Trust Fusion with Choquet-Based Interaction Aggregation in Federated Learning

1st Alicja Rachwał
Dept. of Computational Intelligence
Lublin University of Technology
Lublin, Poland
alicja.rachwal@pollub.pl

2nd Albert Rachwał
Dept. of Computational Intelligence
Lublin University of Technology
Lublin, Poland
a.rachwal@pollub.pl

3rd Paweł Karczmarek
Dept. of Computational Intelligence
Lublin University of Technology
Lublin, Poland
p.karczmarek@pollub.pl

***Abstract*—Federated learning (FL) enables collaborative model training without centralizing raw data, but its performance can degrade under non-identically distributed (non-IID) client data and unreliable client updates. This paper introduces Φ-FL, a server-side aggregation strategy that augments Federated Averaging with trust fusion based on multiple reliability signals computed on a small server proxy set, robust magnitude control via interquartile-range (IQR) norm clipping, and interaction-sensitive reweighting using a 2-additive Choquet-inspired mechanism to promote complementary trusted updates. We evaluate Φ-FL on a suite of benchmark tabular classification datasets under Dirichlet-partitioned non-IID settings, using repeated paired runs. Compared to FedAvg, Φ-FL improves the mean test accuracy from 69.56% to 74.74%, achieving higher mean accuracy on 9 out of 12 datasets. In relation to more advanced baseline model, FLTrust, the proposed method Φ-FL achieves average accuracy higher by 1.7 percentage point, while the median accuracy increases by 3.7 percentage point. The significance of the obtained results is confirmed by statistical tests.**

***Index Terms*—Federated Learning, Information Fusion, Choquet Integral, Agent Trust.**

I. Introduction

Modern machine learning systems increasingly rely on large and diverse datasets to achieve high predictive performance. At the same time, many high-impact application domains are inherently data-sensitive: medical records, biometric signals, financial transactions, behavioral traces, and industrial telemetry often contain information that is personal, proprietary, or regulated. In such settings, privacy is not only a desirable feature but a practical requirement shaped by legal constraints, organizational policies, and the simple fact that data owners may be unwilling or unable to centralize raw data in a single location.

From an engineering perspective, this tension exposes a fundamental bottleneck of classical centralized training. While centralized pipelines offer straightforward optimization and monitoring, they require data movement, long-term storage, and broad access privileges, which increase exposure to breaches, misuse, or unintended disclosure. This is particularly problematic when data resides across multiple institutions or devices and cannot be pooled due to compliance constraints or governance barriers. Consequently, the machine learning community has sought training paradigms that preserve the utility of collaborative learning while reducing the need for raw data exchange.

Federated learning (FL) addresses this need by enabling distributed model training in which clients keep raw data locally and share only model parameters or updates with a coordinating server [1]. As a result, FL has become relevant to a wide range of real-world deployments where distributed data ownership and privacy constraints are the norm rather than the exception.

Federated learning has already been explored in several practically relevant settings where data are naturally fragmented across devices or institutions. It has been used for on-device next-word prediction in mobile keyboards, demonstrating that user-generated text data can remain on personal devices while still supporting collaborative model improvement [2]. In healthcare, FL has been investigated both as a general paradigm for multi-institutional digital health systems [3] and in concrete predictive tasks based on electronic health records, such as mortality prediction across multiple hospitals [4]. It has also been adopted in federated medical imaging workflows, for example in the FeTS initiative for brain tumor segmentation across geographically distributed sites [5]. Outside healthcare, federated learning is increasingly considered in other domains characterized by strong data-governance constraints and heterogeneous local data sources. Recent studies report FL-based approaches for collaborative credit card fraud detection across financial institutions [6], predictive maintenance and quality inspection in industrial environments [7], and intrusion detection in IoT ecosystems [8].

FL introduces its own set of challenges that are less pronounced in centralized learning [9]. Client data are typically non-identically distributed, often imbalanced, and affected by variable data quality and heterogeneous acquisition processes. A recent benchmark study systematically evaluates personalized federated learning methods across diverse datasets and client settings, and reports that standard FedAvg combined with fine-tuning can remain

highly competitive under strong heterogeneity [10].

However, work [11] shows that uncertainty-aware aggregation in federated learning, including Choquet-integral fusion and entropy-based weighting, improves robustness in medical applications. In other studies, Smooth OWA is applied in Federated Learning with quadrature-based smoothing of OWA input values to stabilize and adapt the aggregation of client updates beyond simple averaging [12]. The paper [13] proposes Adaptive Robust Clipping (ARC), an adaptive gradient clipping strategy that dynamically sets clipping thresholds to improve robustness of federated and distributed learning, particularly in heterogeneous and adversarial settings. Recent work [14] also highlights that static clipping thresholds can be suboptimal in federated optimization, motivating adaptive clipping strategies that improve stability across tasks and settings.

Beyond simple averaging, aggregation can be formalized using fuzzy integrals that explicitly model interactions among criteria, with the Choquet integral being a classical and well-established tool in multicriteria aggregation [15]. Learning general fuzzy measures is computationally challenging due to the exponential number of coefficients and monotonicity constraints, hence recent work explores complexity-reduction principles such as k-interactivity with linear-programming-based learning [16]. Alongside the classical Choquet integral, various extensions and alternatives have been proposed for interaction-aware aggregation; notably, NEUR-HCI shows that hierarchical MCDM models composed of 2-additive Choquet aggregators can be identified automatically while enforcing Choquet-model constraints in an end-to-end neural architecture [17]. Quadrature-inspired generalizations of the Choquet integral replace the standard term under the integral sign (e.g., product/t-norm) with Newton–Cotes-like constructions to better reflect numerical integration and improve aggregation fidelity [18].

