# Redefining Contributions: Shapley-Driven Federated Learning

**Nurbek Tastan**, **Samar Fares**, **Toluwani Aremu**, **Samuel Horvath**, **Karthik Nandakumar**
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
{nurbek.tastan, samar.fares, toluwani.aremu, samuel.horvath, karthik.nandakumar}@mbzuai.ac.ae

## Abstract

Federated learning (FL) has emerged as a pivotal approach in machine learning, enabling multiple participants to collaboratively train a global model without sharing raw data. While FL finds applications in various domains such as healthcare and finance, it is challenging to ensure global model convergence when participants do not contribute equally and/or honestly. To overcome this challenge, principled mechanisms are required to evaluate the contributions made by individual participants in the FL setting. Existing solutions for contribution assessment rely on general accuracy evaluation, often failing to capture nuanced dynamics and class-specific influences. This paper proposes a novel contribution assessment method called ShapFed for fine-grained evaluation of participant contributions in FL. Our approach uses Shapley values from cooperative game theory to provide a granular understanding of class-specific influences. Based on ShapFed, we introduce a weighted aggregation method called ShapFed-WA, which outperforms conventional federated averaging, especially in class-imbalanced scenarios. Personalizing participant updates based on their contributions further enhances collaborative fairness by delivering differentiated models commensurate with the participant contributions. Experiments on CIFAR-10, Chest X-Ray, and Fed-ISIC2019 datasets demonstrate the effectiveness of our approach in improving utility, efficiency, and fairness in FL systems. The code can be found at https://github.com/tnurbek/shapfed.

## 1 Introduction

Federated Learning (FL) is a machine learning (ML) paradigm that enables training powerful models while respecting data privacy and decentralization. In traditional ML, data centralization poses significant privacy concerns and logistical hurdles. FL, on the other hand, flips this paradigm by allowing multiple participants or edge devices to collaborate without sharing their raw data [McMahan *et al.*, 2017; Li *et al.*, 2020; Tastan and Nandakumar, 2023; Wei *et al.*, 2020]. Since only model updates are exchanged in FL, the privacy of local data sources is preserved. This approach has found applications in domains such as healthcare and finance.

In a typical FL environment, all the participants are assumed to collaborate honestly and contribute equally. The convergence and utility of a global model in FL can be hindered when this assumption is not met. When some participants intentionally or unintentionally introduce biases or skewed data distributions into the training process, it negatively impacts the overall model's performance. To address this issue, current solutions [Jiang *et al.*, 2023; Lv *et al.*, 2021; Shi *et al.*, 2022; Siomos and Passerat-Palmbach, 2023; Wang *et al.*, 2020; Xu *et al.*, 2021] primarily rely on evaluating each participant's individual accuracy on an auxiliary validation set to assess their contributions to the system.

However, these solutions often fail to detect subtle but crucial variations in how each party affects the model's performance, especially in scenarios where imbalanced data or class disparities exist. Furthermore, they fail to capture the nuanced dynamics of participant influence on specific class predictions. As a result, there is a pressing need for more refined evaluation methods that go beyond general accuracy assessment and provide a granular understanding of participant contributions to class-specific model performance, ensuring the fairness and effectiveness of FL systems.

In this work, we present a novel approach to appropriately value participant contributions within the context of FL and subsequently allocate rewards. The proposed method employs Shapley values (SVs) from cooperative game theory, integrating them with class-specific performance metrics. Existing contribution assessment methods [Xu *et al.*, 2021; Lyu *et al.*, 2020] utilize gradients from each participant and compute the directional alignment between these gradients as a proxy for contribution evaluation. We show that large-size gradient comparisons exhibit limitations in accurately assessing contributions, particularly in the context of deep neural networks. We make two key assumptions to address these limitations and refine the data valuation process. First, instead of evaluating the cosine similarity across all gradients or model parameters, we focus solely on the gradients of the last layer, providing computational advantages and a more accurate approximation of true Shapley values. Second, rather than just computing the marginal contributions, we calculate class-specific contributions, defining them as a measure of

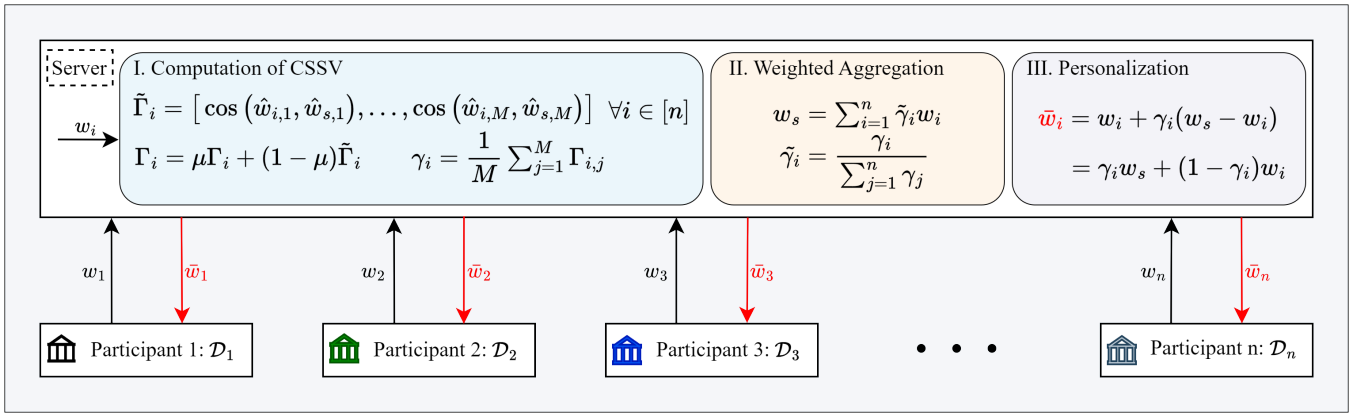



Figure 1: **Overview of our proposed ShapFed algorithm:** Each participant $i$ transmits their locally computed iterates $w_i$ to the server. The server then, (i) computes class-specific Shapley values (CSSVs) using the last layer parameters (gradients) $\hat{w}$ (as illustrated in Figure 2), (ii) aggregates the weights by employing normalized contribution assessment values $\tilde{\gamma}_i$ for each participant $i$, and (iii) broadcasts the personalized weights $\bar{w}_i$ to each participant, using their individual, not-normalized contribution values $\gamma_i$.

