

# An Experimental Study of Class Imbalance in Federated Learning

1<sup>st</sup> Chenguang Xiao

*School of Computer Science*

*University of Birmingham*

Edgbaston, Birmingham B15 2TT

Email: cxx075@student.bham.ac.uk

2<sup>nd</sup> Shuo Wang

*School of Computer Science*

*University of Birmingham*

Edgbaston, Birmingham B15 2TT

Email: S.Wang.2@bham.ac.uk

**Abstract**—Federated learning is a distributed machine learning paradigm that trains a global model for prediction based on several local models at clients while local data privacy is preserved. Class imbalance is believed to be one of the factors that degrades the global model performance. However, there has been very little research on if and how class imbalance can affect the global performance in various imbalance scenarios. Class imbalance in federated learning is much more complex than that in traditional non-distributed machine learning, due to different class imbalance situations at local clients. Class imbalance needs to be re-defined in distributed learning environments, so that corresponding solutions can be proposed. In this paper, first, we propose two new metrics to define class imbalance – the global class imbalance degree (MID) and the local difference of class imbalance among clients (WCS). Class imbalance is categorized into four scenarios under the definition. Then, we conduct extensive experiments to analyze the impact of class imbalance on the global performance in various scenarios. Our results show that a higher MID and a larger WCS degrade more the performance of the global model. Besides, WCS is shown to slow down the convergence of the global model by misdirecting the optimization.

**Index Terms**—class imbalance, federated learning, multiclass classification

## I. INTRODUCTION

As the rapid development of advanced computing hardware and machine learning algorithms, edge computing and ubiquitous computing systems have sprung forth. Local devices, such as mobile phones and wearable devices, have become a major source of data [1]–[3]. A large number of devices are interconnected and are equipped with sensors that constantly generate potentially useful data [4]. To learn from such data, local data from clients have to be gathered together. However, these local generated data tend to contain sensitive information, such as end users’ personal information and clients’ medical records. Transmitting data among devices directly can cause privacy leakage and security issues. Traditional machine learning approaches that collect and centralize user data will become legally impossible. Federated learning was thus proposed to learn from data, protect data privacy and improve network security [4].

Federated learning trains a global model at the central server based on a group of local models trained and maintained at the clients [5], [6]. Instead of transmitting data to the

central server, only intermediate local model updates are communicated periodically with the server. In each round of training, the central server selects a group of clients and broadcasts the current global model to them. Then, the selected clients train the received model using their local data and feedback the model updates to the server. Lastly, the central server aggregates the updates from the clients. This iterative training process continues across the network [6].

Federated learning is able to make use of local data for training without violating privacy or breaking data island between clients. Currently, federated learning has been deployed by major service providers and plays a critical role in privacy-sensitive applications [2]. While most research in federated learning focuses on reducing communication loads and protecting data privacy [1], [2], [5], little work has looked into how local data can affect the performance of the global model. Data from different clients are always not independently and identically distributed (Non-IID). Class imbalance is a type of non-iid distributions in federated learning environments. In a classification task, such as fraudulent phone call detection and diagnoses of rare diseases, class imbalance refers to the situation where some classes of data (minority) are significantly under-represented compared to other classes (majority). In traditional non-distributed machine learning, class imbalance can cause great performance degradation, especially the poor accuracy on minority classes [7], [8].

Class imbalance is also common in federated learning. For example, in real-world health data, severe class imbalance is a norm, rather than an exception [9]. However, it is unclear if and how local class imbalance in federated learning can affect the performance of the global model. Compared with the traditional non-distributed learning case, class imbalance in federated learning is much more complex. Local models and their regular updates could be affected by local imbalanced data, but it is unclear how many affected clients or how the severity of local imbalance may cause a global degradation. Meanwhile, the class imbalance status between clients can vary. For example, a client A has a minority class  $c_1$  and a majority class  $c_2$ , but a client B has class  $c_1$  as majority and class  $c_2$  as minority. Therefore, various class imbalance in federated learning should be categorised properly for potential solution to tackle them. This paper aims at providing a full

understanding of the impact of class imbalance in federated learning that will shed lights on suitable solutions of tackling class imbalance in federated learning. We will answer three specific questions in this paper:

- 1) How should we define class imbalance in federated learning? This will include a global class imbalance degree (applicable to multi-class data) and the differences of class imbalance between clients.
- 2) Would the local and global class imbalance affect the performance of the global model and how?
- 3) Would the class imbalance difference between clients affect the performance of the global model and how?