This work focuses on improving the robustness of server-side aggregation while preserving the standard client-side training protocol of FedAvg. We introduce Φ-FL, an aggregation strategy that performs trust fusion by combining multiple reliability signals and integrates interaction-sensitive weighting via a Choquet-based mechanism. The proposed design is motivated by the observation that client updates should not only be weighted by data volume, but also by their behavioral credibility and their complementarity with other trusted clients. To this end, Φ-FL estimates per-client trust on a small proxy set, mitigates extreme update magnitudes using robust clipping, and reweights the remaining updates using a 2-additive interaction model that can reward informative diversity while suppressing unstable or redundant contributions.

The rest of this paper is structured as follows: section II presents mathematical preliminaries, section III contains detailed description of the proposed algorithm, section IV provides the setup and analysis of results from numerical experiments, while section V concludes the research and highlights possible future work directions.

II. Theoretical Background

This section summarizes the mathematical and conceptual preliminaries required for the proposed aggregation mechanism.

A. Federated Learning

Federated learning (FL) considers a set of clients $\{1, \dots, M\}$ that collaboratively train a model without directly pooling raw data. Client k holds a local dataset $\mathcal{D}_k$ with n_k samples, and the global model is parameterized by $w \in \mathbb{R}^d$. A common reference objective is the weighted empirical risk

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \; F(w) &= \sum_{k=1}^{M} \frac{n_k}{n} F_k(w), \\ F_k(w) &= \frac{1}{n_k} \sum_{(x,y) \in \mathcal{D}_k} \ell\big(f_w(x), y\big). \end{aligned} \tag{1}$$

where $n = \sum_k n_k$ and ℓ denotes the training loss.

Training proceeds in communication rounds. At round t, the server broadcasts $w^{(t)}$ to clients, where $w^{(t)}$ means global model parameters in the round t. Each client performs local optimization starting from $w^{(t)}$ and returns either updated parameters $w_k^{(t)}$ or an update vector

$$\Delta w_k^{(t)} = w_k^{(t)} - w^{(t)}. \tag{2}$$

The server aggregates received updates and produces $w^{(t+1)}$. A canonical aggregation rule is Federated Averaging (FedAvg) [19].

B. Robust, Uncertainty, and Interaction Primitives

The proposed Φ-FL aggregation relies on several standard building blocks. First, robust statistics, in particular the median and interquartile range (IQR), are used to control unusually large update norms. Second, cosine similarity is employed to quantify directional agreement or disagreement between client updates. Third, predictive uncertainty and within-class representation dispersion, evaluated on a small proxy set, are used as behavioral indicators of client reliability. Finally, fuzzy membership functions are used to combine heterogeneous trust signals into a single trust score, while a 2-additive Choquet-style interaction model is used to account for pairwise complementarity among the retained client updates. Since these concepts are instantiated directly within the proposed server aggregation routine, their exact formulas are introduced in Section III.

III. Proposed Methodology

The proposed Φ-FL method is designed for non-IID client data and aims to reduce the impact of unreliable, unstable, or redundant updates. A compact overview of the proposed method is shown in Fig. 1. The yellow elements correspond to the conventional federated learning

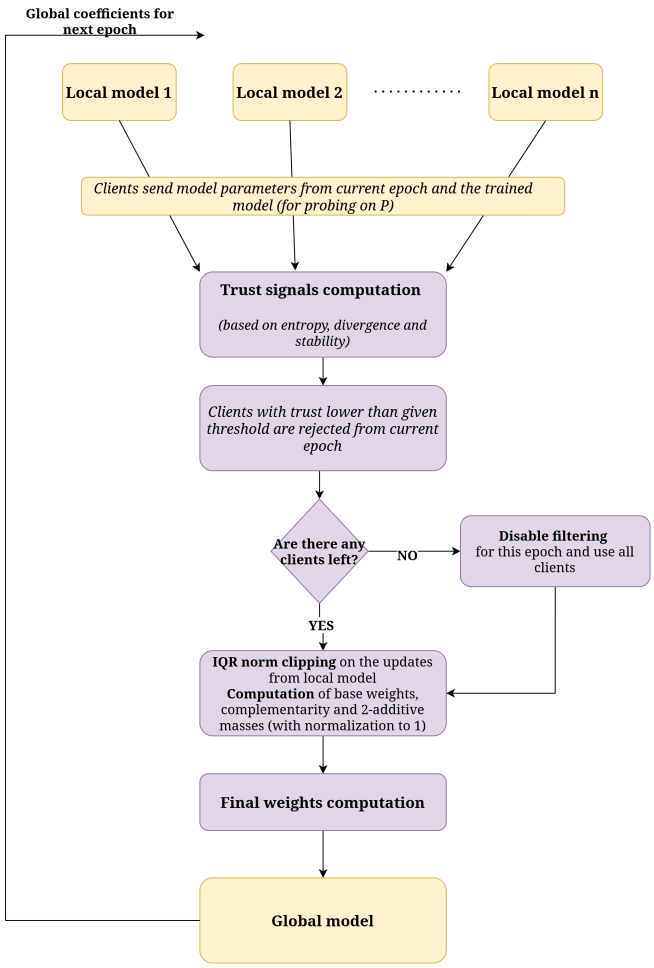


Fig. 1: Compact workflow of the proposed Φ-FL method.

procedure, while the blue elements highlight the additional trust and robust interaction-aware fusion modules introduced by Φ-FL.

