IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier XX.XXXX/ACCESS.2025.xxxxxxx

Interpreting Federated Learning by Aggregating SHAP Explanations

VALERIO BONSIGNORI [1-2] LUCA CORBUCCI [1], FRANCESCA NARETTO [1], RICCARDO GUIDOTTI [1], and ANNA MONREALE [1].

[1]Dipartimento di Informatica, University of Pisa, Italy, Pisa. (e-mail: {valerio.bonsignori, luca.corbucci}@phd.unipi.it; {francesca.naretto, anna.monreale, riccardo.guidotti}@unipi.it)
[2]Scuola Normale Superiore, Italy, Pisa.

Corresponding author: Valerio Bonsignori (e-mail: valerio.bonsignori@phd.unipi.it).

This research was partially supported by: The European Commission under the NextGeneration EU programme – National Recovery and Resilience Plan (PNRR), under agreements: PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", and SoBigData.it – "Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021. The European Union Horizon 2020 program under grant agreement No. 101120763 (TANGO). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). The Italian Project Fondo Italiano per la Scienza FIS00001966 MIMOSA. Neither the European Union nor the granting authority can be held responsible for them.

ABSTRACT Federated Learning trains models without transferring data outside local clients, but usually relies on non-interpretable Neural Networks. While explanations are essential for model adoption and trustworthiness, conventional explainability techniques require centralized data access, which violates Federated Learning principles. We propose a framework for Interpreting Federated Learning by Aggregating SHAP explanations, `iFLASH`, which introduces novel aggregation methods that significantly improve explanation quality compared to naive averaging approaches, while preserving data privacy in federated settings. `iFLASH` enables local Shap explainers on individual clients without exposing raw data by aggregating feature importance values rather than models or gradients. Clients compute and evaluate explanations using performance-based metrics, then send results to the server. The server weighs each client's contribution based on model performance and explanation quality, which aims to produce faithful aggregate explanations. The framework supports various aggregation strategies, adapting to different levels of data imbalance and heterogeneity. Experiments across cross-silo (12-16 clients) and cross-device (50-150 clients) scenarios demonstrate that Faithfulness-based aggregation consistently outperforms uniform averaging in cross-silo settings, E_{Q1} achieves higher Faithfulness than naive averaging in all the cross-silo configurations across all datasets and distributions, while quality-aware methods perform comparably in cross-device environments where the high number of clients provides sufficient averaging effect. `iFLASH` explanations align closely with centralised explanations in feature importance ranking and directionality, sometimes achieving better fidelity. Results demonstrate that `iFLASH` enables accurate, privacy-preserving explanations for domains where data cannot be centralised. We highlight that our proposal has been extensively evaluated also in a cross-device federated setting, a scenario that is overlooked in the explainable AI literature.

INDEX TERMS Cross-device, Cross-silo, Explainable Artificial Intelligence, Faithfulness, Federated Learning, Trustworthy AI.

I. INTRODUCTION

As Machine Learning (ML) advances and provides increasingly accurate solutions for complex scenarios, many ML models still operate as black boxes, giving predictions without exposing their internal reasoning [53]. This lack of transparency undermines user trust and hinders the detection and mitigation of potential biases, raising both ethical and practical concerns [55]. In response, numerous Artificial Intelligence (AI) regulations have been introduced [52, 22, 21, 44], emphasising ethical values such as fairness, privacy, and interpretability. Previous works have investigated these aspects, including the analysis of privacy risks and fairness in machine learning and federated learning systems [48, 25, 14]. Concurrently, the rise of decentralised learning paradigms, such as Federated Learning (FL), has introduced new challenges for interpretability. In FL, data remain distributed across multiple clients and are never shared or cen-

tralised, making this approach appealing in sensitive domains such as healthcare and finance due to its privacy advantages. However, this decentralised nature of FL poses significant challenges to model interpretability, as most existing Explainable AI (XAI) techniques assume central access to either the black-box model or the training data. While the lack of data centralisation is fundamental for respecting user privacy, it limits the applicability of available XAI methods, making it challenging to train explainers on the server side without centrally available data. Despite the growing interest in explainability [41, 50, 28, 29], only a few recent works have begun to address the challenge of providing explanations in federated scenarios [30, 24, 15, 19, 23, 13, 31]. This setting is particularly challenging, as in FL data are stored exclusively at the client side and are never centrally available, and may also be distributed in a non-homogeneous manner across clients. With this work, we contribute to this emerging research area by proposing a framework for post-hoc explainability in federated settings.

For example, when a bank's fraud detection system needs to explain why it flagged a transaction, current FL approaches either compromise privacy by sharing sensitive data or produce unreliable explanations by simply averaging across institutions with different data quality [15]. To overcome these limitations, we propose `iFLASH`, a flexible methodology that enables the use of Shap [41], one of the most widely adopted techniques for feature importance (FI) explanations, to interpret ML models trained using FL, without violating FL requirements on data sharing. `iFLASH` combines the explanations generated by each participant based on local model quality and explanation quality.

While existing approaches treat all federated participants equally, ignoring that some may have poor data quality or model performance, `iFLASH` introduces a quality-aware aggregation framework for federated explanations. Unlike prior work [15], which simply averages explanations across clients, we design aggregation strategies that modulate each client's contribution based on the assessed quality of its explanations. In particular, each client locally evaluates its explanations, and these signals are used at the server to weight, down-weight, or discard less informative contributions.

To validate `iFLASH`, we conducted experiments using five widely used tabular datasets of varying size and characteristics, enabling evaluation in both cross-silo and cross-device FL scenarios [32, 12]. The results show that our explanations are on par with or outperform those produced by existing federated explainability approaches such as SHAP-FL and FedFCM, as well as centralised explainers.

This work extends our preliminary conference paper [15] by considering a larger number of clients, diverse data distributions, and multiple aggregation strategies. The rest of the paper is structured as follows: Section II reviews related work, Section III provides background knowledge, Section IV introduces our methodology, Section V describes the experimental settings, Section VI presents the results, and Section VII concludes the work, highlighting the contributions, open challenges, and future works.

A. MAIN CONTRIBUTIONS

Our work makes the following contributions:

- We propose `iFLASH`, a federated explainability framework that supports multiple quality-aware aggregation strategies for Shap-based explanations,
- We introduce Faithfulness-based aggregation methods that significantly improve explanation quality over naive averaging in cross-silo settings,
- We provide an extensive evaluation across both cross-silo (12, 16 clients) and cross-device (50, 70, 150 clients) scenarios, addressing a gap in the XAI literature for FL,
- We show that federated explanations can match or exceed centralized explanation quality via selective client weighting.

II. RELATED WORK

The growing interest in the XAI research field [26, 6, 40] led to the development of multiple techniques designed to explain models trained in a centralised scenario [41, 50, 28, 29]. More recently, these techniques have been adapted to FL-trained models, especially in domains such as Healthcare [19] and Finance [2]. A review of the current approaches used to explain models trained with FL is presented in [11, 42].

A. FEATURE IMPORTANCE-BASED APPROACHES IN FL

Initially, XAI techniques in FL were used to measure the contribution of the different clients to the final model [49, 61] and to understand clients' importance [60, 64, 66, 39]. Lately, these techniques have been expanded to more traditional XAI goals, such as explaining the reasons behind a prediction. Focusing on neural networks trained using FL, several approaches emerged. A first method has been proposed in [24], where the authors studied the interpretability issues in Horizontal FL [63]. They trained a regression model with FL to predict taxi trip duration within the Brunswick region. To explain the trained model, they derived FI exploiting Integrated Gradients [33]. Their study conducts experiments with a limited number of clients, simulating a cross-silo FL scenario. They did not examine the various methods for aggregating FI

Reference	Method	FLSetting	Cross Silo	Cross Device	Explanation Type	Flexible Aggregation
[24]	Integrated Gradients	Horizontal FL	✓	✗	Features Importance	✗
[15, 18, 8]	Shap	Horizontal FL	✓	✗	Features Importance	✗
[23]	DeepLift	Horizontal FL	✓	✗	Features Importance	✗
[13]	Credible Counterfactual	Vertical FL	✓	✗	Counterfactual	✗
[31]	Decision Trees	Horizontal FL	✓	✗	Rules	✗
iFLASH(Ours)	Shap	Horizontal FL	✓	✓	Features Importance	✓

TABLE 1. A comparison of iFLASH with other methods in the literature.

explanations, choosing the standard mean instead. In contrast, as presented in Section IV, our research extends this work by exploring different aggregation policies for FI, aiming at reaching different objectives, based on the chosen aggregation metric. In a first exploratory work [15], we introduced a Shap-based [41] approach for obtaining FI explanations in FL. In the current paper, we extend the methodology in which each client builds its explainer. When a new explanation is required, the clients generate local Shap values, and their FI are aggregated by the server using a basic mean operation. Another approach using Shap [41] has been introduced in [18]. This method uses Federated Fuzzy C-Means clustering [9] to create a reference set needed to build the Shap explainer. Unlike the approach that we propose in this paper, this method builds the explainer on the server side rather than on the clients participating in the federation. However, it has only been tested in limited cross-silo scenarios, and relying on a single server-side explainer reduces the flexibility of the explanations. Moreover, the sharing of the reference set with the server represents a risk to the client's privacy and breaks FL constraints. Recently, [7] introduced FastSHAP++, a method that adapts FastSHAP [34] to federated settings by training a neural explainer across clients while incorporating Differential Privacy to limit information leakage. In contrast to post-hoc feature attribution methods, this method relies on learning a surrogate model to approximate Shapley values during training. Due to this difference in formulation, we do not include this approach in our empirical evaluation, focusing on methods based on post-hoc Shap explanations. We compare our proposal to [18] in the cross-silo scenario and present the results in Section VI. Our current approach advances beyond these limitations by analysing different forms of aggregation for client computing the Shap values. This introduces a layer of flexibility that improves the quality of aggregated FI explanations compared to those computed on the server side.

The work in [8] presents a federated Shap-based approach for breast cancer classification, where local Shap values are computed by each client and aggregated at the server using simple averaging. However, their study is limited to a basic aggregation strategy with only four clients, and their evaluation focuses primarily on qualitative agreement with centralised explanations without exploring different aggregation policies.

Recently, [23] proposed another Shap [41] and DeepLift [56]-based approach to extract FI explanation from models trained with Horizontal FL. Their approach is heavily privacy-oriented and employs secure multi-party computation to aggregate the explanations computed by the different clients. However, while this privacy-enhancing technology provides stronger security guarantees, it significantly increases the communication overhead required for the aggregation, making its use impractical in scenarios with communication constraints. Additionally, like several previous approaches, the one proposed in [23] has only been evaluated in cross-silo FL settings with a limited number of clients.

