biol. Med. 120 (2020) 103726, http://dx.doi.org/10.1016/j.compbiomed.2020.103726.

[67] Daniel J. Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, et al., Flower: A friendly federated learning research framework, 2020, arXiv preprint arXiv:2007.14390.

[68] Manfredo Perdigão do Carmo, Differential Geometry of Curves and Surfaces, Revised and Updated 2nd ed., Dover Publications, New York, 2016.

[69] Alfred Gray, Modern Differential Geometry of Curves and Surfaces with Mathematica, third ed., CRC Press, Boca Raton, FL, 2006.

[70] Hongda Wu, Ping Wang, Fast-convergent federated learning with adaptive weighting, IEEE Trans. Cogn. Commun. Netw. 7 (4) (2021) 1078–1088.

[71] Jiayu Lin, On the Dirichlet Distribution, vol. 40, Department of Mathematics and Statistics, Queens University, 2016,

[72] Hadi Jamali-Rad, Mohammad Abdizadeh, Anuj Singh, Federated learning with taskonomy for non-IID data, IEEE Trans. Neural Netw. Learn. Syst. (2022).

[73] Frank Nielsen, On the Jensen–Shannon symmetrization of distances relying on abstract means, Entropy 21 (5) (2019) 485.

[74] Adam Davis, Tony Menzo, Ahmed Youssef, Jure Zupan, Earth mover's distance as a measure of CP violation, J. High Energy Phys. 2023 (6) (2023) 1–42.

[75] Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, Aduri Pavan, N.V. Vinodchandran, On approximating total variation distance, 2022, arXiv preprint arXiv:2206.07209.

[76] Ahmed M. Abdelmoniem, Chen-Yu Ho, Pantelis Papageorgiou, Marco Canini, Empirical analysis of federated learning in heterogeneous environments, in: Proceedings of the 2nd European Workshop on Machine Learning and Systems, 2022, pp. 1–9.

[77] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al., Advances and open problems in federated learning, Found. Trends® Mach. Learn. 14 (1–2) (2021) 1–210.

[78] Xiuyu Shen, Jingxu Chen, Siying Zhu, Ran Yan, A decentralized federated learning-based spatial–temporal model for freight traffic speed forecasting, Expert Syst. Appl. 238 (2024) 122302.

[79] Wenjie Fu, Xudong Zhang, Junlong Wang, Di Yang, Yuntong Lv, Yuqing Wang, Zhao Zhen, Fei Wang, A spatiotemporal federated learning based distributed photovoltaic ultra-short-term power forecasting method, in: 2023 IEEE/IAS 59th Industrial and Commercial Power Systems Technical Conference, I&CPS, IEEE, 2023, pp. 1–7.

[80] Xuehan Zhou, Ruimin Ke, Zhiyong Cui, Qiang Liu, Wenxing Qian, Stfl: Spatio-temporal federated learning for vehicle trajectory prediction, in: 2022 IEEE 2nd International Conference on Digital Twins and Parallel Intelligence, DTPI, IEEE, 2022, pp. 1–6.

[81] Dun Zeng, Siqi Liang, Xiangjing Hu, Hui Wang, Zenglin Xu, Fedlab: A flexible federated learning framework, J. Mach. Learn. Res. 24 (100) (2023) 1–7.

[82] Fan Lai, Yinwei Dai, Sanjay Singapuram, Jiachen Liu, Xiangfeng Zhu, Harsha Madhyastha, Mosharaf Chowdhury, Fedscale: Benchmarking model and system performance of federated learning at scale, in: International Conference on Machine Learning, PMLR, 2022, pp. 11814–11827.

[83] Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al., Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings, Adv. Neural Inf. Process. Syst. 35 (2022) 5315–5334.