The contribution of this paper are listed below:

- We propose two new metrics to measure class imbalance in federated learning: Global Imbalance Degree using Multiclass Imbalance Degree (MID) and Local and Global Imbalance Relation using Weighted Cosine Similarity (WCS).
- Based on the new definition, we conduct extensive experiments on real-world datasets to investigate the impact of class imbalance. Four different scenarios are considered. Results show that the global class imbalance degrades the global model performance. The difference of local class imbalance also causes global performance degradation and slows down the model convergence.

The rest of this paper are organized as below. Section II presents related works. We define class imbalance in federated learning in Section III. Section IV presents 4 class imbalance scenarios to be investigated and describes the datasets used in our experiments. Section V provides the experimental analysis. We conclude this paper and discuss the possible future work in Section VI.

## II. RELATED WORKS

There is extensive investigation into class imbalance in non-distributed machine learning [10], [11]. The impact of class imbalance depends on the imbalance level, concept complexity and size of training data [7]. Lack of information caused by small sample size, class overlapping, and small disjuncts within class are main reasons of class imbalance causing performance degradation [11]. To tackle different types of class imbalance, the traditional approaches can be classified into five groups – sampling approaches, re-weighting approaches, feature selection, one class learning, cost-sensitive learning [12] and ensemble learning [13].

The nature of federated learning makes it different from non-distributed machine learning when dealing with class imbalance. In federated learning, class imbalance presented at local data may or may not result in global class imbalance when the central server aggregates the model updates. Therefore, we need to separately define and discuss local class imbalance and global class imbalance. Furthermore, the local and global class imbalance can be totally different [14]. A majority class at some clients can be the minority class at the global level, and vice versa. The privacy protection of

federated learning further increases the difficulty of estimating the class imbalance degree at the central server. As a result, existing class imbalance mitigating approaches can be only used locally, which may not help with global performance.

A few very recent papers [14]–[17] have noticed the negative impact of class imbalance in federated learning and proposed techniques to tackle it. **Fed-Focal Loss** [16] used a modified loss function that down-weights the loss of well-classified samples based on Binary Cross Entropy (BCE) Loss. By doing so, the majority class with a larger number of examples contributes less to the model when it reaches high prediction accuracy. Correspondingly, the minority-class samples contribute more to the local model. **Astraea** [15] added mediators between the central server and clients to re-balance the datasets. Imbalanced clients are rearranged to different balance mediators according to their imbalance levels and label distributions. Within the mediator, the clients perform training sequentially on a single balanced dataset. Then, the mediators communicate with the central server in parallel as clients in original federated learning. **Ratio Loss** [14] employed a monitor scheme on the server to estimate local class imbalance without asking for label distributions. The monitoring scheme uses the relation between the gradient magnitude and the sample quantity to estimate global class imbalance at the server. Then, Ratio Loss based on BCE Loss is deployed at the local training process to strengthen the impact of minority-class examples. Similarly, Yang et al. [17] proposed a local class imbalance estimator based on gradient magnitude. Then clients selection is used to achieve class balance globally.

The aforementioned papers proposed new techniques to tackle class imbalance in specific setting, which shows the necessity of studying the class imbalance issue in federated learning. However, they all treated a small part of class imbalance scenario as the full picture of class imbalance in federated learning. Astraea was only valid with slightly global imbalance, while Ratio Loss can do nothing with a balanced global dataset consisted of imbalance local datasets. Ratio Loss addressed the impact of mismatch between local and global imbalance, while the conclusion is drawn based on particular experiments setting without considering more general case with other imbalance degree. In addition, Metrics used in those works include imbalance ratio, cosine similarity cannot fully reflect the imbalance states. Likelihood-ratio imbalanced degree (LRID) [18] failed to measure the multiclass imbalance degree as well. The comparison between those metrics and ours will be discussed in Section III. This paper thus aims at an in-depth understanding of class imbalance in various federated learning scenarios, which will help to develop the most suitable solutions in the future.

## III. CLASS IMBALANCE DEFINITION

Local imbalance and global imbalance were briefly mentioned in [14] as two types of class imbalance in federated learning. Their experiment on dedicated setting showed that a mismatch between local and global imbalance leads to global model performance degradation. However, there is no clear

definition to measure the global class imbalance degree and the relation between local and global imbalance. In this section, we propose two metrics to define class imbalance status in federated learning environments.

