Handling Non-IID Data in Federated Learning: An Experimental Evaluation Towards Unified Metrics

Marc Haller, Christian Lenz, & Robin Nachtigall
Karlsruhe University of Applied Sciences
Karlsruhe, Germany
{hama1082, lech1028, naro1012}@h-ka.de

Feras M. Awaysheh
Institute of Computer Science, Delta
The University of Tartu
Tartu, Estonia
feras.awaysheh@ut.ee

Sadi Alawadi
Department of Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
sadi.alawadi@bth.se

***Abstract*—Recent research has demonstrated that Non-Identically Distributed (Non-IID) data can negatively impact the performance of global models constructed in federated learning. To address this concern, multiple approaches have been developed. Nonetheless, previous research lacks a cohesive overview and fails to uniformly assess these strategies, resulting in challenges when comparing and choosing relevant options for real-world scenarios. This study presents a structured survey of cutting-edge techniques for handling the Non-IID data, accompanied by proposing a metric to develop a standardized approach for assessing data skew and its harmony with the appropriate approach. The findings affirm the metric's suitability as a heuristic for assessing data skew in distributed datasets without having insight into client data, serving both scientific and practical purposes and thus supporting the selection of handling strategies. This preliminary research establishes the foundation for discussing standardizing methodologies for evaluating data heterogeneity in federated learning.**

***Index Terms*—Federated Learning, Non-IID, Data Skew Detection, Standardization, Taxonomy**

I. Introduction

The advent of Big Data [1], and the decentralization of data sources has given birth to federated learning (FL). This innovative machine learning paradigm enables multiple clients (nodes or workers) to collaboratively train a global machine learning model locally while preserving individual data privacy and without compromising client's data [2]–[6].

However, ensuring a uniform data distribution among clients is uncommon and not guaranteed in real scenarios. Indeed, data across clients usually exhibits distinct characteristics known as Non-Independent and Identically Distributed (Non-IID) [7], indicating variations in statistical attributes such as uneven feature distributions, imbalance classes, or differences in data quantity. The empirical evidence presented in the literature highlights the significant impact of Non-IID data on model performance within the FL context [8]–[10]. This phenomenon introduces unique challenges to the FL landscape.

A recent study by Hsieh et al. shed light on the significant detrimental effects of Non-IID data characteristics on the efficacy of the global FL model [8]. Therefore, numerous techniques have been proposed to tackle the issues arising from data heterogeneity. However, the current research landscape needs to be more cohesive, with many strategies developed in isolated contexts and requiring a unified evaluation criterion. This situation challenges researchers and practitioners as they seek to understand, compare, and select the most appropriate approach for their specific use cases. Moreover, there remains a pressing need for a holistic research perspective, given the piecemeal strategies often developed in siloed contexts, each crying out for a unified evaluation benchmark.

Considering the research gap in this area, our study aims to provide an overview of the existing techniques to tackle non-IID data in the FL context, with the intention of bridging knowledge gaps. Moreover, and more significantly, this research proposes a novel metric to measure data skewness consistently. Using simulated data sets, we validate this metric's effectiveness and applicability, aiming to present a heuristic tool for assessing data heterogeneity without the need for direct access to client-specific data sets.

Our main vision motivating this work extends beyond clarifying the challenges of non-IID data and their implications on the ML model performance. It aims to pave the way for a standardized approach to evaluate data heterogeneity within FL frameworks. This standardization is of essential importance for Non-IID data new approaches progress within FL settings, facilitating greater consistency in model performance and providing practitioners with the guidance needed to make informed decisions regarding data handling strategies based on the characteristics of their distributed data sets.

The structure of this paper is as follows: We start by draft the problem statement and motivation scenario in Section II. In Section III, we present a comprehensive taxonomy that categorizes the existing literature. the proposed approach and the experimental settings described in IV. Next, both the validation of the proposed approach and the findings are discussed in Section V. Finally, in Section VI, and we draw our conclusions and future work in Section VI.

II. Problem Formulation

This paper focuses on analyzing the cutting-edge techniques for handling the Non-IID data in FL settings by classifying them in a taxonomy and experimentally evaluating them to propose best practices and develop a standardized approach

for assessing data skew and its harmony with the appropriate approach.

Non-IID data distributions can introduce multiple challenges in FL solution. Among the various facets of Non-IID, the most challenging form of data skew is typically label distribution skew or class imbalance across edge clients. many factors cause this, including the following:

Diverse Real-world Scenarios: Different clients (like mobile devices or sensors) can have data from vastly different distributions in many real-world applications. For example, a health monitoring device from an elderly person will generate a different set of health data compared to that from a teenager.

Model Convergence: Models can struggle to converge or might converge to a suboptimal solution when a significant class imbalance exists across clients. Some clients might have samples from only a subset of classes, making the global model biased if not handled appropriately.

Model Performance: In imbalanced client classes, the global model's performance might be excellent for frequently represented classes but poor for under-represented classes.

Client Participation Bias: If only a subset of clients with a particular data skew participate more frequently in the federated learning rounds, it can introduce further bias into the model.

To understand this issue better, consider the following scenarios on the challenge presented by label distribution skew in FL. Suppose there are C clients, K classes, and the data at each client c is represented by D_c, and the overall dataset by D. The distribution of class k at client c is denoted by $p_{c,k}$.

Consider that for a certain client c and class k, $p_{c,k} = 0$, meaning that client c has no samples of class k.

In the FL paradigm, each client trains a model on its local data and subsequently sends the model updates to the server. The server updates the global model by averaging these updates (using the FedAVG algorithm). Mathematically, the global model M is given by:

$$M = \frac{1}{C} \sum_{c=1}^{C} M_c$$

where M_c is the model trained on the data of client c.