B. OTHER EXPLAINABILITY APPROACHES IN FL

Several researchers have proposed alternatives to Shap [41] based approaches. In [37], the author envisioned a method for building rule sets through an iterative decision tree induction process, where features are removed at each step after selection in the previous step. This approach considered the design of incorporating user input, enabling hybrid decision-making. The authors of [65] introduced a novel method for automatically determining the appropriate logical connector (AND or OR) when aggregating rules extracted by clients at the server level. They developed concepts of *Diagonality* and *Exclusivity* for aggregating a local rule's set, which is enriched each round. At each round, the model's federated training is followed by the derivation of a local rule set by each client. The authors of [10] develop a system where fuzzy rules are generated locally on clients and then aggregated by the server, which sends them back to the clients. Their work

includes the development of a comprehensive interface and a streaming service to support this process. One recently introduced approach is GLOR-FLEX [31], which builds local Trepan Decision Trees [16] to capture the local behaviour of the black-box model. The rules from the decision trees, alongside a synthetic data partition, are sent to the server that merges the local rules into a single rule set that captures the behaviour of the global model. An approach specific to Vertical FL was proposed in [13]. In this case, the type of explanation provided by this approach is a federated counterfactual explanation aiming to explain a sample by altering it to change the output of the model. The authors of [54] introduced a Bottleneck Concept Learner to learn and represent concepts within the input space. Clients learn these concepts, which are then filtered by the server to account for potential malicious clients. They select concepts that effectively represent the input by employing an attention mechanism and a quantised loss. These concepts are subsequently used for prediction, making the model inherently interpretable. Our present method is based on Shap [41] and supports Horizontal FL. It differs fundamentally from these rule-based or counterfactual approaches, making direct comparisons unfeasible.

C. EXPLAINABILITY FOR ROBUSTNESS AND SECURITY IN FL

Explainable AI techniques have also been used to explain the misbehaviour of models trained using FL. The authors of [30] focused on the incorrect predictions of an FL model, as these predictions could be signs of an attack. To monitor changes in the model behaviour, the nodes involved during the computation explain the trained model at each FL round. The change of the FI between two consecutive rounds exceeding a predetermined threshold determines the attacker's presence. In a related direction, [57] proposed Zero-Trust Agentic Federated Learning (ZTA-FL) for securing Industrial IoT intrusion detection systems. Their framework employs Shap-weighted aggregation to detect Byzantine poisoning attacks under non-IID conditions, combining hardware-based agent authentication with privacy-preserving adversarial training. While this work demonstrates the utility of Shap values in a security-oriented aggregation context, its objective is attack detection rather than producing faithful, flexible FI explanations for end users.

To our knowledge, no previous work addressed the problem of interpretability in Horizontal FL by designing a FI aggregation approach that provides flexible aggregation strategies while maintaining data on the client side. In fact, unlike previous methods, our approach does not require clients to share their training data or reference datasets with the central server or other clients.

Table 1 summarises the comparison between `iFLASH` and existing federated explainability approaches. Unlike previous methods, limited to specific FL settings, that rely on fixed aggregation strategies, `iFLASH` is the only framework that supports and experiments in both cross-silo and cross-device scenarios while offering flexible aggregation techniques for improved explanation quality.

D. RESEARCH GAPS

While existing federated explainability work has made important contributions, several limitations remain: (i) most methods focus exclusively on cross-silo settings with limited client counts, leaving cross-device scenarios unexplored; (ii) aggregation strategies typically employ simple averaging without considering explanation quality metrics; (iii) privacy-preserving approaches either require additional federated processes (e.g., fuzzy C-means in [9]) or incur significant communication overhead (e.g., secure multi-party computation in [23]); (iv) evaluation is often limited to qualitative comparison with centralized baselines without rigorous statistical testing. Our work addresses these gaps by introducing flexible aggregation strategies evaluated extensively across both cross-silo and cross-device settings with statistical validation.

III. BACKGROUND

In this section, the key methods and approaches that are relevant to our methodology are presented. We begin with an overview of FL in Section III-A, followed by a discussion on Explainable AI in Section III-B. In particular, we focus on state-of-the-art methods, with a detailed examination of Shap, which we employ in our experiments. The consistent notation is explicitly listed in Table 2.

A. FEDERATED LEARNING

FL [47] enables the training of a ML model by coordinating a set of C distributed clients, each client c has a private local dataset D_c. Iteratively, at each round r, the central server S orchestrates the training process by selecting a subset of clients $C_r \subseteq C$. The clients receive a shared global model θ_r from the server, train it locally on their data for a fixed number of epochs e, and then send their updated models back to the server. The server aggregates the received updates to produce a new global model θ_{r+1}, which is used in the next round. This process is repeated for a fixed number of rounds or until convergence. The aggregation strategy employed by the server is directly responsible for model convergence and performance. The main algorithms for the aggregation of the local contributions of updates' models are Federated-SGD [47] (Fed-SGD) and Federated-Average [47] (Fed-AVG). Fed-SGD

Symbol	Meaning	Symbol	Meaning
C	FL Clients	R	FL Rounds
D_c	Dataset of client $c \in C$	n_c	Size of the dataset D_c
e	Local FL epochs	g_c	Gradients of client c
S	Server	C_r	Subset of clients for FL Round r
θ	Neural model	c_i	i-th client
ϕ^j_*	Shap values of j-th sample	ϕ^j_c	Shap value of j-th sample from client c
b	black box	f_c	Shap Explainer of of client $c \in C$
N	Number of labels	$x, \hat{y}$	a sample and its prediction

TABLE 2. Table of notations used in the following section.

is an aggregation algorithm that requires each client to perform a single training step on their training dataset before sharing the computed gradient g_c with the server. The server receives $|C_r|$ gradients, and aggregates them updating the global model $\theta_{r+1} = \theta_r - \eta \sum_{c=1}^{|C_r|} \frac{n_c}{n} g_c$ where n_c is the size of the train set of client c, $n = \sum_{c=1}^{|C_r|} n_c$ and η the learning rate. The constraint of sharing the gradient after each step is the main limitation of this algorithm because of the slowing down of training time due to numerous communications.

On the contrary, when using Fed-AVG, the selected clients can train the model received by the server for multiple epochs e before sharing it with the server. The server collects $|C_r|$ models and aggregate them in $\theta_{r+1} = \sum_{c=1}^{|C_r|} \frac{n_c}{n} \theta_c$. In this paper, we used Fed-AVG because of the improved performance with respect to Fed-SGD. Depending on the number of clients involved in the training and the availability of those, we can distinguish between two types of FL: cross-silo and cross-device ([38] presents an analysis of these two settings from several viewpoints). In the cross-silo scenario, a few clients (10-50) are always available during the training. On the contrary, in the cross-device scenario, a higher order of devices is involved in the computation; however, they are available only under certain conditions, e.g., the device must be connected to power and have a stable internet connection. We conducted extensive experimentation of our proposal in both the cross-silo and the cross-device FL scenarios.

B. EXPLAINABLE AI

The black-box nature of the deep learning models prevents the ability to receive an explanation of the reasoning behind a prediction. In recent years, the XAI community has developed several techniques to mitigate this issue and to open up these black-box models [6, 27].

Explanation methods can be categorised into three aspects [6]. The first aspect contrasts *model-specific* vs *model-agnostic* approaches, depending on whether the explanation method exploits knowledge about the internals of the black box. The second aspect contrasts *local* vs *global* approaches, based on whether the explanation is provided for a specific instance or the overall logic of the black-box model. Additionally, we can distinguish between *post-hoc* vs *ante-hoc* methods, depending on whether they are designed to explain a pre-trained model or are inherently interpretable by design. Among the various XAI approaches, we focus on feature importance explanations due to their interpretability and compatibility with federated constraints. Feature importance (FI) is a popular type of explanation returned by local methods [6]. In this case, the explainer assigns an importance value to each feature, representing the impact of that particular feature on the final prediction. Given a record x, an explainer $f(\cdot)$ produces a FI explanation as a vector $\phi = \{\phi_1, \phi_2, \ldots, \phi_f\}$, in which the value $\phi_i \in \phi$ is the importance of the i^{th} feature for the decision made by the black box model $b(x)$. To understand the contribution of each feature, the sign and the magnitude of each value ϕ_i are considered. Considering the sign, if $\phi_i < 0$, the feature contributes negatively to the outcome $\hat{y}$; otherwise, if $\phi_i > 0$, the feature contributes positively. The magnitude instead represents how great the feature contribution is to the final prediction $\hat{y}$. In particular, the greater the value of $|\phi_i|$, the greater its contribution. Hence, when $\phi_i = 0$, it means that the i^{th} feature is showing no contribution. An example of a feature based explanation is $\phi = \{(age, 0.8), (income, 0.0), (education, -0.2)\}, \hat{y} = deny$. Here, *age* is the most important feature for the decision *deny*, *income* does not affect the outcome, and *education* has a negative contribution.

In this paper, we present a method based on FI, and for our experiments, we rely on Shap, one of the most widely used FI explanation methods [41]. Shap, computes features importance using Shap-

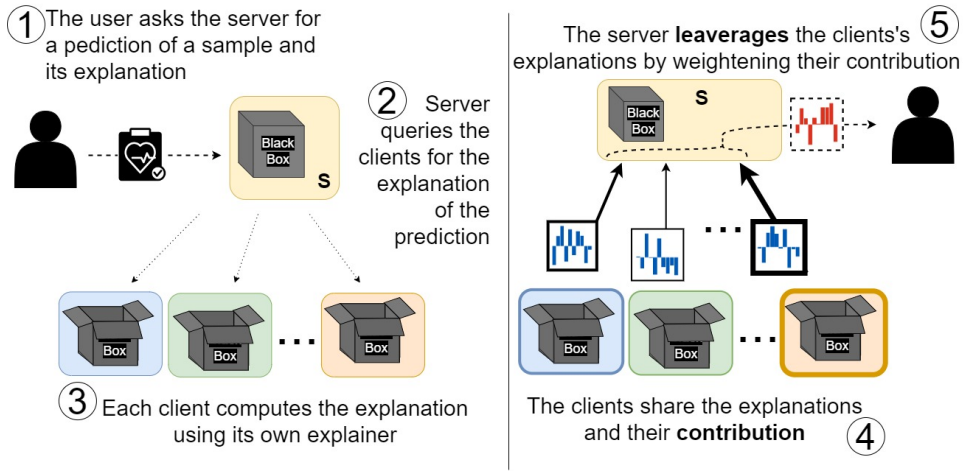

FIGURE 1. An overview of our Explanation methodology. An external user who needs a prediction and an explanation interacts with the server. The server computes the prediction using the black box model trained using FL. The server then queries the clients that participated in the FL process: each client uses their local explainer to compute the Shap values for the sample. Finally, the clients send the Shap values to the Server that aggregates them in a collaborative explanation provided to the user.

ley values[1], a concept from cooperative game theory. The explanations are *additive feature attributions* and respect the following definition: $f(z') = e_0 + \sum_{i=1}^{F} e_i z'_i$, where f is the explanation model, $z' \sim x$, the original record to explain in the range $z' \in \{0,1\}^F$ with F the number of simplified input features and $\phi_i \in \mathbb{R}$ the effects assigned to each feature. Shap has three properties: *(i) local accuracy*, meaning that $f(x)$ matches the black-box model b in the vicinity of the record to explain; *(ii) missingness*, which allows for features $x_i = 0$ to have no attributed impact on the Shap values; *(iii) consistency*, meaning that if a model changes so that the marginal contribution of a feature value increases (or stays the same), the Shap value also increases (or stays the same). There are several variants to compute the FI with Shap [41, 34]. The main difference among them is that they approximate the Shap values. In particular, some methods exploit the structure of the learning model to provide either an exact computation of the Shap values or an ad-hoc approximation to speed up the execution. However, throughout this work, we refer to KernelShap, the model-agnostic version presented in the original paper [41]. However, this work focuses on KernelShap, the model-agnostic version presented in the original paper. This choice ensures consistent approximation across models regardless of their architecture, maintaining methodological generality.