### A. Federated Learning Problem Formulation

Assume there are  $P$  clients with local dataset  $D_1, \dots, D_P$  in a size of  $n_1, \dots, n_P$  respectively. If merging them together, the global dataset  $D$  has  $C$  classes and  $N$  samples in total. At global time  $t$ , the global model is denoted as  $w^t$ . The selected client  $p$  performs local training to derive a new local model  $w_p^{t+1}$  by:

$$w_p^{t+1} = w_p^t - \nabla L(w_p^t, D_p) \quad (1)$$

where  $L(w^t, D_p)$  denotes the loss of model  $w^t$  on dataset  $D_p$ . The update global model following the FedAvg [4] will be:

$$w^{t+1} = w^t - \sum_{i=1}^P \frac{n_i}{N} \nabla L(w^t, D_i) \quad (2)$$

### B. Global Imbalance Degree.

To measure the global imbalance degree, we are inspired by (LRID) [18]. It was designed to measure class imbalance level in multi-class datasets. The commonly used Imbalance Ratio (denoted as  $\Gamma$  below), referred to as the ratio between the numbers of the majority and minority classes, cannot fully describe the imbalance status in multi-class data as only two classes are considered. For a dataset with  $N$  data samples and  $C$  possible classes, the number of samples with label  $c$  is  $n_c$ . The class imbalance level according to [18] is defined as:

$$LRID = -2 \sum_{c=1}^C n_c \ln \frac{N}{C n_c} \quad (3)$$

The  $LRID$  of an absolutely class balanced dataset is 0. The larger  $LRID$  is, the more class imbalanced the dataset is. However,  $LRID$  is sensitive to the size of datasets. For two datasets  $D_1$  and  $D_2$  where  $D_2$  contains exactly  $k$  times the samples of  $D_1$  for each class,  $LRID_2$  based on Equation (3) becomes  $-2k \sum_{c=1}^C n_c \ln \frac{N}{C n_c}$ , which is  $k$  times larger than that of  $D_1$  as Equation (3) even though the proportion of each class remains the same.

Given an extreme case where a  $C$ -class dataset contains  $[N, 0, \dots, 0]$  samples for each class,  $LRID_{extreme} = 2N \log C$  according to Equation (3). This  $LRID$  value changes with the total sample number  $N$ , which is misleading as a measure of class imbalance. Therefore we improve LRID and propose Multiclass Imbalance Degree (MID):

$$MID = \frac{LRID}{LRID_{extreme}} = \sum_{c=1}^C \frac{n_c}{N} \log_C \frac{C n_c}{N} \quad (4)$$

MID eliminates the impact of the size of dataset and ranges between 0 and 1. MID equal to 0 implies a strictly balanced dataset. The larger the MID, the more imbalanced the dataset is. In our experiments, we use  $MID$  to express how class imbalanced the global dataset  $D$  is.

### C. Local and Global Imbalance Relation.

Mean cosine similarity (MCS) has been used to evaluate the mismatch between local and global imbalance [14]. Cosine similarity of vector  $A$  and  $B$  is defined as

$$similarity(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \quad (5)$$

where  $\theta$  is the angle between  $A$  and  $B$  and  $\|A\|$  denotes the L2 norm of vector  $A$ . Label distribution vector of client  $j$  is  $l_j = [n_j^1, \dots, n_j^c, \dots, n_j^C]$  where  $n_j^c$  is the number of samples with label  $c$ . The Global label distribution vector is  $L = [\sum_{i=1}^P n_i^1, \dots, \sum_{i=1}^P n_i^c, \dots, \sum_{i=1}^P n_i^C]$ . Mean cosine similarity averages the similarity of global label distribution vector  $L$  and local label distribution vector  $l$  as below:

$$MCS = \frac{1}{P} \sum_{i=1}^P \frac{L \cdot l_i}{\|L\|_2 \|l_i\|_2} \quad (6)$$

It treats all clients equally and does not consider the sample size, which can be misleading. For example, a two-client federated dataset with label distribution vectors  $l_1 = [100, 99]$  (client1) and  $l_2 = [0, 1]$  (client2) has  $MCS = 1/2(\frac{L \cdot l_1}{\|L\|_2 \|l_1\|_2} + \frac{L \cdot l_2}{\|L\|_2 \|l_2\|_2}) = 0.853$  according to Equation (6). However, the similarity of global and local class imbalance should be nearly 1 as client2 contributes little to the global model. In this case, the small local dataset with extreme class imbalance leads to a biased estimation of the local and global imbalance relation when using MCS. Therefore we propose Weighted Cosine Similarity (WCS) to measure the relationship between local and global imbalance that considers the contribution of local datasets. WCS is defined as:

$$\begin{aligned} WCS &= \sum_{i=1}^P \frac{\|l_i\|_1}{\|L\|_1} similarity(L, l_i) \\ &= \sum_{i=1}^P \frac{\|l_i\|_1 L \cdot l_i}{\|L\|_1 \|L\|_2 \|l_i\|_2} \\ &= \frac{1}{\|L\|_1 \|L\|_2} \sum_{i=1}^P \frac{\|l_i\|_1}{\|l_i\|_2} L \cdot l_i \end{aligned} \quad (7)$$

$\|l_i\|_1$  denotes the total number of samples of client  $j$ , the same as  $\sum_{i=1}^C n_j^i$ .