Performance Implications of Skew Given that client c lacks representation for class k, the model M_c might perform poorly on class k. This performance gap is integrated into the global model during averaging, leading to suboptimal performance for class k even if other clients have data from that class.

To quantify this, let $\text{acc}_{c,k}$ be the accuracy for class k for client c and acc_k be the corresponding accuracy for the global model. Owing to the absence of data for class k at client c, $\text{acc}_{c,k}$ could be substantially lower than the average accuracy across clients. Consequently, the aggregated global model's accuracy for class k may suffer:

$$\text{acc}_k < \frac{1}{C} \sum_{c=1}^{C} \text{acc}_{c,k}$$

This inequality signifies a performance decrement for class k in the global model, especially when there's a large discrepancy in $\text{acc}_{c,k}$ values among clients.

Moreover, the issue of statistical divergence between clients.This issue can be measured using the Kullback-Leibler (KL) divergence. Given two clients c_1 and c_2, the divergence due to label distribution skew is defined as:

$$D_{KL}(p_{c_1} || p_{c_2}) = \sum_k p_{c_1,k} \log \left(\frac{p_{c_1,k}}{p_{c_2,k}} \right) \tag{1}$$

This equation quantifies how the distribution of one client diverges from another.

Also, the error introduced during the model aggregation phase due to skew can be defined. Let's denote the true global model as M^* and the aggregated model as M. The error, represented by E, due to data skew in the aggregation process is:

$$E = \|M - M^*\| \tag{2}$$

This error E quantifies the difference between the ideal global model and the model aggregated considering the data skew.

Data skew affects the test accuracy. Consider an IID test dataset D_{test}. The test accuracy of the global model is expected to decrease with increasing skew. Let the accuracy of model M on D_{test} be denoted by $\text{acc}(M, D_{test})$. It can be postulated that:

$$\text{acc}(M, D_{test}) \text{ decreases with increasing data skew.} \tag{3}$$

This mathematical exposition underscores the challenge presented by label distribution skew in FL. Models from clients with skewed distributions can adversely affect the performance of the global model. Addressing these skews may require sophisticated aggregation methods or data resampling strategies.

Despite the literature efforts, addressing data skew remains a primary concern in ensuring the robustness and fairness of FL models. Moreover, the literature lacks standardized criteria to handle this challenge. This research gap was the main motivation behind our study. In the next section, we will classify the proposed solutions in the literature in a taxonomy followed by experiments to evaluate different datasets using several aggregation algorithms.

III. STATE-OF-THE-ART TAXONOMY

Several existing techniques are proposed in the literature to handle Non-IID data, with some directly tailored to Non-IID scenarios and others having indirect applications. This diversity can pose challenges for researchers seeking to select an appropriate strategy and understand the landscape comprehensively. Therefore, our main contribution in this paper is the development of a comprehensive taxonomy rooted in the current state-of-the-art techniques for managing Non-IID data. This structured taxonomy was established through a systematic literature review classification, employing the snowball sampling approach that began with two foundational overview papers and extended to encompass subsequent relevant literature.

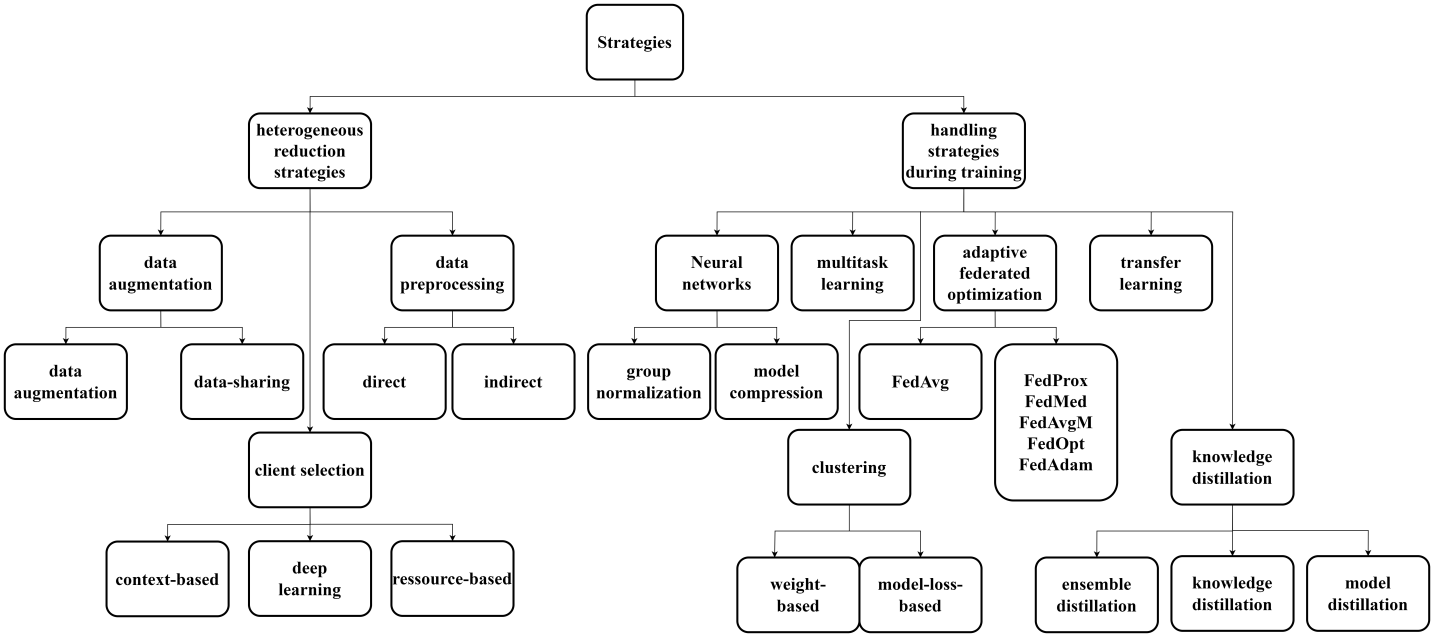

Fig. 1. Taxonomy of strategies for handling Non-IID data

Figure 1 illustrates the taxonomy, elucidating the procedures of individual methods, referencing their original contributions, and highlighting potential challenges associated with their implementation.

