



## **Beyond Leave-One-Out: Private and robust contribution evaluation in Federated Learning**

Downloaded from: <https://research.chalmers.se>, 2026-10-13 03:28 UTC

Citation for the original published paper (version of record):

Jaramillo Velez, D., Biczók, G., Graell Amat, A. et al (2026). Beyond Leave-One-Out: Private and robust contribution evaluation in Federated Learning. *Intelligent Systems with Applications*, 32. <http://dx.doi.org/10.1016/j.iswa.2026.200725>

N.B. When citing this work, cite the original published paper.



## Beyond *Leave-One-Out*: Private and robust contribution evaluation in Federated Learning

Delio Jaramillo-Velez <sup>a,1</sup>, Gergely Biczók <sup>b,c,1</sup>, Alexandre Graell i Amat <sup>d,1</sup>, Balázs Pejó <sup>b,c,1,\*</sup>

<sup>a</sup> Dpto. de Matemáticas, Estadística e Investigación Operativa, Instituto Universitario de Matemáticas y Aplicaciones (IMAULL), Universidad de La Laguna, Apartado de Correos 456, La Laguna, Tenerife, 38200, Spain

<sup>b</sup> Laboratory of Cryptography and System Security, Dept. of Networked Systems and Services, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Műgyetem rkp. 3, Budapest, 1111, Hungary

<sup>c</sup> HUN-REN-BME Information Systems Research Group, HUN-REN Hungarian Research Network, Office for Supported Research Groups, Piarista utca 4, Budapest, 1052, Hungary

<sup>d</sup> Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, 41296, Sweden

### ARTICLE INFO

#### Keywords:

Federated Learning  
Secure aggregation  
Contribution evaluation  
Shapley value  
Fairness  
Robustness

### ABSTRACT

Cross-silo Federated Learning enables multiple organizations to collaboratively train machine learning models without sharing raw data, but client updates may still leak sensitive information through inference attacks. Secure Aggregation protects privacy by concealing individual updates; however, it also complicates contribution evaluation, which is essential for fair reward allocation and identifying low-quality or malicious participants. Existing marginal-contribution approaches, such as the Shapley value, are incompatible with Secure Aggregation, while practical alternatives, such as *Leave-One-Out*, are limited and rely on self-evaluation.

We introduce two marginal difference based contribution scores compatible with Secure Aggregation. Fair-Private satisfies standard fairness axioms, while Everybody-Else eliminates self-evaluation and provides resistance to direct manipulation, addressing an overlooked vulnerability in contribution evaluation. We establish theoretical guarantees for fairness, privacy, robustness, and computational efficiency, and evaluate our methods on multiple medical image datasets and CIFAR10 in cross-silo settings. Our scores consistently outperform existing baselines, more closely approximate Shapley-based client rankings, and improve downstream model performance and misbehavior detection. On ISIC2019, for example, they reduce the ranking error relative to the Shapley value by an order of magnitude compared to *Leave-One-Out* (0.11 vs. 1.06) and improve the Spearman correlation from 0.82 to 0.94. These results demonstrate that fairness, privacy, robustness, and practical utility can be jointly achieved in federated contribution evaluation, providing a principled solution for real-world cross-silo deployments.

### 1. Introduction

Federated Learning (FL) (Kairouz et al., 2021) is a distributed machine learning approach in which multiple parties jointly train a machine learning model without exchanging their raw data. Instead of centralizing the training data, FL delegates computation to the clients and communicates only model updates (e.g., gradients) with a central server. This decentralized approach enhances privacy by design, as sensitive data never leaves the clients.

In addition to privacy preservation, Contribution Evaluation (CE) (Siomos & Passerat-Palmbach, 2023) is a critical component of FL. Fairly assessing client contributions is essential for incentivizing honest participation, ensuring equitable reward distribution, and detecting

under-performing or malicious actors. CE schemes are specifically designed to assign a *score* to each participant that reflects their importance toward the collaborative process.

While FL provides built-in privacy protection by keeping raw data local, individual model updates can still leak sensitive information through inference attacks (Wang et al., 2019). To mitigate this risk, additional privacy-preserving mechanisms such as Differential Privacy and Secure Aggregation have been introduced. In particular, Secure Aggregation (SA) (Zhou et al., 2022) has emerged as a standard technique for concealing individual client updates, making it harder to perform inference attacks. Based on lightweight secure multiparty computation (Cramer et al., 2015) and secret sharing (Shamir, 1979), secure

\* Corresponding author at: Laboratory of Cryptography and System Security, Dept. of Networked Systems and Services, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Műgyetem rkp. 3, Budapest, 1111, Hungary.

E-mail address: [pejo@crysys.hu](mailto:pejo@crysys.hu) (B. Pejó).

<https://doi.org/10.1016/j.iswa.2026.200725>

Received 11 May 2026; Received in revised form 26 July 2026; Accepted 25 August 2026

Available online 14 September 2026

2667-3053/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

aggregation ensures that individual client updates remain hidden from the server while still enabling correct computation of the aggregated global model. SA therefore poses a fundamental tension with CE, as most prior methods depend on individual client updates (Liu et al., 2022; Sim et al., 2022) and are incompatible with the SA mechanism.

One of the most widely studied methods for contribution evaluation is the Shapley value (Shapley, 1953), a marginal-contribution-based metric rooted in cooperative game theory. It is uniquely characterized by four classic properties: efficiency, symmetry, null player, and linearity (although the latter is arguably a structural mathematical property rather than an explicit fairness condition (Kamil, 2005)). While this framework establishes a principled foundation for equitable reward allocation, its strong theoretical appeal is constrained by computational complexity; exact calculation is intractable beyond a small number of clients due to the need to train an exponential number of models. Although it has been extensively applied in machine learning and several approximation methods have been proposed, these only alleviate its computational burden (Chu et al., 2025; Song et al., 2019). More fundamentally, its reliance on marginal contributions makes it incompatible with privacy-enhancing mechanisms such as secure aggregation. Under SA, only limited coalition information is accessible, whereas Shapley and its approximations require evaluations over a rich set of coalitions, including subsets not observable in the SA setting.

The Leave-One-Out mechanism (Black & Fredrikson, 2021; Evgeniou et al., 2004) remains the only practical marginal difference based CE method compatible with SA. However, this approach is severely limited, yielding a crude approximation. Furthermore, it lacks the efficiency property and requires self evaluation when SA is applied, making it unreliable in the presence of selfish clients who seek to manipulate scores.

Existing literature in the contribution evaluation domain has focused almost exclusively on the algorithmic challenges of CE, primarily proposing heuristics to speed up the computation of metrics like the Shapley value (Chu et al., 2025; Song et al., 2019). In contrast, establishing rigorous, theoretical fairness guarantees remains exceedingly rare in the CE field. Our work addresses this critical gap. Rather than relying on unverified approximations, we provide a principled framework backed by provable axiomatic guarantees that formally certify the fairness, privacy, and robustness of our evaluation scores. This theoretical foundation strengthens the role of contribution evaluation in secure federated learning by establishing formal properties that go beyond heuristic approaches and enable rigorous analysis of fairness and manipulation resistance.

**Our contribution.** In this work, we address the fundamental problem of designing CE methods for cross-silo FL that simultaneously satisfy fairness guarantees (e.g., efficiency and symmetry), privacy (via compatibility with SA), and robustness against selfish clients. We consider a threat model in which clients are strategic but non-colluding, and secure aggregation is treated as a fundamental system requirement. Our contributions are the following.

- We propose two novel CE scores that are compatible with SA: Fair-Private (FP) and Everybody-Else (EE). FP satisfies fairness properties, such as efficiency, symmetry, and null player, but still relies on self evaluation. On the contrary, EE deliberately relaxes the null player property to provide robustness against selfish clients.
- We establish theoretical guarantees for the proposed CE methods, covering fairness, privacy, computational complexity, and resistance to manipulation.
- We empirically evaluate the proposed scores on three medical image datasets and CIFAR10. Our methods consistently outperform the baseline and better approximate the client contributions induced by the ground truth.

## 2. Related work

The privacy-utility trade-off in CE is mostly studied in the literature from two angles. The first examines how existing CE methods amplify privacy leakage during learning, while the second focuses on designing CE schemes that yield meaningful and reliable scores while preserving privacy. The former has received considerably more attention. A recent survey (Lin et al., 2024) highlights that most existing work at the intersection of CE and privacy focuses on how different CE metrics may increase the risk of inference attacks. In contrast, the latter problem, how to construct CE methods that are intrinsically privacy-preserving, remains mostly neglected.

Only a few works have explicitly addressed privacy-preserving CE. The authors of Zheng et al. (2023) proposed a multi-server solution based on encryption, which incurs significant computational overhead. Another paper (Watson et al., 2022) used Differential Privacy (Pej6 & Desfontaines, 2022) by injecting noise to client updates before evaluation, which severely degrades utility. Finally, in Ma et al. (2021), the researchers proposed a blockchain-based framework that requires a drastic architectural change. Critically, none of these approaches considers SA, despite its widespread adoption in practical FL systems.

To the best of our knowledge, only three prior works address contribution evaluation under secure aggregation: Federated Group Testing (FedGT) (Xhemrishi and Östman, et al., 2025), Quality Inference (QI) (Pej6 & Bicz6k, 2023), and their combination, Detect&Score (Xhemrishi, Amat, & Pej6, 2025). However, these approaches do not target fair contribution evaluation and thus provide no built-in fairness guarantees for the participating clients. Furthermore, they rely on different privacy assumptions, participation models, and threat models than ours, preventing a direct comparison. As summarized in Table 1, FedGT (Xhemrishi and Östman, et al., 2025) and Detect&Score (Xhemrishi, Amat, & Pej6, 2025) require revealing sub-coalitions of clients, which reduces the anonymity guarantees provided by secure aggregation, while QI (Pej6 & Bicz6k, 2023) assumes partial client participation, which is incompatible with the cross-silo setting considered in this work. Consequently, directly benchmarking against these methods would either require relaxing the strict secure aggregation assumptions that motivate our approach or ignoring their different operating regimes and objectives. Instead, we compare against methods that satisfy the same strict secure aggregation requirements, namely the Cosine Similarity mechanism (COS) (Xu et al., 2021) and Leave-One-Out (LOO) (Evgeniou et al., 2004). We additionally report the Shapley value approximation (MR-SV) (Song et al., 2019) and the exact Shapley value as information-unlimited upper bounds, since these methods are not constrained by privacy.

## 3. Preliminaries

**Notation.** Our notation is summarized in Table 2. We denote by  $M_0$  and  $M$  the global model before and after a training round, respectively, and by  $M_i$  the local model of client  $i \in [N] = \{1, \dots, N\}$ . The local update of client  $i$  is defined as the pseudogradient  $U_i = \frac{1}{N}(M_i - M_0)$ . We assume that aggregation is performed by simple addition, i.e.,  $M = M_0 + \sum_{i=1}^N U_i$ . This setting encompasses the standard aggregation technique *Federated Averaging* (FedAvg) (Kairouz et al., 2021).

