Federated SHAP: Privacy-Preserving and Consistent Post-hoc Explainability in Federated Learning

Pietro Ducange[1]†, Francesco Marcelloni[1]†,
Giustino Claudio Miglionico[1]†, Alessandro Renda[2]†,
Fabrizio Ruffini[1*]†

[1]Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, Pisa, 56122, Italy, Italy.
[2]Department of Engineering and Architecture, University of Trieste, Via Valerio 6/1, Trieste, 34127, Italy, Italy.

*Corresponding author(s). E-mail(s): fabrizio.ruffini@unipi.it;
Contributing authors: pietro.ducange@unipi.it;
francesco.marcelloni@unipi.it; giustino.miglionico@phd.unipi.it;
alessandro.renda@dia.units.it;
†These authors contributed equally to this work.

Abstract

The widespread adoption of Artificial Intelligence in everyday activities highlights a growing and urgent need for trustworthiness. Designing trustworthy AI systems requires addressing key technical challenges, including ensuring data privacy and model explainability. Federated Learning (FL) is a widely adopted paradigm to preserve data privacy in collaborative learning scenarios, while post-hoc methods are commonly applied to enhance the explainability of opaque AI-based models. In this paper, we propose a novel approach, called Federated SHAP, to simultaneously address privacy and explainability. Specifically, we leverage the SHapley Additive exPlanations (SHAP) method to provide post-hoc explanations of Neural Networks trained through FL. SHAP relies on a representative background dataset; however, constructing such a dataset in the FL setting is particularly challenging since raw data distributed across multiple clients cannot be shared directly due to strict privacy requirements. To address this challenge, we propose two tailored strategies depending on the data type: for tabular data, we adopt a Federated Fuzzy C-Means clustering algorithm to collaboratively summarize the distributed datasets into a suitable background dataset; for image data, we introduce a Federated Generative Adversarial Network (GAN) to synthesize representative background instances. A comprehensive experimental evaluation demonstrates the effectiveness and robustness of our proposed approaches, comparing them against several baseline and alternative strategies in terms of both representativeness and quality of generated explanations. Compared to baselines employing randomly generated representative background datasets, our approach reduces the discrepancy of SHAP explanations by up to three times on tabular data and two times on image data (depending on the test case involved), when measured against the centralized SHAP values computed using the full training set as background dataset.

Keywords: Federated Learning, Explainable Artificial Intelligence, SHAP, Fuzzy Clustering, Fuzzy C-Means, GAN

1 Introduction

The rapid and widespread adoption of Artificial Intelligence (AI) and Machine Learning (ML) systems has raised significant concerns regarding the trustworthiness of their outcomes. In response, various initiatives have been undertaken to promote trust, spanning from ethical frameworks to the development of new laws and regulations. Notably, in 2024, the European Union approved the world's first comprehensive AI legislation: the AI Act[1]. Initially proposed in 2021, the Act builds upon earlier efforts such as the EU's Ethics Guidelines for Trustworthy AI [1]. These documents reflect the EU's commitment to fostering the responsible development and use of AI for the benefit of society. In this context, trustworthy AI is defined through three key dimensions: lawfulness, ethical alignment, and technical robustness. According to the guidelines, AI systems must meet several essential requirements, among them, privacy, data governance, and transparency are particularly critical, especially when AI-driven decisions have significant implications for individuals, such as in healthcare or security applications.

Data privacy represents a critical concern in ML, primarily due to the high data demand for model training. To address privacy issues, in 2016, Google introduced a decentralized learning paradigm named Federated Learning (FL) [2]. FL allows the collaborative training of a global ML model without requiring participants to share data. In practice, each party (client) locally trains a model on its data and contributes model updates, rather than the data itself, to collaboratively build a shared global model. The FL paradigm can manage horizontally partitioned data (clients have distinct instances but identical features), vertically partitioned data (clients have identical instances but different features), or hybrid scenarios. Neural Networks (NN) are particularly suited for training in a FL setting; however, they are often regarded as opaque and difficult to interpret.

[1] https://artificialintelligenceact.eu/the-act/, accessed March 2024

Alongside privacy, transparency is an equally critical requirement for trustworthiness. Transparency, as described in [1], includes traceability, communication clarity, and explainability. Specifically, explainability refers to "*the ability to explain the technical processes of an AI system*" [1] and is the central focus of Explainable AI (XAI) [3, 4]. Explainability can be pursued through interpretable-by-design models or via post-hoc methods that provide explanations for otherwise opaque models, such as NN and ensemble models.