heterogeneity. Our contributions are as follows:

- We introduce a novel **contribution assessment** method (ShapFed) that precisely quantifies each participant's impact on the global model, capturing both their overall contribution and class-specific influence.
- Building upon our contribution assessment approach, we propose a new **weighted aggregation** method (ShapFed-WA) that outperforms the conventional federated averaging algorithm ([McMahan *et al.*, 2017]).
- To enhance the collaborative fairness of the system, we **personalize** the updates sent from the server to clients based on their contributions. This ensures that clients making substantial contributions receive better updates than those with minimal or no contributions.

## 2 Literature Review

Data valuation is a well-studied topic in ML. The "Data Shapley" framework proposed in [Ghorbani and Zou, 2019] quantifies the value of individual training data points in model learning. This approach claims to offer more detailed insights into the significance of each data point, compared to traditional leave-one-out scores. Monte Carlo approximation methods have been proposed to estimate Shapley values (SVs) by sampling random permutations of data points and determining their marginal contributions. Further, [Jia *et al.*, 2019] present algorithms designed to approximate the Shapley value with fewer model evaluations, thereby facilitating more efficient information sharing across different evaluations. These algorithms make specific assumptions about the utility function, including the stability of the learning algorithm and the characteristics of smooth loss functions. Class-specific SVs were proposed in [Schoch *et al.*, 2022] for more fine-grained data valuation. However, none of the above methods are applicable in the FL context because they require access to the raw data, which is not available at the orchestration server in FL.

The work in [Sim *et al.*, 2020] is the first to propose a collaborative learning scheme that considers incentives based on model rewards. They propose a data valuation method using information gain on model parameters given the data, by injecting Gaussian noise into aggregated data, to mitigate the need for a validation dataset agreed upon by all parties. Yet, even this method cannot be applied to FL directly because it assumes raw data sharing. Majority of the contribution assessment methods in FL require an auxiliary validation dataset, which introduces considerable time overhead in evaluating the contributions of participants. This issue is exemplified in the study by [Song *et al.*, 2019], where the authors introduce a novel metric known as "contribution index". This metric aims to assess the contribution of each data provider. The contribution index can be calculated using two proposed gradient-based methods (one round and multi rounds). These methods are designed to estimate confidence intervals, thereby providing a more refined and reliable measure of data contribution. However, the reliance on additional validation datasets and the time-intensive nature of these methods pose challenges to their practical implementation in real-world scenarios.

Gradient-based methods [Liu *et al.*, 2022; Xu *et al.*, 2021] have recently emerged as a practical approach for calculating Shapley values in FL. In [Liu *et al.*, 2022], the authors introduce the GTG-Shapley (guided truncation gradient Shapley) method. This approach combines between-round and within-round truncations to significantly reduce training costs. Between-round truncation is utilized to eliminate entire rounds of Shapley value calculations when the remaining marginal gain is small. Conversely, within-round truncation is applied to skip the evaluation of the remaining sub-models in permutations if the expected marginal gain is negligible, enhancing computational efficiency.

Meanwhile, the study in [Xu *et al.*, 2021] proposed cosine gradient Shapley value (CGSV), which quantifies the contribution of a participant to the overall coalition based on the cosine similarity between the individual gradients and the aggregated gradient. This approach has been widely adopted in recent literature [Wu *et al.*, 2024; Lin *et al.*, 2023]. Furthermore, [Wang *et al.*, 2020] proposed another dimension to SV

approximation in FL: a sampling-based approximation and a group testing-based approximation. Works related to fairness in FL such as FedFAIM [Shi *et al.*, 2022] and FedCE [Jiang *et al.*, 2023] also rely on gradient space to estimate the client contributions. Despite these advancements, the primary challenge lies in developing a strategy that effectively assesses contributions considering cost, practical feasibility, and per-class valuation of client data.

## 3 Problem Formulation

In this work, the high-level goal is to solve the following standard cross-silo federated learning optimization problem:

$$f^\star \coloneqq \min_{w\in\mathbb{R}^d}\left[f(w)\coloneqq\frac{1}{n}\sum_{i=1}^{n}f_i(w)\right], \qquad f_i(w)\coloneqq\mathbb{E}_{\xi\sim\mathcal{D}_i}\big[F_i(w,\xi)\big] \tag{1}$$

where the components $f_i:\mathbb{R}^d\to\mathbb{R}$ are distributed among $n$ local participants and are expressed in a stochastic format $f_i(w)\coloneqq\mathbb{E}_{\xi\sim\mathcal{D}_i}\big[F_i(w,\xi)\big]$, where $\mathcal{D}_i$ represents the distribution of $\xi$ on participant $i\in\mathcal{N}$, where $\mathcal{N}=\{1,2,\ldots,n\}$. This problem encapsulates standard empirical risk minimization as a particular case when each $\mathcal{D}_i$ consists of a finite number $m_i$ of elements $\{\xi_1^i,\ldots,\xi_{m_i}^i\}$. In such cases, $f_i$ simplifies to $f_i(w)=\frac{1}{m_i}\sum_{j=1}^{m_i}F_i(w,\xi_j^i)$. Our approach does not impose any restrictive assumptions on the data distribution $\mathcal{D}_i$, and therefore our work covers the case of heterogeneous (non-i.i.d.) data where $\mathcal{D}_i\neq\mathcal{D}_j,\forall\, i\neq j$ and the *local minima* $w_i^\star\coloneqq\arg\min_{w\in\mathbb{R}^d}f_i(w)$ might significantly differ from the global minimizer of the objective function (1).