### 3.1. Secure aggregation in Federated Learning

To ensure this work is self-contained, we briefly review the foundational secure aggregation protocol originally introduced by Google (Bonawitz et al., 2017). In our cross-silo federated learning setting, under the honest-but-curious, non-colluding trust assumptions, SA cryptographically guarantees that individual client updates  $U_i$  remain completely hidden from the central server. The server is structurally blinded to individual inputs, yet mathematically capable of computing the valid aggregated global model.

**Table 1**  
Positioning of CE methods by the information they require.

| Method                                     | Setting  | Anonymity | Fairness         |
|--------------------------------------------|----------|-----------|------------------|
| FedGT (Xhemrishi and Östman, et al., 2025) | C-silo   | Reduced   | None             |
| Det.&Sco. Xhemrishi, Amat, and Pejó (2025) | C-silo   | Reduced   | None             |
| QI (Pejó & Biczók, 2023)                   | C-device | Maximal   | None             |
| COS (Xu et al., 2021)                      | C-silo   | Maximal   | None             |
| LOO (Black & Fredrikson, 2021)             | C-silo   | Maximal   | Null, Sym.       |
| FP (ours)                                  | C-silo   | Maximal   | Eff., Null, Sym. |
| EE (ours)                                  | C-silo   | Maximal   | Eff., Sym.       |

**Table 2**

Notation.

| Symbol     | Meaning                                             |
|------------|-----------------------------------------------------|
| $N$        | Number of clients                                   |
| $M_0$      | Initial model before the training round             |
| $M$        | Aggregated model after the training round           |
| $U_i$      | Client $i$ local update within the round            |
| $v(\cdot)$ | Evaluation function (negative loss, accuracy, etc.) |

**Masking with one-time pads.** The core protocol achieves perfect secrecy by masking local updates using pairwise random vectors. By establishing a strict total order on the clients  $i \in [N]$ , every unique pair of clients  $(i, j)$  where  $i < j$  agrees on a shared random vector  $s_{i,j}$ . When preparing their local updates, client  $i$  and  $j$  add and subtract this secret vector, respectively. Formally, each client  $i$  computes its masked update  $Y_i$  over a sufficiently large field size  $R$ :

$$Y_i = U_i + \sum_{j \in [N]: i < j} s_{i,j} - \sum_{j \in [N]: i > j} s_{j,i} \pmod{R}$$

Each client then transmits only its masked update  $Y_i$  to the server. Because the masks are indistinguishable from random noise, the server learns absolutely zero information about any individual  $U_i$ . The server aggregates these received vectors by computing:

$$Z = \sum_{i=1}^N Y_i \pmod{R} = \sum_{i=1}^N \left( U_i + \sum_{j: i < j} s_{i,j} - \sum_{j: i > j} s_{j,i} \right) \pmod{R} \quad (1)$$

Because every random vector  $s_{i,j}$  is added exactly once (by client  $i$ ) and subtracted exactly once (by client  $j$ ), all pairwise masks mutually cancel out during the server-side summation. This leaves the true accumulation intact:

$$Z = \sum_{i=1}^N U_i \pmod{R}$$

**Remark 1.** Although the basic masking scheme originally proposed by Google (Bonawitz et al., 2017) elegantly demonstrates how individual client updates remain completely hidden from the server, a naive implementation suffers from quadratic communication overhead and a vulnerability to client drop-outs (where a disconnected user leaves uncanceled masks that corrupt the final sum). Production-grade deployments resolve these limitations seamlessly: Diffie–Hellman key agreement combined with Pseudo-Random Generators (PRGs) reduces communication overhead, while Shamir’s  $t$ -out-of- $n$  secret sharing enables the server to securely reconstruct the missing masks of dropped-out clients. This reconstruction is strictly restricted to the cryptographic masks themselves; at no point during drop-out handling is any individual client update  $U_i$  ever exposed to the server, preserving privacy throughout the entire aggregation process.

### 3.2. Contribution evaluation

To establish an objective reference point for contribution assessment, most CE schemes employ an external dataset to evaluate the aggregated models. This approach is based on the premise that the value of a dataset correlates with the performance of a model trained

on it, thereby creating an intrinsic coupling between data valuation and performance validation (Liu et al., 2022; Song et al., 2019). Other methods attempt to estimate contributions by measuring the structural similarity between the global model and individual client updates, such as using Cosine Similarity (COS) (Xu et al., 2021). However, this similarity-driven approach suffers from fundamental limitations: COS captures only directional alignment while ignoring differences in gradient magnitude, feature importance, and decision boundaries (Draganov et al., 2024). Consequently, two models may appear highly aligned under this metric while exhibiting drastically different generalization behaviors. Moreover, using the aggregated global model as the baseline introduces additional bias, as it may incorporate malicious or low-quality client updates, rendering it an unreliable anchor (Sim et al., 2020).

In this work, we align with the mainstream marginal-difference-based CE literature (Ghorbani & Zou, 2019; Song et al., 2019) by assuming the availability of an external, representative global test dataset that reflects the overall population. Relying on an external global test set is the standard assumption in CE because it provides an objective, unbiased reference point for accurately quantifying client utility. While certain evaluation-free techniques exist that estimate contributions solely from model parameters or update trajectories, without a global test set, they fundamentally lack axiomatic fairness properties. Without an external global test anchor, these evaluation-free heuristics cannot map client updates to actual generalization performance, rendering them incapable of satisfying core game-theoretic principles such as efficiency or symmetry.

Thus, a representative global test set remains a necessary prerequisite for mathematically guaranteed fairness. In practical cross-silo FL deployments, such a dataset can be securely constructed. For instance, clients can leverage Generative Adversarial Networks (GANs) to generate representative, synthetic local test sets (Odena et al., 2017). Crucially, this generation process can be infused with Differential Privacy to protect local data boundaries (Venugopal et al., 2022; Yale et al., 2020), allowing the synthetic sets to be safely transmitted and combined at the server to create a secure, privacy-preserving global test set (Li et al., 2024).

**Shapley value.** In FL, the most widely used CE methods relying on a global test set are approximations of the *Shapley value* (SV) (Rozenberg et al., 2022). The SV of a client is defined as the weighted sum of its marginal contributions to all possible coalitions  $S \subseteq [N]$ . It relies on a utility function  $v(\cdot)$ , which assigns a value to each coalition of clients. In FL,  $v$  typically measures model performance (e.g., accuracy or negative loss); that is, for each coalition of clients  $S \subseteq [N]$ ,  $v(S)$  is the performance of the aggregate model of clients in  $S$  on the global test set.

**Definition 1 (Shapley Value (Shapley, 1951)).** The Shapley value of client  $i$  for a given utility function  $v$  is formalized below.

$$SV(i) = \frac{1}{N} \sum_{S \subseteq [N] \setminus \{i\}} \binom{N-1}{|S|} \cdot (v(S \cup \{i\}) - v(S)). \quad (2)$$

The SV is the only mechanism satisfying four fundamental properties (*linearity*, *efficiency*, *null player*, and *symmetry*), widely accepted as

collectively characterizing fair reward allocation. The linearity property concerns the utility functions  $v$  employed in the Shapley value, asserting that the allocation under the combined utility function  $v_1 + v_2$  equals the sum of the individual Shapley values computed under  $v_1$  and  $v_2$  independently. Unlike efficiency, symmetry, and the null player property, linearity is an algebraic constraint rather than a reflection of a client's relative contribution to a coalition (Kamil, 2005). Consequently, the primary design goal for establishing a fair CE method is to satisfy the remaining three core fairness properties.

**Property 1 (Efficiency).** *The total value generated by the grand coalition must be fully distributed among all players.*

**Property 2 (Null Player).** *If a player contributes no additional value to any coalition, its score should be zero.*

**Property 3 (Symmetry).** *If two players contribute equally to every possible coalition, they should receive the same score.*

Exact computation of the SV requires executing the entire federated learning process from scratch for every possible subset of clients  $S \subseteq [N]$ . For each newly formed coalition, the model is trained over multiple rounds, after which the aggregated model's performance is evaluated on a global test set. This exhaustive procedure is widely referred to as the *retraining game*. This retraining approach is computationally infeasible because it requires training an exponential number of models. To eliminate the need for retraining entirely, an approximation method proposed by Song et al. (2019) evaluates client contributions using the per-round gradient updates generated during the training of a single global model. Instead of launching new training pipelines for each subset, this method analyzes all possible combinations of client updates within each isolated communication round to compute client scores, which are then aggregated across the entire training process. We refer to this tracking-based approximation as the multi-round Shapley value (MR – SV).

A major drawback of these approaches is their incompatibility with secure aggregation, as they require access to individual client updates to construct the aggregated models for each coalition of clients. Under SA, only a limited set of submodels is available for evaluation at each round. The server has access only to  $M_0$  and  $M$ , while client  $i$  can compute two additional coalitions:  $M_0 + U_i$  and  $M - U_i$ . Consequently, only clients can compute marginal difference based CE scores using this restricted set of submodels. One approach to approximate the SV based on these submodels is *Leave-One-Out* (LOO) (Black & Fredrikson, 2021; Evgeniou et al., 2004), which evaluates a client's contribution by comparing the performance of the model of the grand coalition  $S = [N]$  against that of the model where the local update of the client is omitted,  $S' = [N] \setminus \{i\}$ .

**Definition 2 (Leave-One-Out).** The Leave-One-Out score of client  $i$  is defined as  $LOO(i) = v(M) - v(M - U_i)$ .

Importantly, only client  $i$  can evaluate both  $v(M_0 + U_i)$  and  $v(M - U_i)$ , since its update  $U_i$  is hidden from other clients and the server. This requires clients to have access to the global evaluation dataset shared by the server (the contribution test set in our experiments) to perform these evaluations. More problematically, it turns LOO into a *self evaluation* technique in which each client computes and reports its own score to the server, thereby becoming sensitive to manipulation by selfish clients.

### 3.3. Note on semivalues

As alternatives to the standard Shapley value, the broader family of *semivalues* has been increasingly utilized to address data valuation problems in machine learning. Formally, a semivalue is a function in the space of cooperative games that satisfies the symmetry, null player,

and linearity properties, but does not necessarily satisfy the efficiency property (Dubey et al., 1981). For instance, Beta Shapley (Kwon & Zou, 2021) and Data Banzhaf (Wang & Jia, 2023) are prominent semivalues that have demonstrated strong performance in tasks such as noisy-label detection, where the traditional Shapley value can be suboptimal (Kwon & Zou, 2021; Wang & Jia, 2023). Both the Shapley value and the LOO score discussed previously can be formalized within or related to this family (Dubey et al., 1981).