One of the most popular post-hoc explanation techniques is SHapley Additive exPlanations (SHAP) [5], which explains individual AI model predictions by quantifying input feature importance using Shapley values from game theory [6]. SHAP requires three key elements: the trained model, the instance to explain, and a representative "background (reference) dataset" drawn from the same distribution as the original training data. The background dataset plays a crucial role in SHAP. In fact, feature attributions are computed by evaluating model predictions across possible subsets of input features, and the values of missing features are replaced using samples from the background dataset. This dataset acts as a reference distribution and is essential to simulate realistic feature combinations, making it a key component for generating accurate and meaningful explanations. Usually, the training dataset itself serves as the background, possibly undersampled to reduce SHAP's computational complexity.

The design of an appropriate background dataset becomes even more challenging in FL scenarios, where explanations must simultaneously satisfy privacy, consistency, and accuracy requirements. Due to privacy constraints inherent in FL, the background dataset cannot consist of the scattered local datasets, as local data cannot be centralized. A naive solution would be to use local data separately for each client; however, this approach exhibits several shortcomings. First, local datasets typically have heterogeneous distributions, which may lead to inconsistent explanations among clients for identical predictions [7], thus undermining explanation reliability. Second, this solution could also lead to privacy violations, particularly in realistic scenarios where the instance to be explained resides either on a single client or on an external client that did not participate in the federated training. Therefore, such an approach may not be feasible in practice. In general, explanations generated in an FL setting should ideally approximate as closely as possible those obtained in a centralized scenario.

This work builds upon recent advances at the intersection of FL and XAI (labeled as Fed-XAI [8]), combining their strengths to develop trustworthy AI solutions that simultaneously ensure privacy and explainability. Specifically, our research extends our previous work [9], in which we proposed Federated SHAP, an approach for generating representative background datasets in FL settings using a Federated Fuzzy C-Means clustering algorithm [10] tailored for tabular data. While our previous contribution demonstrated promising results, it was limited by the exclusive focus on tabular datasets and lacked extensive comparison with alternative baseline strategies.

To overcome these limitations, we have significantly extended the experimental analysis for tabular data scenarios, introducing new datasets and developing additional baseline methods, including a novel approach based on local differential privacy principles [11]. Furthermore, recognizing the broader applicability of FL, we expanded our

Federated SHAP approach to the critical domain of image data, leveraging Federated Generative Adversarial Networks (GANs) [12] to collaboratively generate synthetic yet representative background instances. This extension is particularly relevant, given that computer vision represents one of the main application fields of FL [13]. This relevance stems from the widespread distribution of image data across multiple clients, particularly in sensitive fields such as healthcare, surveillance, and autonomous driving, where data sharing restrictions are stringent and collaboration among different entities is often essential. The core technical contribution of this work is the proposal of a unified framework that bridges explainability and privacy by enabling SHAP-based explanations in FL. Specifically, we describe two modality-aware strategies for constructing representative background datasets, namely FedFCM for tabular data and FedGAN for images, thus making SHAP explanations feasible and reliable in decentralized settings.

Through extensive experimental validation, encompassing diverse classification and regression datasets, we demonstrate the effectiveness, privacy preservation, accuracy, and consistency of our proposals compared to baseline and alternative strategies, thereby confirming their suitability for practical deployment in trustworthy AI scenarios.

The remainder of this paper is structured as follows. Section 2 discusses recent literature on FL approaches to XAI models, with particular emphasis on methods adopting SHAP as a post-hoc explanation technique. Section 3 introduces the formal statement of the addressed problem, followed by a detailed presentation of our Federated SHAP approaches for tabular and image datasets. Section 4 outlines the experimental setup, describes the datasets and models considered, and presents the baseline strategies used to generate background datasets for comparison with our proposed approaches. The experimental results and related discussions are presented in Section 5. Finally, Section 6 provides concluding remarks and highlights possible future directions.

2 Related Works

Fed-XAI has recently attracted substantial attention due to the increasing demand for trustworthy AI models that combine interpretability with strong data privacy guarantees [8, 14, 15]. Existing Fed-XAI approaches can be broadly categorized into two main groups. The first group includes inherently interpretable-by-design methods, such as decision trees and rule-based systems, trained directly within federated settings [16–18]. While these approaches offer built-in transparency, they often come at the cost of reduced predictive performance. The second group consists of post-hoc interpretability methods applied to complex opaque models (e.g., neural networks), which aim to explain model behavior without compromising predictive accuracy. Among post-hoc techniques, SHAP has emerged as one of the most robust and widely adopted tools due to its effective feature-level explanations.