For example, a federated learning network has 3 clients with label distribution vectors  $l_1 = [2, 0, 0]$ ,  $l_2 = [0, 4, 0]$ , and  $l_3 = [0, 0, 6]$ . As shown in Fig. 1,  $L = [2, 4, 6]$ . Follow Equation (7),  $L = [2, 4, 6]$ . Let  $\alpha_i$  be the angle between  $L$  and  $l_i$ , then we have

$$WCS = \frac{2 \cos \alpha_1 + 4 \cos \alpha_2 + 6 \cos \alpha_3}{2 + 4 + 6} = 0.62$$

Consider the extreme circumstance when the all distribution vectors  $l_1, l_p$  are in same direction, we have  $\alpha_i = 0 (i \in \{1, \dots, P\})$  and  $WCS = 1$ . When the label distribution vectors are completely in different directions  $l_i l_j = 0 (i \neq j)$  and same length,  $WCS = \frac{1}{\sqrt{C}}$ , which is also the minimum of  $WCS$ .

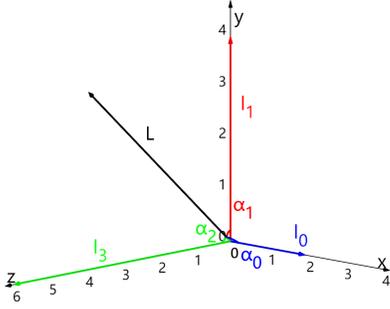


Fig. 1. Weighted Cosine Similarity

#### IV. CLASS IMBALANCED SCENARIOS AND DATA GENERATION

We use two metrics to define the class imbalance degree in Section III –  $MID$  and  $WCS$ . Based on the definitions, there exist 4 kinds of class imbalance scenarios.

- 1)  $MID = 0, WCS = 1$  (scenario 1): the global data is strictly class balanced; all of the local label distribution vectors follow the same direction..
- 2)  $MID > 0, WCS = 1$  (scenario 2): the global data presents to be class imbalanced; all of the local label distribution vectors follow the same direction.
- 3)  $MID = 0, WCS < 1$  (scenario 3): the global data is strictly class balanced; the local label distribution vectors present discrepancy in directions.
- 4)  $MID > 0, WCS < 1$  (scenario 4): the global data presents to be class imbalanced; the local label distribution vectors present discrepancy in directions.

When the local label distribution vectors presents discrepancy in directions, it means that some local datasets are class imbalanced. In the following experiments, we will discuss how the global model performs in these four scenarios, in particular the last three scenarios when either global or local data are class imbalanced.

##### A. Dataset Description and Preprocessing

We select three popular datasets to simulate the three class imbalanced scenarios (scenarios 2-4) as identified above.

**MNIST** [19] is a handwriting digits dataset with 60000 train samples and 10000 test samples. Each data point consists of  $28 \times 28$  gray pixels with a label range from 0 to 9. This dataset is split to 100 clients.

**FEMNIST** [20] is a federated version of MNIST dataset with 341873 samples from 3383 writers which is clients in federated learning. With distinct writing styles, the data from different writers is non-iid. The box plot in Fig. 2 shows the number of samples from different classes. As shown in this figure, the FEMNIST dataset is nearly globally class balanced, while local imbalance exists among clients due to the outliers. Each client in original FEMNIST dataset contains around 100 samples from 10 classes, which limits the experiments on extreme class imbalance when the local class imbalance ratio  $\Gamma$  which is the ratio of majority and minority classes

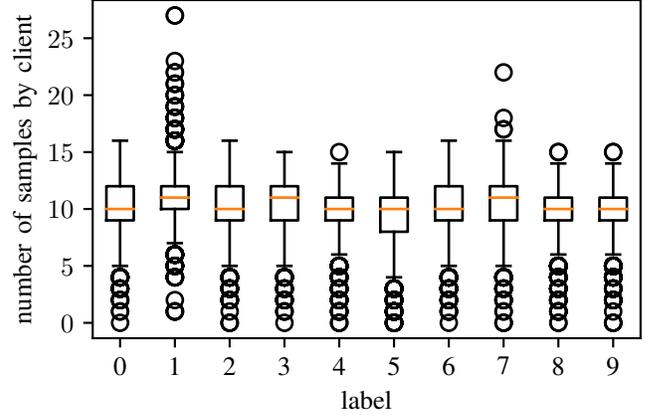


Fig. 2. Nature of FEMNIST Dataset by Class

exceeds 10 : 1. Therefore, 3383 clients of FEMNIST dataset are divided into 100 clients with a larger size local dataset that allows much more extreme local class imbalance degree.