As the main criteria to categorize the strategies, the position in the ML process was chosen as proposed by H. Xuming et al. [11]. These strategies are applied before or during the initialization of the training to prepare the data set and are labeled as "heterogeneous reduction strategies." Meanwhile, more robust methods to deal with Non-IID during training are called "handling strategies during training" [11]. In the second level, they are clustered by their approach to setting up a robust training environment.

The first cluster of heterogeneous reduction strategies is "data augmentation." Duan et al. propose an augmentation method that enriches client data with client-specific synthesized data to reduce the degree of heterogeneity [12], whereas other data-sharing strategies use one globally-shared data set that contains all labels with homogeneous distribution. The global dataset, with the size of 1-20% of all client data depending on the specific strategy, is used to train the initial model and distributed to all clients [13]. The distribution of data raises privacy concerns [14], and distributing the potentially very high amount of data produces a communication overhead that could overwhelm clients, especially in a cross-device setting.

Client selection-based methods are chosen based on different criteria for training in each communication round that work well together based on different metrics concerning the composition of their data. Context-based client selection chooses different clients in each communication round based on information about their respective data composition, which raises privacy concerns [15]. Deep learning-based client selection does not need this a priori information but can result in communication overhead [11]. Resource-based client selection does not necessarily aim for better handling of Non-IID data but can help in this regard [16], [17]. All methods have in common that maintaining one global model tends to produce a bias and overfitting towards "good" clients. A combination with other methods like data-sharing is recommended.s

The initial model gets trained with a benchmark data set and gets forwarded with the associated validation score by direct data preprocessing. In the following training, clients only use the data which produces a similar loss with the forwarded model. This approach results in solid bias, and the model quality heavily depends on the benchmark data set more than data-sharing. Indirect data preprocessing is, essentially, context-based client selection while using encrypted information about the client's data composition [11].

Handling strategies during training aim to be more robust against Non-IID data than FedAvg without additional measurements. The first cluster is concentrated on neural networks and provides methods that alter the behavior of the local training of neural networks. With group normalization, K. Hsieh et al. offer an alternative to the highly vulnerable batch normalization [8]. Meanwhile, S. Wiedemann et al. uses model compression, which tends to be more robust against Non-IID data as a side effect [18]. Especially when combined with other strategies, these methods can produce improved results.

Clustering strategies give up the premise that the result of FL is one global model but aim to deliver multiple models that better fit each client group than one universal global model could ever offer. The presented clustering strategies have in common that they work without the necessity to have insight into the client data composition in contrast

to some client selection methods. Weight-based clustering groups clients based on the weights they send after the initial communication rounds as a metric for the similarity of their data. Hierarchical clustering can be applied, which merges the most similar clusters iteratively to reduce the number of clusters until a set number of clusters or the maximum dissimilarity between clients in one cluster is reached [19]. Model-loss-based clustering works similarly, but instead of using all weights for clustering uses only the communicated model-loss [11]. After the initial clustering, each cluster gets trained independently and receives its global model. Neither privacy nor communication overhead are issues, as the latter is only of significance during the clustering itself. Unfortunately, for future deployment, every user has to be placed in the fitting cluster for which he needs a sufficient personal data set.

Under "adaptive federated learning," alternatives to the benchmark aggregating function "FedAvg" [20] are listed. "FedProx" [21] is listed as the most popular to limit the weight divergence using a regularisation term [22] and by that ensure convergence of the training [23]. To obtain a better representation through a wider variety of aggregation functions the common alternatives "FedMedian" [24], "FedAvgM" [25], "FedOpt" [26] and "FedAdam" [27], [28] were used in the later experiment as well. On the other hand, Federated Multitask Learning lets the clients share their respective models and continue developing them with local updates before sending them to the aggregator. This strategy reduces weight divergence between the clients, and the communication rounds can be lowered. However, the proposed methodology with clustering tends to isolate heterogeneous clients, which can produce models with a substantial divergence in quality [11].

Knowledge distillation also works with local updates but uses one or multiple teacher clients to carry out the local updates more controlled. Model distillation can achieve satisfying improvements as well. W. Ouyang et al. promote transfer learning in cases with a strong label-distribution skew [29]–[32].

Based on this overview, we further analyze which strategy performs best for a given level of data skew. Nevertheless, as this paper presents early-stage research, only limited methods were conducted.

IV. METHODOLOGY

As shown previously, there is a variety of different strategies for dealing with Non-IID data. Previous studies have not evaluated these methods with a consistent framework, making it difficult to compare and select them for practical real-world situations. In addition, differences in the degree of heterogeneity of the data have not been considered. Therefore, in the following, the authors examine, how well-selected strategies perform compared to each other and for different levels of heterogeneity in data sets.