Several studies have integrated SHAP into FL, primarily focusing on tabular data. Briola et al. [19] used SHAP for federated breast cancer classification, highlighting data privacy benefits, but did not adequately address consistency among federated clients. Sandeepa et al. [20] applied SHAP in cybersecurity to identify data poisoning attacks, but their analysis of privacy implications related to background data generation was limited. Corbucci et al. [21] proposed averaging local SHAP explanations, effectively ensuring consistency and approximation of centralized results; however, their method requires unrealistic assumptions regarding shared test data, thus weakening privacy guarantees. Additional contributions, such as those from Asiri et al. [22], Saad et al. [23], Kalakoti et al. [24], Sarker et al. [25], Abtahi et al. [26], and Fatema et al. [27] either lack systematic analysis of privacy preservation, inter-client consistency, or do not thoroughly evaluate the representativeness of background datasets used for explanations.

While all these works primarily focus on tabular data, scenarios involving image data introduce additional complexities, such as high dimensionality and spatial structure, which pose significant challenges for generating privacy-preserving and representative background datasets suitable for SHAP explanations.

Recent literature emphasizes FL as an effective privacy-preserving paradigm for distributed training of image-based models, especially in sensitive domains such as medical diagnostics [28, 29]. Concurrently, the integration of post-hoc explainability techniques like SHAP and LIME (introduced in [30]) into deep image models has been recognized as crucial for enhancing interpretability and user trust [31, 32]. However, Fed-XAI, specifically designed for image data, remains relatively unexplored. Existing methods typically apply post-hoc explanations independently of the federated setting, neglecting the potential impact of federated architectures on the quality, reliability, and consistency of generated explanations. Moreover, quantitative metrics for assessing inter-client consistency of visual explanations have rarely been investigated.

To address the aforementioned limitations, this work introduces a novel multi-modal Fed-XAI approach, specifically designed to support both tabular and image datasets. For tabular data, we propose the use of a Federated Fuzzy C-Means (Fed-FCM) clustering algorithm [10] to collaboratively generate a representative and privacy-preserving reference dataset for SHAP explanations. This method effectively balances privacy constraints, explanation accuracy, and consistency, addressing gaps identified in prior studies.

For image data, we introduce an innovative approach that leverages Federated Generative Adversarial Networks (GAN) to generate synthetic reference images [12, 33]. These synthetic images are used as a background for SHAP-based explanations, addressing the critical challenge of constructing suitable reference datasets under strict privacy constraints. Although federated GAN frameworks have been proposed to handle data heterogeneity and maintain privacy in image synthesis tasks [34–36], their application for systematically constructing reference datasets to support SHAP-based explainability is, to the best of our knowledge, unprecedented.

In summary, previous Fed-XAI approaches present considerable limitations regarding inter-client consistency, privacy, and explanation accuracy, especially in real-world federated environments. Our multi-modal approach, labeled as *FedSHAP-FCM* for tabular data and *FedSHAP-GAN* for image data, provides a comprehensive and robust solution explicitly addressing these gaps. Through extensive experimental validation against various baseline methods, our proposals significantly advance the state of Fed-XAI research.

3 Federated SHAP

In this section, we first describe the problem addressed in this work and then outline the proposed approaches to use SHAP in an FL setting. We collectively refer to them as Federated SHAP (FedSHAP). Our methodology focuses on developing techniques to generate representative background datasets in a privacy-preserving manner. In devising the methodology, we also considered the well-known problem of the relevant computational effort needed to calculate the Shapley values. The proposed techniques are tailored to specific data types, namely tabular and image datasets.

3.1 Problem statement

We consider a scenario where M clients ($m = 1, \ldots, M$) collaborate to create an AI model for classification or regression tasks. The training phase is orchestrated by an independent entity (e.g., a server). To reproduce a realistic situation commonly found in decentralized settings, we simulated non-IID scenarios where the data are horizontally partitioned: each client's dataset has a distribution that does not follow the overall data distribution and is also different from the other clients' ones. For a client m, the training data is indicated with $(\mathbf{X}^m, \mathbf{Y}^m)$, where $\mathbf{X}^m$ indicates an $N_m \times F$ matrix of N_m instances described by F input features, and $\mathbf{Y}^m$ indicates the vector of the N_m associated target values. The generic instance of the training data for a client m is indicated by $\mathbf{x}_i^m$, with $i \in [1, N_m]$.