### 3.1 Preliminaries

Let $\mathcal{M}_w:\mathcal{X}\to\mathcal{Y}$ be a supervised classifier parameterized by $w$, where $\mathcal{X}\subseteq\mathbb{R}^d$ and $\mathcal{Y}=\{1,2,\ldots,M\}$ denote the input and label spaces, respectively, $d$ is the input dimensionality, and $M$ is the number of classes. We set $F_i(w,\xi)=\mathcal{L}(\mathcal{M}_w(x),y)$, where $\mathcal{L}$ is the loss function and $\xi\coloneqq(x,y)$ is a training sample such that $x\in\mathcal{X}$ and $y\in\mathcal{Y}$.

**Shapley Values.** Shapley values, a concept derived from cooperative game theory, offer a principled approach to attributing the value of a coalition to its individual members [Winter, 2002]. In the context of federated learning, Shapley values can be instrumental in quantifying the contribution of each participant to the learning of the global model. This is particularly relevant in FL, where diverse participants collaboratively train a model while keeping their data localized. For a federated learning system with $n$ participants, the Shapley value (contribution value) of the $i^{\text{th}}$ participant is defined as:

$$\phi_i(\nu)=\sum_{S\subseteq\mathcal{N}\setminus\{i\}}\frac{|S|!(n-|S|-1)!}{n!}\Big(\nu\big(S\cup\{i\}\big)-\nu(S)\Big) \tag{2}$$

where $\mathcal{N}$ is the set of all participants, $S$ is a subset of participants excluding $i$, and $\nu(S)$ is the utility function that measures the performance of the subset $S$. The contribution estimation, $\phi_i(\nu)$, represents the average marginal contribution of participant $i$ over all possible coalitions. In FL, the utility function $\nu(S)$ can be defined as the performance of the global model trained using data from a subset of participants $S$. This can be measured in various ways, such as improvement in model accuracy or loss reduction. The challenge lies in the computational complexity of calculating Shapley values, as it requires evaluating the contribution of each participant across all possible subsets. To address this issue, we come up with an approximation method and describe it in Section 4.1.

**Algorithm 1** ShapFed algorithm

**input:** global weight initialization $w_s^1$, local learning rate $\eta$, no. of communication rounds $T$, no. of local epochs $K$, momentum factor $\mu$, $\mathcal{N}=[1,2,\ldots,n]$

Initialize $\gamma_i\leftarrow 1/n,\forall i\in\mathcal{N}$
**for** $t=1,\ldots,T$ **do**
  **if** $t=1$ **then**
    Broadcast $w_s^t$ to participants: $\bar{w}_i^t\leftarrow w_s^t,\forall i\in\mathcal{N}$
  **else**
    Send $\bar{w}_i^t=\gamma_i w_s^t+(1-\gamma_i)w_i^t$ to each party $i$
  **end if**
  **for** participant $i\in\mathcal{N}$ in parallel **do**
    Initialize local model $w_{i,0}^t\leftarrow\bar{w}_i^t$
    **for** $k=1,\ldots,K$ **do**
      Sample $\xi_{i,k}^t$, compute $\nabla F_i\big(w_{i,k}^t,\xi_{i,k}^t\big)$
      $w_{i,k+1}^t\leftarrow w_{i,k}^t-\eta\nabla F_i\big(w_{i,k}^t\big)$
    **end for**
  **end for**
  $w_s^{t+1}\leftarrow\sum_{i=1}^n\tilde{\gamma}_i w_{i,K}^t$
  Compute $\tilde{\Gamma}$ using Equation 5
  $\Gamma\leftarrow\tilde{\Gamma}$ if $t=1$ else $\Gamma=\mu\Gamma+(1-\mu)\tilde{\Gamma}$
  Update $\gamma_i$ using Equation 6
**end for**

**Federated averaging.** A common approach to solving (1) in the distributed setting is FedAvg [McMahan *et al.*, 2017]. This algorithm involves the participants performing several local steps of SGD (local epochs) and communicating with the server over multiple communication rounds (i.e., every communication round consists of some local number of epochs). In each communication round, the updates from the participants are averaged on the server and sent back to all participants. For a local epoch $t$ and participant $i\in\mathcal{N}$, the local iterate is updated according to:

$$w_i^{t+1}=w_i^t-\eta_i^t\nabla F_i\big(w_i^t,\xi_i\big), \tag{3}$$

where $\xi_i=\{\xi_1^i,\ldots,\xi_{m_i}^i\}$. For a communication round, the update will be:

$$w_s^{t+1}=\frac{1}{n}\sum_{i=1}^{n}\Big(w_i^t-\eta_i^t\nabla F_i\big(w_i^t,\xi_i\big)\Big). \tag{4}$$

The server then broadcasts the updated model $w_s^{t+1}$ to all participants, which is used as $w_i^{t+1}$ for the next local epoch.