A. EVALUATION OF HETEROGENEITY

To attain the aforementioned objective, a specific metric is needed. Existing research has presented very few ideas for assessing the degree of heterogeneity in federated environments. Of the few concepts that have been presented, no standard has yet been established, which makes the comparability of data sets used to evaluate Non-IID handling strategies almost impossible. Therefore, we adapt an existing concept at first to subsequently present it as a standard for determining the degree of heterogeneity. For this purpose, we adapt the "model traveling" [8] approach presented by Hsieh et al. to develop a metric for determining data skew in yet unknown federated environments. For the model traveling, a predefined initial model is trained on a randomly selected client for a sufficiently high number of epochs. Subsequently, the trained model is transferred to each client. The clients evaluate the trained model on their local data and report the achieved accuracy back to the server. Our contribution to this is that based on the transmitted accuracies a key metric (data skew) is calculated. The metric enables comparability between Non-IID data sets and allows evaluation of the degree of heterogeneity present in federated environments without requiring insight into the clients' data, which is a necessity in the context of Federated Learning. For the purpose of this paper we introduce data skew as the following metric:

$$DATASKEW = \frac{max(\Delta Accuracy_{\text{ pairwise}})}{\frac{1}{n}\sum_{i=0}^{n}(Accuracy_{\text{ Client}_i})}$$

As a first input variable the maximum pairwise deviation of the accuracy score over all clients (including the client on which to initial model was trained) is consulted. For strongly heterogeneous data sets this variable can approximate 1. For example, in the case of pathological Non-IID data sets where every client has strictly distinct labels which no other client shares the initial model might get an accuracy-score close to 1 on the client it was trained on while it may reach an accuracy of nearly 0 on another client. The maximum pairwise deviation is being consulted so the accuracies in between don't matter in this case. To put this into relation the average accuracy over all clients is used otherwise there would be a strong bias towards environments with a high average accuracy. To give an example with numbers: Let the pairwise deviation of the accuracies in two given setups be 0.2. But in the first environment the average accuracy over all clients is 0.6 while in the second setup it is 0.1. With the presented metric the second environment gets evaluated with a data skew value of 2, while the first achieves a much lower value of 0.33. The experiment later on will show that even a value of 0.33 is high enough to benefit from strategies to reduce the effect of the heterogeneous distribution of the data. The metric serves the purpose of this paper as an indicator of the level of data skew to evaluate which of the presented strategies is appropriate for a given scenario.

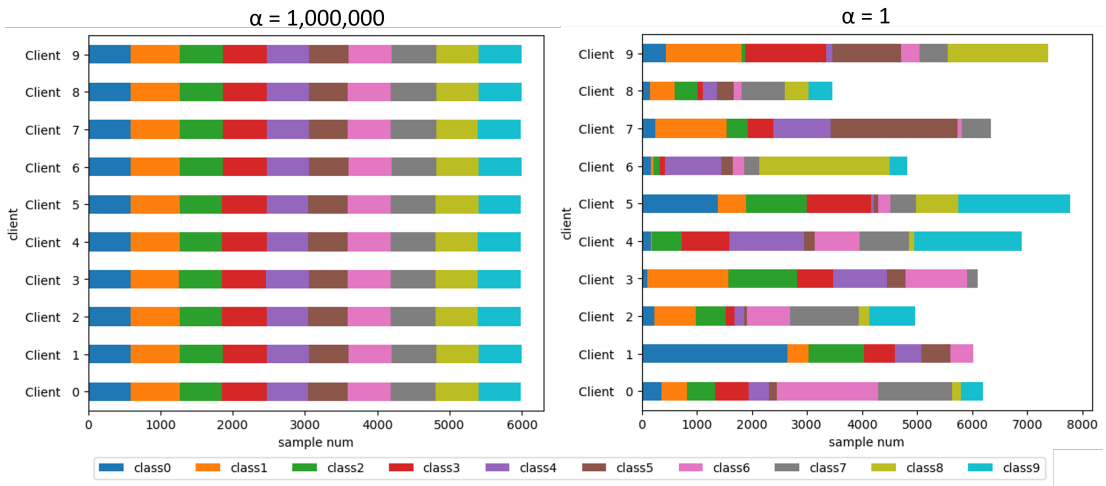

Fig. 2. Data partitioning of MNIST using different alpha-values for inhomogeneity simulation

B. EXPERIMENTAL SETUP

Our study has associated different degrees of data skew with each aggregation strategy to validate the proposed approach within FL settings. We conducted these experiments using three well-known datasets widely recognized in the community: **CIFAR10**, **CIFAR100**, and **MNIST**. For our experiments, we utilized a Convolutional Neural Network (CNN) model as described in Table I.

The results obtained in this study serve as a benchmark and a valuable reference for researchers, offering guidance on selecting an appropriate skew strategy to handle non-IID data grounded in a thorough dataset evaluation effectively. All experiments in this study were conducted on a machine equipped with the following specifications: **Intel(R) Core(TM) i7-10870H**, **CPU @ 2.20GHz 2.21 GHz**, **16 GB RAM**, and **NVIDIA GeForce RTX 3060 Laptop GPU**.

We employed a random data partitioning approach across ten clients to simulate a federated learning environment. This partitioning scheme encompasses two distinct cases: (1) near homogeneous partitioning (IID), and (2) heterogeneous partitioning (Non-IID). Following the methods proposed by Yurochkin et al. and Wang et al., we utilized the Dirichlet distribution to allocate data to each client, with the allocation being determined by the dataset and a corresponding α-value [33] [34]. It's worth noting that higher α-values result in more homogeneous data distribution. For the first near-homogeneous case, we set α to 1,000,000. For the second case, we employed smaller α-values ranging from 0.01 to 10, generating varying degrees of data skew within the datasets. This data partitioning approach primarily targets skew in label and sample volume distribution, ensuring that each client retains the entirety of the features without introducing skew in feature distribution.