While semivalues are highly effective in centralized settings, the practical and conceptual shift introduced by federated learning brings new valuation challenges. In centralized data valuation, the objective is primarily descriptive: to measure which individual data points improve model performance the most. Because there is no requirement for a strictly fair, budget-balanced division of a reward, semivalues fit this description perfectly.

However, federated learning inherently demands *allocation-aware* scores. In FL, contribution evaluation is intimately tied to resource allocation decisions — such as distributing financial incentives, computational quotas, or communication priorities — which typically operate under a fixed, finite budget (Sim et al., 2020). Because FL requires dividing a fixed resource among participants, the *efficiency* property ceases to be optional and becomes a natural, practically necessary requirement.

**Efficiency vs linearity.** This necessity introduces a fundamental trade-off between efficiency and linearity. In contrast to the indispensable role of efficiency in federated learning, linearity can be safely relaxed. In cooperative game theory, linearity primarily serves as a technical and algebraic condition to establish uniqueness results, rather than as a foundational fairness property. Specifically, the classic proof for the uniqueness of the Shapley value relies on decomposing any arbitrary cooperative game into a unique linear combination of simple unanimity games. The core fairness properties (efficiency, symmetry, and the null player) uniquely pin down the allocation rule only within these simple basis games. Linearity is then invoked strictly as a mathematical tool to linearly extend this unique allocation from the basis games to the entire space of all possible games (Winter, 2002). Consequently, linearity does not dictate how rewards are equitably distributed among competing players; it merely ensures mathematical generalization across the game space. This purely structural role is further underscored by Kamil (2005), which explicitly points out that linearity lacks meaningful justification or sense when restricted to specific families of games. Therefore, relaxing linearity does not compromise the fundamental fairness of the evaluation method.

This tension between axiomatic properties is further exacerbated when privacy requirements are introduced. For example, under secure aggregation, the space of computable semivalues effectively collapses due to the limited number of observable marginal contributions. In fact, it is mathematically impossible to compute a contribution score that simultaneously satisfies linearity, symmetry, null player, efficiency, and SA-compatibility, because the only score satisfying Properties 1–3 and linearity is the Shapley value (Shapley, 1953), whose formula considers all coalitions and thus violates compatibility with secure aggregation (formalized as Property 4).

Ultimately, the choice of axioms underlying a data valuation score cannot be decoupled from the structure of the underlying machine learning task. While centralized settings may favor descriptive semivalues that relax efficiency, federated environments necessitate a departure from pure semivalues. Enforcing efficiency — even at the cost of linearity — is practically indispensable for driving incentive mechanisms under fixed-budget and strict-privacy constraints.

### 3.4. Threat model and goal

We study cross-silo Federated Learning with secure Aggregation. Strict SA compatibility is essential for preventing leakage of sensitive

**Table 3**  
Threat model summary.

| Entity                 | Assumptions                                                         | Capabilities                                                                    | Goals                                                        |
|------------------------|---------------------------------------------------------------------|---------------------------------------------------------------------------------|--------------------------------------------------------------|
| Server                 | Honest-but-curious; follows protocol; not malicious                 | Receives masked local updates; aggregates local update; aggregates local scores | Coordinate training; learn aggregated model                  |
| Client (training)      | Honest-but-curious; follows training protocol; uses true local data | Holds local dataset; sends masked local update to server                        | Aims to achieve best possible global model                   |
| Client (post training) | Selfish-but-non-malicious; does not disrupt training                | Knows local and aggregated update; estimates marginal differences               | Tries to maximize own contribution score or minimize others' |

training data via white-box attacks such as Deep Leakage from Gradients (DLG) (Zhu et al., 2019). In regulated environments (e.g., GDPR), exposing individual gradients is often legally prohibited. While some attacks target aggregated gradients, SA ensures that any potential leakage cannot be traced back to specific participants, making it a critical Privacy Enhancing Technology (PET) for cross-silo consortia (Heyndrickx et al., 2023).

Our trust assumptions are summarized in Table 3. The server is *honest-but-curious*: it faithfully executes the training protocol but may attempt to infer additional information from aggregated updates. Clients are modeled as *selfish*, aiming to maximize their individual utility. During training, utility is aligned with global model quality, and thus, honest local training is the dominant strategy. However, contribution evaluation, performed as a post-processing step after each round, introduces additional incentives. Clients may attempt to maximize their contribution scores either directly, by inflating reports, or indirectly, by reducing others' measured contributions.

We focus on a cross-silo federated learning setting, where participants are typically known organizations collaborating within established partnerships, rather than anonymous devices in an open participation environment. In such settings, as exemplified by real-world collaborations such as MELLODDY (Heyndrickx et al., 2023), participants have strong incentives to follow the agreed protocol, while strategic behavior related to resource allocation and contribution rewards may still arise. Accordingly, we exclude fully malicious behavior such as protocol deviation, Sybil attacks, or collusion, and treat clients as independent, non-colluding actors. These assumptions are consistent with prior work (Anada et al., 2025). More broadly, we adopt a *selfish-but-non-malicious* adversary model: clients may strategically optimize their outcomes within the protocol but do not arbitrarily deviate from its execution. This assumption does not imply robustness against arbitrary adversarial behavior. Attacks involving poisoned updates, fabricated data, identity-based attacks, collusion, or adversarial manipulation of validation information require additional defenses beyond contribution evaluation and are outside the scope of this work.

We treat Secure Aggregation as a black-box primitive: although we describe the standard masking construction in Section 3 for completeness, we do not propose, modify, or analyze the security of specific cryptographic protocols. Our contribution is orthogonal — ensuring CE mechanisms remain compatible with SA and rely only on the coalitions it exposes. In particular, CE must operate without access to individual updates and rely only on information available under SA (see Section 3.1).

Given these assumptions, we aim to design CE mechanisms that preserve privacy while remaining compatible with SA and robust against strategic manipulation of contribution reports.

#### 4. Private contribution evaluation

We focus on contribution scoring in a cross-silo setting, specifically evaluating contributions within a single, representative training round, rather than aggregating round-wise scores as in Song et al. (2019). Under SA, the only models accessible to all clients are the grand

coalition  $S = [N]$  and the empty coalition  $S = \emptyset$ , i.e.,  $M$  and  $M_0$ , respectively. However, client  $i$  individually has access to the model corresponding to the singleton coalition  $S = \{i\}$ , given by  $M_0 + U_i$ . Moreover, by removing its own update from the global model, client  $i$  can reconstruct the model corresponding to  $S = [N] \setminus \{i\}$ , i.e.,  $M - U_i$  (see Section 3.1).

Building on the above, we formalize the additional privacy property that a CE method must have to be compatible with SA, alongside the efficiency, null player, and symmetry properties.

**Property 4 (SA Compatibility).** A CE scheme is compatible with SA if the score assigned to each client depends only on the models corresponding to the coalitions  $[N]$ ,  $\emptyset$ ,  $[N] \setminus \{i\}$ , and  $\{i\} \forall i \in [N]$ .

##### 4.1. Self evaluation

In this section, we introduce a novel CE score, termed *Fair-Private* (FP). The key idea is to leverage the singleton coalition  $\{i\}$  available to client  $i$  to construct a CE score more expressive than the standard L00 score.

**Definition 3 (Fair-Private).** The Fair-Private score of client  $i$  is defined in Eq. (3), where  $\alpha$  is defined in (4).

$$FP(i) = \left( \frac{[\alpha(i)]^+}{\sum_{j=1}^N [\alpha(j)]^+} \right) \cdot v(M), \quad (3)$$

$$\alpha(i) = \frac{[v(M) - v(M - U_i)] + [v(M_0 + U_i) - v(M_0)]}{2}. \quad (4)$$

The first term in (4) is the Leave-One-Out score, while the second is its symmetric counterpart *Include-One-In* (IOI): instead of removing  $i$  from the grand coalition, we add it to the empty set. The combination of these marginal differences mirrors the SV, where coalitions of complementary sizes receive identical weights. The operator  $[\cdot]^+$  ( $\max(x, 0)$ ) clips negative marginals to zero, so that a harmful or non-contributing client receives no reward and the resulting shares remain non-negative. Our construction of the FP score further draws inspiration from the normalized Banzhaf power index (Dubey & Shapley, 1979). By applying this normalization in Eq. (3), we enforce the efficiency property while simultaneously preserving the other fairness and privacy properties, as established in the following theorem.

**Theorem 1 (Properties of Fair-Private).** FP satisfies Properties 1, 2, 3, and 4.

**Proof.** If client  $i$  is a null player, meaning it does not contribute to any coalition, then  $v(M_0 + U_i) = v(M_0)$  and  $v(M - U_i) = v(M)$ , implying  $\alpha(i) = 0$ ; hence  $[\alpha(i)]^+ = 0$  and  $FP(i) = 0$ . This means FP satisfies Property 2. Similarly, if two clients  $i$  and  $j$  contribute equally to all coalitions, then  $v(M - U_i) = v(M - U_j)$  and  $v(M_0 + U_i) = v(M_0 + U_j)$ , which implies

$$\alpha(i) = \frac{1}{2}((v(M) - v(M - U_i)) + (v(M_0 + U_i) - v(M_0)))$$

$$= \frac{1}{2}((v(M) - v(M - U_j)) + (v(M_0 + U_j) - v(M_0))) = \alpha(j),$$

and therefore  $[\alpha(i)]^+ = [\alpha(j)]^+$ , so  $\text{FP}(i) = \text{FP}(j)$  and FP satisfies the symmetry [Property 3](#). Finally, [Properties 1](#) and [4](#) follow directly from the definitions of FP and  $\alpha(i)$ , since  $\sum_{i=1}^N \text{FP}(i) = v(M)$  and the scores depend only on the models corresponding to the permitted coalitions.  $\square$

**Remark 2.** We account for the edge case where the denominator in Eq. (3) is zero (i.e.,  $[\alpha(j)]^+ = 0$  for all  $j$ ). In such a scenario we substitute  $\alpha(i)$  with either  $v(M) - v(M - U_i)$  or  $v(M_0 + U_i) - v(M_0)$ . This substitution preserves the axiomatic properties of the score while preventing a zero denominator. Additionally, note that the raw scores can be arbitrarily scaled and may take negative value, for instance, when the utility is defined as a negative loss function. To address this, ranking and reward allocation operate on the non-negative fractional shares, which remain well-defined and correctly ordered even when  $v(M)$  is negative (e.g., when the utility is a negative loss), while the factor  $v(M)$  certifies efficiency; for cross-method comparison we additionally apply the normalization of [Remark 3](#).

While FP enables privacy-preserving contribution evaluation, it does not, by itself, enforce truthful reporting of client-provided evaluation values. Ensuring incentive compatibility and verifiability is an orthogonal challenge that can be addressed through complementary mechanisms: peer prediction approaches ([Witkowski & Parkes, 2012](#)) provide incentives for truthful reporting without ground truth; cryptographic commitments and verification mechanisms have been successfully integrated into federated learning systems such as martFL ([Li et al., 2023](#)); finally, verifiable computation frameworks such as Pinocchio ([Parno et al., 2016](#)) enable proofs of correct computation without revealing private inputs. These techniques can be applied on top of FP to mitigate strategic reporting while preserving its privacy guarantees.