## 4 Proposed Method

### 4.1 Contribution Assessment: ShapFed

Suppose that we define the utility function $\nu$ as class-wise performance assessed using a specific validation set. This

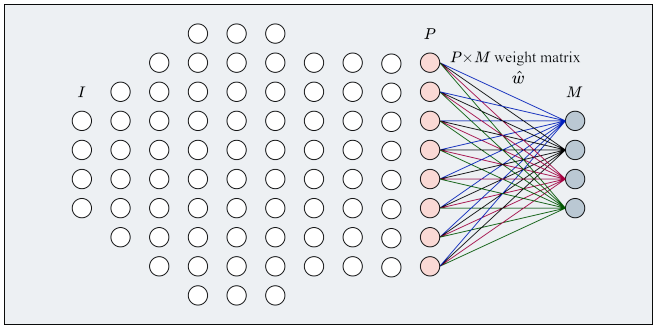



Figure 2: Illustration of the specific weight segments utilized for evaluating class-wise contributions and participant heterogeneity.

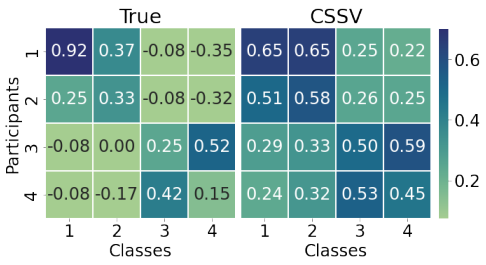



Figure 3: Experimental demonstration comparing performance of utility functions $\nu$ on a synthetic dataset. The experiment contrasts the true Shapley value approach, using validation accuracy across all coalitions, with our proposed approximation method (Section 4.1).

framework enables the server to calculate *Class-Specific Shapley Values* (CSSV), as delineated in Equation 2. Since $\phi_i(\nu)$ is no longer a scalar, but a $M$-dimensional vector, we use the notation $\Gamma_i$ to denote CSSV. Although conceptually straightforward, this approach encounters practical challenges due to the necessity of a server-side validation set. This is problematic because users cannot share their data samples in FL due to privacy concerns. Furthermore, establishing a validation set that fairly evaluates all distributions $\mathcal{D}_i$, which could be extremely non-iid, presents a significant challenge. Some methods have tried to circumvent this by using publicly available datasets or creating synthetic samples [Li *et al.*, 2023]. However, these solutions still require the server to perform a large number of inferences ($2^n - 1$ calls) to compute CSSVs accurately. To overcome this, subsequent works [Xu *et al.*, 2021; Jiang *et al.*, 2023] have proposed using the directional alignment of gradients as an alternative utility measure.

Motivated by this idea, we introduce an algorithm to evaluate CSSVs by utilizing the directional alignment between the gradients or network parameters of the last layer in the classifier $\mathcal{M}_w$. Conceptually, the classifier $\mathcal{M}_w$ can be viewed as a composition of a feature extractor that maps an input to a $P$-dimensional embedding ($\mathcal{X} \to \mathbb{R}^P$) and a linear classification layer (parameterized by $\hat{w} \in \mathbb{R}^{P \times M}$) that maps the feature embedding to a $M$-dimensional logits space ($\mathbb{R}^P \to \mathbb{R}^M$). Figure 2 illustrates an example network architecture, highlighting $\hat{w}$. We divide this matrix $\hat{w}$ into $M$ column vectors, with each vector corresponding to one of the output classes.

In each communication round, the server collects the parameter updates or gradients ($w_i^t$) from all participants ($i \in \mathcal{N}$) and aggregates them to obtain $w_s^{t+1}$. For simplicity of notation, we drop the index $t$ in the subsequent discussion. Let $\hat{w}_i \subset w_i$ be the subset of updates corresponding to the last linear layer from participant $i$ and let $\hat{w}_s \subset w_s$ be the corresponding subset of the aggregated update computed by the server. Furthermore, let $\hat{w}_{i,j}$ ($\hat{w}_{s,j}$) denote the $j$-th column vector of $\hat{w}_i$ ($\hat{w}_s$), where $j \in \mathcal{Y}$. We define the contribution (CSSV) $\Gamma_i$ of participant $i$ as:

$$\Gamma_i \coloneqq \Big[\cos\left(\hat{w}_{i,1}, \hat{w}_{s,1}\right), \cos\left(\hat{w}_{i,2}, \hat{w}_{s,2}\right), \ldots, \cos\left(\hat{w}_{i,M}, \hat{w}_{s,M}\right)\Big] \tag{5}$$

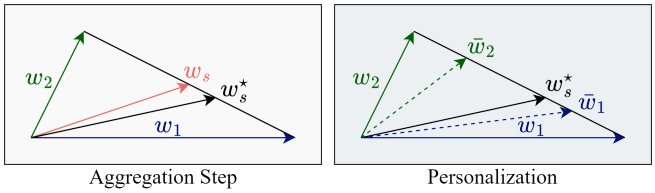
The $j$-th element of $\Gamma_i$ can be viewed as the contribution of participant $i$ ($i \in \mathcal{N}$) to the learning of the $j$-th class ($j \in \mathcal{Y}$), with higher values indicating stronger contribution. By the end of this process, the server obtains a matrix $\Gamma \coloneqq [\Gamma_1, \Gamma_2, \ldots, \Gamma_n] \in [-1, 1]^{n \times M}$, representing the heterogeneity and class-wise contribution of each participant. This matrix not only aids in understanding the individual contributions but also provides insights into the diverse nature of the data and learning across different classes. The CSSV value can be updated after every communication round with a momentum factor $\mu$ to facilitate smoother convergence.