At each round t, Φ-FL executes a four-step server routine: trust signal extraction on a proxy set, fuzzy trust inference with filtering, robust update clipping, and Choquet-inspired interaction reweighting. The server samples a small proxy subset $\mathcal{P}$ from the training split (before client partitioning). This set is used solely to probe each locally trained client model in a consistent way. For each client k in round t, we compute three signals:

(i) Predictive entropy. Let $p_k(\cdot \mid x)$ be the softmax distribution of the locally trained client model. We define the mean predictive entropy over $\mathcal{P}$ as

$$H_k^{(t)} = \frac{1}{|\mathcal{P}|} \sum_{x \in \mathcal{P}} \Big(-\sum_{c=1}^{C} p_k(c \mid x) \log p_k(c \mid x) \Big), \quad (3)$$

where C is the number of classes. Higher values of $H_k^{(t)}$ indicate that the predictive distribution is less concentrated, i.e., the model assigns more similar probabilities to multiple classes and is therefore less certain about its decisions on the proxy set. Since predictive entropy is widely used as a compact uncertainty descriptor in machine learning, we interpret high entropy as a sign that the current local model is less reliable from the server's perspective [20], [21].

(ii) Update divergence. We quantify abrupt directional changes in the client update trajectory by the cosine-based divergence

$$D_k^{(t)} = \begin{cases} 0, & t = 0, \\ \max\left(0,\, 1 - \cos(\Delta w_k^{(t)}, \Delta w_k^{(t-1)})\right), & t > 0, \end{cases} \quad (4)$$

which penalizes erratic update directions relative to the previous round.

(iii) Representation stability. Let $h_k(x)$ be the penultimate representation of the client model for x. For each class c with proxy subset $\mathcal{P}_c$, we compute within-class dispersion

$$\begin{aligned} \mathrm{Var}_{k,c}^{(t)} &= \frac{1}{|\mathcal{P}_c|} \sum_{x \in \mathcal{P}_c} \big\| h_k(x) - \mu_{k,c} \big\|_2^2, \\ \mu_{k,c} &= \frac{1}{|\mathcal{P}_c|} \sum_{x \in \mathcal{P}_c} h_k(x). \end{aligned} \quad (5)$$

and map the aggregated dispersion to a bounded stability score

$$S_k^{(t)} = \frac{1}{1 + \frac{1}{C} \sum_{c=1}^{C} \mathrm{Var}_{k,c}^{(t)}}. \quad (6)$$

Intuitively, low stability means that proxy samples belonging to the same class are mapped to more scattered penultimate-layer representations, which indicates weaker intra-class compactness and a less coherent class structure. This interpretation is aligned with the broader representation-learning literature, where compact within-class features are commonly associated with more effective and better-generalizing embeddings [22], [23].

We convert the triple $(H_k^{(t)}, D_k^{(t)}, S_k^{(t)})$ into a single trust value $T_k^{(t)} \in [0,1]$ via membership functions and multiplicative fusion (Section II). Concretely, we use a triangular membership for entropy, a decreasing sigmoid for divergence, and an increasing sigmoid for stability:

$$\begin{aligned} \mu_H(H) &= \mathrm{Tri}(H; a, b, c), \\ \mu_D(D) &= \frac{1}{1 + \exp\big(\kappa_D (D - x_{0D})\big)}, \\ \mu_S(S) &= \frac{1}{1 + \exp\big(-\kappa_S (S - x_{0S})\big)}. \end{aligned} \quad (7)$$

The entropy triangle is parameterized relative to the maximum entropy scale: we set $b = \rho \log(C)$ with $\rho \in (0,1)$ and choose a, c as low/high fractions of $\log(C)$, which avoids tying the trust mechanism to a single dataset-specific regime. The final trust is

$$T_k^{(t)} = \mathrm{clip}_{[0,1]}\Big(\mu_H(H_k^{(t)}) \cdot \mu_D(D_k^{(t)}) \cdot \mu_S(S_k^{(t)}) \Big). \quad (8)$$

Clients with $T_k^{(t)} < \tau$ are excluded from aggregation in the current round, and we define $\mathcal{K}^{(t)} = \{k : T_k^{(t)} \geq \tau\}$.

If $\mathcal{K}^{(t)} = \emptyset$, we disable filtering and use all the clients in that round. In our setting, the trust score $T_k^{(t)}$ reflects the reliability of the current client update, not a fixed long-term property of client k. Under non-IID data partitions, the behavioral quality of local updates may vary across communication rounds. As a result, a client that is temporarily excluded in one round may be re-admitted in later rounds if its update becomes more reliable again.

Even among the retained clients, update magnitudes can vary substantially. To make the aggregation more robust to unusually large updates, we apply norm clipping based on the interquartile range (IQR). Let $\{\|\Delta w_k^{(t)}\| : k \in \mathcal{K}^{(t)}\}$ denote the set of update norms in round t. We define the median $\mathrm{med}(\|\Delta w^{(t)}\|)$ and the interquartile range

$$\mathrm{IQR}(\|\Delta w^{(t)}\|) = Q_3(\|\Delta w^{(t)}\|) - Q_1(\|\Delta w^{(t)}\|), \tag{9}$$

where Q_1 and Q_3 are the first and third quartiles of the norm distribution. A robust clipping threshold is then set as

$$\theta^{(t)} = \mathrm{med}(\|\Delta w^{(t)}\|) + \gamma \cdot \mathrm{IQR}(\|\Delta w^{(t)}\|), \tag{10}$$

where $\gamma > 0$ controls the clipping strength. Each retained client update is then rescaled only if its norm exceeds this threshold:

$$\widetilde{\Delta w}_k^{(t)} = \begin{cases} \Delta w_k^{(t)}, & \|\Delta w_k^{(t)}\| \le \theta^{(t)}, \\ \Delta w_k^{(t)} \cdot \dfrac{\theta^{(t)}}{\|\Delta w_k^{(t)}\|}, & \text{otherwise.} \end{cases} \tag{11}$$

In the implementation, a small constant ε may be added to the denominator as a numerical safeguard; however, it is not required by the mathematical definition above. Since the second branch is applied only when $\|\Delta w_k^{(t)}\| > \theta^{(t)} \ge 0$, the denominator is strictly positive.