IV. PROPOSED METHODOLOGY

In this paper, we present `iFLASH`, a federated explainability framework based on feature importance (FI). In our setting, we rely on Shap [41] to generate local explanations at each client. Our design choice is driven by the fact that Shap requires access to data distributions, which are only available locally in FL. While Shap is typically applied in centralized settings where the full dataset is available, this assumption does not hold in FL, where data are distributed across clients. As a result, explanations must be computed locally, making it challenging to obtain consistent and reliable explanations at the global level. In addition, due to data heterogeneity, each client may provide explanations of varying quality, depending on how well its local data distribution represents the instance to be explained.

To address these challenges, `iFLASH` introduces aggregation strategies that combine local explanations without requiring data sharing, while accounting for their quality. In this way, clients that provide more reliable explanations contribute more to the final result, while less informative contributions are down-weighted.

Shap computes FI using Shapley values, a concept from cooperative game theory [41], described in Section Section III. In this work, we adopt KernelShap, the model-agnostic variant of Shap, as it ensures a consistent approximation of feature attributions across heterogeneous models and clients.

We consider a FL scenario with $C = c_1, \ldots, c_n$ clients, where each client $c \in C$ has its own local and private dataset $D_c \subset \mathbb{R}^m$, with m denoting the number of features. All clients collaborate to train an ML model $b : \mathbb{R}^m \to \mathbb{Y}$, where $\mathbb{Y}$ is the label space. It is important to note that, depending on the specific FL setting, external clients who do not participate in the training phase may still query the final trained model. Once the black-box model b has been trained, `iFLASH` is used to generate FI explanations in a federated manner.

Each client $c \in C$ builds a *local explainer* f_c using its private dataset D_c^{train}, which remains local and is not shared. Once the explainers are ini-

[1]We refer the interested reader to: https://christophm.github.io/interpretable-ml-book/Shapley.html

tialised, the *explanation generation* step produces an explanation for a given input instance x. An overview is provided in Figure Figure 1:

1) A user queries the server with an instance x,
2) The server distributes x to all clients,
3) Each client computes its local explanation $\phi_c = f_c(x)$,
4) Each explanation is locally evaluated using quality metrics[2], and the pair $(\phi_c, \text{quality}_c)$ is sent to the server,
5) The server performs the aggregation with `iFLASH` and returns the final explanation.

The evaluation step is central to our method, as data distributions vary across clients. As a result, depending on the input, some local explainers may provide more accurate explanations than others.

In the remainder of this section, we detail `iFLASH`. We first describe the explanation generation step, followed by the aggregation strategies on the server side, which exploit the evaluation metrics provided by the clients.

A. EXPLANATION GENERATION

`iFLASH` adopts Shap to define a local explainer $f_c(\cdot)$ at each client, built using the local dataset D_c^{train}. This design is required since no federated version of the explainer exists, and reference data cannot be shared across clients due to FL constraints.

Given an input instance x, the server first computes the prediction $\hat{y} = b(x)$ and then queries all clients for explanations. Each client produces a local explanation $\phi_c^x = f_c(x)$, which is sent to the server. The server collects the set of local explanations $\phi_{c\,c\in C}^x$ and aggregates them according to the strategies defined in `iFLASH` to obtain a single explanation ϕ^x for x.

Instantiation of Shap explainers' client

KernelShap requires a reference background dataset to compute feature importance. The quality and representativeness of this background data directly impact explanation accuracy. However, using the entire training dataset is computationally prohibitive and may not capture the most informative data distribution. Building upon our preliminary work [15] that used the limited dataset means as background, we propose an *advanced nested clustering strategy* that yields a compact yet representative background dataset. In real-world federated deployments, practitioners typically employ clustering techniques rather than using entire datasets as background data, both for computational efficiency and to capture meaningful data structure. We employ a two-stage nested clustering strategy that generates compact, representative background datasets locally on each client.

The procedure is executed locally by every client and consists of the following steps:

1) The client performs a first clustering procedure using k-means [43], incrementally adjusting k until they identifies the optimal elbow point. The optimal k is determined by computing silhouette scores for each candidate k, calculating the first-order differences between consecutive scores, and selecting the k that yields the maximum difference.
2) For each cluster, the client performs a secondary k-means clustering operation using a single-feature approach based on the centroid distance. This second clustering step is necessary to have representative samples for each of the clusters found in the previous step. The client chooses a random sample within each sub-cluster.

This nested approach balances two objectives: the first clustering captures medium-grain diversity across the local dataset, while the second clustering ensures fine-grain representation within each cluster. This procedure generates compact yet representative background datasets suitable for reliable SHAP computation while maintaining computational efficiency.

B. AGGREGATIONS OF THE EXPLANATIONS

The core contribution of `iFLASH` is the design of aggregation strategies that improve the quality of federated explanations compared to existing approaches. These strategies aim to produce explanations that are consistent with those obtained in centralized settings, while preserving Faithfulness to the black-box model.

Regarding federated explanations, the simplest approach consists of averaging the FI explanations across clients [15]. However, this approach ignores the heterogeneity of clients in FL. Since data distributions may differ, some clients may be better suited to explain a given instance, depending on how well it is represented in their local data.

To account for this heterogeneity, we design aggregation strategies that weight the contribution of each client based on evaluation metrics, allowing the server to prioritise more reliable explanations.

In practice, each client builds a local explainer using the available training data to generate FI explanations. Each client then sends its explanation, together with the corresponding Faithfulness and F1 score, to the server. The server aggregates the received explanations according to the strategies defined in the following and returns a final explanation to the practitioner.

[2]Examples include predictive performance or explanation Faithfulness.

1) Weights for client's contributions

To implement the proposed aggregation strategies, we assume that each client provides its local Shap explanation together with a set of evaluation metrics. These metrics are used by the server to weight or filter the contribution of each client in the aggregation process.

In particular, we consider two types of information: (i) the predictive performance of the local model, and (ii) the quality of the local explanation. We adopt the Weighted F1 and Macro F1 scores as measures of model performance, and Faithfulness as a measure of explanation quality. In the following, we describe the metrics.

Generalisation capabilities: To measure the generalisation performance of the black-box model on each client, we adopt the Weighted F1 score, which for N classes is defined as:

$$\text{Weighted F1} = \sum_{i=1}^{N} w_i \cdot \text{F1 Score}_i \qquad (1)$$

where w_i is the proportion of samples of the i-th class, and F1 Score_i is the F1 score of the i-th class.

This metric accounts for class imbalance and is suitable for heterogeneous federated settings, where data distributions may differ across clients.

In case we want to capture subtle variations in the performance of the ML model across all classes, regardless of their frequency, we propose to use the Macro F1 score, which is defined as the unweighted mean of the F1 scores across all N classes:

$$\text{Macro F1} = \frac{1}{N} \sum_{i=1}^{N} \text{F1 Score}_i \qquad (2)$$

We highlight that, unlike the weighted F1, which is heavily influenced by the data distribution across clients, the Macro F1 gives equal importance to all classes, providing a more holistic view of the model's behaviour across diverse data distributions. It is important to note that both F1 metrics produce a single value per client model, reflecting the overall performance of each client on its own local data.

Faithfulness as a quality measure: To evaluate the quality of local FI explanations, we adopt the Faithfulness metric [1], which measures how well the importance assigned to each feature reflects the impact on the model's prediction. Algorithm 1 has the pseudocode to compute Faithfulness.

In practice, the metric performs multiple queries to the black-box model. Given an input x, the model f is evaluated while progressively masking input features according to the ranking provided by the explanation ϕ. At each step, the feature selected by the ranking is replaced with a baseline value, and the output probability for the predicted class $\hat{y} = f(x)$ is stored in $\mathbf{p}$. The metric then computes the correlation between the importance values in ϕ and the vector of probabilities $\mathbf{p}$ that records the changes in the model's confidence.

Algorithm 1 The pseudo algorithm to compute the Faithfulness.

Require: x sample to be explained, ϕ feature-importance explanation, f black box, $\hat{y}$ the predicted class for x by f.

1: $\pi \leftarrow \text{argsort}(-|\phi|)$ ▷ sort by decreasing magnitude
2: $\mathbf{p} \leftarrow zeros$ ▷ Init. vector of prob. (same size as π)
3: **for** $i \in \pi$ **do**
4: $\quad \mathbf{x}' \leftarrow \mathbf{x}$
5: $\quad \mathbf{x}'_i \leftarrow \mathbf{b}_i$ ▷ replace feature i^{th} with base value
6: $\quad p_i \leftarrow f(\mathbf{x}')_{\hat{y}}$ ▷ prob. of initial predicted class
7: **end for**
8: **return** $-\text{corr}(\phi, \mathbf{p})$ ▷ Pearson Correlation

In `iFLASH`, Faithfulness is used to ensure that explanations reflect the model's decision process. Compared to performance-based metrics such as the F1 score, Faithfulness provides a sample-level measure that enables more fine-grained weighting of client contributions, helping to mitigate biases introduced by global performance indicators. In our setting, we use the feature-wise mean computed over the client's local dataset as the masking baseline $\mathbf{b}$ (line 5 of Algorithm 1). This choice avoids out-of-distribution artifacts, respects federated privacy constraints, and is theoretically aligned with distributional baseline approaches [58]. An ablation study comparing alternative baseline choices is reported in Appendix III.

2) Aggregation Strategies

In the following, we introduce a set of aggregation strategies within `iFLASH`, designed to account for different aspects of client contributions. These strategies rely on different weighting schemes, based on model performance, explanation quality, or a combination of both, and are later compared in our experimental section.

Let us assume we are computing the explanation for a sample j; let ϕ_c^j denote the explanation produced by client c for the j-th sample, and let $|C|$ be the number of clients.

Average ($\mathbf{E}_{Avg}$): The Shap values produced by the clients are averaged across all participants. This strategy treats all clients equally and provides a uniform aggregation.

$$E_{Avg}^j = \frac{1}{|C|} \sum_{c \in C} \phi_c^j \qquad (3)$$

Weighted Average ($\mathbf{E}_{F1}$): The Shap values produced by the clients are weighted and averaged using the local F1 score. This strategy assigns a higher weight to clients with better predictive performance, allowing their contributions to have a greater impact on the aggregated explanation.

$$E_{F1}^j = \sum_{c \in C} \left(\frac{F1_c}{\sum_{c \in C} F1_c} \right) \phi_c^j \qquad (4)$$

where $\mathrm{F1}_c$ is the F1 weighted score of the client c.

Weighted Average ($\mathrm{E}_{F1_{macro}}$): The Shap values produced by the clients are weighted and averaged using the local macro F1 score. Unlike the previous strategy, this approach considers the average performance across all classes. It is particularly suitable in the presence of class imbalance, as clients with more balanced class distributions tend to achieve higher macro F1 scores and therefore receive greater weight in the aggregation.

$$E^j_{F1_{macro}} = \sum_{c \in C} \left(\frac{F1_{Mc}}{\sum_{c \in C} F1_{Mc}} \right) \cdot \phi^j_c \tag{5}$$

where $F1_{Mc}$ is the macro F1 score of the client c.