**CIFAR10** [21] dataset contains 50000  $32 \times 32$  colour images in 10 classes. Each class has 5000 images. Similarly to MNIST, CIFAR10 was originally a centralized dataset. We split it to 100 clients randomly for federated learning.

All the three datasets are globally balanced with 10 classes (labels 0 to 9). For MNIST and CIFAR10 datasets, each client contains the same number of samples for every class, i.e. locally balanced. FEMNIST dataset are nearly class balanced as each class contains around 10 samples per client.

##### B. Global Class Imbalance – Scenario 2

To investigate the impact of global class imbalance, we let all the local datasets have the same class imbalance degree. To simulate more general case as in Focal Loss [16], 4 classes are randomly selected as the minority classes (0, 1, 3, 6 classes). Given a local imbalance ratio  $\Gamma$ , we randomly samples  $\frac{1}{\Gamma}$  of data points of classes 0, 1, 3, 6 and keep all the data from classes 2, 4, 5, 7, 9. For the FEMNIST dataset,  $\Gamma$  is set to 10 : 1, 30 : 1, and 100 : 1 respectively. When  $\Gamma = 10 : 1$ , every local dataset has only 3 data samples for classes 0, 1, 3, and 6. To make sure the local datasets contain samples from all the classes,  $L$  is set to 10 : 1, 20 : 1, 60 : 1 for MNIST dataset, and 10 : 1, 20 : 1, 50 : 1 for CIFAR10 dataset.

##### C. Local imbalance – Scenario 3

To simulate local class imbalance for MNIST and CIFAR10 datasets, We randomly select  $S$  classes out of 10 for every client and evenly assign samples from the selected classes to that client. For example, CIFAR10 has 5000 training samples for each class. If  $S$  is set to 2, every client contains 2 classes with 250 data points from each class. For FEMNIST, it is globally class balanced and contains class imbalanced local datasets as shown in Fig. 2. Instead of assigning all the samples to the clients, samples for classes that has not been selected are dropped from the balanced FEMNIST dataset.

TABLE I  
SIMULATED CLASS IMBALANCED DATA FOR SCENARIOS 1-4

Dataset	Scenario	$\Gamma$	S	LRID	MID	WCS		
MNIST	1	1:1	10	171	0	1		
		10:1	10	22650	0.13	1		
		20:1	10	27758	0.17	1		
	2	60:1	10	32458	0.2	1		
		1:1	5	171	0	0.71		
		1:1	2	171	0	0.45		
	3	1:1	1	171	0	0.32		
		10:1	2	22650	0.13	0.49		
		20:1	2	27758	0.17	0.49		
	4	60:1	2	32458	0.2	0.5		
		CIFAR10	1	1:1	10	0	1	
				10:1	10	19352	0.13	1
20:1	10			24696	0.17	1		
2	50:1		10	27123	0.19	1		
	1:1		5	0	0	0.71		
	1:1		2	0	0	0.45		
3	1:1		1	0	0	0.32		
	10:1		2	19352	0.13	0.48		
	20:1		2	23647	0.17	0.5		
4	50:1		2	27123	0.19	0.49		
	FEMNIST		1	1:1	10	761	0	1
				10:1	10	129292	0.13	1
30:1		10		171270	0.18	1		
2		100:1	10	193238	0.21	1		
		1:1	5	392	0	0.71		
		1:1	2	147	0	0.45		
3		1:1	1	98	0	0.32		
		10:1	2	26479	0.13	0.48		
		30:1	2	34253	0.18	0.5		
4		100:1	2	38650	0.21	0.51		

#### D. Global and Local Imbalance – Scenario 4

It is more common to have class imbalance at both local and global levels. We combine the data simulation steps from both Scenarios 2 and 3. Given an imbalance ratio  $\Gamma$  between majority and minority classes, for MNIST and CIFAR10, data samples from classes 0, 1, 3, 6 are sampled at a rate of  $\Gamma$  while data from the other classes are all kept. Then the obtained imbalanced data is distributed to 100 clients with  $S$  classes at each client. For FEMNIST, We randomly select  $S$  classes for each client. For every client, if the selected class belong to one of the minority classes 0, 1, 3 and 6, the data is downsampled at the rate of  $\frac{1}{\Gamma}$ . The samples from non-selected classes are dropped.

#### E. Summary of Generated Class Imbalanced Data

In summary, we have generated 10 datasets that covers all four class imbalanced scenarios, based on each of the MNIST, FEMNIST and CIFAR10 datasets. Table I summarizes all the cases.  $\Gamma$  presents the global class imbalance.  $MID$  and  $WCS$ , as defined in the previous section, show the global class imbalance degree and the mismatch between local and global imbalance.