The data splits for the two respective α-values are illustrated in Figure 2, wherein the left panel, α=1,000,000 represents the near homogeneous edge case. In contrast, in the right panel, α=1 represents a more realistic distribution for real-world applications. Moreover, in Figure 2, each label (10 lables) of the MNIST dataset is represented in a distinct color, and the corresponding number of data samples assigned to each client is depicted.

With the synthetically generated Non-IID datasets in place, we examined and compared various non-IID data handling strategies in terms of kappa, accuracy and F1 measure to validate their effectiveness in tackling this issue in FL context. To ensure comparability and applicability to other research, we have adopted commonly used model parameters frequently used in FL studies [35] . In our conducted experiment, we simulate a federated learning scenario where a vanilla CNN model is trained independently locally by ten clients. We executed 10 communication rounds and 5 local epochs throughout the training process. It's important to note that our computational resources imposed these limitations on the study.

Subsequently, we will use several existing non-IID data handling strategies in the literature. Specifically, we will examine the performance of **FedProx** [21], **FedAdam** [27], [28], **FedAvgM** [25], **FedOpt** [26], and **FedMedian** [24], utilizing the default parameters as provided by Flowers framework [36], across various α-values and datasets. It's important to note that the FedAvg strategy will also be used as a benchmark to provide a comparative baseline, disregarding the non-IID data issue. Table II reports the model Accuracy, F1-Score and Kappa values obtained from the experiments, which will serve as a reference for identifying the most effective strategies in addressing different levels of data heterogeneity.

V. RESULTS AND ANALYSIS

The experimental results validate the efficacy of our proposed metric, *Dataskew*, in gauging the heterogeneity prevalent within distributed datasets. Intriguingly, while reliably assessing heterogeneity, our method operates without needing direct insight into client-specific data. Thus, both from a

TABLE I
THE CONVOLUTIONAL NEURAL NETWORK (CNN) ARCHITECTURE USED IN THIS STUDY TO VALIDATE THE PROPOSED APPROACH.

Layer	Type	Input dimension	Output dimension	Activation
Conv1	Conv2d	(32, 3/1*, 32/28*, 32/28*)	(32, 6, 28/24*, 28/24*)	ReLU
MaxPool1	MaxPool2d	(32, 6, 28/24*, 28/24*)	(32, 6, 14/12*, 14/12*)	None
Conv2	Conv2d	(32, 6, 14/12*, 14/12*)	(32, 16, 10/8*, 10/8*)	ReLU
MaxPool2	MaxPool2d	(32, 16, 10/8*, 10/8*)	(32, 16, 5/4*, 5/4*)	None
Flatten	View	(32, 16, 5/4*, 5/4*)	(32, 400/256*)	None
Fully Connected (FC1)	Linear	(32, 400/256*)	(32, 120)	ReLU
Fully Connected (FC2)	Linear	(32, 120)	(32, 84)	ReLU
Fully Connected (FC3)	Linear	(32, 84)	(32, 10/100*)	None

* It's worth noting that the configuration of the output layer varies depending on the number of labels present in the dataset, which includes CIFAR10, CIFAR100, and MNIST.

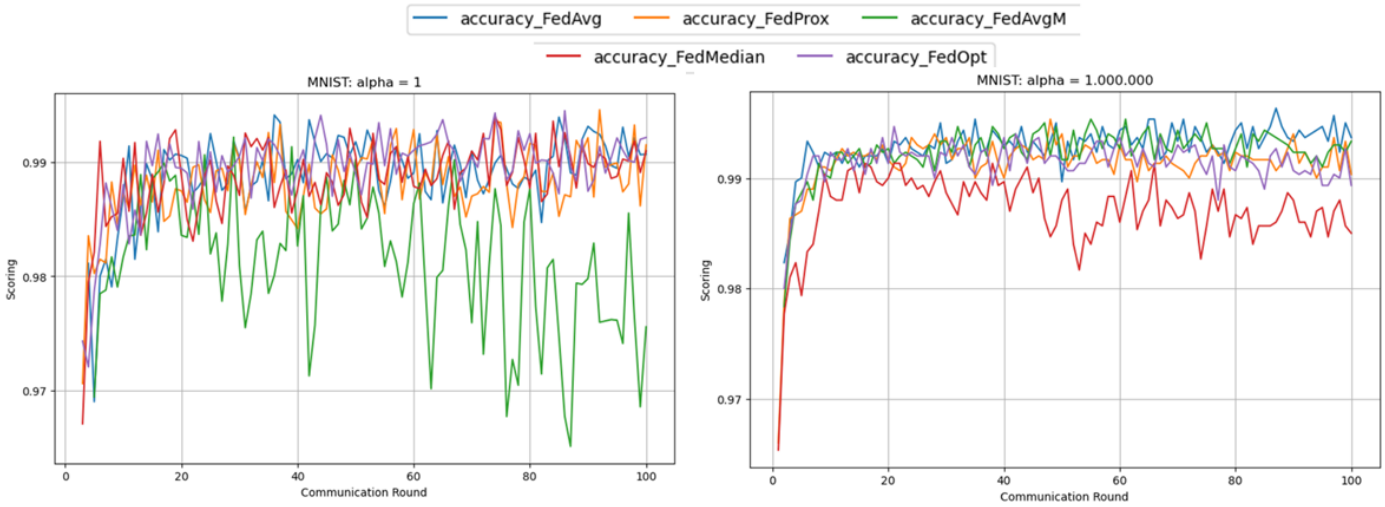

Fig. 3. Performance of the used aggregation strategies for MNIST

practical and theoretical standpoint, *Dataskew* emerges as a compelling heuristic for this context.