Faithfulness Weighted Average E_{faith}: The explanations produced by the clients are weighted and averaged using Faithfulness [1].

$$E^j_{faith} = \sum_{c \in C} \left(\frac{faith^j_c}{\sum_{c \in C} faith^j_c} \right) \cdot \phi^j_c \tag{6}$$

where $faith^j_c$ is the Faithfulness of the explanation of the client c for the sample j. Since Faithfulness measures the alignment between FI explanations and the model's behaviour, weighting clients accordingly improves the aggregated explanation[3]

Threshold-based Average E^0_{Avg}: This method filters explanations based on Faithfulness before aggregation, discarding those with values less than or equal to zero[4], and averaging the remaining ones:

$$E^0_{Avg} = \sum_{c \in C} \phi^j_c \cdot \mathbb{I}(faith^j_c > 0) \tag{7}$$

where $\mathbb{I}(\cdot)$ is the indicator function.[5]

Mixed Threshold E_{Mix}: This method combines thresholding and weighted averaging. It first discards explanations with Faithfulness values less than or equal to zero. If any explanations remain, they are weighted by their Faithfulness scores. Otherwise, if no explanation meets this criterion, macro F1 score weighting is applied. The combined explanation ${E_{Mix}}^j$ is computed as:

$${E_{Mix}}^j = \begin{cases} E^j_{faith^+} & \text{if } \exists c.\, faith^j_c > 0 \\ E^j_{F1_{macro}} & \text{otherwise} \end{cases} \tag{8}$$

[3]We apply the normalization $\mathrm{faith}^j_c \leftarrow \frac{\mathrm{faith}^j_c + 1}{2}$ before computing weights, ensuring non-negative contributions while preserving relative quality differences across clients. Note that this mapping sends $0 \mapsto 0.5$: an explanation with neutral Faithfulness still receives a positive weight rather than being discarded. This is intentional: E_{faith} is a soft aggregation that down-weights poor explanations without fully excluding them, in contrast to the hard-threshold methods E^0_{Avg} and E_{Mix}.

[4]This method applies threshold filtering *before* averaging. It excludes all explanations with normalized Faithfulness ≤ 0.5 (corresponding to raw Faithfulness ≤ 0), then averages the remaining explanations using uniform weights.

[5]Note that for E^0_{Avg} and E_{Mix}, no normalization is applied to Faithfulness prior to filtering: the threshold $faith^j_c > 0$ is evaluated on the raw $[-1, 1]$ scale, so only explanations with a genuinely positive alignment with the model are retained.

where $E^j_{faith^+}$ denotes the Faithfulness weighted average E_{faith} computed only with the explanation with positive Faithfulness values, i.e., $faith^j_c > 0$. By combining both strategies, this method prioritises high-quality explanations when available, and falls back on model performance otherwise.

Mixed Threshold-F1 $\mathrm{E}_{Mix_{F1}}$: Similar to E_{Mix}, this method first discards explanations with Faithfulness less than or equal to zero, and then weights the remaining ones using the corresponding clients' macro F1 scores:

$$E^j_{Mix_{F1}} = \sum_{c \in C} \left(\frac{F1_{Mc} \cdot \mathbb{I}(faith^j_c > 0)}{\sum_{c \in C} F1_{Mc} \cdot \mathbb{I}(faith^j_c > 0)} \right) \cdot \phi^j_c \tag{9}$$

Top Quartile Average E_{Q1}: This aggregation strategy averages the explanations from the top 25% most faithful clients:

$$E^j_{Q1} = \sum_{c \in C} \left(\frac{\mathbb{I}(faith^j_c \geq Q1(faith^j_C))}{\sum_{c \in C} \mathbb{I}(faith^j_c \geq Q1(faith^j_C))} \right) \cdot \phi^j_c \tag{10}$$

where $Q1(faith^j_C)$ is the first quartile of the Faithfulness values across all the explanations' clients C for the j-th sample.[6]

A summary of the aggregation strategies is reported in Table 5, in Section IV-B2.

C. PRIVACY AND COMMUNICATION CONSIDERATIONS

The objective of `iFLASH` is to generate explanations that reflect the contributions of all clients involved in the training of the federated model. Our design is motivated by the observation that, similarly to federated learning, where model training relies on locally available data, explanation methods, such as Shap, also require access to data distributions to produce reliable feature attributions.

From the literature, it is well known that the output of Shap is directly influenced by the data used during the explanation process. For this reason, in our setting, we design a procedure that ensures the representation of different regions of the data space across clients, described in the Section IV-A.

Given such dependence on local data, each client in our framework generates explanations locally, and the aggregation step combines these contributions to improve the quality of the final explanation with respect to the global FL model.

In this context, from a privacy perspective, `iFLASH` follows the standard FL paradigm: no raw data are shared across clients. Instead, only explanation vectors and associated quality metrics are transmitted to the server. While this approach avoids direct data sharing, we acknowledge

[6]Similarly, the quartile threshold $Q1(faith^j_C)$ in E_{Q1} is computed on raw Faithfulness values, before any normalization, ensuring that the ranking of clients reflects their true relative Faithfulness.

that explanation vectors may still encode information about local data distributions. Therefore, our method does not provide additional formal privacy guarantees, but rather preserves the same privacy assumptions as standard federated learning.

Regarding communication cost, our approach exchanges explanation vectors instead of model updates. Since explanations are generated on demand for each input instance and no global explainer is available at the server, this process is repeated for each record. However, for each instance, only a single communication step is required from each client to the server. As a result, the communication cost depends on both the number of features and the number of explained instances, rather than on the number of model parameters. This can be beneficial for complex models, but it may become more demanding for high-dimensional data or large evaluation sets. In practice, the communication cost remains manageable and can be reduced using techniques such as feature selection.

V. EXPERIMENTAL SETUP

This section details the experimental settings. Section V-A describes the configurations used in the experiments, including the datasets, the number of clients, and the data distribution across clients. Section V-B reports the training configuration performed. Section V-C describes the evaluation metrics. Section V-D introduces the baseline methods used for comparison.

A. DATASETS AND PARTITIONING STRATEGIES

In our study, we conduct extensive experiments using five different datasets:

- Adult [3]: a tabular dataset with $48,842$ samples containing demographic and employment related information. The objective is to predict whether an individual's income exceeds 50K per year.
- Diamond [62]: a tabular dataset with $53,134$ records describing the prices and characteristics of diamonds. The objective is to predict the diamond cut quality with five possible classes.
- Dry bean [35]: a tabular dataset composed of $13,611$ samples, aimed at classifying dry beans into seven possible types.
- Dutch [59]: A tabular dataset containing information about $60,420$ individuals. The objective is to determine whether an individual's salary exceeds 50K per year.
- ACS Income (ACSI) [17]: a tabular dataset derived from the American Community Survey (ACS), naturally partitioned into 51 subsets corresponding to U.S. states and Puerto Rico. The task is to predict whether an individual's salary exceeds 50K per year.

To allow a diverse experimental analysis, we consider both *cross-silo* and *cross-device* FL settings. Adult, Diamond, and Dry bean are used for cross-silo experiments, while Dutch and ACSI are employed in the cross-device setting due to their larger data availability. Furthermore, to simulate realistic FL scenarios, we apply both IID and non-IID partitioning strategies to selected datasets, while keeping the natural partitioning of the ACSI dataset. The data are distributed across different numbers of clients, and we evaluate the performance of the FL algorithms under these settings.

An overview of the datasets is provided in Table 3. We employ the following partitionings:

- **IID Partitioning.** To preserve the original data distribution, we divide the datasets into 12 and 16 partitions for the cross-silo setting using the Adult, Diamond, and Dry bean datasets. For the cross-device setting, we split the Dutch dataset across 50, 70, and 150 clients.
- **Non-IID Partitioning.** To generate skewed data partitions, we use a Latent Dirichlet Allocation (LDA) sampling scheme. We consider two levels of heterogeneity: *high skew* ($\alpha = 0.99$), which produces highly imbalanced partitions, and *moderate skew* ($\alpha = 5.0$), which results in less pronounced imbalance while maintaining a similar number of samples across clients.
- **Natural Partitioning.** For the ACSI dataset, we use its natural partitioning into 51 states to represent real-world data distribution across geographical regions.

Prior to training, categorical features were encoded using target encoding for high-cardinality features and one-hot encoding for low-cardinality ones. Continuous features were normalized to a common scale. These preprocessing steps were applied consistently within each dataset.

B. TRAINING OF THE FL MODELS

To train the models to be explained, we simulate a FedAvg FL training process using *Flower* [4], a widely adopted FL framework [51].

We perform a systematic hyperparameter search to optimise the models in each setting. Hyperparameters are selected to maximise the centralized weighted F1 score on a validation set. For cross-silo settings, validation is performed on local client data, while for cross-device settings, a subset of clients is used for validation.

The set of hyperparameters that maximises the weighted F1 score across federated clients is selected and used to train the black-box model on the full development set (training and validation).

The samples used for explanation are not seen during training. In the cross-silo setting, they are taken from the union of the clients' test sets, while

Cross-Silo						
Dataset	**Samples**	**Clients**	**Distributions**	**# Features**	**Task**	**Classes**
Adult	48,842	12, 16	$\alpha = 0.99$, $\alpha = 5.0$, IID	14	Income prediction	Binary
Diamond	53,134	12, 16	$\alpha = 0.99$, $\alpha = 5.0$, IID	10	Cut quality classification	5 classes
Dry bean	13,611	12, 16	$\alpha = 0.99$, $\alpha = 5.0$, IID	16	Bean type classification	7 classes

Cross-Device						
Dataset	**Samples**	**Clients**	**Distributions**	**# Features**	**Task**	**Classes**
Dutch	60,420	50, 70, 150	$\alpha = 0.99$, $\alpha = 5.0$, IID	14	Income prediction	Binary
ACSI	1,664,500	51	*Natural*	11	Income prediction	Binary

TABLE 3. Overview of datasets used in our experiments, including their respective sample sizes, number of clients, and data distributions. For each dataset, experiments were conducted across all listed client configurations and distributions, ensuring comprehensive evaluation of the impact of different data splits (IID and non-IID with α values of 0.99 and 5.0) on the performance of the system.

Cross-Silo				
	#	**IID**	**α = 5.0**	**α = 0.99**
Dry bean	12	0.94 ± 0.01	0.95 ± 0.01	0.88 ± 0.04
	16	0.94 ± 0.01	0.93 ± 0.01	0.88 ± 0.06
Adult	12	0.78 ± 0.01	0.78 ± 0.02	0.66 ± 0.10
	16	0.78 ± 0.01	0.77 ± 0.02	0.69 ± 0.08
Diamond	12	0.76 ± 0.01	0.72 ± 0.02	0.70 ± 0.03
	16	0.77 ± 0.01	0.73 ± 0.01	0.61 ± 0.08
Cross-Device				
	#	**IID**	**α = 5.0**	**α = 0.99**
Dutch	50	0.83 ± 0.03	0.47 ± 0.01	0.53 ± 0.16
	70	0.83 ± 0.06	0.45 ± 0.01	0.57 ± 0.18
	150	0.83 ± 0.03	0.55 ± 0.17	0.64 ± 0.20
ACSI	51	–	0.75 ± 0.08	–

TABLE 4. Mean Macro-F1 scores and standard deviations for different datasets and distributions across Cross-Silo and Cross-Device settings. The distributions include IID and two levels of non-IID (5.0 and 0.99). As for ACSI, we use the natural partitioning of the dataset into 51 states.

in the cross-device setting, they are sampled from the datasets of the training clients.