#### 4.2. Evaluating others

Note that all previous schemes (e.g., L00, FP) directly used  $U_i$  to assign a score to client  $i$ . Thus, those scores can only be computed by the respective client. Such self-evaluation scores have an inherent drawback: a selfish client can misreport or manipulate its model's performance, thereby gaining an unfair advantage and undermining the integrity of the CE process. To overcome this limitation, we introduce an alternative CE score termed *Everybody-Else* (EE) that, in addition to being compatible with SA, does not rely on self-evaluation, thereby providing resistance to direct manipulation. Here, manipulation resistance specifically refers to preventing direct self-assessment bias, in which a client can influence their own score through self-evaluation. This property is considered under the cross-silo FL setting defined in our threat model, where clients follow the prescribed protocol while strategically optimizing their utility.

**Property 5 (Manipulation Resistance).** A CE method is manipulation-resistant if no client can directly manipulate their own score.

To satisfy [Property 5](#), we introduce the idea of having clients evaluate each other, rather than themselves. This ensures that no client can directly influence its own score.

**Definition 4 (Everybody-Else).** The Everybody-Else score is defined in Eq. (5), where  $\beta$  and  $\gamma$  are defined in Eq. (6).

$$\text{EE}(i) = \left( \frac{[\beta(i)]^+ + [\gamma(i)]^+}{\sum_{j=1}^N \frac{[\beta(j)]^+ + [\gamma(j)]^+}{2}} \right) \cdot v(M), \quad (5)$$

$$\beta(i) = \sum_{j \neq i} \frac{[v(M) - v(M_0 + U_j)]}{(N-1)^2}; \quad \gamma(i) = \sum_{j \neq i} \frac{[v(M - U_j) - v(M_0)]}{(N-1)^2}. \quad (6)$$

Eq. (5) contains the same rescaling as before, while Eq. (6) formalizes our idea of having clients evaluate each other. The score of client  $i$  is determined collectively by the remaining clients  $j \in [N] \setminus \{i\}$ . Direct computation of client  $i$ 's score via marginal difference based evaluation is infeasible, as no other client  $j$  can compare two coalitions that differ solely in the presence or absence of client  $i$ . Instead, our key insight is that the contribution of client  $i$  can be inferred indirectly by assessing the collective contribution of the other clients. Specifically, each client  $j \neq i$  computes marginal differences involving itself.  $\beta(i)$  measures the effect of comparing the grand coalition  $[N]$  against  $\{j\}$ , while  $\gamma(i)$  measures the effect of comparing  $[N] \setminus \{j\}$  against  $\emptyset$ . Aggregating these evaluations across all  $j \neq i$ , we obtain an estimate of client  $i$ 's score without requiring self evaluation. As in FP, the operator  $[\cdot]^+$  clips negative channels to zero, which keeps the resulting shares non-negative.

Across all  $j \neq i$ , client  $i$  is the only participant that appears in every coalition considered in the summations for  $\beta(i)$  and  $\gamma(i)$ , while the other clients appear less frequently. Since client  $i$  is systematically present, the aggregation of these marginal differences over  $j \neq i$  yields a quantity that predominantly captures the contribution of client  $i$ . Hence,  $\text{EE}(i)$  can be interpreted as an indirect estimate of client  $i$ 's contribution, even though it is constructed solely from the evaluations of all other clients.

In line with the principles of the SV, each marginal difference in the summations of  $\beta(i)$  and  $\gamma(i)$  is weighted equally, since all of them are computed over coalitions of the same size. The normalization by  $(N-1)^2$  serves two purposes. First, the summation aggregates contributions from  $N-1$  clients (excluding client  $i$ ) and dividing by  $N-1$  yields their average. Second, each term corresponds to a coalition of  $N-1$  clients (excluding client  $j$ ), and we distribute its value uniformly across the remaining  $N-1$  clients. This normalization ensures that the resulting scores are on the same scale as the SV, as confirmed in our experiments.

For clarity, we next illustrate the definition with an example, followed by our theoretical results.

**Example 1 (Three Clients).** Consider a scenario with three clients:  $A$ ,  $B$ , and  $C$ . The score of client  $A$  is determined from the evaluations provided by clients  $B$  and  $C$ . Specifically, client  $B$  computes  $\gamma(A)$  as the marginal difference between the coalition  $\{A, C\}$  and the empty set  $\emptyset$ , and computes  $\beta(A)$  by comparing  $\{B\}$  with the grand coalition  $\{A, B, C\}$ . In this way, client  $B$  effectively evaluates the set  $\{A, C\}$ . Similarly, client  $C$  performs the same type of evaluations to measure the contribution of  $\{A, B\}$ . Combining these results in  $\text{EE}(A)$  produces a score that reflects both  $\{A, C\}$  and  $\{A, B\}$ . In these coalitions, all clients are included at least once, but only client  $A$  appears in every coalition. Because client  $A$  is systematically present, aggregating the differences across clients  $B$  and  $C$  yields a quantity that primarily reflects client  $A$ 's contribution.

**Theorem 2 (Properties of Everybody-Else).** EE satisfies [Properties 1, 3, 4](#), and [5](#).

**Proof.** We focus on the parts of the metric. [Property 4](#) follows directly from the definitions of  $\beta(i)$  and  $\gamma(i)$ , since these terms are composed of the available coalitions. For [Property 5](#), note that the only values a client  $i$  could potentially misreport are  $v(M - U_i)$  and  $v(M_0 + U_i)$ . However, these terms do not appear in  $\beta(i)$  or  $\gamma(i)$ , nor in their clipped forms. Therefore, EE satisfies it. To prove symmetry, assume that clients  $i$  and  $j$  contribute equally to all possible coalitions, i.e.,  $v(M - U_i) = v(M - U_j)$  and  $v(M_0 + U_i) = v(M_0 + U_j)$ . Then, based on (6):

$$\begin{aligned} \beta(i) &= \frac{\sum_{k \notin \{i,j\}} [v(M) - v(M_0 + U_k)] + [v(M) - v(M_0 + U_j)]}{(N-1)^2} \\ &= \frac{\sum_{k \notin \{i,j\}} [v(M) - v(M_0 + U_k)] + [v(M) - v(M_0 + U_i)]}{(N-1)^2} = \beta(j), \end{aligned}$$

$$\gamma(i) = \frac{\sum_{k \notin \{i,j\}} [v(M - U_k) - v(M_0)] + [v(M - U_j) - v(M_0)]}{(N-1)^2}$$

$$= \frac{\sum_{k \notin \{i,j\}} [v(M - U_k) - v(M_0)] + [v(M - U_i) - v(M_0)]}{(N-1)^2} = \gamma(j).$$

Since  $\beta(i) = \beta(j)$  and  $\gamma(i) = \gamma(j)$ , it follows that  $[\beta(i)]^+ = [\beta(j)]^+$  and  $[\gamma(i)]^+ = [\gamma(j)]^+$ , and thus  $EE(i) = EE(j)$ ; that is, EE satisfies [Property 3](#). Finally, [Property 1](#) is also satisfied, as the same normalizing step is used as in FP.  $\square$

Note that, as with FP, the same edge case wherein the denominator of Eq. (5) vanishes may also arise. This issue can be resolved similarly to the approach in [Remark 2](#) by substituting either  $\beta(i)$  or  $\gamma(i)$ . Furthermore, we apply the same normalization step described in [Remark 3](#) to EE to scale the scores and prevent negative values. As a consequence, note that [Property 5](#) holds only for the unnormalized quantities; the efficiency normalization couples all scores through the common denominator, re-introducing an indirect influence channel, which we quantify and discuss (together with its mitigation) in Section 5. Finally, we show that EE does not satisfy the null player axiom.

**Example 2.** Consider a scenario with three players where the coalition values are given by  $v(\emptyset) = 0$ ,  $v(\{A\}) = 0$ ,  $v(\{B\}) = 1$ ,  $v(\{C\}) = 2$ ,  $v(\{A, B\}) = 1$ ,  $v(\{A, C\}) = 2$ ,  $v(\{B, C\}) = 3$ , and  $v(\{A, B, C\}) = 3$ . Here, client A is clearly a null player, as adding A to any coalition does not increase its value. However, the score computed by EE is positive: the raw quantity  $(\beta(A) + \gamma(A))/2$  and the normalized score  $EE(A)$  of Eq. (5) both equal  $3/4$  in this instance, demonstrating that EE does not satisfy [Property 2](#).

Beyond linearity, our framework explicitly relaxes the null player property, a decision we defend as a conscious, strategic trade-off rather than an algorithmic oversight. In classic cooperative game theory, a null player is a passive actor contributing exactly zero marginal value. While a passive free-rider in federated learning might fit this definition by simply submitting zero gradients, such passive behavior is inherently non-problematic; it does not degrade the system, nor does it actively exploit the evaluation.

The critical vulnerability in FL lies not in passive free-riding, but in active strategic manipulation of the reward mechanism itself. In a self-reported setup, adversaries can easily manipulate, inflate, or misreport the local evaluations of their models to claim unfair rewards. Rigidly enforcing the null player property binds the framework to a naive assumption of passivity, leaving it entirely blind to some adversarial dynamics. To counter this, our method derives client scores exclusively from peer evaluations, meaning a client's reward is determined entirely by how the rest of the network evaluates them, not by what they report. We deliberately sacrifice the null player property to achieve this robust manipulation resistance, ensuring that dishonest participants cannot game their metrics. This relaxation is a practical necessity for secure FL.

#### 4.3. Overview

To facilitate understanding of our proposed schemes, their properties, and their interconnections, we provide a summary in [Table 4](#). The table not only indicates which scheme satisfies each axiom but also highlights the underlying building blocks on which each scheme relies. We then analyze the computational complexity of each scoring method.