Across three diverse datasets, the behavior of the metric aligns with our expectations. The metric's value gravitates towards zero in scenarios with near-uniform data distributions. Conversely, the metric exhibits an upward trend as the dataset's heterogeneity intensifies. A notable observation from Table 2 is the discernibly lower *Dataskew* for the CIFAR10 dataset at an α value of 100 in comparison to 1,000,000. This can be attributed to the marginal variance that the data split introduces at these particular α values. This observed higher *Dataskew* at α=100, relative to α=1,000,000, is likely a consequence of random data fluctuations. The synthesis parameter, α, used to generate data splits, is available in our study's context. However, this parameter remains elusive in real-world applications where client data remains inaccessible. The *Dataskew* metric, given its commendable correlation with the α-value, emerges as a potent tool to bridge this informational chasm.

As elucidated in Figure 3 for the MNIST dataset at alpha-values of 1 and 1,000,000, the convergence dynamics of different aggregation methodologies furnish valuable insights. Given MNIST's computationally amenable nature, we ramped up the communication rounds to a century to guarantee convergence. For the sake of clarity, the initial communication rounds, which demonstrate homogenous convergence trends across strategies, have been eschewed. Similarly, FedAdam, which paradoxically showcased deteriorating accuracies with escalating communication rounds—possibly a fallout of leveraging built-in default parameters—is excluded. A juxtaposition of the convergence patterns for disparate alpha-values vividly underscores the repercussions of Non-IID data. As heterogeneity escalates, palpable fluctuations in the global model's quality become evident. Notably, FedAdam and FedAvgM are particularly susceptible to this volatility. However, other aggregation strategies exhibit resilience and consistency, even under pronounced Non-IID conditions. To render a more holistic evaluation of aggregation strategies, especially from a long-term perspective, we advocate for an augmentation in communication rounds, surpassing the 100 mark.

All four algorithms perform well when there is no data skew (alpha = 1.000.000). However, as the data skew increases, the performance of FedAvg and FedProx deteriorates. FedMedian and FedOpt, on the other hand, are able to maintain good performance even when the data skew is high.

TABLE II

EXPERIMENTAL RESULTS AFTER 100 COMMUNICATION ROUNDS, BLUE VALUES SHOWING MAXIMUM RESULTS, RED VALUES SHOWING MINIMUM RESULTS.

Data set	α	dataskew	Achieved Accuracy [%] F1-Score Kappa							
			FedAvg	FedProx			FedAdam	FedAvgM	FedMedian	FedOpt
				μ=0.01	μ=0.1	μ=0.5				
CIFAR10	0.01	3.899	38.0%	**39.0%**	**21.2%**	24.7%	25.2%	33.1%	24.9%	28.1%
			0.496	0.513	0.285	0.271	0.044	0.472	0.248	0.405
			0.190	0.224	0.110	0.067	0.005	0.183	0.137	0.163
	0.1	2.245	39.6%	**48.0%**	37.7%	20.6%	**16.6%**	46.6%	44.5%	47.3%
			0.456	0.547	0.371	0.185	0.132	0.542	0.465	0.534
			0.246	0.317	0.251	0.064	0.035	0.296	0.267	0.311
	1	0.725	55.0%	58.9%	49.4%	**14.0%**	21.6%	54.5%	**60.0%**	55.6%
			0.567	0.617	0.519	0.053	0.199	0.565	0.622	0.579
			0.481	0.524	0.419	0.0	0.138	0.472	0.529	0.485
	10	0.188	58.3%	59.4%	50.1%	**9.8%**	50.8%	56.0%	**61.1%**	58.9%
			0.583	0.597	0.495	0.018	0.397	0.566	0.613	0.588
			0.535	0.546	0.443	0.0	0.343	0.509	0.565	0.541
	100	0.088	**61.0%**	60.6%	47.3%	**9.6%**	35.9%	57.4%	58.9%	**61.0%**
			0.608	0.602	0.462	0.034	0.326	0.577	0.589	0.608
			0.567	0.562	0.413	-0.007	0.288	0.526	0.543	0.566
	1,000,000	0.194	**61.2%**	59.7%	**10.1%**	11.1%	43.7%	56.4%	61.0%	60.2%
			0.611	0.596	0.021	0.039	0.425	0.565	0.609	0.601
			0.569	0.552	0.0	0.012	0.375	0.515	0.566	0.557
CIFAR100	0.01	8.907	8.9%	9.4%	**11.6%**	**0.1%**	1.1%	8.0%	5.2%	8.7%
			0.100	0.124	0.147	0.0	0.015	0.110	0.053	0.103
			0.075	0.081	0.101	-0.001	0.007	0.065	0.041	0.074
	0.1	3.827	17.0%	**20.2%**	18.3%	**0.8%**	6.2%	15.0%	14.9%	17.2%
			0.178	0.243	0.218	0.001	0.075	0.187	0.149	0.203
			0.156	0.187	0.169	0.001	0.051	0.136	0.135	0.159
	1	0.800	22.2%	24.8%	17.5%	**0.4%**	4.8%	19.3%	23.9%	**25.6%**
			0.223	0.258	0.167	0.0	0.033	0.197	0.238	0.266
			0.213	0.239	0.166	0.0	0.038	0.183	0.229	0.247
	10	0.442	24.2%	23.4%	1.2%	**1.1%**	7.3%	22.1%	23.7%	**26.1%**
			0.238	0.227	0.001	0.0	0.069	0.220	0.233	0.252
			0.234	0.226	0.004	0.0	0.063	0.213	0.229	0.253
	100	0.374	24.9%	24.4%	**0.7%**	1.3%	10.1%	21.6%	24.4%	**25.8%**
			0.239	0.231	0.0	0.001	0.089	0.214	0.233	0.249
			0.241	0.236	0.0	0.005	0.092	0.208	0.236	0.250
	1,000,000	0.318	25.5%	**27.4%**	**1.0%**	1.2%	12.5%	20.9%	24.5%	25.0%
			0.245	0.256	0.002	0.007	0.115	0.209	0.237	0.240
			0.247	0.266	-0.001	0.002	0.116	0.201	0.237	0.242
MNIST	0.01	10	52.5%	**83.3%**	39.6%	**20.0%**	30.8%	67.6%	40.7%	37.4%
			0.607	0.901	0.503	0.254	0.414	0.731	0.463	0.438
			0.040	0.325	0.094	0.042	0.008	0.203	0.002	0.004
	0.1	2.005	92.6%	97.4%	91.3%	**87.7%**	90.0%	96.1%	**98.3%**	97.9%
			0.942	0.983	0.947	0.916	0.880	0.976	0.990	0.987
			0.863	0.951	0.815	0.775	0.793	0.906	0.947	0.954
	1	0.668	98.2%	**98.9%**	97.0%	**89.2%**	94.9%	98.5%	98.6%	98.5%
			0.983	0.989	0.971	0.895	0.950	0.986	0.986	0.985
			0.978	0.987	0.963	0.869	0.938	0.982	0.982	0.981
	10	0.029	98.8%	98.8%	97.8%	**12.0%**	97.5%	98.8%	99.0%	**99.2%**
			0.988	0.988	0.978	0.057	0.975	0.988	0.990	0.992
			0.987	0.986	0.975	0.045	0.972	0.987	0.989	0.991
	100	0.024	99.0%	99.1%	97.9%	**92.7%**	97.5%	99.0%	**99.2%**	99.0%
			0.990	0.991	0.979	0.927	0.975	0.990	0.992	0.990
			0.989	0.990	0.977	0.919	0.972	0.989	0.991	0.989
	1,000,000	0.025	98.9%	98.9%	98.1%	**11.9%**	97.2%	98.7%	98.8%	**99.0%**
			0.989	0.989	0.981	0.046	0.972	0.987	0.988	0.990
			0.988	0.987	0.978	0.023	0.969	0.985	0.986	0.988