We then define base weights that blend data volume and trust:

$$\bar{w}_k^{(t)} = n_k \, T_k^{(t)}, \qquad w_k^{(t)} = \frac{\bar{w}_k^{(t)}}{\sum_{j \in \mathcal{K}^{(t)}} \bar{w}_j^{(t)}}. \tag{12}$$

with a size-only fallback if the normalization becomes numerically degenerate.

Base weights treat clients independently. To incorporate simple interaction effects, we compute a complementarity score between clipped updates:

$$C_{ij}^{(t)} = \mathrm{clip}_{[0,1]}\big(1 - \cos(\widetilde{\Delta w}_i^{(t)}, \widetilde{\Delta w}_j^{(t)})\big), \tag{13}$$

where larger values indicate more directionally distinct updates.

Using the 2-additive construction (Section II), we define singleton and pairwise masses for a mixing parameter $\lambda \in [0, 1]$:

$$m_i = (1-\lambda)\, w_i^{(t)}, \qquad m_{ij} = \lambda\, w_i^{(t)} w_j^{(t)} C_{ij}^{(t)}, \quad i < j, \tag{14}$$

followed by normalization so that $\sum_i m_i + \sum_{i<j} m_{ij} = 1$. Final aggregation weights are obtained via the Shapley-style induced contributions

$$\alpha_i^{(t)} = m_i + \frac{1}{2} \sum_{j \ne i} m_{ij}, \qquad \sum_{i \in \mathcal{K}^{(t)}} \alpha_i^{(t)} = 1, \ \alpha_i^{(t)} \ge 0. \tag{15}$$

The server update rule becomes

$$w^{(t+1)} = w^{(t)} + \eta_s \sum_{k \in \mathcal{K}^{(t)}} \alpha_k^{(t)} \, \widetilde{\Delta w}_k^{(t)}. \tag{16}$$

When $\lambda = 0$, the interaction terms vanish and $\alpha^{(t)}$ reduces to the base weights $w^{(t)}$. For $\lambda > 0$, clients that are both trusted and complementary gain additional influence.

Algorithm 1 Φ-FL Server Aggregation for Round t

Require: Global parameters $w^{(t)}$, proxy set $\mathcal{P}$, client set $\{1, \ldots, M\}$.

1: Broadcast $w^{(t)}$ to all clients.
2: Each client k trains locally and returns $(\Delta w_k^{(t)}, n_k)$; the server also receives the trained model for trust probing.
3: For each client k, compute trust signals on $\mathcal{P}$: entropy $H_k^{(t)}$, divergence $D_k^{(t)}$, stability $S_k^{(t)}$.
4: Compute trust $T_k^{(t)}$ via Eq. (8); set $\mathcal{K}^{(t)} = \{k : T_k^{(t)} \ge \tau\}$, fallback to all clients if $\mathcal{K}^{(t)} = \emptyset$.
5: Clip updates $\Delta w_k^{(t)} \mapsto \widetilde{\Delta w}_k^{(t)}$ for $k \in \mathcal{K}^{(t)}$ (Eq. (11)).
6: Compute base weights $w_k^{(t)}$ on $\mathcal{K}^{(t)}$ (Eq. (12)).
7: Compute complementarity $C_{ij}^{(t)}$ (Eq. (13)).
8: Compute 2-additive masses m_i, m_{ij} (Eq. (14)) and normalize total mass to 1.
9: Compute final weights $\alpha_i^{(t)}$ (Eq. (15)).
10: Update global model $w^{(t+1)}$ (Eq. (16)).

For clarity, Algorithm 1 summarizes the server routine executed at each communication round. The client-side optimizer and training loop remain identical to FedAvg; all modifications occur on the server.

IV. Numerical Experiments

This section presents the numerical experiments conducted to evaluate the proposed Φ-FL method. First, the experimental setup is described, including the datasets, model configuration, and evaluation protocol. Next, the predictive performance of Φ-FL is compared against the considered baseline methods using accuracy-based analyses and statistical tests. Finally, a sensitivity analysis is provided to assess the robustness of the method with respect to its main hyperparameters.

A. Experimental setup

To simulate statistical heterogeneity, the training set is partitioned among M clients using Dirichlet label partitioning with concentration $\alpha_{\mathrm{dir}} = 0.3$. If any client receives an empty partition, we repair the split by transferring a single sample from the currently largest partition to the

empty one, ensuring that every client contributes at least one local update.

We compare two baselines, FedAvg and FLTrust, against the proposed Φ-FL in a simulated cross-device setting with $M = 10$ clients. FLTrust is a Byzantine-robust federated learning method that bootstraps trust from a small clean server-side reference dataset; in each round, it computes a server update on this trusted data, assigns trust scores to client updates according to their directional alignment with the server update, and normalizes client update magnitudes before weighted averaging [24]. It is a particularly relevant baseline in our setting because, similarly to Φ-FL, it relies on a small trusted server-side subset for reliability assessment, while differing in the way trust is modeled and incorporated into aggregation. The choice of $M = 10$ reflects a compromise between simulating client-level decentralization and preserving sufficiently informative local partitions, especially for the smaller datasets considered in this study. For datasets such as Glass (214 samples) or Ionosphere (351 samples), increasing the number of clients further would, after the train-test split and proxy-set extraction, produce very small Dirichlet-based local subsets, which could make both local optimization and trust estimation unstable.