Hyperparameter search is performed using Bayesian optimisation. We use the WandB MLOps platform [5] to conduct the search. Through systematic Bayesian hyperparameter optimization (detailed in Appendix I-B), we identify optimal architectures for each dataset and distribution. Architectures typically consist of 2-3 hidden layers with 64-256 units, ReLU activations, and dropout regularization (0.1-0.5). We use the Adam optimizer with learning rates in $[10^{-4}, 10^{-2}]$ and batch sizes of 32-128. Additional details about the hyperparameter tuning process can be found in Appendix I.

Table 4 reports the performance evaluations of the different datasets and settings considered, with good prediction performance overall, even in highly imbalanced scenarios.

C. EVALUATION METRICS

We adopt the following metrics to evaluate the quality of explanations at both the client level and after aggregation.

First, we use Faithfulness [1] to evaluate explanation quality. This metric plays a dual role in our framework: during aggregation, to assess the quality of local explanations and to weight client contributions Algorithm 1, and after aggregation, to evaluate the quality of the final federated explanations.

In addition, extending the preliminary validation in our previous work [15], we incorporate proximity-based metrics from [46], referred to as *ShapGap*. These metrics are used to compare Shapley values across different sources. In particular, *ShapGap* computes the mean of a chosen distance function applied to the Shapley values of two models for the same instance:

$$ShapGap\ (D, d) = \frac{1}{|D|} \sum_{j=1}^{|D|} d(\phi_A^j, \phi_B^j)$$

where D is the explanation set, d is the chosen distance function, ϕ^j is the Shapley value of the j-th sample, A and B are, in our case, the aggregation strategies to compare. Following [46], we adopt two

distance functions: the L2 norm and cosine similarity. Therefore, the *ShapGap* measure is rewritten as:

$$ShapGap_{L2} = \frac{1}{|D|}\sum_{j=1}^{|D|} ||\phi_A^j - \phi_B^j||_2$$

and

$$ShapGap_{Cos} = \frac{1}{|D|}\sum_{j=1}^{|D|} 1 - \frac{\phi_A^j \cdot \phi_B^j}{||\phi_A^j||_2||\phi_B^j||_2}$$

$ShapGap_{L2}$ quantifies the magnitude of the difference between the Shapley values; it is zero when the values are identical and increases with greater divergence. In contrast, $ShapGap_{Cos}$ measures the coherence of the direction between the Shapley values, bounded between -1 and 1.

Name	Aggregation
E_{Avg}	Average
E_{F1}	Weighted mean on weighted F1 score
$E_{F1_{macro}}$	Weighted mean on macro F1 score
E_{faith}	Weighted mean on Faithfulness
E^0_{Avg}	Threshold on Faithfulness
E_{Mix}	Threshold and Weighted mean
$E_{Mix_{F1}}$	Threshold on Faith. and Weighted mean
E_{Q1}	Threshold on the Faith. first percentile

TABLE 5. Aggregation strategies and Their Notations

We use *ShapGap* to compare the quality of iFLASH explanations and those obtained via an Audit procedure. The Audit explanations are computed centrally on the server, assuming access to the union of all local client datasets $D_{audit} = \cup_c^C D_c$. We remind the readers that the generation of the explanations by the Audit violates the federated conditions for sharing the data, but we consider those explanations as the baseline for our comparison. For consistency, we apply the nested clustering strategy to the D_{audit} set[7], and use the resulting clusters as background data for the audit f_{audit} explainer. This setup allows us to assess the system under controlled conditions.

Faithfulness Normalization: The Faithfulness metric, as computed by Algorithm 1, returns values in the range $[-1, 1]$, where negative values indicate anti-correlation between feature importance and model predictions. For aggregation purposes, we normalize these values to $[0, 1]$ using an affine transformation: $\text{faith}_{\text{norm}} = \frac{\text{faith}+1}{2}$. This transformation ensures that (i) all client contributions receive non-negative weights during aggregation, (ii) a Faithfulness value of 0 (no correlation) maps to 0.5, representing a neutral contribution, and (iii) negative Faithfulness values (anti-correlation) are assigned proportionally lower weights without being excluded entirely.

[7]We remind the reader that the data unification is not permissible in a standard FL deployment.

D. COMPETITORS AND BASELINE

To evaluate the effectiveness of iFLASH, we compare against two federated explainability competitors and one centralized baseline, described below. Both competitors are evaluated solely in the cross-silo setting, as neither was designed or evaluated for cross-device scenarios; including them in cross-device experiments would be an unfair comparison.

FedFCM [18] aims to generate Shap explanations in federated settings via a different strategy: it runs a federated fuzzy C-means clustering process to construct a reference dataset, which is then sent to the server for a centralized Shap computation. This requires sharing reference data with the server, introducing a privacy risk not present in iFLASH. Furthermore, relying on a single server-side explainer reduces flexibility compared to our client-distributed approach.

SHAP-FL [20] proposes a method for federated Shap-based explainability where each client independently synthesizes a background dataset using an adaptive sampling strategy, and the local Shap explanations are aggregated at the server by simple averaging. Unlike iFLASH, SHAP-FL uses a fixed averaging aggregation without quality-aware weighting. We use SHAP-FL as a representative of sampling-based background construction, allowing us to assess the robustness of our nested clustering strategy against an alternative design choice.

Audit** denotes a centralized Shap explainer trained on the union of all client datasets. It is used as an upper-bound reference, not as a deployable method; gathering all data in one place violates FL's core privacy constraint. The ** notation in tables and figures signals this status.

VI. RESULTS

In this section, we present the numerical results obtained in both cross-silo and cross-device settings. In cross-silo scenarios, we optimize over a comprehensive parameter space including number of layers (1-3), hidden units (16-128), batch size (4-128), learning rate (10^{-4} to 0.5), local epochs (1-5), and dropout (0.004-0.180). For cross-device settings, we focus on batch size, learning rate, local epochs, and optimizer selection, reflecting the constraints of resource-limited client populations. All architectures use ReLU activations throughout, with LogSoftmax in the final layer for numerical stability, and Adam optimizer with categorical cross-entropy loss. The number of communication rounds is fixed at 20 to stabilize training.

The optimal hyperparameters vary considerably across datasets and data distributions. Skewed distributions generally require more local epochs for convergence when compared to IID settings, and careful tuning of learning rates and dropout. Dataset sensitivity also varies: Dry bean demonstrates ro-

bustness with stable performance across distributions, while Diamond shows higher sensitivity to distributional changes, requiring significant hyperparameter adjustments. Complete hyperparameter configurations and detailed sensitivity analysis are provided in Appendix I.

As described in Section IV, iFLASH generates local FI explanations at the client level before aggregating them server-side. Our experiments consistently identified 8-10 representative centroids across all client partitions, with each centroid selecting 4 additional samples, resulting in compact reference datasets of 32-40 samples for the Shap explainer. In such a scenario, averaging all FI may not be the optimal choice: some clients may produce better Shap explanations depending on the record being explained and the local data distribution.

For this reason, our iFLASH methodology supports the use of different aggregation strategies that focus on the prediction performance of the client and the Faithfulness of the explanations, i.e., a metric designed to evaluate the quality of the explanations provided.

The complete Faithfulness results across all datasets, aggregation strategies, and distributions are reported in Table 8 in Appendix II. In cross-silo settings, E_{Q1} emerges as the best aggregation strategy for the majority of the datasets and distributions considered, significantly outperforming simple averaging (E_{Avg} [15]). On the simpler binary classification task (Adult), FedFCM and SHAP-FL are statistically equivalent to E_{Q1}, all three occupy the top rank group, reflecting the low inherent difficulty of that task. On the more complex multi-class datasets (Dry bean, Diamond), the advantage of E_{Q1} over both competitors becomes clear and consistent. In the cross-device setting, multiple aggregation strategies show comparable performance, suggesting that the optimal choice depends more on specific deployment constraints.

Negative Faithfulness Scores: Negative Faithfulness values appear in some configurations (illustrated in Figure 5 and in Table 8 of Appendix II), particularly for Dry bean. These indicate anti-correlation between feature importance rankings and model behaviour: features ranked as important by Shap do not consistently affect model predictions as expected. This can occur when: (i) the background dataset does not adequately represent the distribution of test samples, (ii) feature interactions dominate individual feature effects, or (iii) the model behaves non-linearly in certain regions.[8]

Given the different settings analysed, we discuss the outcomes of the aggregation strategies starting with the cross-silo setting in Section VI-A, then addressing the cross-device setting in Section VI-C.

[8]Our Faithfulness-based methodologies improve the aggregate Faithfulness.

A. iFLASH IN CROSS-SILO

In this section, we evaluate the effectiveness of iFLASH in the cross-silo setting, where we have fewer clients with larger data portions. We conducted experiments using Dry bean, Adult, and Diamond datasets. We focus our discussion on Dry bean, setting with 12 clients, considering all the 3 data distributions, while results for other datasets are presented in Appendix II.

Figure 2 reports the results of the different aggregation strategies for Shap explanations with 12 clients. The figure shows the Faithfulness distributions across three data settings (IID, moderately imbalanced, and highly imbalanced), and the corresponding Critical Difference (CD) diagrams to compare the statistical performance of the methods.

Our analysis reveals a clear pattern: Faithfulness-based aggregation strategies dominate across all distribution scenarios. In particular, E_{Q1} achieves the best performance, with E_{faith} and E_{Mix} following closely. These methods form a statistically distinct group, outperforming all other approaches. On Dry bean, both FedFCM and SHAP-FL rank in the middle group with standard averaging (E_{Avg}), behind the Faithfulness-based strategies. Similar trends are observed across other cross-silo settings. On Diamond, the gap is even more pronounced: E_{Q1} ranks first with a large margin, while FedFCM falls to the lower half and SHAP-FL ranks last among all methods. A different picture emerges on Adult, where the binary nature of the task allows SHAP-FL and FedFCM to achieve statistical equivalence with E_{Q1}, with all three methods sharing the top rank group. However, this parity is specific to simple binary classification tasks. On more complex multi-class datasets, E_{Q1} clearly and statistically separates itself from both competitors. This pattern indicates that quality-aware aggregation is most beneficial in complex, multi-class settings, where background data construction and client selection have a stronger impact on explanation quality.

Results from the cross-silo setting show that using Faithfulness to weight client contributions provides a more reliable indicator of explanation quality than the alternative aggregation strategies we designed. This is particularly important in federated settings, where heterogeneous data distributions influence the quality of local explainers.

In this context, the quality of client explanations may vary depending on the input instance, as each client relies on its own local dataset. As a result, some clients may be better suited to explain a given record than others. Unlike other strategies, our approach operates at the instance level, allowing client selection to adapt to each input.