**Theorem 3 (Complexity).** *Computing FP and EE for all clients requires  $\mathcal{O}(N)$  model evaluations.*

**Proof.** We prove the statement for each individual building block of the metrics, ensuring that their combination also inherits this validity. From [Definition 3](#), it is clear that computing  $\alpha(i)$  for a single client requires  $\mathcal{O}(1)$  model evaluations. Thus, computing the scores for all  $N$  clients requires  $\mathcal{O}(N)$  model evaluations overall. Similarly, according

**Table 4**

The building blocks for the scores and the properties they satisfy. For example,  $LOO(i) = v(M) - v(M - U_i)$ , which depends only on  $M$  and  $U_i$ .

| Cont. Eval. | $M_0$ | $M_0+U_i$ | $M_0+U_j$ | $M-U_j$ | $M-U_i$ | $M$ | Property | 1 | 2 | 3 | 4 | 5 |
|-------------|-------|-----------|-----------|---------|---------|-----|----------|---|---|---|---|---|
| $LOO(i)$    |       |           |           |         | •       | •   |          |   | ✓ | ✓ | ✓ | ✓ |
| $FP(i)$     | •     | •         |           |         |         | •   | ✓        | ✓ | ✓ | ✓ | ✓ | ✓ |
| $EE(i)$     | •     |           | •         | •       |         | •   | ✓        |   |   | ✓ | ✓ | ✓ |

**Table 5**

Overview of the experimental setup.

|         | ISIC2019    | PCam         | Brain-MRI   | CIFAR       |
|---------|-------------|--------------|-------------|-------------|
| Task    | Multi-class | Binary-class | Multi-class | Multi-class |
| Pred.   | Melanoma    | Metastatic   | Tumor       | Image       |
| Model   | ResNet50    | ResNet18     | Simple CNN  | Simple CNN  |
| Size    | 23,247      | 327,680      | 7023        | 60,000      |
| L-Epoch | 3           | 5            | 5           | 5           |

to Eq. (6), the complexity of computing the terms  $\beta(i)$  and  $\gamma(i)$  for a single client is  $\mathcal{O}(N)$ . However, the evaluations  $v(M_0+U_j)$  and  $v(M-U_j)$  for each client  $j$  are shared when computing the scores for different clients  $i$ , since each evaluation depends only on client  $j$ , not on the client being scored. Therefore, the overall computational cost remains  $\mathcal{O}(N)$ .  $\square$

## 5. Experiments

**Setup.** We evaluate our CE schemes in a cross-silo FL scenario using four image datasets: ISIC2019 ([Bcn20000: Dermoscopic lesions in the wild, 2024](#)), PatchCamelyon ([Veeling et al., 2018](#)), Brain-MRI ([Nickparvar, 2021](#)), and CIFAR10 ([Krizhevsky et al., 2009](#)). In contrast to cross-device FL, cross-silo deployments typically involve a limited number of known organizational participants, and our experiments, primarily with 6 and 9 clients (with scaling experiments up to 15), reflect this target setting. Extending CE mechanisms to large-scale cross-device scenarios with partial participation introduces additional challenges and is outside the scope of this work.

We use the ResNet architecture for ISIC2019 and PatchCamelyon. For Brain-MRI and CIFAR10, we use a 4-layer CNN: two convolutional layers, Conv2d(3,16,3) and Conv2d(16,32,3), followed by two linear layers; the final layer has dimension equal to the number of classes (4 for Brain-MRI, 10 for CIFAR10). Max pooling is used only for Brain-MRI. We utilized the Adam optimizer with both learning rate and weight decay of 0.001. Further details are provided in [Table 5](#). Throughout the paper, the utility function  $v$  is defined as the negative average loss evaluated on the contribution test set.

For each dataset, we construct a non-IID baseline by splitting the data across clients according to a Dirichlet distribution with parameter  $\mu = 0.5$ . All results are reported across 10 runs, and we report both mean and variance. Following prior work, we assume a global test set for CE that contains the same number of samples per class and is available to all clients ([Liu et al., 2022](#); [Song et al., 2019](#)). We use two disjoint held-out test sets: a contribution set, used solely to compute the marginal evaluations behind the scores, and a separate test set, used to measure global model performance and convergence. It is worth noting that the specific nature of the contribution test directly impacts client evaluation scores. However, a rigorous and exhaustive analysis of this relationship poses a highly complex challenge in its own ([Jahagirdar et al., 2024](#)). A full characterization of this issue lies outside the scope of our current assumptions and is preserved as a critical direction for future work.

The privacy-preserving score estimation follows a star topology, in which clients do not communicate directly with each other. Each client computes the marginal differences for its available coalitions and reports them to the server. The server then aggregates these reported

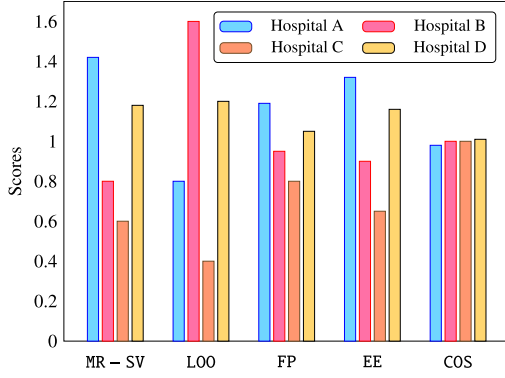


Fig. 1. Various normalized CE scores for a scenario with 4 clients ( $\mu = 0.5$ ) at the 10th communication round using the ISIC2019 dataset.

differences to obtain the final scores. To support reproducibility, we have open-sourced our implementation: [https://github.com/Delio090418/Fair\\_Private\\_Contribution\\_Evaluation](https://github.com/Delio090418/Fair_Private_Contribution_Evaluation).

**Evaluation metrics.** We compute all scores at a single training round, set to the 10th by default, as the models have reached convergence at this stage. Consequently, the proposed metrics capture a meaningful representation of each client's true contribution. The primary baseline CE score is the Multi-Round approximation of the Shapley value (MR – SV) as described in Song et al. (2019, Algorithm 2). Because computing the true Shapley value is computationally infeasible, MR – SV represents a fully adequate benchmark in this domain; it is widely recognized for effectively capturing dynamic, round-by-round client marginal contributions without requiring model retraining. However, to provide even more rigorous experimental validation, we include a dedicated case in which we explicitly compute the exact, true Shapley value as a definitive ground truth. We also compute the cosine similarity (COS), which (unlike the other scores) is accumulated across all training rounds to improve performance.

In CE, the relative ranking of clients is often more relevant, particularly for identifying the most and least valuable contributors. We therefore report the Spearman correlation coefficient  $\phi$  (Zar, 2005), the Kendall rank  $\kappa$  (Abdi, 2007), and the Pearson value  $\rho$  (Cohen et al., 2009). In our experiments, all three metrics showed consistent results, so we primarily focus on the Spearman coefficient. It measures how closely our schemes' rankings match the ground truth. A value of  $\phi = 1$  indicates perfect alignment with the baseline ranking,  $\phi = -1$  indicates a perfectly inverted ranking, and values near zero suggest no correlation.

**Remark 3.** Naturally, the absolute values of the scores are also important. Since CE scores can be arbitrarily scaled and may take negative values, we normalize them in two steps before comparison: (i) subtracting the minimum score to ensure non-negativity, and (ii) dividing by the mean to account for scale differences. In this normalized space, the  $L_2$  distance effectively captures the spatial proximity of the scores, which is critical for fair reward allocation. For instance, a score vector of  $[0.01, 0.02, 0.97]$  implies drastically different financial or resource incentives compared to  $[0.32, 0.33, 0.35]$ , even though the two vectors share the same ranking (imagine distributing \$10,000 proportional to those scores).

### 5.1. Comparative results

**Comparison to the multi-round shapley value.** Fig. 1 compares the proposed privacy-preserving normalized CE scores with the normalized MR – SV. Although the absolute values differ across schemes, the relative

Table 6

The various metrics of the privacy-preserving marginal difference based contribution scores, along with COS for a setting with 9 clients ( $\mu = 0.5$ ) at the 10th communication round.

|       |          | L00              | EE                                 | FP                                 | COS              |
|-------|----------|------------------|------------------------------------|------------------------------------|------------------|
| ISIC  | $L_2$    | $1.057 \pm .432$ | $0.339 \pm .159$                   | <b><math>0.112 \pm .168</math></b> | $2.028 \pm .807$ |
|       | $\phi$   | $0.821 \pm .042$ | <b><math>0.942 \pm .054</math></b> | $0.933 \pm .045$                   | $0.002 \pm .667$ |
|       | $\rho$   | $0.833 \pm .051$ | <b><math>0.964 \pm .019</math></b> | $0.961 \pm .034$                   | $0.349 \pm .398$ |
|       | $\kappa$ | $0.783 \pm .082$ | $0.901 \pm .024$                   | <b><math>0.912 \pm .022</math></b> | $0.098 \pm .235$ |
| Brain | $L_2$    | $0.612 \pm .261$ | $0.163 \pm .074$                   | <b><math>0.032 \pm .021</math></b> | $0.964 \pm .363$ |
|       | $\phi$   | $0.793 \pm .111$ | <b><math>0.891 \pm .063</math></b> | <b><math>0.891 \pm .063</math></b> | $0.083 \pm .065$ |
|       | $\rho$   | $0.831 \pm .022$ | $0.911 \pm .012$                   | <b><math>0.916 \pm .013</math></b> | $0.211 \pm .032$ |
|       | $\kappa$ | $0.782 \pm .112$ | <b><math>0.871 \pm .113</math></b> | <b><math>0.871 \pm .113</math></b> | $0.001 \pm .052$ |
| PCam  | $L_2$    | $0.667 \pm .096$ | $0.422 \pm .093$                   | <b><math>0.396 \pm .053</math></b> | $1.690 \pm .206$ |
|       | $\phi$   | $0.546 \pm .149$ | <b><math>0.946 \pm .035</math></b> | $0.944 \pm .024$                   | $0.211 \pm .376$ |
|       | $\rho$   | $0.645 \pm .002$ | $0.965 \pm .004$                   | <b><math>0.968 \pm .002</math></b> | $0.308 \pm .024$ |
|       | $\kappa$ | $0.476 \pm .133$ | $0.901 \pm .086$                   | <b><math>0.901 \pm .055</math></b> | $0.148 \pm .087$ |
| CIFAR | $L_2$    | $1.607 \pm .533$ | $0.077 \pm .045$                   | <b><math>0.033 \pm .014</math></b> | $1.234 \pm .106$ |
|       | $\phi$   | $0.684 \pm .203$ | <b><math>0.904 \pm .042</math></b> | $0.904 \pm .055$                   | $0.004 \pm .322$ |
|       | $\rho$   | $0.821 \pm .114$ | $0.951 \pm .024$                   | <b><math>0.955 \pm .022</math></b> | $0.012 \pm .303$ |
|       | $\kappa$ | $0.691 \pm .082$ | <b><math>0.875 \pm .043</math></b> | $0.869 \pm .046$                   | $0.013 \pm .273$ |

Table 7

The  $L_2$  error and coefficient  $\phi$  of our privacy-preserving marginal difference based scores compared to the true Shapley value for a setting with 9 clients ( $\mu = 0.5$ ) at the 10th communication round.

|       |        | MR – SV          | L00              | EE                                 | FP                                 |
|-------|--------|------------------|------------------|------------------------------------|------------------------------------|
| ISIC  | $L_2$  | $0.024 \pm .029$ | $1.384 \pm .051$ | $0.649 \pm .011$                   | <b><math>0.459 \pm .018</math></b> |
|       | $\phi$ | $0.990 \pm .002$ | $0.741 \pm .004$ | <b><math>0.867 \pm .062</math></b> | $0.861 \pm .073$                   |
| Brain | $L_2$  | $0.191 \pm .053$ | $1.660 \pm .410$ | $0.775 \pm .070$                   | <b><math>0.205 \pm .023</math></b> |
|       | $\phi$ | $0.982 \pm .012$ | $0.517 \pm .142$ | $0.971 \pm .023$                   | <b><math>0.975 \pm .023</math></b> |
| PCam  | $L_2$  | $0.161 \pm .028$ | $1.843 \pm .079$ | $0.308 \pm .049$                   | <b><math>0.304 \pm .061</math></b> |
|       | $\phi$ | $0.966 \pm .038$ | $0.749 \pm .051$ | <b><math>0.837 \pm .013</math></b> | <b><math>0.837 \pm .013</math></b> |
| CIFAR | $L_2$  | $0.025 \pm .006$ | $1.850 \pm .387$ | $0.352 \pm .053$                   | <b><math>0.137 \pm .029</math></b> |
|       | $\phi$ | $0.996 \pm .004$ | $0.483 \pm .319$ | <b><math>0.841 \pm .035</math></b> | $0.831 \pm .045$                   |

client rankings remain largely consistent. This demonstrates that our linear-time methods, despite considering far fewer coalitions than MR – SV (that is incompatible with SA), can still effectively capture the ranking of clients. Table 6 provides a more detailed comparison of the proposed CE scores, using both absolute error ( $L_2$ ) and ranking correlation metrics (Spearman, Kendall, and Pearson). Both FP and EE yield rankings that closely align with those obtained with the baseline MR – SV, whereas L00 exhibits a noticeably weaker correlation.