From Table I, when the majority and minority classes are fixed,  $MID$  follow the trend the global class imbalance ration  $\Gamma$ .  $MID$  is more informative than  $\Gamma$ , especially When there are not only majority and minority classes but also other classes

TABLE II  
GLOBAL PERFORMANCE OF SCENARIOS 1 (BASELINE) AND 2

Dataset	MID	Accuracy	F1
MNIST	0	0.9893	0.9892
	0.13	0.9778	0.9777
	0.17	0.9686	0.9682
	0.2	0.9350	0.9326
CIFAR10	0	0.6254	0.6224
	0.13	0.4781	0.4247
	0.17	0.4389	0.3491
FEMNIST	0.19	0.4274	0.3264
	0	0.9916	0.9916
	0.13	0.9850	0.9849
	0.18	0.9772	0.9769
	0.21	0.9539	0.9532

TABLE III  
GLOBAL PERFORMANCE OF THE 2<sup>nd</sup> CASE IN SCENARIO 3 (BASELINE) AND ALL CASES IN SCENARIO 4

Dataset	MID	WCS	Accuracy	F1
MNIST	0	0.45	0.9704	0.9703
	0.13	0.49	0.9312	0.9300
	0.17	0.49	0.9241	0.9227
	0.2	0.5	0.8598	0.8532
CIFAR10	0	0.45	0.4164	0.3880
	0.13	0.48	0.3315	0.2673
	0.17	0.5	0.3083	0.2268
	0.19	0.49	0.3066	0.2223
FEMNIST	0	0.45	0.8380	0.8323
	0.13	0.48	0.8247	0.8209
	0.18	0.5	0.6635	0.6337
	0.21	0.51	0.5375	0.4661

between them. Compared with  $LRID$ ,  $MID$  is more stable with varies size of datasets.

## V. EXPERIMENTS

We adopt FedAvg [4] algorithm to train a convolution neural networks (CNN) as the global model for each dataset. As MNIST, FEMNIST and CIFAR10 are all image datasets, we use the same CNN structure setting for all cases – two convolutional layers followed by two dense layers. At each round of training, 10 out of 100 clients are randomly selected to participate [16]. SGD is used in the local training as the optimizer with local learning rate equal to 0.1. The server’s learning rate is set to 1.0 following the recommended value of TensorFlow Federated framework. Each global model is trained for 50 iterations with 5 local epochs every iteration at a batch size of 128. All three datasets can converge with this setting in scenario 1. In our experiment, the whole training process is repeated 15 times to get an average result. As a class imbalance problem, we add F1-Score as an implement metrics to the overall accuracy to measure the model performance.

#### A. Impact of Global Class Imbalance Degree

Table II and Fig. 3 compare the performance of the global model in scenario 1 (baseline case) and scenario 2, where  $WCS = 1$  and  $MID$  varies between  $[0, 0.21]$ . Table II shows the final predictive accuracy and F1-Score. Fig. 3 shows the