**Example 1 (Identification of Class-specific Data Heterogeneity).** Consider a scenario with $n = 4$ and $M = 4$ (as depicted in Figure 3), where the first two participants exclusively contain data belonging to the first two classes, and a similar splitting scenario applies to the remaining participants and classes. Ideally, our algorithm should generate block matrices in which the values of the diagonal blocks $\{\Gamma_{1:2,1:2}, \Gamma_{3:4,3:4}\}$ are significantly higher than those in the off-diagonal blocks $\{\Gamma_{1:2,3:4}, \Gamma_{3:4,1:2}\}$. As shown in Figure 3, our contribution assessment method effectively discerns the heterogeneity among participants, offering a viable alternative to the resource-intensive computation of Shapley values. Additionally, our approach eliminates the need for a server-side validation set, a considerable advantage over methods that depend on such datasets. Finally, ShapFed requires only $\mathcal{O}(n+1)$ inferences, a stark contrast to the $\mathcal{O}(2^n - 1)$ calls necessitated by exact SV computation (which becomes practically infeasible when $n \gg 0$.

It must be emphasized that the work of [Xu *et al.*, 2021] (CGSV) has already established that cosine similarity between gradients is a good approximation of SV (marginal contribution of each participant) and also provides a theoretical bound on this approximation error. ShapFed builds upon this result and *extends CGSV to the class-specific setting*. The main difference with CGSV lies in how the overall SV is computed based on cosine similarity between gradients - while CGSV treats the complete model gradient as a whole and computes a single similarity value, we first estimate class-specific influences based on the gradients/weights of the last layer and average them, thereby providing a more granular understanding of the differences between the participants.

## 4.2 Weighted Aggregation: ShapFed-WA

While the FedAvg algorithm [McMahan *et al.*, 2017] has become the standard approach in FL, it either treats all par-




Aggregation Step Personalization

Figure 4: **Weighted Aggregation:** The optimal weights $w_s^\star$ are derived using Equation 7, while $w_s$ represents the result of applying equal weights (FedAvg). **Personalization:** Rather than distributing a uniform global model to all users, we provide personalized weights $\bar{w}_i$, which are $\gamma_i$ combinations of individual user weights $w_i$ and the optimally aggregated weight $w_s^\star$.

ticipants uniformly or weights them based on their self-reported training set size. However, this aggregation approach sometimes results in decreased accuracy [Kairouz *et al.*, 2021], particularly with non-iid (heterogeneous) data [Li *et al.*, 2020; Zhao *et al.*, 2018] and when a large number of local epochs are employed [Kairouz *et al.*, 2021]. The discrepancy between the local data distribution of each participant and the overall global distribution can cause the local objectives to diverge from the global optimum. Consequently, simple averaging may result in a sub-optimal solution or a slow convergence to the optimal solution.

To overcome this problem, we propose a weighted aggregation method called ShapFed-WA based on CSSV. Specifically, we first normalize CSSV between 0 and 1 and compute the weight assigned to each participant ($\gamma_i$) by averaging the normalized CSSV across all classes. We further normalize the participant weights such that they sum up to 1. These operations are summarized in Equation 7.

$$\gamma_i = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1+\Gamma_{i,j}}{2}\right), \quad \tilde{\gamma}_i = \frac{\gamma_i}{\sum_{k=1}^{n}\gamma_k}. \tag{6}$$

The final ShapFed-WA aggregation rule is given by:

$$w_s = \sum_{i=1}^{n}\tilde{\gamma}_i w_i. \tag{7}$$

Figure 4 shows a scenario where $w_s$ derived using standard FedAvg, assigning equal importance to each update ($w_s = 0.5w_1 + 0.5w_2$), leads to a sub-optimal outcome compared to homogeneous (iid) setting. In contrast, $w_s^\star$ represents the outcome of our proposed weighted aggregation method, which aligns more closely with the global optimum.

### 4.3 Personalization

Ensuring fairness within a FL system necessitates an incentive mechanism that appropriately rewards participants according to their contributions to the collaborative effort. In this work, we employ the following widely-used definition of collaborative fairness [Lyu *et al.*, 2020; Xu *et al.*, 2021; Jiang *et al.*, 2023]: *In a federated setup, a high-contribution participant should be rewarded with a better-performing local model than a low-contribution participant. Mathematically, fairness can be quantified using Pearson's correlation coefficient between the standalone accuracies of participants and their respective final model accuracies.*

To achieve collaborative fairness, we propose a personalization approach wherein the server transmits participant-specific updates instead of identical aggregated update. This technique is illustrated in Figure 4 and can be calculated as:

$$\bar{w}_i = w_i + \gamma_i(w_s - w_i) = \gamma_i w_s + (1-\gamma_i)w_i. \tag{8}$$

The above personalization method ensures that participants contributing less to the global model will progress towards the optimal solution $w^\star$ at a slower rate compared to those with higher contributions. Our strategy is akin to [Gasanov *et al.*, 2022], where the problem formulation involves an explicit mixture of local models and a global model (Equation 3 of [Gasanov *et al.*, 2022]), and the mixture weight is a fixed hyperparameter. In contrast, we dynamically determine the mixture weights using our contribution assessment, which measures the quality of the updates received from each participant. Furthermore, the concept of personalization serves as a form of penalization in our method. Generally, low-contribution participants can detrimentally affect the performance of the global model, thereby potentially reducing the collaboration gain for high-contribution participants. Moreover, the low-contribution participants could be either free-riders, who seek to benefit from the global model without making any meaningful contribution, or even Byzantines, who intentionally send random updates without using any computational resources, thereby undermining the collaborative learning process. Such participants must be penalized with models having very low accuracy.