An additional interesting finding regards the ranking of the Audit. Despite having access to the complete dataset, the Audit does not achieve the

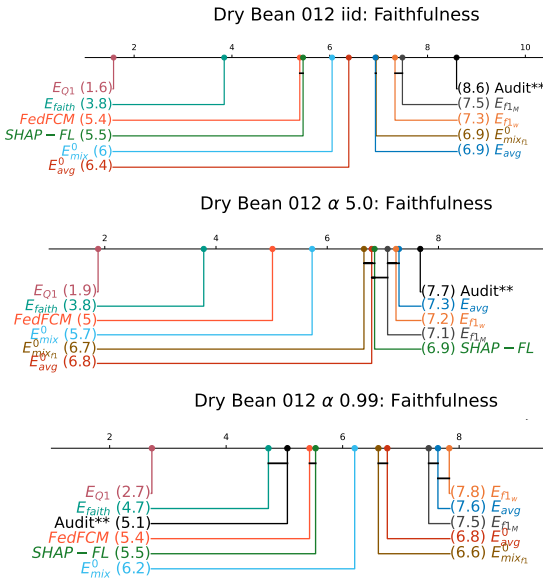

FIGURE 2. Comparative analysis of Faithfulness scores across aggregation strategies for the Dry bean dataset with 12 clients. Each subplot presents CD diagrams (bottom) for different data distributions: (i) IID, (ii) moderately imbalanced ($\alpha = 5.0$), and (iii) highly imbalanced ($\alpha = 0.99$). The CD diagrams rank strategies from best (left) to worst (right), with connected lines indicating statistically equivalent performance.

highest Faithfulness scores. This counter-intuitive scenario is explained by data distribution effects: certain clients with specialised (even imbalanced) datasets provide more accurate explanations for specific records that closely match their distribution. This record-level variability means that selectively aggregating FI explanations from the most relevant clients can outperform even a globally *trained* explainer, highlighting a fundamental advantage of our federated explanation approach.

To better analyse the results obtained in terms of explanations, Figure 3 reports the average FI values extracted for all records in the setting of 12 clients with $\alpha = 5.0$, grouped by class. For each class, the results obtained using the best aggregation strategy, E_{Q1}, are shown with red bars, while the Audit explanations are plotted in blue.

As illustrated, there is a general agreement between the FI values derived from the best aggregation strategy and those from the Audit. This is evident from the direction of the lines for all features, indicating alignment in identifying important features, i.e., the FI for both E_{Q1} and Audit are in the same direction for all of the features. The differences observed are mainly in the magnitude of the values: while they generally align, slight differences may occur. For instance, in class 2, the Audit assigns higher importance to attributes *SF4*, *SF1* (*Shape Form* 4 and 1), and *Rnd* (*Roundness*) compared to the values provided by E_{Q1}. This suggests that, although both methods agree on the importance of certain features, they may differ in the degree of importance assigned, leading to slight variations in the overall ranking of important features. However, this variation typically does not affect the most important features and is more pronounced among those of lower importance. The complete numerical Faithfulness results are reported in Table 8in Appendix II. We note that Dry bean exhibits negative average Faithfulness across all configurations. This reflects anti-correlation between the Shap ranking and the masking response of the model, a known behaviour in multiclass settings with complex feature interactions, where the local masking baseline does not fully capture the model's decision logic [1]. The relative ordering of aggregation strategies remains meaningful: E_{Q1} consistently achieves the least negative values, confirming the benefit of quality-aware selection even when absolute Faithfulness is low.

B. FAITHFULNESS COMPARISON BY DISTRIBUTION

Across our experiments, we observe that federated aggregation strategies, in particular E_{Q1}, can match or exceed the Faithfulness of the centralized Audit explainer, despite the latter having access to the full union of client data. We further analyse the phenomenon to understand this behaviour.

This result may appear counterintuitive, but it can be explained by two complementary mechanisms. First, *local-expert specialization*: each client's explainer is trained on a homogeneous data subset. In skewed settings, a client whose local distribution closely matches a given test instance can produce explanations that are more faithful to that instance than those obtained from an explainer trained on the heterogeneous union D_{audit}, which must account for all distributions at the same time. This observation is consistent with findings in FL showing that local models can outperform global ones on locally representative samples [45].

Second, *sample-wise client selection*: unlike E_{Avg}, which treats all clients equally, E_{Q1} dynamically selects the top-quartile most faithful clients *per record*. This means that for each sample, the aggregation is driven by whichever clients are most locally expert for that specific instance, to a granularity that a single centralized explainer cannot replicate. The combination of these two effects explains why federated quality-aware aggregation can be competitive with, or superior to, a centralized baseline that nominally has more information.

Figure 5 analyses the sensitivity of Faithfulness scores to the choice of background dataset construction, comparing our nested clustering approach (described in Section IV) with independent binning-based sampling [20]. The choice of background data is a well-known challenge in Shap, as it directly

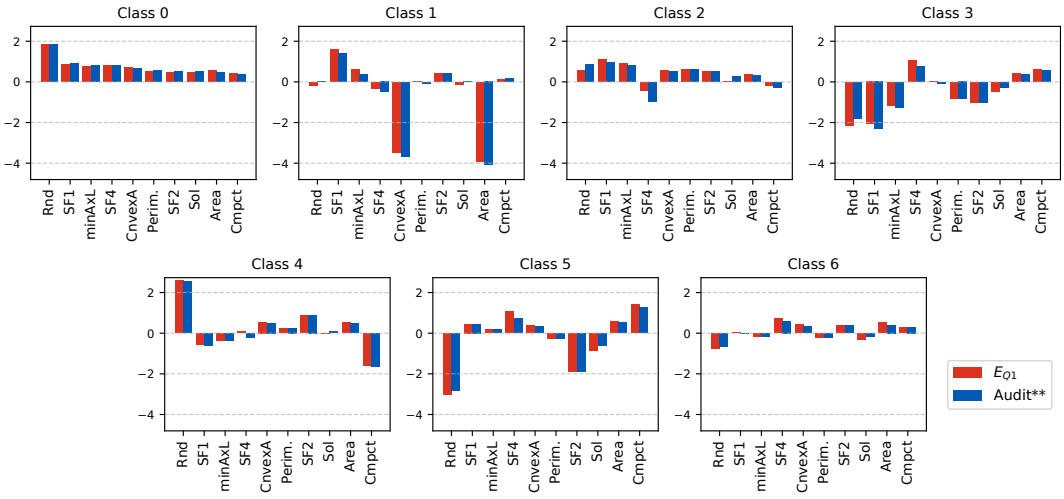

FIGURE 3. **FI comparison between E_{Q1} aggregation (red) and Audit (blue) for the Dry bean dataset with 12 clients and moderate imbalance ($\alpha = 5.0$). Each subplot represents a different bean class, showing average Shap values for samples classified in that class. We show only the first top 10 features in magnitude for readability reasons.**

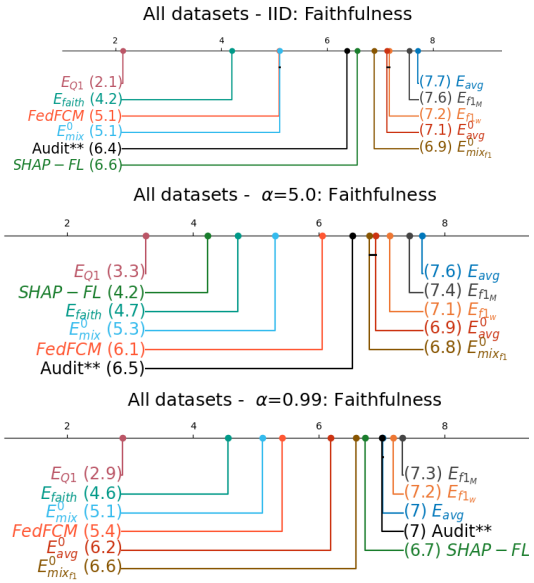

FIGURE 4. **Comparative analysis of Faithfulness scores across aggregation strategies for ALL the dataset with 12 and 16 clients. Each subplot presents CD diagrams for different data distributions: (i) IID, (ii) moderately imbalanced ($\alpha = 5.0$), and (iii) highly imbalanced ($\alpha = 0.99$). The CD diagrams rank strategies from best (left) to worst (right), with connected lines indicating statistically equivalent performance.**

influences the resulting explanations. In this work, we focus on analysing this aspect in the specific context of FL, and the results are shown for Adult, Diamond, and Dry bean with $\alpha = 5$ and 12 clients.

Overall, the distributions of Faithfulness scores across aggregation strategies remain stable under both approaches, indicating that our methods are robust to this design choice.

For Adult, where the task is relatively simple, both construction strategies yield similar results across all methods. For Diamond and Dry bean, differences are more pronounced: clustering-based explanations tend to resemble cluster-based competitors, while sampling-based explanations are closer to sampling-based ones.

Despite these variations, E_{Q1} achieves the best Faithfulness scores, while competitor methods such as Audit and SHAP-FL show greater variability across datasets and settings. This suggests that the advantage of E_{Q1} derives from its quality-aware client selection, rather than from the specific choice of background dataset construction.

C. `iFLASH` IN CROSS-DEVICE

We now analyse `iFLASH` in the cross-device setting, characterised by a large number of clients with smaller local datasets, which presents different challenges compared to the cross-silo setting. We conduct experiments on the Dutch and ACSI datasets, focusing the discussion on Dutch (results for ACSI are reported in Appendix II).

For Dutch, we analyse three settings: 50, 70 and 150 clients; results for the 50 and 150 client configurations are reported in Appendix II. Figure 6 presents the results of the different aggregation strategies for Shap explanations in the setting with 70 clients. The figure is composed of three plots, each corresponding to one of the analysed distributions: (i) an IID setting; (ii) $\alpha = 5.0$, a moderately skewed scenario; and (iii) $\alpha = 0.99$, a highly skewed setting. For each scenario, a CD plot ranks all aggregation strategies.

In the cross-device setting, which introduces additional complexity, we observe a slightly different behaviour compared to the cross-silo one. In the IID case, E_{Q1} ranks first by a large margin, clearly outperforming all other aggregation strategies, while Audit appears in fifth position, with a substantial gap. In the moderately imbalanced setting, E_{Q1} drops to the third cluster, preceded by

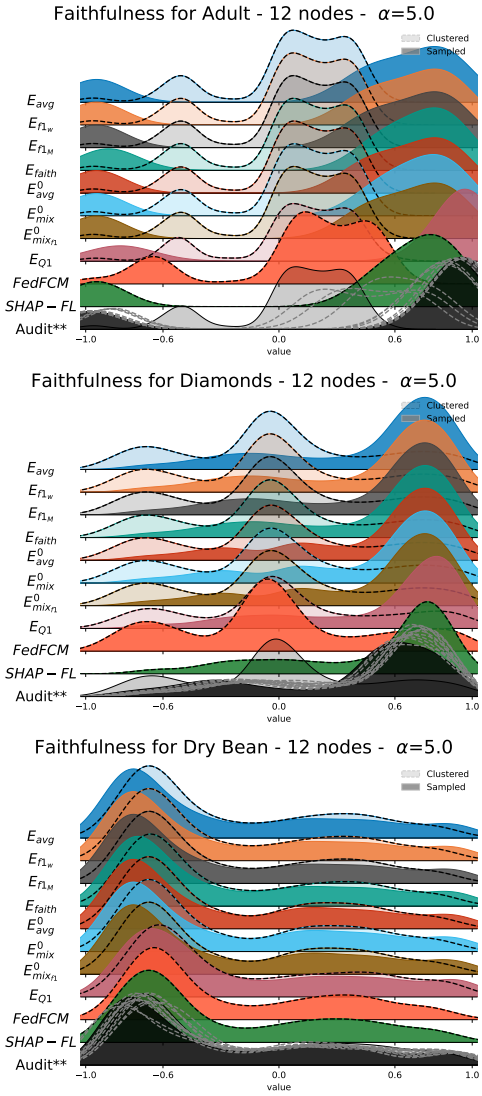

FIGURE 5. Distribution of Faithfulness scores across iFLASH aggregation strategies and the Audit baseline, for Adult, Diamond, and Dry bean at $\alpha = 5.0$ with 12 clients (cross-silo). Each ridge shows *two overlapping distributions* per strategy: one using our nested clustering background dataset construction (described in Section IV, in solid) and one using a local adaptive sampling strategy analogous to the approach in [20], which builds a dynamic lattice subset of the local training set as the Audit does on the full data. The degree of overlap within each strategy indicates robustness to background dataset construction; differences *across* strategies reflect the effect of the aggregation policy. Faithfulness lies in the [-1,1] domain.