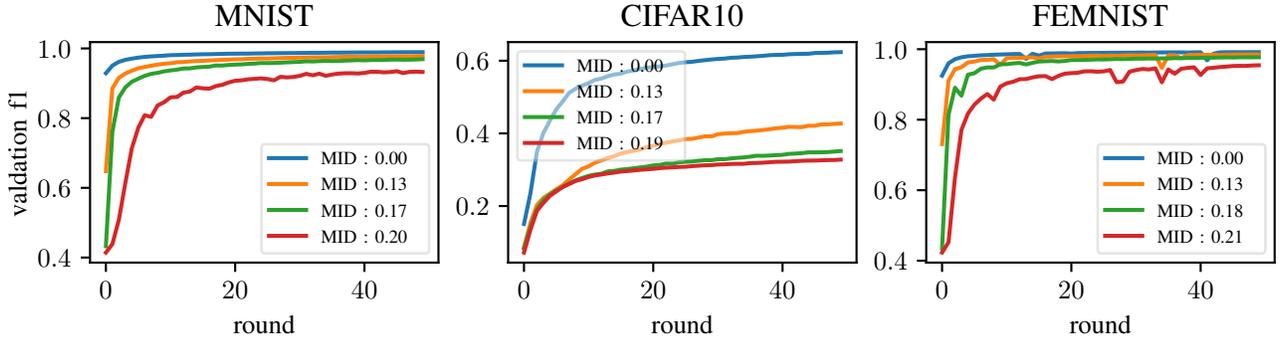


Fig. 3. Validation F1-Score of Scenarios 1 (baseline) and 2

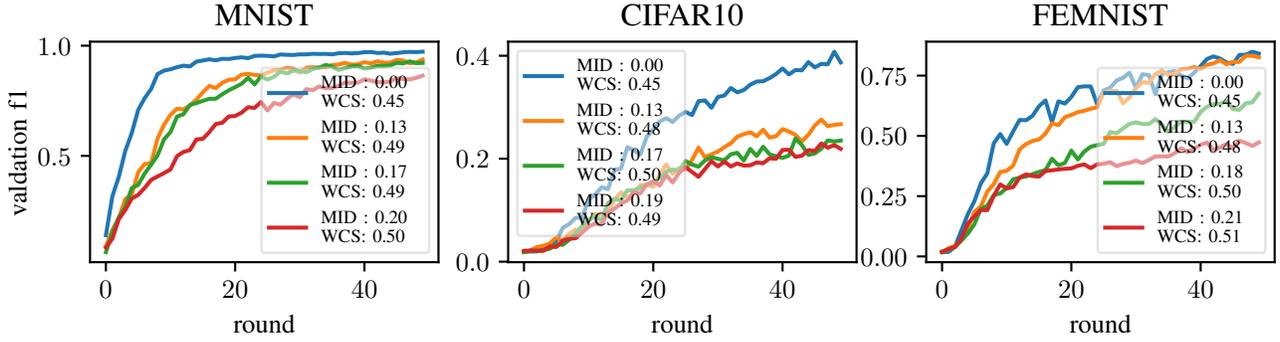


Fig. 4. Validation F1-Score of the 2<sup>nd</sup> case in scenario 3 (baseline) and all cases in Scenario 4

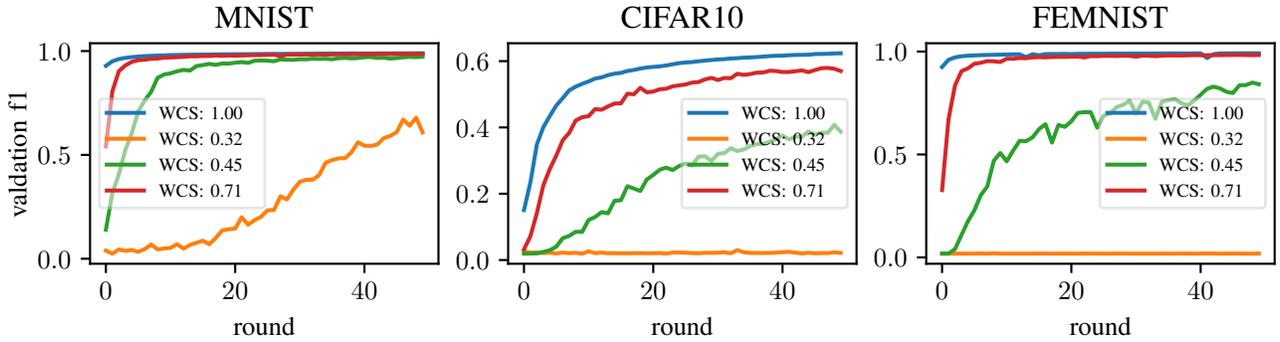


Fig. 5. Validation F1-Score of Scenarios 1 (baseline) and 3

F1-Score curves along with training. The accuracy curves are very similar to the F1-Score ones, so they are omitted from the figure for the space reason. The results tell us how the global class imbalance degree impacts the global performance. We can observe a decrease of accuracy and F1-Score with the increase of global class imbalance degree  $MID$  on all three datasets. Table III shows the accuracy and F1-Score of the global model in the second case from scenario 3 (baseline case) and all cases from scenario 4 with  $S = 2$ , where  $WCS \approx 0.5$  and  $MID$  varies between  $[0, 0.21]$ . The corresponding F1-Score curves are presented in Fig. 4. They

also show that a larger global class imbalance degree reduces the global performance significantly.

### B. Impact of Local Class Imbalance Difference

The difference of local class imbalance is a feature in federated learning that distinguishes itself from class imbalance problems in centralized machine learning. Table IV shows the accuracy and F1-Score of the global model in scenario 1 (baseline case) and scenario 3, where  $MID$  remains 0 and  $WCS$  varies between  $[0.32, 1]$ . The corresponding F1-Score curves are presented in Fig. 5. We observe that a smaller similarity between local class imbalance results in degradation