**Comparison to cosine similarity.** For completeness, we also include a comparison with an alternative privacy-preserving method based on cosine similarity, a widely used approach that does not rely on marginal differences (Evans et al., 2024; Wu & Wang, 2021; Xu et al., 2021). In this method, each client is assigned a score based on the similarity between its local gradient and the aggregated global gradient, i.e.,  $\text{COS}(i) = \text{CosSim}(M_0 + U_i, M)$ . This method also operates only with the coalitions available with SA. The last column of Table 6 presents the performance of the COS score with respect to MR – SV across all evaluation metrics. Both our proposed scoring schemes consistently outperform COS across all metrics, typically by a wide margin.

**Comparison to the true shapley value.** In Table 7, we also compare against the true Shapley value, which is based on the retraining game, where each coalition of clients trains a model from scratch using FL, rather than aggregating gradients at each round. The proposed CE scores exhibit a high correlation ( $\phi$  close to 1) with the true Shapley value, demonstrating that they effectively capture client contributions across diverse settings, including non-IID distributions and varying client counts. This indicates that our approximations preserve the relative importance of clients even when exact Shapley computation would be infeasible, providing a practical yet principled alternative for real-world deployments.

**Table 8**

Paired Wilcoxon signed-rank  $p$ -values on the per-run Spearman  $\phi$  (agreement with MR–SV) across the 10 runs (9 clients,  $\mu = 0.5$ , 10th round). Entries  $< 0.05$  indicate a statistically significant difference between the two methods.

|         | FP vs. EE | FP vs. L00 | EE vs. L00 | FP vs. COS | EE vs. COS |
|---------|-----------|------------|------------|------------|------------|
| ISIC    | 0.010     | 0.004      | 0.004      | 0.002      | 0.002      |
| Brain   | 0.921     | 0.037      | 0.049      | 0.002      | 0.002      |
| PCam    | 0.508     | 0.002      | 0.002      | 0.002      | 0.002      |
| CIFAR10 | 0.846     | 0.002      | 0.002      | 0.002      | 0.002      |

**Statistical significance.** The comparisons above rank each method by how closely its scores agree with the reference MR – SV, measured per run by the Spearman coefficient  $\phi$ . To verify that the observed gaps are not artifacts of run-to-run variance, we test the methods *against each other* on these per-run  $\phi$  values using a paired Wilcoxon signed-rank test across the 10 runs. The test is paired because all methods share the same data partition and seed within a run, so pairing cancels the cross-run variance reflected in the  $\pm$  intervals and isolates the method-to-method difference. (MR – SV is the reference embedded in  $\phi$ , not a competitor, and is therefore not part of the test.) Results are reported in Table 8. FP and EE outperform L00 and COS with  $p < 0.05$  on all datasets, confirming the gains are statistically significant rather than noise. The difference between FP and EE is not significant on most datasets ( $p > 0.05$ ), indicating that the two proposed scores are statistically comparable in their agreement with MR–SV, consistent with their shared construction.

### 5.2. Computational overhead

Beyond the asymptotic  $\mathcal{O}(N)$  analysis of Theorem 3, Table 9 reports the empirical wall-clock execution times per round on the ISIC2019 dataset. Here, the No CE row establishes the absolute baseline duration for a standard federated learning round (comprising global validation on the server and local training across clients). For all subsequent rows, the entries represent the *additional overhead* introduced by the contribution evaluation mechanism.

We decouple this overhead by the operational side: *Server time* measures the computational cost of executing weight or marginal aggregations across client subsets, whereas *Client time* quantifies the duration of running extra inference passes ( $v(\cdot)$ ). The linear-time schemes (COS, L00, FP, and EE) require virtually no server intervention but incur a client-side overhead that scales predictably with the number of evaluated combinations, matching the  $\sim 1.2$  s required per sequential model evaluation. Concretely, L00 requires two additional model evaluations per client ( $v(M)$  and  $v(M - U_i)$ ), while FP and EE require four (adding  $v(M_0)$  and  $v(M_0 + U_i)$ ); EE’s cross-scoring is subsequently performed by the server on the reported values at negligible cost. Conversely, MR – SV completely frees the clients by transferring the entire computational load to the server, which sequentially processes all  $2^9 = 512$  subset combinations ( $\sim 614.4$  s of evaluation plus  $\sim 0.6$  s of aggregation).

**Remark 4 (Retraining Game Complexity).** These overheads remain exceptionally low because they operate under an evaluation-based paradigm utilizing pre-existing models. If a full retraining game were implemented instead — requiring models to be trained from scratch for all  $2^9$  coalitions — the computational overhead would scale multiplicatively against the client baseline ( $\sim 1520$  s per training cycle). This would catastrophically shift the total execution time from minutes to days of continuous compute, demonstrating the necessity of evaluation-only metrics.

**Table 9**

Server- and client-side wall-clock overhead per round on the ISIC2019 dataset. Values represent the mean and variance overhead (in seconds) over 10 runs with 9 clients ( $\mu = 0.5$ ) at the 10th communication round, measured relative to a baseline with no CE computation. All experiments were conducted on an NVIDIA A40 GPU.

|         | Server time          | Client time         |
|---------|----------------------|---------------------|
| No CE   | 120.0 $\pm$ 1.8      | 1520.0 $\pm$ 22.5   |
| COS     | 0.000 $\pm$ 0.000    | 0.0003 $\pm$ 0.0001 |
| L00     | 0.0001 $\pm$ 0.00003 | 2.4 $\pm$ 0.12      |
| FP      | 0.0001 $\pm$ 0.00003 | 4.8 $\pm$ 0.21      |
| EE      | 0.0002 $\pm$ 0.00003 | 4.8 $\pm$ 0.21      |
| MR – SV | 615.0 $\pm$ 8.2      | 0.000 $\pm$ 0.000   |

**Table 10**

The Spearman coefficient of MR–SV with L00, EE, and FP with 9 clients ( $\mu = 0.5$ ) at different communication rounds: 5, 10, 15.

|       |    | L00              | EE                      | FP                      |
|-------|----|------------------|-------------------------|-------------------------|
| ISIC  | 5  | 0.735 $\pm$ .022 | <b>0.843</b> $\pm$ .080 | 0.842 $\pm$ .074        |
|       | 10 | 0.821 $\pm$ .042 | <b>0.942</b> $\pm$ .054 | 0.933 $\pm$ .045        |
|       | 15 | 0.905 $\pm$ .006 | <b>0.986</b> $\pm$ .001 | <b>0.986</b> $\pm$ .001 |
| Brain | 5  | 0.731 $\pm$ .199 | <b>0.881</b> $\pm$ .088 | 0.880 $\pm$ .093        |
|       | 10 | 0.793 $\pm$ .111 | <b>0.891</b> $\pm$ .063 | <b>0.891</b> $\pm$ .063 |
|       | 15 | 0.900 $\pm$ .102 | <b>0.986</b> $\pm$ .002 | 0.983 $\pm$ .009        |
| PCam  | 5  | 0.311 $\pm$ .285 | 0.866 $\pm$ .178        | <b>0.869</b> $\pm$ .160 |
|       | 10 | 0.546 $\pm$ .149 | <b>0.946</b> $\pm$ .035 | 0.944 $\pm$ .024        |
|       | 15 | 0.803 $\pm$ .011 | <b>0.952</b> $\pm$ .004 | <b>0.952</b> $\pm$ .004 |
| CIFAR | 5  | 0.606 $\pm$ .034 | 0.784 $\pm$ .147        | <b>0.785</b> $\pm$ .137 |
|       | 10 | 0.684 $\pm$ .203 | <b>0.904</b> $\pm$ .042 | 0.904 $\pm$ .055        |
|       | 15 | 0.829 $\pm$ .164 | <b>0.973</b> $\pm$ .023 | <b>0.973</b> $\pm$ .023 |

**Table 11**

The Spearman coefficient of MR – SV with L00, EE, and FP when the data is split into 6, 9, 12, and 15 clients ( $\mu = 0.5$ ) at the 10th communication round.

|       |    | L00              | EE                      | FP                      |
|-------|----|------------------|-------------------------|-------------------------|
| ISIC  | 6  | 0.862 $\pm$ .001 | <b>0.941</b> $\pm$ .015 | 0.940 $\pm$ .013        |
|       | 9  | 0.821 $\pm$ .042 | <b>0.942</b> $\pm$ .054 | 0.933 $\pm$ .045        |
|       | 12 | 0.733 $\pm$ .022 | 0.882 $\pm$ .020        | <b>0.890</b> $\pm$ .011 |
|       | 15 | 0.682 $\pm$ .031 | 0.864 $\pm$ .078        | <b>0.870</b> $\pm$ .051 |
| Brain | 6  | 0.932 $\pm$ .012 | 0.950 $\pm$ .050        | <b>0.961</b> $\pm$ .024 |
|       | 9  | 0.793 $\pm$ .111 | <b>0.891</b> $\pm$ .063 | <b>0.891</b> $\pm$ .063 |
|       | 12 | 0.745 $\pm$ .083 | 0.844 $\pm$ .103        | <b>0.854</b> $\pm$ .141 |
|       | 15 | 0.612 $\pm$ .243 | <b>0.841</b> $\pm$ .028 | 0.841 $\pm$ .051        |
| PCam  | 6  | 0.737 $\pm$ .005 | 0.980 $\pm$ .006        | <b>0.981</b> $\pm$ .005 |
|       | 9  | 0.546 $\pm$ .149 | <b>0.946</b> $\pm$ .035 | 0.944 $\pm$ .024        |
|       | 12 | 0.528 $\pm$ .032 | 0.905 $\pm$ .077        | <b>0.906</b> $\pm$ .035 |
|       | 15 | 0.512 $\pm$ .006 | <b>0.866</b> $\pm$ .033 | 0.866 $\pm$ .054        |
| CIFAR | 6  | 0.793 $\pm$ .151 | <b>0.955</b> $\pm$ .023 | 0.948 $\pm$ .021        |
|       | 9  | 0.684 $\pm$ .203 | <b>0.904</b> $\pm$ .042 | 0.904 $\pm$ .055        |
|       | 12 | 0.601 $\pm$ .094 | <b>0.892</b> $\pm$ .072 | <b>0.892</b> $\pm$ .072 |
|       | 15 | 0.431 $\pm$ .006 | 0.862 $\pm$ .005        | <b>0.868</b> $\pm$ .004 |