## 5 Experiments and Results

### 5.1 Datasets

**CIFAR-10** [Krizhevsky *et al.*, 2009]: This dataset comprises 60,000 RGB images, each with dimensions of $32 \times 32$ pixels, spanning 10 different classes. It is divided into a training set of 50,000 images and a testing set of 10,000 images. These images represent a diverse collection of objects, providing a comprehensive resource for evaluating image recognition models. **Chest X-Ray** [Rahman *et al.*, 2020]: The Tuberculosis (TB) Chest X-ray Database is a comprehensive collection of chest X-ray images containing 700 publicly accessible TB-positive images and 3500 normal images. **Fed-ISIC2019** [Ogier du Terrail *et al.*, 2022]: This dataset is an amalgamation of the ISIC 2019 challenge dataset and the HAM1000 database, presenting a total of 23,247 dermatological images of skin lesions (8 classes). It encompasses data from six distinct centers, with respective data distributions of 9930/2483, 3163/791, 2961/672, 1807/452, 655/164, and 351/88. As a preprocessing step, we resized images to the same shorter side of 224 pixels while maintaining their aspect ratio, and normalized images' brightness and contrast, as specified in [Ogier du Terrail *et al.*, 2022].

### 5.2 Experimental Setup

**CIFAR-10.** We leverage the ResNet-34 architecture trained using the SGD optimizer with a fixed learning rate of 0.01. For FL, we use 50 communication rounds.

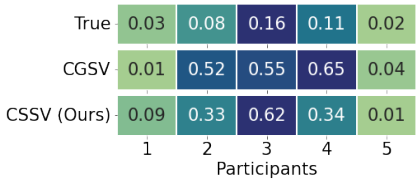



Figure 5: Comparison of our proposed contribution assessment algorithm (CSSV) with CGSV and true Shapley value computations using ResNet-34 architecture on Chest X-Ray dataset.

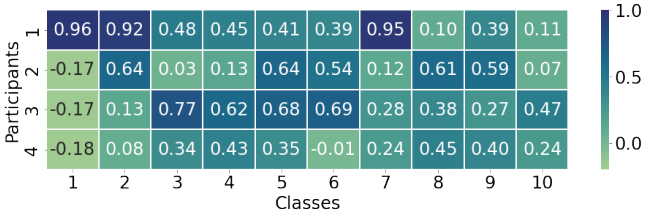



Figure 6: Heatmap visualization of class-specific Shapley values for heterogeneous setting (explained in Section 5.2) evaluated on CIFAR-10 dataset.

**Chest X-Ray.** We employ a custom CNN architecture with three convolution layers followed by three fully connected layers. All the layers (except the last) are followed by ReLU activation. We use SGD optimizer with a learning rate of $0.01$, momentum of $0.9$, and weight decay of $5 \cdot 10^{-4}$. The models are trained for $50$ rounds.

**Fed-ISIC2019.** We employ the EfficientNet_B0 model and use the same training settings as in CIFAR-10 with $200$ communication rounds. Due to the heterogeneity of this dataset, we use the balanced accuracy metric.

Our experimental design involves dividing the training dataset among participants using two distinct strategies:

- **Imbalanced partitioning**: Here, the data is distributed unevenly among participants. We control the class distribution by manipulating two parameters: the major allocation probability and the number of classes designated as the majority group. In our experiments, we set the first class as the majority, with a higher allocation probability of 0.7, while the remaining classes are assigned equal probabilities of 0.1 each. This configuration allows us to investigate the model's behaviour in scenarios where data distribution is skewed.
- **Heterogeneous partitioning**: This strategy introduces a more complex and varied class allocation among participants. Each participant receives a unique distribution of probabilities across all classes. This scenario is designed to mimic real-world conditions where data distribution can be highly irregular and participant-specific. The Fed-ISIC2019 dataset fits into this purpose. For CIFAR-10, we implement label skew partitioning, where the first class is exclusively owned by the first participant, and the remaining nine classes are partitioned equally among all participants. For the Chest X-Ray, we adopt an equal major allocation strategy with a class variant among participants. Specifically, the first class is divided among 5 participants with probabilities 40%, 30%, 20%, 10%, and 0%, respectively. The second class is similarly divided among 5 participants with probabilities 0%, 10%, 20%, 30%, and 40%, respectively.

## 5.3 Contribution Assessment

Figures 5 and 6 provide heatmap visualizations of CSSVs for Chest X-Ray and CIFAR-10 datasets, respectively, using ResNet-34 architecture as outlined in Section 5.2.

- Figure 5 offers a comparative analysis of the true estimation of Shapley values (with the utility function being validation accuracy on an auxiliary set), CGSV, and our proposed contribution assessment approach, CSSV. In the given data distribution scenario, where each participant holds equal data volumes, $P_3$ is expected to have a higher utility and contribution measure. Conversely, $P_1$ and $P_5$ are predicted to score the lowest in terms of contribution, with $P_2$ and $P_4$ expected to have moderate contributions. The obtained empirical results align with these expectations for rows 1 and 3 (true estimation and our approach), demonstrating the accuracy of our method. However, the CGSV approach deviates by assigning a disproportionately high score to $P_4$ and comparable scores to $P_2$ and $P_5$. This discrepancy stems from CGSV's reliance on all network layers for Shapley value computation, which becomes a limitation in the context of large models such as ResNet-34. Our method, on the other hand, effectively identifies the importance of each participant and closely tracks the true Shapley value while being computationally efficient.
- Figure 6 showcases a heatmap representing contributions in the CIFAR-10 dataset under a heterogeneous setting. Here, $P_1$ possesses all entries for class 1, while the remaining classes are evenly distributed among all four participants. Consistent with expectations, participants $P_2$, $P_3$ and $P_4$ exhibit the lowest contribution scores for class 1. This outcome highlights the effectiveness of our approach in identifying and quantifying participant contributions in a scenario where data is heterogeneous.