TABLE IV  
GLOBAL PERFORMANCE OF SCENARIOS 1 (BASELINE) AND 3

Dataset	WCS	Accuracy	F1
MNIST	1	0.9893	0.9892
	0.71	0.9863	0.9862
	0.45	0.9704	0.9703
	0.32	0.6718	0.6416
CIFAR10	1	0.6254	0.6224
	0.71	0.5830	0.5757
	0.45	0.4164	0.3880
	0.32	0.1026	0.0214
FEMNIST	1	0.9916	0.9916
	0.71	0.9833	0.9831
	0.45	0.8380	0.8323
	0.32	0.1012	0.0184

of the global model performance. Besides, the decrease of  $WCS$  results in a larger fluctuation on the performance curves especially in CIFAR10 and FEMNIST datasets as shown in Fig. 5 and Fig. 4 in comparison with Fig. 3. As a result, the convergence of the global model is significantly slowed down. In scenario 3 when  $WCS = 0.32$  for CIFAR10 and FEMNIST (in the middle and right plots of Fig. 5), the global model cannot even converge.

When comparing Scenario 4 (where  $MID > 0$  and  $WCS < 1$ ) with Scenario 2 (where  $WCS = 1$ ), we can see that a large difference of local class imbalance not only causes the degradation of the global model performance, but also introduces performance fluctuation and slows down the convergence of the global model. In short, by reducing  $WCS$  (i.e. a larger difference), the global model becomes more difficult to converge and suffers worse prediction accuracy.

## VI. CONCLUSIONS

This paper investigates the impact of class imbalance in federated learning. We focus on three research questions: Q1. define class imbalance in federated learning. Q2. explore the impact of global class imbalance on the global model. Q3. explore the impact of imbalance differences between local clients on the global model.

For Q1, we proposed two new metrics – MID and WCS. MID measures the global class imbalance degree. It improves the traditional Imbalance Ratio and LRID, which is suitable to multi-class data and is insensitive to the size of datasets. WCS is specifically designed for federated learning that measures the class imbalance differences among local clients and considers the contributions of local datasets. Based on MID and WCS, we looked into 4 class imbalanced scenarios to answer Q2 and Q3. For Q2, we found that a larger MID leads to more significant degradation of the global performance in terms of prediction accuracy and F1-Score. For Q3, we showed that a large difference of class imbalance degree among local datasets not only reduces the global performance, but also slows down the convergence by introducing fluctuation in optimization.

This work suggests that 1) class imbalance in federated learning should be studied separately considering MID and

WCS, 2) global class imbalance should be studied and tackled appropriately in federated learning for better global model performance, and 3) the differences of local class imbalance should also be treated seriously that could affect the global performance and convergence speed.

## REFERENCES

- [1] V. Smith, C. K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," in *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Nips, 2017, pp. 4425–4435.
- [2] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.
- [3] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: vision, hype and reality for data privacy and protection," *arXiv preprint arXiv:1907.09693*, 2019.
- [4] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, vol. 54, 2017.
- [5] 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," *arXiv preprint arXiv:1912.04977*, 2019.
- [6] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.
- [7] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.
- [8] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem," *Int. J. Advance Soft Compu. Appl*, vol. 5, no. 3, 2013.
- [9] O. Choudhury, Y. Park, T. Salonidis, A. Gkoulalas-Divanis, I. Sylla *et al.*, "Predicting adverse drug reactions on distributed health data using federated learning," in *AMIA Annual symposium proceedings*, vol. 2019. American Medical Informatics Association, 2019, p. 313.
- [10] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, vol. 4, pp. 192–201, 2008.
- [11] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *International Journal of Advances in Soft Computing and its Applications*, vol. 7, no. 3, pp. 176–204, 2015.
- [12] C. X. Ling and V. S. Sheng, "Cost-sensitive learning and the class imbalance problem," *Encyclopedia of machine learning*, vol. 2011, pp. 231–235, 2008.
- [13] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539–550, 2009.
- [14] L. Wang, X. Wang, S. Xu, and Q. Zhu, "Towards class imbalance in federated learning," *arXiv*, 2020.
- [15] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan, and L. Liang, "Self-Balancing Federated Learning with Global Imbalanced Data in Mobile Systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 1, pp. 59–71, 2021.
- [16] D. Sarkar, A. Narang, and S. Rai, "Fed-focal loss for imbalanced data classification in federated learning," *arXiv*, 2020.
- [17] M. Yang, A. Wong, H. Zhu, H. Wang, and H. Qian, "Federated learning with class imbalance reduction," *arXiv preprint arXiv:2011.11266*, 2020.
- [18] R. Zhu, Z. Wang, Z. Ma, G. Wang, and J. H. Xue, "LRID: A new metric of multi-class imbalance degree based on likelihood-ratio test," *Pattern Recognition Letters*, vol. 116, pp. 36–42, 2018. [Online]. Available: <https://doi.org/10.1016/j.patrec.2018.09.012>
- [19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [20] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.
- [21] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009.