On each client, we train a lightweight three-layer MLP for tabular classification with 128 hidden units. Local optimization uses SGD with learning rate 0.05, momentum 0.9, weight decay 10^{-4}, and batch size 32. Training proceeds for $T = 20$ global communication rounds, and in each round every participating client performs $E = 1$ local epoch before returning the update vector $\Delta w_k^{(t)}$ (Eq. (2)). The server applies aggregated updates with step size $\eta_s = 1.0$;

In Φ-FL, clients with trust below 0.05 are excluded from aggregation (unless this would exclude all clients in a round). To improve robustness, client updates are clipped using an interquartile-range rule with multiplier 2.5, and interaction-aware fusion is performed via a 2-additive Choquet weighting with parameter $\lambda = 0.35$. After each communication round, we evaluate test accuracy and test log-loss. To quantify run-to-run variability, the full experimental pipeline is repeated over 100 independent runs for each dataset and method.

B. Comparison of accuracy against baseline models

Table I presents detalied comparison of test accuracy for all datasets and methods. It can be observed that Φ-FL attains higher mean test accuracy than FedAvg on 9 out of 12 datasets and higher than FLTrust on 6 out of 12 datasets, indicating that the proposed trust-aware aggregation improves performance for the majority of the considered benchmark problems. Relative to FedAvg, the improvements are statistically significant for all datasets marked with a superscript asterisk in Table I. Relative to FLTrust, Φ-FL yields statistically significant improvements on 5 datasets, as indicated by the subscript asterisks.

TABLE I: Comparison of test accuracy (mean ± std. dev. [%]) for FedAvg, FLTrust and Φ-FL. A superscript asterisk indicates that the result of Φ-FL is significantly better than FedAvg, whereas a subscript asterisk indicates that it is significantly better than FLTrust, according to the paired Wilcoxon test with $\alpha = 0.05$.

Dataset	FedAvg	FLTrust	Φ-FL
Zoo	68.16 ± 9.53	74.60 ± 7.87	$83.88 \pm 4.94^{*}_{*}$
Segment	78.34 ± 3.51	90.74 ± 2.05	$89.38 \pm 1.66^{*}$
Wine	86.82 ± 9.67	96.71 ± 2.94	$96.40 \pm 3.04^{*}$
Iris	79.24 ± 10.39	76.61 ± 9.49	$88.71 \pm 5.74^{*}_{*}$
Glass	53.56 ± 9.97	65.43 ± 8.68	$61.96 \pm 11.32^{*}$
Letter Recognition	78.39 ± 1.06	84.18 ± 1.92	$83.51 \pm 0.95^{*}$
Lung Cancer	37.57 ± 12.62	39.14 ± 14.16	$42.43 \pm 15.26^{*}$
EEG Eye State	72.87 ± 6.69	70.32 ± 12.53	$75.09 \pm 9.98^{*}_{*}$
Ionosphere	78.72 ± 10.18	87.61 ± 4.41	$80.52 \pm 10.25^{*}$
Winequality Red	56.33 ± 1.80	55.68 ± 3.22	55.32 ± 4.40
Winequality White	53.41 ± 1.82	50.12 ± 3.94	$51.71 \pm 4.53_{*}$
Isolet	91.36 ± 1.35	85.20 ± 3.19	$87.99 \pm 6.28_{*}$

The largest improvements of Φ-FL over FedAvg were observed for Zoo, Segment, Wine, Iris, and Glass datasets. In particular, the gain reached 15.72 percentage points for Zoo, 11.04 for Segment, 9.58 for Wine, 9.47 for Iris, and 8.40 for Glass. The clearest gains over FLTrust are observed on Zoo and Iris, where Φ-FL improves the mean test accuracy by approximately 9.28 and 12.10 percentage points, respectively. On the remaining datasets where Φ-FL outperforms FLTrust, the gains are typically more moderate, usually within the range of about 1–5 percentage points. At the same time, pronounced degradations relative to FLTrust are rare. The most visible decreases are observed on Glass and Ionosphere, whereas on the remaining datasets the results of both methods remain broadly comparable. For Winequality Red, the observed difference is not statistically significant.

TABLE II: Global descriptive statistics of test accuracy [%] aggregated over all paired runs.

Statistic	FedAvg	FLTrust	Φ-FL
Minimum	0	0	14.29
25-th percentile	55.75	57.41	57.37
Mean	69.56	73.03	74.74
Median	73.51	78.95	82.63
75-th percentile	81.58	87.51	88.64
Maximum	100	100	100
Standard deviation	17.20	18.41	18.43

Table II complements this per-dataset analysis by reporting global descriptive statistics of test accuracy. The average accuracy increases from 69.56 (FedAvg) to 74.74 (Φ-FL), corresponding to a mean gain of approximately 5.18. Consistent with the dataset-level comparison, the global mean accuracy is higher for Φ-FL than for FedAvg by 5.18 percentage point, and the median increases from 73.51% for FedAvg to 82.63% for *Phi*-FL. At the same time, the distributional summary reveals substantial variability across tasks, reflected by the relatively large stan-