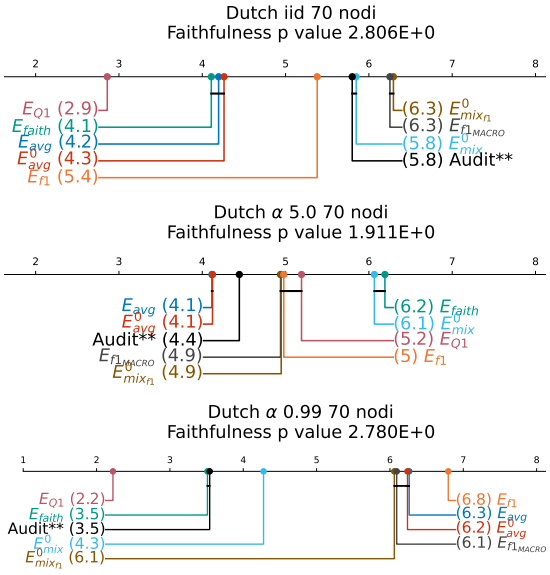

FIGURE 6. Comparative analysis of Faithfulness scores across aggregation strategies for the Dutch dataset with 70 clients. Each subplot presents CD diagrams for different data distributions: (i) IID, (ii) moderately skewed ($\alpha = 5.0$), and (iii) highly skewed ($\alpha = 0.99$). Connected lines indicate statistically equivalent performance.

several methods based on predictive performance and also the Audit. In the highly imbalanced setting, however, E_{Q1} again ranks first.

Although the pattern is less clear than in the cross-silo scenario, E_{Q1} still shows an overall advantage. In particular, even in the most challenging setting, i.e., under strong imbalance, E_{Q1} remains the top-performing method.

This behaviour differs from the cross-silo case, suggesting that the optimal aggregation strategy depends not only on data distribution but also on the number of clients and the amount of local data. With a larger number of clients, even simple averaging approaches can perform well, but cross-silo settings benefit more consistently from Faithfulness-based selection. This can be explained by a variance reduction effect: when aggregating over many clients, random noise in individual explanations tends to cancel out, reducing the relative advantage of selective weighting. As a consequence, the absolute differences between strategies in the cross-device regime are small and nearly indistinguishable in distribution plots. This is precisely why we rely on CD diagrams based on the Friedman test rather than visual inspection alone: statistical analysis confirms that E_{Q1} retains a consistent, if modest, advantage even where direct comparison of distributions would not support a clear conclusion when instead compared to Figure 7 that examines the distribution of Faithfulness scores across aggregation strategies for the Dutch dataset at $\alpha = 5.0$ with 70 clients. A high amount of contributions cancels out the advantages of smart aggregations, suggesting that quality-aware aggregation is most impactful in data-scarce, low-client-count settings, where noise in individual explanations is harder to average away.

Similar to the cross-silo analysis, Figure 8 reports FI comparisons between our best strategy and Audit. The consistency in both direction and magnitude shows that our federated approach preserves explanation fidelity while respecting federated constraints, which is particularly important in the cross-device setting, where clients may operate on sensitive user data.

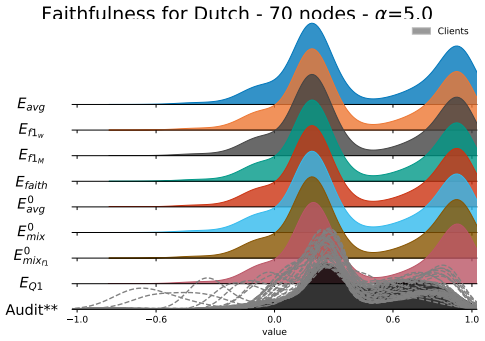

FIGURE 7. Distribution of Faithfulness scores across aggregation strategies for the Dutch dataset, 70 clients, $\alpha = 5.0$, cross-device setting.

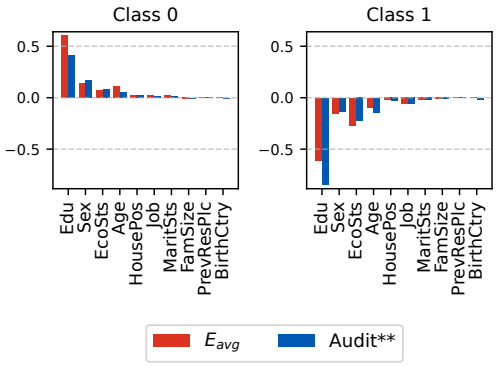

FIGURE 8. The global explanation by class for the dataset Dutch when having 70 clients with moderate skew distributions ($\alpha = 5.0$). The FI of the audit is plotted as a blue bar; meanwhile, the aggregation E_{Avg} is represented as red bars. E_{Avg} has the highest rank among the aggregations in the plot of 6) We show only the first top 10 features in magnitude for readability reasons.

VII. CONCLUSION AND FUTURE WORK

This work investigates the impact of aggregation strategies on the quality of FI explanations in federated learning settings, introducing `iFLASH` as a methodology to systematically analyse and combine local explanations. First, we show that aggregating client-side explanations can match, and in some cases surpass, the quality of explanations obtained from centralized models. This result challenges the common assumption that explainability degrades under federated constraints, suggesting instead that federated settings can enhance explanation quality when aggregation is properly designed.

Second, our results indicate that aggregation strategies based on explanation quality outperform the alternatives. In particular, E_{Q1} emerges as the most effective approach overall, demonstrating that weighting client contributions according to the quality of their explanations is a reliable strategy for improving federated explainability. This finding highlights that predictive performance alone is not sufficient to guide aggregation and that explanation quality provides a more informative signal.

Third, the cross-device analysis shows that the optimal aggregation strategy depends on the number of clients and the data distribution. While simpler approaches such as averaging or filtering may perform well in balanced settings, quality-aware strategies remain robust in more challenging scenarios, especially under strong data heterogeneity. Future work could investigate more granular aggregation strategies; for instance, the isochronous temporal metric [36] provides a complementary perspective by identifying interpretable feature relations based on 'attribute velocity' on higher-dimensional curved surfaces, which could refine how `iFLASH` weights individual client contributions. `iFLASH` focuses on tabular data and relies on FI-based explanations, which may be applied to other types of data and models. Our evaluation is based on Faithfulness, which does not capture aspects of explanation quality, as no human-centered evaluation is considered. Furthermore, while our approach follows the standard federated learning paradigm, it does not provide formal privacy guarantees, and explanation vectors may still encode information about local data distributions.

These limitations open several directions for future work. First, extending the evaluation to larger-scale and real-world federated environments is an important next step to assess scalability and robustness under realistic conditions. Second, extending `iFLASH` to other data, such as images, text, and time series, would require domain-specific aggregation strategies. Third, future work should investigate robustness against adversarial clients and potential information leakage through aggregated explanations. Finally, user studies are needed to evaluate the practical usefulness of federated explanations beyond quantitative metrics.

References

[1] D. Alvarez-Melis et al. "Towards Robust Interpretability with Self-Explaining Neural Networks". In: *Advances in Neural Information Processing Systems*. 2018.

[2] T. Awosika et al. "Transparency and Privacy: The Role of Explainable AI and Federated Learning in Financial Fraud Detection". In: *IEEE Access* 12 (2024). DOI: 10 . 1109 / access.2024.3394528.

[3] B. Becker et al. *Adult*. 1996.

[4] D. J. Beutel et al. *Flower: A Friendly Federated Learning Research Framework*. 2020.

[5] L. Biewald. *Experiment Tracking with Weights and Biases*. 2020.

[6] F. Bodria et al. "Benchmarking and Survey of Explanation Methods for Black Box Models". In: *Data Mining and Knowledge Discovery* 37.5 (2023), pp. 1719–1778. DOI: 10.1007/s10618-023-00933-9.

[7] V. Bonsignori et al. "Differentially Private FastSHAP for Federated Learning Model Explainability". In: *International Joint Conference on Neural Networks (IJCNN)*. 2025, pp. 1–8. DOI: 10.1109/ijcnn64981.2025.11227553.

[8] Briola et al. "A Federated Explainable AI Model for Breast Cancer Classification". In: *Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference*. EICC '24. Xanthi, Greece: Association for Computing Machinery, 2024, pp. 194–201. ISBN: 979-8-4007-1651-5. DOI: 10.1145/3655693.3660255.

[9] J. L. C. Bárcena et al. "A Federated Fuzzy C-Means Clustering Algorithm." In: *WILF*. 2021.

[10] J. L. C. Bárcena et al. "Enabling Federated Learning of Explainable AI Models within Beyond-5G/6G Networks". In: *Computer Communications* 210 (2023), pp. 356–375. ISSN: 0140-3664. DOI: 10.1016/j.comcom.2023.07.039.

[11] J. L. C. Bárcena et al. "Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models". In: *XAI.it@AI*IA*. CEUR Workshop Proceedings. 2022.

[12] D. Chen et al. "Fs-Real: Towards Real-World Cross-Device Federated Learning". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023. DOI: 10.1145/3580305.3599829.

[13] P. Chen et al. "EVFL: An Explainable Vertical Federated Learning for Data-Oriented Artificial Intelligence Systems". In: *Journal of Systems Architecture* 126 (2022), p. 102474. ISSN: 1383-7621. DOI: 10.1016/j.sysarc.2022.102474.

[14] L. Corbucci et al. "Enhancing Privacy and Utility in Federated Learning: A Hybrid P2p and Server-Based Approach with Differential Privacy Protection". In: *Proceedings of the Int. Conference on Security and Cryptography*. Vol. 1. Science and Technology Publications, Lda, 2024. DOI: 10.5220/0012863600003767.

[15] L. Corbucci et al. "Explaining Black-Boxes in Federated Learning". In: *Explainable Artificial Intelligence*. Springer Nature, 2023. DOI: 10.1007/978-3-031-44067-0_8.

[16] M. Craven et al. "Extracting Tree-Structured Representations of Trained Networks". In: *Advances in Neural Information Processing Systems*. Vol. 8. MIT Press, 1995.