Hence, Figure 3 demonstrates that data skew can significantly impact the performance of FL algorithms. It also shows that there are algorithms that can be used to mitigate the effects of data skew.

Moreover, our benchmarking experiments in Table II offer pivotal insights, especially pertinent for datasets exhibiting extremities in *Dataskew* values. For instances with *Dataskew* values ranging between 2 and 10, FedProx, particularly at lower μ values, consistently outstrips FedAvg, reiterating its superiority in high heterogeneity contexts. Meanwhile, for datasets with minimal skewness (values below 0.5), FedAvg produces commendable results, albeit marginally overshadowed by FedOpt, hinting at the latter's potential robustness in low heterogeneity settings. The performance spectrum for *Dataskew* values nestled between 0.5 and 2, especially with respect to FedMedian, warrants more exhaustive exploration to discern overarching patterns. As we advance in this research trajectory, we remain optimistic about our metric's sustained correlation with actual data heterogeneity. However, more empirical evaluations, especially with real-world datasets, are imperative to cement its universality.

Overall, the table shows that FedMedian and FedOpt are good choices for FL applications with a data skew risk. If computational efficiency is essential, then FedMedian is a good choice. If performance is the most crucial consideration, then FedOpt is a good choice.

In future work, we anticipate that the metric will continue to correlate well with the actual heterogeneity in the data set. Further test series with other handling strategies and other data sets, especially real world data sets, are needed to determine if the selection of the best strategy correlates adequately with the introduced metric or whether additional parameters are necessary to make a final decision. Table 2 contains the preliminary results.

VI. CONCLUSIONS AND FUTURE WORK

Over recent years, federated learning (FL) has emerged as a promising paradigm for privacy-preserving distributed machine learning. However, the challenge posed by non-IID data has consistently undermined its potential in various applications. In addressing this issue, this paper has made a significant contribution by proposing a novel metric to quantify data heterogeneity in distributed datasets, thereby filling an existing void in the literature.

Our systematic review and state-of-the-art taxonomy of non-IID data handling strategies offer a valuable resource for researchers and practitioners. This foundation assists in the informed selection of appropriate strategies contingent on the level of data skewness in specific datasets. By providing clarity on this front, our work facilitates more pragmatic decision-making when navigating the complexities of non-IID data in real-world federated learning deployments.

This research signifies a substantial advancement towards a systematic approach for assessing and managing non-IID data in FL, further deepening our comprehension of its ramifications and offering means to counteract its effects. Notwithstanding these contributions, areas for future exploration remain evident. There is potential in broadening the spectrum of strategies evaluated, refining the proposed metric for broader applicability, and investigating clustered aggregation and data-sharing nuances. One limitation worth addressing in subsequent studies is our model's susceptibility to single outliers, especially in expansive cross-device settings. Such endeavors will undeniably fortify the robustness and reliability of FL models in the face of data heterogeneity.

REFERENCES

[1] F. M. Awaysheh, M. Alazab, S. Garg, D. Niyato, and C. Verikoukis, "Big data resource management & networks: Taxonomy, survey, and future directions," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2098–2130, 2021.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[3] F. M. Awaysheh, S. Alawadi, and S. AlZubi, "Fliodt: A federated learning architecture from privacy by design to privacy by default over iot," in *2022 Seventh International Conference on Fog and Mobile Edge Computing (FMEC)*. IEEE, 2022, pp. 1–6.