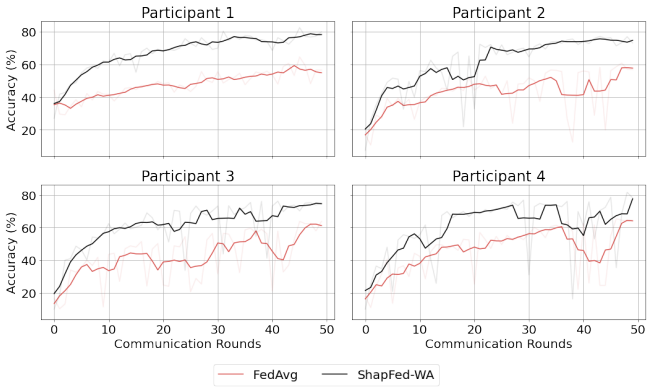



Figure 7: Comparing FedAvg and ShapFed-WA on CIFAR10 under an imbalanced split scenario: insights into the balanced accuracy of four individual participants.

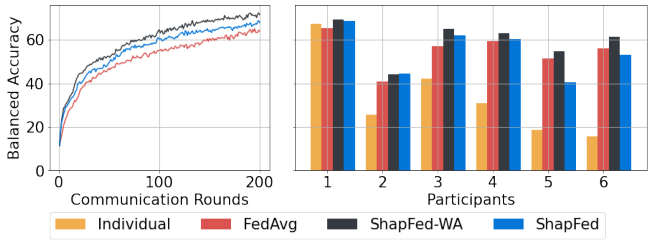



Figure 8: (**Left**) The balanced accuracy of our methods (ShapFed-WA & ShapFed) vs FedAvg. (**Right**) Per-participant accuracy using all methods evaluated on Fed-ISIC2019 dataset.

### 5.4 Weighted Aggregation

**CIFAR-10.** We evaluate the performance of ShapFed-WA and FedAvg on the CIFAR-10 dataset across the two split scenarios. The results in Figure 7 demonstrate the superiority of our method over FedAvg in terms of (balanced) validation accuracy for both the global model and individual participant models (in imbalanced partitioning).

**Fed-ISIC2019.** Our method consistently outperformed the FedAvg method as shown in Figure 8. This improvement can be attributed to ShapFed-WA's ability to dynamically adjust based on the varying contributions of individual clients. Comparative effectiveness of our approach in terms of participant-wise balanced accuracy is clearly illustrated in histogram plot, in Figure 8.

### 5.5 Personalization

We assessed the personalization approach by observing individual participant validation accuracy under three distinct training conditions: first, when each participant trained exclusively on its data, second, when trained collaboratively with other participants using our methods: ShapFed-WA and ShapFed, and third when trained collaboratively with other participants using FedAvg method. We also report the collaborative fairness (Pearson's correlation coefficient) of these methods.

**CIFAR-10.** In Table 1, we report the balanced accuracy for Individual, FedAvg, CGSV and ShapFed approaches for each participant. In imbalanced partitioning, faster convergence

| Dataset / Partition | | Setting | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | Corr. |
|---|---|---|---|---|---|---|---|---|
| ChestXRay | Het. | Individual | 50.0 | 64.7 | 62.0 | 53.7 | 50.0 | — |
| | | FedAvg | 50.0 | 55.8 | 61.9 | 54.2 | 50.0 | 0.82 |
| | | ShapFed | 50.0 | 65.2 | 69.5 | 58.5 | 50.0 | **0.93** |
| CIFAR-10 | Imb. | Individual | 75.8 | 45.4 | 48.6 | 31.6 | — | — |
| | | FedAvg | 56.6 | 56.8 | 63.8 | 64.2 | — | −0.60 |
| | | CGSV | 57.2 | 59.0 | 58.8 | 60.4 | — | −0.98 |
| | | ShapFed | 81.4 | 78.2 | 71.8 | 73.6 | — | **0.74** |
| | Het. | Individual | 75.2 | 68.8 | 66.8 | 69.0 | — | — |
| | | FedAvg | 74.6 | 70.2 | 70.2 | 76.0 | — | 0.53 |
| | | CGSV | 55.0 | 55.8 | 57.2 | 52.6 | — | −0.26 |
| | | ShapFed | 79.8 | 75.4 | 69.0 | 75.0 | — | **0.90** |

Table 1: Performance and fairness comparison with our method and FedAvg. We use Pearson's correlation (↑) as a fairness metric on CIFAR-10. The red highlight indicates a negative gain from collaboration.

| Setting | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_6$ | Corr. |
|---|---|---|---|---|---|---|---|
| Individual | 67.2 | 25.7 | 42.3 | 31.0 | 18.5 | 15.6 | — |
| FedAvg | 65.4 | 40.9 | 57.2 | 59.3 | 51.5 | 56.2 | 0.63 |
| ShapFed-WA | 69.3 | 44.3 | 65.0 | 63.1 | 54.8 | 61.2 | 0.62 |
| ShapFed | 68.5 | 44.4 | 61.9 | 60.4 | 40.6 | 53.2 | **0.84** |

Table 2: Performance and fairness comparison using Pearson's correlation (↑) as a fairness metric on Fed-ISIC2019. The red highlight indicates a negative gain from collaboration.

was observed for the first participant model, which held a significant 70% of the data. Conversely, participants with lower contributions exhibited slower progress towards the optimal solution. This implies that the participants receive distinct updates based on their contributions to the collaborative learning process. On the other hand, models that are trained using the FedAvg and CGSV algorithms, exhibit close accuracy, implying unfair treatment. CGSV demonstrated a negative correlation - lower collaborative fairness of $-0.98$. A similar pattern is observed in heterogeneous partitioning, where higher correlation values with individual accuracies imply a fairer consideration of diverse client contributions.