[17] F. Ding et al. "Retiring Adult: New Datasets for Fair Machine Learning". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual*. 2021.

[18] P. Ducange et al. "Federated SHAP: Privacy-Preserving and Consistent Post-Hoc Explainability in Federated Learning". In: *Machine Learning* 115.2 (2026), p. 24. DOI: 10.1007/s10994-025-06956-1.

[19] P. Ducange et al. "Federated Learning of XAI Models in Healthcare: A Case Study on Parkinson's Disease". In: *Cognitive Computation* 16.6 (2024). DOI: 10.1007/s12559-024-10332-x.

[20] C. Düsing et al. "SHAP-FL: Improving Explainability for Multicentric Sepsis Onset Prediction Through Background Dataset Synthesis". In: *Int. Conf. on Artificial Intelligence in Medicine*. Springer, 2025, pp. 120–129. DOI: 10.1007/978-3-031-95838-0_12.

[21] European Commission. *Ethics Guidelines for Trustworthy AI*. Publications Office, 2019. DOI: 10.2759/177365.

[22] Expert group on how AI principles should be implemented. "AI Governance in Japan". In: (2023).

[23] A. A. Fahliani et al. "Privacy-Preserving Federated Interpretability". In: *IEEE Int. Conf. on Big Data, BigData 2024*. IEEE-Institute of Electrical and Electronics Engineers Inc., 2024.

[24] J. Fiosina. "Explainable Federated Learning for Taxi Travel Time Prediction". In: *VEHITS*. SCITEPRESS, 2021. DOI: 10.5220/0010485606700677.

[25] M. Fontana et al. "Monitoring Fairness in HOLDA". In: *Frontiers in Artificial Intelligence and Applications*. Vol. 354. IOS Press, 2022, pp. 246–248. DOI: 10.3233/faia220205.

[26] A. A. Freitas. "Comprehensible Classification Models: A Position Paper". In: *SIGKDD Explor.* 15.1 (2013). DOI: 10.1145/2594473.2594475.

[27] F. Giannotti et al. "Explainable for Trustworthy AI". In: *Discovery Science*. Vol. 13692. Lecture Notes in Computer Science. Springer, 2023. DOI: 10.1007/978-3-031-24349-3_10.

[28] R. Guidotti et al. "Factual and Counterfactual Explanations for Black Box Decision Making". In: *IEEE Intelligent Systems* 34.6 (2019), pp. 14–23. DOI: 10.1109/mis.2019.2957223.

[29] R. Guidotti et al. "Stable and Actionable Explanations of Black-Box Models through Factual and Counterfactual Rules". In: *Data Min. Knowl. Discov.* 38.5 (2024). DOI: 10.1007/s10618-022-00878-5.

[30] R. Haffar et al. "Explaining Predictions and Attacks in Federated Learning via Random Forests". In: *Appl. Intell.* (2022). ISSN: 1573-7497. DOI: 10.1007/s10489-022-03435-1.

[31] R. Haffar et al. "GLOR-FLEX: Local to Global Rule-Based EXplanations for Federated Learning". In: *2024 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*. 2024. DOI: 10.1109/fuzz-ieee60900.2024.10611878.

[32] C. Huang et al. "Cross-Silo Federated Learning: Challenges and Opportunities". In: *arXiv preprint arXiv:2206.12949* (2022). DOI: 10.48550/arXiv.2206.12949. arXiv: 2206.12949.

[33] D. Janzing et al. "Feature Relevance Quantification in Explainable AI: A Causal Problem". In: *The 23rd Int. Conf. on Artificial Intelligence and Statistics*. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2907–2916.

[34] N. Jethani et al. "FastSHAP: Real-Time Shapley Value Estimation". In: *The Tenth Int. Conf. on Learning Representations, ICLR*. 2022.

[35] M. Koklu et al. "Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques". In: *Computers and Electronics in Agriculture* (2020). DOI: 10.1016/j.compag.2020.105507.

[36] A. K. Kumar et al. "Isochronous Temporal Metric for Neighbourhood Analysis in Classification Tasks". In: *SN Computer Science* 4.6 (2023), p. 807. DOI: 10.1007/s42979-023-02351-6.

[37] A. Kusiak. "Federated Explainable Artificial Intelligence (fXAI): A Digital Manufacturing Perspective". In: *Int. journal of production research* 62.1-2 (2024). DOI: 10.1080/00207543.2023.2238083.

[38] Q. Li et al. "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection". In: *arXiv e-prints* (2019). DOI: 10.48550/arXiv.1907.09693.

[39] Z. Liu et al. "GTG-Shapley: Efficient and Accurate Participant Contribution Evaluation in Federated Learning". In: *ACM Trans. Intell. Syst. Technol.* 13.4 (2022). ISSN: 2157-6904. DOI: 10.1145/3501811.

[40] L. Longo et al. "Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions". In: *CD-MAKE*. Vol. 12279. Lecture Notes in Computer Science. Springer, 2020. DOI: 10.1007/978-3-030-57321-8_1.

[41] S. M. Lundberg et al. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. 2017, pp. 4765–4774.

[42] R. López-Blanco et al. "Federated Learning of Explainable Artificial Intelligence (FED-XAI): A Review". In: *Distributed Computing and Artificial Intelligence, 20th Int. Conf.* Cham: Springer Nature, 2023. ISBN: 978-3-031-38333-5.

[43] J. MacQueen. "Some Methods for Classification and Analysis of Multivariate Observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. University of California press, 1967.

[44] T. Madiega. "Artificial Intelligence Act". In: *European Parliament: European Parliamentary Research Service* (2021).

[45] Y. Mansour et al. *Three Approaches for Personalization with Applications to Federated Learning*. 2020. DOI: 10.48550/arxiv.2002.10619.

[46] E. Mariotti et al. "Beyond Prediction Similarity: ShapGAP for Evaluating Faithful Surrogate Models in XAI". In: *Explainable Artificial Intelligence*. Springer Nature, 2023. DOI: 10.1007/978-3-031-44064-9_10.

[47] B. McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: *Proceedings of the 20th Int. Conf. on Artificial Intelligence and Statistics, AISTATS*. Vol. 54. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1273–1282.

[48] F. Naretto et al. "Prediction and Explanation of Privacy Risk on Mobility Data with Neural Networks". In: *ECML PKDD 2020 Workshops*. Vol. 1323. Communications in Computer and Information Science. Springer, Cham, 2020. DOI: 10.1007/978-3-030-65965-3_34.

[49] T. Nishio et al. "Estimation of Individual Device Contributions for Incentivizing Federated Learning". In: *2020 IEEE Globecom Workshops (GC Wkshps*. IEEE, 2020. DOI: 10.1109/gcwkshps50303.2020.9367484.

[50] M. T. Ribeiro et al. "" Why Should i Trust You?" Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[51] W. Riviera et al. "FeLebrities: A User-Centric Assessment of Federated Learning Frameworks". In: (2023). DOI: 10.1109/access.2023.3312579.

[52] H. Roberts et al. "The Chinese Approach to Artificial Intelligence: An Analysis of Pol-

icy, Ethics, and Regulation". In: *AI & society* (2021). DOI: 10.1007/s00146-020-00992-2.

[53] C. Rudin. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *Nature machine intelligence* 1.5 (2019), pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

[54] J. Shen et al. "Explaining Federated Learning Through Concepts in Image Classification". In: *Int. Conf. on Algorithms and Architectures for Parallel Processing*. Springer, 2023. DOI: 10.1007/978-3-031-46311-2_22.

[55] D. Shin. "The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI". In: *Int. J. of Human-Computer Studies* 146 (2021). ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2020.102551.

[56] A. Shrikumar et al. "Learning Important Features Through Propagating Activation Differences". In: *Proceedings of the 34th Int. Conf. on Machine Learning, ICML 2017*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017.

[57] S. K. Singh et al. *Zero-Trust Agentic Federated Learning for Secure IIoT Defense Systems*. 2025.

[58] P. Sturmfels et al. "Visualizing the Impact of Feature Attribution Baselines". In: *Distill* (2020). DOI: 10.23915/distill.00022.

[59] P. Van der Laan. "The 2001 Census in the Netherlands: Integration of Registers and Surveys". In: 2001.

[60] G. Wang et al. "Measure Contribution of Participants in Federated Learning". In: *2019 IEEE Int. Conf. on Big Data*. IEEE, 2019, pp. 2597–2604. DOI: 10.1109/bigdata47090.2019.9006179.

[61] J. Wang et al. "Efficient Participant Contribution Evaluation for Horizontal and Vertical Federated Learning". In: *2022 IEEE 38th Int. Conf. on Data Engineering (ICDE)*. IEEE, 2022. DOI: 10.1109/icde53745.2022.00073.

[62] H. Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer, 2016. ISBN: 978-3-319-24277-4.

[63] Q. Yang et al. *Federated Machine Learning: Concept and Applications*. 2019.

[64] L. Zhang et al. "Intrinsic Performance Influence-Based Participant Contribution Estimation for Horizontal Federated Learning". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 13.6 (2022). DOI: 10.1145/3532612.

[65] Y. Zhang et al. "LR-XFL: Logical Reasoning-Based Explainable Federated Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 2024. DOI: 10.1609/aaai.v38i19.30179.

[66] M. K. Zuziak et al. "Amplified Contribution Analysis for Federated Learning". In: *Advances in Intelligent Data Analysis XXII*. Cham: Springer Nature, 2024. DOI: 10.1007/978-3-031-58553-1_6.

VALERIO BONSIGNORI is a Ph.D. graduate in Artificial Intelligence from the University of Pisa. His research interests include explainable artificial intelligence, interpretable Machine Learning, and Federated Learning. His doctoral work focused on explainable federated models and their application in real-world scenarios.

LUCA CORBUCCI is a Postdoctoral Researcher at Fondazione Bruno Kessler, where he is part of the Machine Translation unit. He holds a PhD in Computer Science from the University of Pisa. During his PhD, he focused on Fair and Private-Preserving Federated Learning. As part of his PhD, he visited the Telefónica Discovery Scientific Research Laboratory in Barcelona.

FRANCESCA NARETTO is an Assistant Professor at the University of Pisa, Italy. She has a Ph.D. in Data Science from Scuola Normale Superiore, Italy, with a thesis on the synergies and tensions between Data Privacy and Explainability, awarded as Best Ph.D. thesis at the ITADATA 2024 conference. Her research interests are in explainable artificial intelligence, data privacy, fairness, and federated learning.

RICCARDO GUIDOTTI is an Associate Professor at the Department of Computer Science at the University of Pisa and a member of the Knowledge Discovery and Data Mining Laboratory (KDDLab), a research group in collaboration with ISTI-CNR of Pisa and with Scuola Normale Superiore. His research interests include explainable artificial intelligence, interpretable machine learning, personal data mining, clustering, and the analysis of transactional data and time series.

ANNA MONREALE is an Associate Professor at the Computer Science Department of the University of Pisa and a member of the Knowledge Discovery and Data Mining Laboratory (KDD-Lab), a joint research group with the ISTI-CNR, in Pisa. Her research interests include big data analytics, responsible AI. She earned her Ph.D. in computer science in 2011, and her dissertation was about privacy-by-design in data mining.