### 5.3. Ablation study

We next analyze the impact of three factors on CE: the training round at which the evaluation is performed, the number of clients, and the distribution of data across clients. Table 10 reports the correlation of our proposed scores with MR – SV when the evaluation is carried out five rounds earlier or later than in the default setting. Table 11 shows the correlation for varying numbers of clients. Finally, Table 12 presents the performance of scores relative to L00 for different values of  $\mu$  and the non-IID Dirichlet partitioning of the data. These results indicate that our scoring schemes remain reliable under realistic variations in training dynamics, client populations, and data heterogeneity, confirming their robustness for practical cross-silo deployments.

As expected, performance improves in later rounds for nearly all methods, since contributions are easier to assess once the model has

**Table 12**

The Spearman coefficient of MR – SV with L00, EE, and FP with 9 clients ( $\mu \in \{0.1, 0.5, 1.0\}$ ) at the 10th communication round.

|       |     | L00              | EE                      | FP                      |
|-------|-----|------------------|-------------------------|-------------------------|
| ISIC  | 0.1 | 0.868 $\pm$ .013 | <b>0.981</b> $\pm$ .004 | 0.980 $\pm$ .006        |
|       | 0.5 | 0.821 $\pm$ .042 | <b>0.942</b> $\pm$ .054 | 0.933 $\pm$ .045        |
|       | 1.0 | 0.822 $\pm$ .022 | <b>0.916</b> $\pm$ .032 | 0.916 $\pm$ .056        |
| Brain | 0.1 | 0.801 $\pm$ .032 | <b>0.945</b> $\pm$ .037 | 0.935 $\pm$ .039        |
|       | 0.5 | 0.793 $\pm$ .111 | <b>0.891</b> $\pm$ .063 | <b>0.891</b> $\pm$ .063 |
|       | 1.0 | 0.677 $\pm$ .147 | <b>0.817</b> $\pm$ .139 | 0.816 $\pm$ .123        |
| PCam  | 0.1 | 0.757 $\pm$ .013 | 0.980 $\pm$ .023        | <b>0.981</b> $\pm$ .012 |
|       | 0.5 | 0.546 $\pm$ .149 | <b>0.946</b> $\pm$ .035 | 0.944 $\pm$ .024        |
|       | 1.0 | 0.534 $\pm$ .076 | 0.924 $\pm$ .038        | <b>0.931</b> $\pm$ .021 |
| CIFAR | 0.1 | 0.880 $\pm$ .014 | 0.963 $\pm$ .033        | <b>0.965</b> $\pm$ .033 |
|       | 0.5 | 0.684 $\pm$ .203 | <b>0.904</b> $\pm$ .042 | 0.904 $\pm$ .055        |
|       | 1.0 | 0.648 $\pm$ .021 | <b>0.850</b> $\pm$ .042 | 0.848 $\pm$ .046        |

**Table 13**

The relative influence (i.e., the columns sum up to 1) of a single client on another's EE score using the Brain dataset with 6 clients ( $\mu = 0.5$ ) at the 10th communication round. We only show the mean as the variance is always below 0.01.

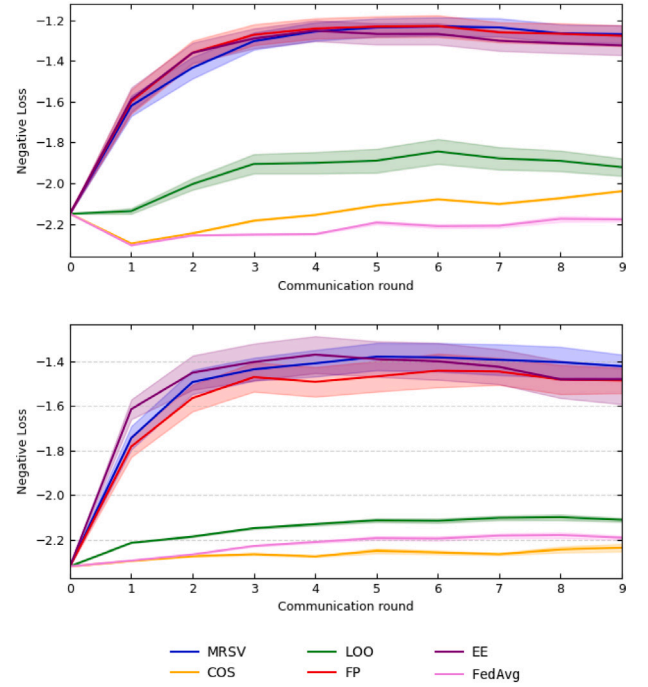
| No. clients | 1     | 2     | 3     | 4     | 5     | 6     |
|-------------|-------|-------|-------|-------|-------|-------|
| 1           | 0.000 | 0.145 | 0.151 | 0.169 | 0.150 | 0.167 |
| 2           | 0.129 | 0.000 | 0.132 | 0.147 | 0.131 | 0.145 |
| 3           | 0.170 | 0.167 | 0.000 | 0.194 | 0.172 | 0.192 |
| 4           | 0.273 | 0.268 | 0.280 | 0.000 | 0.278 | 0.309 |
| 5           | 0.164 | 0.161 | 0.167 | 0.187 | 0.000 | 0.185 |
| 6           | 0.262 | 0.257 | 0.268 | 0.300 | 0.266 | 0.000 |

converged. In contrast, increasing the number of clients leads to reduced correlation, reflecting the widening gap between the exponentially complex Shapley value and our linear-time approximations. Nevertheless, both FP and EE stay stable, maintaining strong performance even under more challenging conditions such as early-round evaluation or larger client populations. Regarding data distribution, higher values of  $\mu$  lead to reduced performance of the scoring schemes. This behavior is expected, since larger  $\mu$  produces distributions closer to IID, making clients harder to distinguish. Nevertheless, FP and EE remain robust across all three aspects studied.

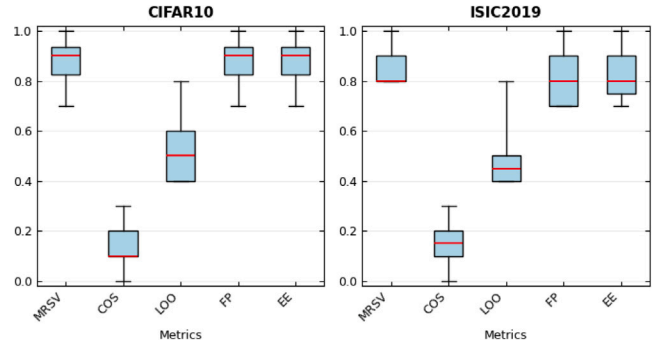
#### 5.4. Robustness

Next, we analyze the impact of selfish behavior in the post-processing phase, focusing on scenarios where a client misreports its marginal contribution to gain an advantage. While FP and L00 are vulnerable to this type of manipulation, recall that clients cannot influence their own scores in the EE scheme. On the other hand, according to Eq. (5), client  $i$  can influence the score of others to the extent of  $(v(M) - v(M_0 + U_i)) + (v(M - U_i) - v(M_0))$ . We present the relative, percentage-wise influence in Table 13.

The main diagonal of the table is zero, since no client contributes to its own score. This demonstrates that EE is resistant to selfish behavior. Note that it remains vulnerable to strategic misreporting: a client could artificially deflate its reported value to diminish other clients' scores. However, such malicious downvoting is effectively a “kamikaze” attack, as recently demonstrated in Pejo et al. (2026). Furthermore, this vulnerability can be mitigated by augmenting the averaging step in EE's definition in Eq. (6) with anomaly detection techniques or by adopting Byzantine-resilient aggregation rules (e.g., substituting the summation with a median operator). A similar attack scenario under analogous trust assumptions was explored in the context of decentralized FL in Anada et al. (2025). Indeed, the reported scores are all in the rather conservative range ( $\pm 0.1$  from the average 0.2), and the clients tend to agree as the scores in each row are rather consistent. Consequently, a deviation can be easily flagged.



**Fig. 2.** The negative loss of the global model on the CIFAR10 (top), and ISIC2019 (bottom) datasets. Training is for 9 clients with 10 communication rounds; mean and variance are reported over 10 runs.



**Fig. 3.** Misbehavior detection score (the probability that the attacker gets the lowest score) on the CIFAR10 (left) and ISIC2019 (right) datasets. Training is for 9 clients with 10 communication rounds; results are aggregated over 10 runs.

#### 5.5. Downstream tasks

**Performance gain.** To demonstrate practical utility beyond ranking alignment, we evaluate global model performance when aggregating client updates weighted by their CE scores. For the evaluation, we use a global test different from the contribution test set used to compute the scores, which contains the same number of samples per class. We assess the effectiveness of our proposed scoring schemes when clients contribute data of varying quality, as in Pejó and Biczók (2023). We consider IID clients with different levels of label noise: labels in each client's dataset are randomized with a linearly increasing probability before training. With 9 clients, the label-flipping rates are  $[0.00, 0.11, 0.22, \dots, 0.89, 1.00]$ . During aggregation, each update is weighted by the current value of its normalized score, so that top-performing and bottom-performing participants receive weights above and below 1, respectively.

In Fig. 2, it is visible that both FP and EE are superior to FedAvg (naive uniform aggregation) and to the baselines COS and L00. In fact,

their performance almost reaches that of MR – SV. These results imply that 1) even the limited information available with SA could be used beneficially, and 2) the potential highest gain of client weighting is almost fully reachable, even with privacy in mind. Hence, both FP and EE successfully identify and upweight high-quality contributors while suppressing noise, translating their valuation information into tangible gains in model performance.

**Misbehavior detection.** Another frequently utilized benchmarking downstream task is the identification of dishonest participants. Again, we consider honest IID clients with a single exception: one is a Byzantine client executing an adversarial attack (label flipping).