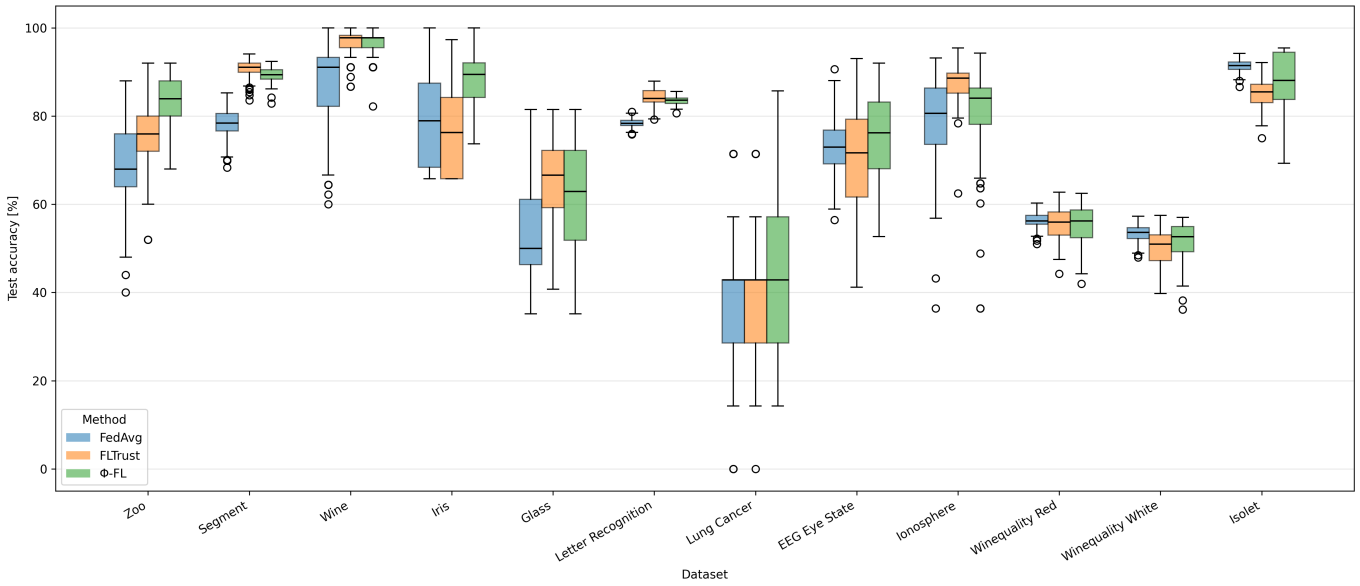


Fig. 2: Per-dataset boxplots of test accuracy comparing FedAvg, FLTrust and Φ-FL across paired runs.

dard deviations and the spread between quartiles, which is expected given the heterogeneity of the benchmark suite.

The results indicate that Φ-FL also outperforms FLTrust at the global level. *Phi*-FL is the only method for which the minimum observed accuracy is greater than 0%, reaching 14.29%. The 25th percentile is nearly identical for FLTrust and Φ-FL, but clearer differences emerge in the subsequent statistics. In particular, the mean accuracy is higher for Φ-FL by 1.71 percentage points, the median is higher by 3.68 percentage points, and the 75th percentile is higher by 1.13 percentage points. Both methods achieve the same maximum of 100%, while their standard deviations are practically identical.

The overall significance of the differences among the three compared methods was confirmed by the Friedman test with statistic value $\chi^2 = 220.463$ and $p < 0.001$. This result was further supported by the post-hoc Nemenyi and Conover tests with Holm correction, reported in Tables IIIa and IIIb. All pairwise comparisons were found to be statistically significant. Taken together with the descriptive statistics and per-dataset results, these findings support the ranking of the compared methods in ascending order of accuracy as follows: FedAvg, FLTrust, and Φ-FL.

Figure 2 summarizes the test accuracy distributions for FedAvg, FLTrust and Φ-FL separately for each dataset in the considered benchmark suite. Across most datasets, the Φ-FL boxplots are visibly shifted upwards relative to FedAvg and, in several cases, also relative to FLTrust, which indicates that the performance gains are not limited to isolated runs but are observed repeatedly across random seeds.

The clearest improvements of Φ-FL are visible for datasets such as Zoo and Iris. Favorable distributions

TABLE III: Post-hoc pairwise comparison p-values for the considered methods.

(a) Nemenyi test

	FedAvg	FLTrust	Φ-FL
FedAvg	1	< 0.001	< 0.001
FLTrust	< 0.001	1	< 0.001
Φ-FL	< 0.001	< 0.001	1

(b) Conover test with Holm correction

	FedAvg	FLTrust	Φ-FL
FedAvg	1	< 0.001	< 0.001
FLTrust	< 0.001	1	< 0.001
Φ-FL	< 0.001	< 0.001	1

can also be observed for Isolet and Winequality White. For Lung Cancer, the median remains similar to that of FLTrust, but the overall distribution suggests a slightly more favorable tendency for Φ-FL across runs. Small deteriorations relative to FLTrust can be observed for Segment and Ionosphere, while for Glass the decrease is less visually pronounced despite the lower mean value reported in Table I.

C. Sensitivity Analysis

To assess the robustness of the proposed Φ-FL method with respect to its main hyperparameters, a one-factor-at-a-time sensitivity analysis was conducted around the baseline configuration defined in the experimental setup. In each examined case, exactly one parameter was modified, while all remaining settings were kept identical to the baseline. For each configuration, 10 paired experimental runs were performed. In these 10 runs, the baseline configuration achieved an accuracy of 68.0%.

Table IV reports the changes in mean test accuracy with respect to the baseline configuration. Overall, the observed differences were small, usually remaining within approximately ± 1 percentage point. Only two of the twelve examined cases yielded statistically significant differences according to the paired Wilcoxon test at $\alpha = 0.05$. The best result was obtained for $\lambda = 0.75$, which improved the mean accuracy by approximately 0.94 percentage points relative to the baseline. In contrast, the weakest results, approximately 1.1-1.2 percentage points below the baseline, were observed for Proxy size 0.1, Threshold and of values 0.1 or 0.2. Among these, only the difference for Threshold $= 0.1$ was statistically significant.