**Chest X-Ray.** As anticipated, our proposed ShapFed outperformed FedAvg in this scenario. Additionally, even though the five participants have equal amounts of data split among them, their class distributions vary. Notably, the third participant experienced the most substantial gain from the collaboration, resulting in higher accuracy.

**Fed-ISIC2019.** As shown in Table 2, our method exhibits a strong correlation (0.84) with individual client performances, indicating a high degree of fairness. In contrast, FedAvg demonstrates a significantly lower correlation (0.63), implying a more uniform distribution of updates regardless of individual client contributions. As such, FedAvg has a negative collaboration gain for $P_1$ (highlighted in red). These findings confirm that our methods effectively recognize and incorporate the distinct contributions of each client in a collaborative learning environment with complex data distributions.

## 6 Summary

This work proposes Class-Specific Shapley Values (CSSVs) to quantify participant contributions at a granular level. The contributions of this work include a novel method to deepen the understanding of participant impact and improve fairness analysis. Evaluation against FedAvg shows superior performance and additional experiments reveal enhanced fairness by personalizing client updates based on contributions. Overall, the approach aims to achieve a more equitable distribution of benefits in FL. In future, we plan to conduct an in-depth theoretical analysis aimed at identifying the specific characteristics that contribute to an effective estimation of Shapley values. This analysis will enhance our understanding of the factors that influence the accuracy and reliability of Shapley value approximations. Furthermore, an investigation into what makes our approximation of cosine similarity from the last layer a robust indicator of contributions will be explored.

## References


[Gasanov *et al.*, 2022] Elnur Gasanov, Ahmed Khaled, Samuel Horvath, and Peter Richtarik. Flix: A simple and communication-efficient alternative to local methods in federated learning. In *AISTATS 2022: 25th International Conference on Artificial Intelligence and Statistics*. arXiv, 2022.

[Ghorbani and Zou, 2019] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

[Jia *et al.*, 2019] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019.

[Jiang *et al.*, 2023] Meirui Jiang, Holger R Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair federated medical image segmentation via client contribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16302–16311, 2023.

[Kairouz *et al.*, 2021] 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. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Li *et al.*, 2020] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[Li *et al.*, 2023] Bo Li, Yasin Esfandiari, Mikkel N Schmidt, Tommy S Alstrøm, and Sebastian U Stich. Synthetic data shuffling accelerates the convergence of federated learning under data heterogeneity. *arXiv preprint arXiv:2306.13263*, 2023.

[Lin *et al.*, 2023] Xiaoqiang Lin, Xinyi Xu, See-Kiong Ng, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Fair yet asymptotically equal collaborative learning. In *International Conference on Machine Learning*, pages 21223–21259. PMLR, 2023.

[Liu *et al.*, 2022] Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM Trans. Intell. Syst. Technol.*, 13(4), may 2022.

[Lv *et al.*, 2021] Hongtao Lv, Zhenzhe Zheng, Tie Luo, Fan Wu, Shaojie Tang, Lifeng Hua, Rongfei Jia, and Chengfei Lv. Data-free evaluation of user contributions in federated learning. In *2021 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*, pages 1–8, 2021.

[Lyu *et al.*, 2020] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. *Federated Learning: Privacy and Incentive*, pages 189–204, 2020.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Ogier du Terrail *et al.*, 2022] 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. *Advances in Neural Information Processing Systems*, 35:5315–5334, 2022.

[Rahman *et al.*, 2020] Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F. Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, and etal. Mahbub. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020.

[Schoch *et al.*, 2022] Stephanie Schoch, Haifeng Xu, and Yangfeng Ji. CS-Shapley: class-wise Shapley values for data valuation in classification. In *Advances in Neural Information Processing Systems*, pages 34574–34585, 2022.

[Shi *et al.*, 2022] Zhuan Shi, Lan Zhang, Zhenyu Yao, Lingjuan Lyu, Cen Chen, Li Wang, Junhao Wang, and Xiang-Yang Li. Fedfaim: A model performance-based fair incentive mechanism for federated learning. *IEEE Transactions on Big Data*, 2022.

[Sim *et al.*, 2020] Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. Collaborative machine learning with incentive-aware model rewards. In *International conference on machine learning*, pages 8927–8936. PMLR, 2020.

[Siomos and Passerat-Palmbach, 2023] Vasilis Siomos and Jonathan Passerat-Palmbach. Contribution evaluation in federated learning: Examining current approaches, 2023.

[Song *et al.*, 2019] Tianshu Song, Yongxin Tong, and Shuyue Wei. Profit allocation for federated learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2577–2586. IEEE, 2019.

[Tastan and Nandakumar, 2023] Nurbek Tastan and Karthik Nandakumar. Capride learning: Confidential and private decentralized learning based on encryption-friendly distillation loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8084–8092, 2023.

[Wang *et al.*, 2020] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled ap-

proach to data valuation for federated learning. *Federated Learning: Privacy and Incentive*, pages 153–167, 2020.

[Wei *et al.*, 2020] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

[Winter, 2002] Eyal Winter. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054, 2002.

[Wu *et al.*, 2024] Zhaoxuan Wu, Mohammad Mohammadi Amiri, Ramesh Raskar, and Bryan Kian Hsiang Low. Incentive-aware federated learning with training-time model rewards. In *The Twelfth International Conference on Learning Representations*, 2024.

[Xu *et al.*, 2021] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. *Advances in Neural Information Processing Systems*, 34:16104–16117, 2021.

[Zhao *et al.*, 2018] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.