In Fig. 3, we measure the detection rate of the scoring schemes by counting how often the attacker received the lowest score (out of ten experiments). It is visible that both FP and EE are superior to L00 and C0S, almost reaching the upper bound of MR – SV. These results imply that privacy-preserving misbehavior detection is practical.

## 6. Conclusion

We study contribution evaluation in cross-silo Federated Learning under secure aggregation, where individual client updates must remain hidden while reward allocation and accountability still require meaningful per-client scoring. This creates a fundamental mismatch: classical marginal-contribution methods require access to many coalitions and individual updates, whereas secure aggregation exposes only the empty and grand coalitions. To bridge this gap, we propose two privacy-preserving marginal difference scores fully compatible with secure aggregation: Fair-Private (FP), which leverages the limited coalitions available to each client to satisfy standard fairness axioms, and Everybody-Else (EE), which avoids self evaluation and provides manipulation resistance through indirect cross-scoring. We establish formal guarantees for fairness, privacy, and computational efficiency, and empirically demonstrate across multiple medical image datasets and CIFAR10 that FP and EE substantially outperform existing secure-aggregation-compatible baselines, closely approximating Shapley-induced rankings while remaining practical at scale. Quantitatively, across datasets FP and EE attain Spearman correlations of 0.89 – 0.95 with the multi-round Shapley value, versus 0.55 – 0.82 for Leave-One-Out, and reduce the normalized  $L_2$  ranking error by up to an order of magnitude.

Overall, FP and EE preserve Shapley-based fairness principles while producing stable, interpretable client contributions under challenging conditions, including non-IID data and larger client populations within the cross-silo regime. Meaningful client valuations can thus be obtained efficiently and privately, supporting reward allocation and misbehavior detection without costly architectural changes or added overhead.

Future work includes extending private contribution evaluation to client sampling, churn, and asynchronous training, and strengthening manipulation resistance by combining EE-style cross-evaluation with robust aggregation or anomaly detection.

## Declaration of competing interest

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

## Acknowledgments

Project no. 145832, implemented with the support provided by the Ministry of Innovation and Technology from the NRDI Fund, financed under the PD\_23 funding scheme. Balázs Pejó was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00148/26).

## Data availability

Public data was used.

## References

- Abdi, H. (2007). The Kendall rank correlation coefficient. In *Encyclopedia of measurement and statistics*. Sage.
- Anada, H., Kaneko, T., & Takamaeda-Yamazaki, S. (2025). Measuring participant contributions in decentralized federated learning. *ArXiv Preprint*: 2505.23246 [Cs.LG].
- (2024). Bcn20000: Dermoscopic lesions in the wild. *Scientific Data*.
- Black, E., & Fredrikson, M. (2021). Leave-one-out unfairness. In *Proc. ACM conf. fairness, accountability, and transparency (fAccT)*.
- Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In *Proc. ACM SIGSAC conf. computer and communications security*.
- Chu, X., Jiang, X., Qiu, R., Gao, J., & Zhao, J. (2025). MODEL SHAPLEY: Find your ideal parameter player via one gradient backpropagation. In *The thirty-ninth annual conference on neural information processing systems (neurIPS)*.
- Cohen, I., Huang, Y., Chen, J., & Benesty, J. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer.
- Cramer, R., Damgård, I. B., & Nielsen, J. B. (2015). *Secure multiparty computation and secret sharing*. Cambridge University Press.
- Draganov, A., Vadgama, S., & Bekkers, E. J. (2024). The hidden pitfalls of the cosine similarity loss. In *ICML workshop: high-dimensional learning dynamics 2024: the emergence of structure and reasoning*.
- Dubey, P., Neyman, A., & Weber, R. J. (1981). Value theory without efficiency. *Mathematics of Operations Research*.
- Dubey, P., & Shapley, L. S. (1979). Mathematical properties of the banzhaf power index. *Mathematics of Operations Research*.
- Evans, N. J., Mills, G. B., Wu, G., Song, X., & McWeeney, S. (2024). Data valuation with gradient similarity. *ArXiv Preprint*: 2405.08217 [Cs.LG].
- Evgeniou, T., Pontil, M., & Elisseeff, A. (2004). Leave one out error, stability, and generalization of voting combinations of classifiers. *Mach. Learn.*.
- Ghorbani, A., & Zou, J. (2019). Data Shapley: Equitable valuation of data for machine learning. In *Proc. int. conf. on mach. learn.*.
- Heyndrickx, W., Mervin, L., Morawietz, T., Sturm, N., Friedrich, L., Zalewski, A., Pentina, A., Humbeck, L., Oldenhof, M., Niwayama, R., et al. (2023). Melloddy: Cross-pharma federated learning at unprecedented scale unlocks benefits in qsar without compromising proprietary information. *Journal of Chemical Information and Modeling*.
- Jahagirdar, H., Wang, J. T., & Jia, R. (2024). Data valuation in the absence of a reliable validation set. *Transactions on Machine Learning Research*.
- Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., et al. (2021). Advances and open problems in federated learning. *Found. Trends Mach. Learn.*.
- Kamil, A. A. (2005). The controversy of linearity in Shapley's axioms. *Matematika*.
- Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.
- Kwon, Y., & Zou, J. (2021). Beta Shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*.
- Li, Q., Li, X., Liu, Z., & Qi, H. (2024). Pfedce: Personalized federated learning based on contribution evaluation. In *Proc. int. conf. artif. intell., robotics, and communication*.
- Li, Q., Liu, Z., Li, Q., & Xu, K. (2023). Martfl: Enabling utility-driven data marketplace with a robust and verifiable federated learning architecture. In W. Meng, C. D. Jensen, C. Cremers, & E. Kirda (Eds.), *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*. ACM, <http://dx.doi.org/10.1145/3576915.3623134>.
- Lin, H., Wan, S., Xie, Z., Chen, K., Zhang, M., Shou, L., & Chen, G. (2024). A comprehensive study of Shapley value in data analytics. *ArXiv Preprint*: 2412.01460 [Cs.DB].
- Liu, Z., Chen, Y., Yu, H., Liu, Y., & Cui, L. (2022). GTG-Shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM Trans. Intell. Syst. Technol.*.
- Ma, S., Cao, Y., & Xiong, L. (2021). Transparent contribution evaluation for secure federated learning on blockchain. In *Proc. int. conf. data eng. workshops*.
- Nickparvar, M. (2021). Brain tumor MRI dataset. <https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset>. (Accessed 30 April 2025).
- Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier GANs. In *Proc. int. conf. mach. learn.*.
- Parno, B., Howell, J., Gentry, C., & Raykova, M. (2016). Pinocchio: nearly practical verifiable computation. *Commun. ACM*, <http://dx.doi.org/10.1145/2856449>.
- Pejó, B., & Biczók, G. (2023). Quality inference in federated learning with secure aggregation. *IEEE Trans. Big Data*.
- Pejó, B., & Desfontaines, D. (2022). *Guide to differential privacy modifications: a taxonomy of variants and extensions*. Springer Nature.

- Pejo, B., Frank, M., Varga, K., Veliczky, P., & Biczok, G. (2026). On the fragility of contribution score computation in federated learning. *IEEE Conference on Secure and Trustworthy Machine Learning*.
- Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., et al. (2022). The Shapley value in machine learning. In *Proc. int. joint conf. artif. intell.*
- Shamir, A. (1979). How to share a secret. *Communications of the ACM*.
- Shapley, L. S. (1951). *Notes on the N-person Game*. Rand Corporation.
- Shapley, L. S. (1953). A value for n-person games. *Contributions To the Theory of Games*.
- Sim, R. H. L., Xu, X., & Low, B. K. H. (2022). Data valuation in machine learning: "ingredients", strategies, and open challenges. In *Proc. int. joint conf. artif. intell.*
- Sim, R. H. L., Zhang, Y., Chan, M. C., & Low, B. K. H. (2020). Collaborative machine learning with incentive-aware model rewards. In *Proc. int. conf. mach. learn.*
- Siomos, V., & Passerat-Palmbach, J. (2023). Contribution evaluation in federated learning: Examining current approaches. ArXiv Preprint: 2311.09856 [Cs.LG].
- Song, T., Tong, Y., & Wei, S. (2019). Profit allocation for federated learning. In *Proc. IEEE int. conf. big data (big data)*.
- Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., & Welling, M. (2018). Rotation equivariant CNNs for digital pathology. ArXiv Preprint: 1806.03962 [Cs.CV].
- Venugopal, R., Shafqat, N., Venugopal, I., Tillbury, B. M. J., Stafford, H. D., & Bourazeri, A. (2022). Privacy preserving generative adversarial networks to model electronic health records. *Neural Networks*.
- Wang, J. T., & Jia, R. (2023). Data banzhaf: A robust data valuation framework for machine learning. In *Proceedings of the 26th international conference on artificial intelligence and statistics*. PMLR.
- Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., & Qi, H. (2019). Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE conference on computer communications*.
- Watson, L., Andreeva, R., Yang, H.-T., & Sarkar, R. (2022). Differentially private Shapley values for data evaluation. ArXiv Preprint: 2206.00511 [Cs.LG].
- Winter, E. (2002). The shapley value. *Handbook of Game Theory with Economic Applications*.
- Witkowski, J., & Parkes, D. C. (2012). Peer prediction without a common prior. In B. Faltings, K. Leyton-Brown, & P. Ipeirotis (Eds.), *Proceedings of the 13th ACM conference on electronic commerce, EC 2012, valencia, Spain, June 4-8, 2012*. ACM, <http://dx.doi.org/10.1145/2229012.2229085>.
- Wu, H., & Wang, P. (2021). Fast-convergent federated learning with adaptive weighting. *IEEE Trans. Cogn. Commun. Netw.*
- Xhemrishi, M., i Amat, A. G., & Pejó, B. (2025). Detect & score: Privacy-preserving misbehaviour detection and contribution evaluation in federated learning. ArXiv Preprint: 2506.23583 [Cs.CR].
- Xhemrishi, M., Östman, J., Wachter-Zeh, A., & i Amat, A. G. (2025). FedGT: Identification of malicious clients in federated learning with secure aggregation. *IEEE Trans. Inf. Forensics Security*.
- Xu, X., Lyu, L., Ma, X., Miao, C., Foo, C. S., & Low, B. K. H. (2021). Gradient driven rewards to guarantee fairness in collaborative machine learning. In *Proc. adv. neural inf. process. syst. (neurIPS)*.
- Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*.
- Zar, J. H. (2005). Spearman rank correlation. *Encyclopedia of Biostatistics*.
- Zheng, S., Cao, Y., & Yoshikawa, M. (2023). Secure Shapley value for cross-silo federated learning. In *Proc. VLDB endowment*.
- Zhou, S., Liao, M., Qiao, B., & Yang, X. (2022). A survey of security aggregation. In *Proc. int. conf. adv. commun. technol.*
- Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. In *Advances in neural information processing systems*.