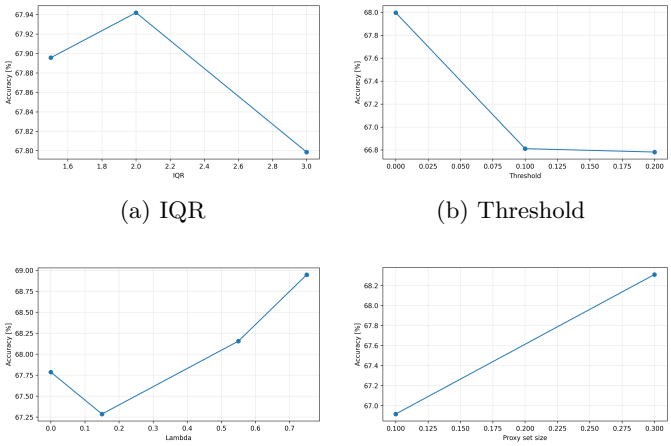
The relationship between individual hyperparameters and the resulting accuracy is illustrated in Fig. 3. For the IQR multiplier, the values 1.5, 2, and 3 were examined. The highest accuracy was obtained for IQR $= 2$, while the lowest was recorded for IQR $= 3$, with a decrease of about 0.15 percentage points. In the case of the λ parameter, the accuracy for $\lambda = 0$ was approximately 67.75%, while for $\lambda = 0.15$ it was lower by about 0.5 percentage points. Subsequently, an increasing trend was observed for $\lambda = 0.55$ and $\lambda = 0.75$, with the clearly best result achieved for $\lambda = 0.75$, reaching nearly 69%.

For the proxy set size, two alternative values, 0.1 and 0.3, were investigated. The larger proxy subset yielded an accuracy higher by approximately 1.2 percentage points than the smaller one, which is consistent with the intuition that a larger proxy set provides more reliable information for trust estimation. Finally, for the trust threshold parameter, the best performance was obtained for the baseline setting, approximately 68%, followed by a substantial decrease for Threshold $= 0.1$ and Threshold $= 0.2$. The results for these two larger thresholds were nearly identical, both amounting to approximately 66.8%.

The sensitivity analysis suggests that Φ-FL is fairly stable with respect to the tested hyperparameter variations. Although some configurations, especially larger values of λ, may provide modest improvements, the method does not appear overly sensitive to parameter tuning, which is a desirable property from the practical perspective.

TABLE IV: Comparison of configuration-wise accuracy differences with respect to the baseline setting. Reported p-values are obtained from the paired Wilcoxon test.

Configuration	Δacc. [%]	p_value	Significance
$\lambda = 0.75$	0.941	< 0.001	Yes
Proxy size $= 0.3$	0.305	0.808	No
$\lambda = 0.55$	0.150	0.059	No
Threshold $= 0$	-0.009	0.184	No
IQR $= 2$	-0.063	0.735	No
IQR $= 1.5$	-0.109	0.199	No
IQR $= 3$	-0.206	0.546	No
$\lambda = 0$	-0.219	0.081	No
$\lambda = 0.15$	-0.718	0.164	No
Proxy size $= 0.1$	-1.089	0.309	No
Threshold $= 0.1$	-1.194	< 0.001	Yes
Threshold $= 0.2$	-1.223	0.013	No



(a) IQR (b) Threshold

(c) Lambda (d) Proxy size

Fig. 3: Sensitivity analysis of selected hyperparameters for accuracy.

V. Conclusion and Future Work

This paper presented Φ-FL, a server-side aggregation strategy with a trust-aware, robust, and complementarity-sensitive fusion mechanism. The method combines fuzzy trust inference from proxy-set behavioral signals with IQR-based norm clipping and a 2-additive Choquet-inspired interaction model, enabling the server to down-weight unreliable or redundant client updates while rewarding complementary contributions. Across benchmark tabular datasets under Dirichlet-partitioned non-IID settings, Φ-FL yields a clear upward shift in pooled test accuracy over all paired runs: the global mean increases from $69.56\% \pm 17.20\%$ (FedAvg) to $74.74\% \pm 18.43\%$ (Φ-FL), and the median rises from 73.51% to 82.63% (with quartiles shifting from $Q_1/Q_3 = 55.75\%/81.58\%$ to $57.37\%/88.64\%$). The benchmark-level paired Wilcoxon analysis further supports that these gains are statistically significant and robust across repeated paired runs with identical splits per run, while the few neutral or slightly negative cases indicate that Φ-FL does not artificially inflate performance when the baseline is already competitive.

The comparison with FLTrust, included as a stronger trust-based baseline, further reinforces this conclusion. Φ-FL achieves higher mean accuracy than FLTrust on half of the considered datasets and also attains better global descriptive statistics, including a higher accuracy mean by 1.7 percentage point and median by 3.7 percentage point. Moreover, the Friedman test and the post-hoc Nemenyi and Conover analyses confirm statistically significant differences among all three compared methods, yielding a consistent ranking from the weakest to the strongest approach: FedAvg, FLTrust, and Φ-FL.

Finally, the sensitivity analysis indicates that the proposed method is relatively stable with respect to moderate changes in its main hyperparameters. For most tested con-

figurations, the accuracy changes remained small, typically within approximately ± 1 percentage point relative to the baseline setting, which suggests that Φ-FL does not rely on overly precise parameter tuning to achieve competitive performance.

The proxy-set assumption will be addressed and practical deployment will be improved in future work. First, proxy-free and label-free trust estimation based on server-side agreement, consistency, and self-supervised uncertainty signals will be studied. Second, comparisons to stronger FL aggregation baselines and a broader range of non-IID levels and client participation patterns will be expanded. Third, sensitivity to key hyperparameters (proxy size, trust threshold, interaction mixing) will be analyzed and dataset-agnostic tuning guidelines will be provided. Finally, the interaction stage will be optimized using sparse or top-K trusted client graphs, and adaptive interaction structures beyond the 2-additive setting will be investigated.