[4] S. Alawadi, K. Alkharabsheh, F. Alkhabbas, V. Kebande, F. M. Awaysheh, and F. Palomba, "Fedcsd: A federated learning based approach for code-smell detection," *arXiv preprint arXiv:2306.00038*, 2023.

[5] A. Ait-Mlouk, S. A. Alawadi, S. Toor, and A. Hellander, "Fedqas: privacy-aware machine reading comprehension with federated learning," *Applied Sciences*, vol. 12, no. 6, p. 3130, 2022.

[6] S. Alawadi, V. R. Kebande, Y. Dong, J. Bugeja, J. A. Persson, and C. M. Olsson, "A federated interactive learning iot-based health monitoring platform," in *European Conference on Advances in Databases and Information Systems*. Springer, 2021, pp. 235–246.

[7] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[8] K. Hsieh, A. Phanishayee, O. Mutlu, and P. B. Gibbons, "The non-iid data quagmire of decentralized machine learning," *CoRR*, vol. abs/1910.00189, 2019. [Online]. Available: http://arxiv.org/abs/1910.00189

[9] Q. Li, Y. Diao, Q. Chen, and B. 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.

[10] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-iid data: A survey," *Neurocomputing*, vol. 465, pp. 371–390, 2021.

[11] H. Xuming, G. Minghan, W. Limin, H. Zaobo, and W. Yanze, "A survey of federated learning on non-iid data," *ZTE Communications*, vol. 20, no. 3, p. 17, 2022.

[12] M. Duan, D. Liu, X. Chen, Y. Tan, J. Ren, L. Qiao, and L. Liang, "Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications," in *2019 IEEE 37th International Conference on Computer Design (ICCD)*, 2019, pp. 246–254.

[13] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *CoRR*, vol. abs/1806.00582, 2018. [Online]. Available: http://arxiv.org/abs/1806.00582

[14] D. Chiaro, E. Prezioso, M. Ianni, and F. Giampaolo, "Fl-enhance: A federated learning framework for balancing non-iid data with augmented and shared compressed samples," *Information Fusion*, vol. 98, p. 101836, 05 2023.

[15] F. M. Awaysheh, "From the cloud to the edge towards a distributed and light weight secure big data pipelines for iot applications," in *Trust, Security and Privacy for Big Data*. CRC Press, 2022, pp. 50–68.

[16] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," *CoRR*, vol. abs/1804.08333, 2018. [Online]. Available: http://arxiv.org/abs/1804.08333

[17] E. Seo, D. Niyato, and E. Elmroth, "Resource-efficient federated learning with non-iid data: An auction theoretic approach," *IEEE Internet of Things Journal*, vol. 9, no. 24, pp. 25 506–25 524, 2022.

[18] S. Wiedemann, K.-R. Müller, and W. Samek, "Compact and computationally efficient representation of deep neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, pp. 772–785, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:44107076

[19] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," *CoRR*, vol. abs/2004.11791, 2020. [Online]. Available: https://arxiv.org/abs/2004.11791

[20] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[21] X. Yuan and P. Li, "On convergence of fedprox: Local dissimilarity invariant bounds, non-smoothness and beyond," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 752–10 765, 2022.

[22] A. K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith, "On the convergence of federated optimization in heterogeneous networks," *CoRR*, vol. abs/1812.06127, 2018. [Online]. Available: http://arxiv.org/abs/1812.06127

[23] P. Qi, D. Chiaro, A. Guzzo, M. Ianni, G. Fortino, and F. Piccialli, "Model aggregation techniques in federated learning: A comprehensive survey," *Future Generation Computer Systems*, vol. 150, 09 2023.

[24] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5650–5659.

[25] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[26] M. Asad, A. Moustafa, and T. Ito, "Fedopt: Towards communication efficiency and privacy preservation in federated learning," *Applied Sciences*, vol. 10, no. 8, p. 2864, 2020.

[27] Z. Huo, Q. Yang, B. Gu, L. C. Huang *et al.*, "Faster on-device training using new federated momentum algorithm," *arXiv preprint arXiv:2002.02090*, 2020.

[28] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," *arXiv preprint arXiv:2003.00295*, 2020.

[29] D. Y. Park, M. Cha, C. Jeong, D. Kim, and B. Han, "Learning student-friendly teacher networks for knowledge distillation," *CoRR*, vol. abs/2102.07650, 2021. [Online]. Available: https://arxiv.org/abs/2102.07650

[30] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *CoRR*, vol. abs/2006.07242, 2020. [Online]. Available: https://arxiv.org/abs/2006.07242

[31] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," *CoRR*, vol. abs/1910.03581, 2019. [Online]. Available: http://arxiv.org/abs/1910.03581

[32] W. Ouyang, X. Wang, C. Zhang, and X. Yang, "Factors in finetuning deep model for object detection with long-tail distribution," 06 2016, pp. 864–873.

[33] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, T. N. Hoang, and Y. Khazaeni, "Bayesian nonparametric federated learning of neural networks," 2019.

[34] H. Wang, M. Yurochkin, Y. Sun, D. S. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," *CoRR*, vol. abs/2002.06440, 2020. [Online]. Available: https://arxiv.org/abs/2002.06440

[35] M. F. Criado, F. E. Casado, R. Iglesias, C. V. Regueiro, and S. Barro, "Non-iid data and continual learning processes in federated learning: A long road ahead," *CoRR*, vol. abs/2111.13394, 2021. [Online]. Available: https://arxiv.org/abs/2111.13394

[36] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão *et al.*, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.