Acknowledgment

This study was conducted using publicly available benchmark datasets. The source code is available from the authors upon reasonable request.

References

[1] B. Pękała, J. Szkoła, K. Dyczkowski, and A. Wilbik, "Federated similarity-based learning with incomplete data," in *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*. The IEEE, 2023, pp. 1–6, 2023 IEEE International Conference on Fuzzy Systems (FUZZ), FUZZ-IEEE 2023; Conference date: 13-08-2023 through 17-08-2023. [Online]. Available: https://2023.fuzz-ieee.org/

[2] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.

[3] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. J. Cardoso, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, p. 119, 2020.

[4] A. Vaid, S. K. Jaladanki, J. Xu, S. Teng, A. Kumar, S. Lee, S. Somani, I. Paranjpe, J. K. D. Freitas, T. Wanyan, K. W. Johnson, M. Bicak, E. Klang, Y. J. Kwon, A. Costa, S. Zhao, R. Miotto, A. W. Charney, E. Böttinger, Z. A. Fayad, G. N. Nadkarni, F. Wang, and B. S. Glicksberg, "Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: Machine learning approach," *JMIR Medical Informatics*, vol. 9, no. 1, p. e24207, 2021.

[5] S. Pati, U. Baid, B. Edwards, S. Thakur, *et al.*, "The federated tumor segmentation (FeTS) tool: An open-source solution to further solid tumor research," *Physics in Medicine & Biology*, vol. 67, no. 20, p. 204002, 2022.

[6] Y. Tang and Y. Liang, "Credit card fraud detection based on federated graph learning," *Expert Systems with Applications*, vol. 256, p. 124979, 2024.

[7] V. Pruckovskaja, A. Weißenfeld, C. Heistracher, A. Graser, J. Kafka, P. Leputsch, D. Schall, and J. Kemnitz, "Federated learning for predictive maintenance and quality inspection in industrial applications," in *2023 Prognostics and Health Management Conference (PHM Paris 2023)*, 2023, pp. 1–6.

[8] B. Olanrewaju-George and B. Pranggono, "Federated learning-based intrusion detection system for the internet of things using unsupervised and supervised deep learning models," *Cyber Security and Applications*, vol. 3, p. 100068, 2025.

[9] B. Pękala, A. Wilbik, K. Dyczkowski, and J. Szkoła, "Unbalanced data supported by Federated Learning with uncertainty by different aggregation methods," in *Uncertainty and Imprecision in Decision Making and Decision Support - New Advances, Challenges, and Perspectives*, K. T. Atanassov, V. Atanassova, J. Kacprzyk, A. Kałuszko, M. Krawczak, J. Owsiński, S. N. Sotirov, E. Sotirova, E. Szmidt, and S. Zadrożny, Eds. Cham: Springer Nature Switzerland, 2026, pp. 65–83.

[10] K. Matsuda, Y. Sasaki, C. Xiao, and M. Onizuka, "Benchmark for personalized federated learning," *IEEE Open Journal of the Computer Society*, vol. 5, pp. 2–13, 2024.

[11] B. Pękala, P. Grzegorzewski, J. Szkoła, and D. Kosior, "Prism-fl-a privacy-aware decision support system for enhanced medical diagnostics based on federated learning," *IEEE Access*, vol. 13, pp. 200 338–200 350, 2025.

[12] A. Rachwał, A. Rachwał, and P. Karczmarek, "Smooth ordered weighted averaging in Federated Learning: A Newton-Cotes quadrature-inspired aggregation framework," *Advances in Science and Technology Research Journal*, vol. 20, no. 1, pp. 89–103, 2026. [Online]. Available: https://doi.org/10.12913/22998624/210058

[13] Y. Allouah, R. Guerraoui, N. Gupta, A. Jellouli, G. Rizk, and J. Stephan, "Adaptive gradient clipping for robust federated learning," 2025. [Online]. Available: https://arxiv.org/abs/2405.14432

[14] J. Yuan and Y. Chen, "Differential private federated learning with per-sample adaptive clipping and layer-wise gradient perturbation," *Computer Networks*, vol. 261, p. 111139, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128625001070

[15] M. Grabisch, "The application of fuzzy integrals in multicriteria decision making," *European Journal of Operational Research*, vol. 89, no. 3, pp. 445–456, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/037722179500176X

[16] G. Beliakov and J.-Z. Wu, "Learning fuzzy measures from data: Simplifications and optimisation strategies," *Information Sciences*, vol. 494, pp. 100–113, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519303573

[17] R. Bresson, J. Cohen, E. Hüllermeier, C. Labreuche, and M. Sebag, "Neural representation and learning of hierarchical 2-additive choquet integrals," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, ser. IJCAI'20, 2021.

[18] P. Karczmarek, M. Dolecki, P. Powroźnik, L. Galka, W. Pedrycz, and D. Czerwiński, "Quadrature-inspired generalized choquet integral," *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–7, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252223693

[19] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *CoRR*, vol. abs/1602.05629, 2016. [Online]. Available: http://arxiv.org/abs/1602.05629

[20] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[21] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[22] T. Kobayashi, "T-vmf similarity for regularizing intra-class feature distribution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6616–6625.

[23] C. Li *et al.*, "A feature optimization approach based on inter-class and intra-class distance for ship type classification," *Sensors*, vol. 20, no. 18, p. 5429, 2020.

[24] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "Fltrust: Byzantine-robust federated learning via trust bootstrapping," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2021.