After the models are trained in an FL fashion, an explainability process is applied to provide local explanations for the model predictions. In this work, we consider two commonly used variants of SHAP, namely KernelSHAP for tabular data and GradientExplainer for images. The background on the two variants of SHAP is reported in Appendix A. We envision that the AI model generated in FL can be used by entities that do not have access to the FL clients' individual training sets. Obviously, these entities may also be clients that participated in the FL and that can have access only to their own local datasets. To simulate this context, we adopt the setup shown in Figure 1, where a unique test set is adopted for the evaluation of both the accuracy of the prediction and of the explanations.

We recall the three desiderata for the explanations, as introduced in Section 1:

- *Privacy*: the explainability process should not violate the privacy of data owners, which is a basic requirement of an FL setting;
- *Accuracy*: explanations obtained in the FL setting should faithfully approximate the explanations that would be obtained in a centralized setting (i.e., in the unrealistic case that the full training set is available at a unique location for the computation of the Shapley values).
- *Consistency*: different entities adopting the AI model generated in FL should obtain the same explanations for identical data instances.

SHAP requires three elements to explain the model prediction $f(\mathbf{x}_i)$: the instance $\mathbf{x}_i$ itself, the model f, and a background dataset representative of the distribution used

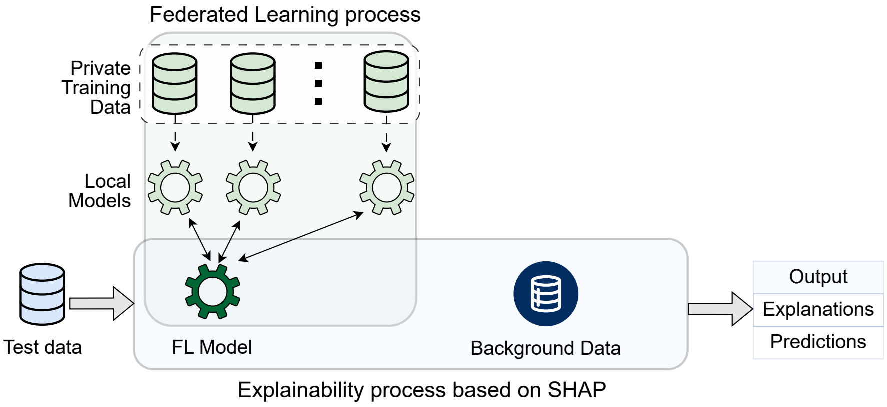


Fig. 1: High-level schema of the accuracy and explanations evaluation setup.

to train the model f. Since we assumed, for data privacy reasons, that the independent entity that deploys SHAP does not have access to the training sets of the clients, a major challenge is represented by the creation of a background dataset that simultaneously allows the accuracy and consistency of the explanations. In addition, the size of the background dataset should be carefully considered to minimize computation time, and its generation process should not introduce significant computational overhead. In the following, we describe the approaches to generate suitable background datasets for tabular (*FedSHAP-FCM*) and image datasets (*FedSHAP-GAN*).

Throughout the FL and explainability process, we consider a privacy model, typically adopted in horizontal FL [10, 18], with honest data owners and *semi-honest* (or *honest-but-curious*) central server: the server may attempt to infer private raw data from the updates provided by data owners, while still adhering to the protocol of the ML algorithm.

3.2 FedSHAP-FCM for tabular data

For tabular datasets, we adopt a strategy that we originally introduced in [9], where the background dataset required by SHAP is constructed through a federated clustering process. In particular, we leverage the FedFCM algorithm [10], which enables the collaborative extraction of cluster centroids across multiple clients holding horizontally partitioned data. In a nutshell, the FedFCM process follows an iterative scheme similar to classical FCM clustering algorithm. After the initial exchange of configuration parameters and cluster centers, each round proceeds as follows, until convergence: clients compute local statistics based on the current cluster centroids and transmit only such aggregated statistics to the server. The server uses such information to update the global cluster centroids and evaluate the stopping condition. The output of FedFCM is the set of K centroids, which act as a compact and privacy-preserving summary of the distributed data. These centroids are then used to compose the SHAP background dataset. Details on FedFCM are provided in Appendix C.

We formally denote the *FedSHAP-FCM* process as follows:

$$\textit{FedSHAP-FCM}_K \leftarrow \text{FedFCM}_K\left(\left(\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^M\right)\right) \tag{1}$$

An overview of the *FedSHAP-FCM* strategy is provided in Figure 2.

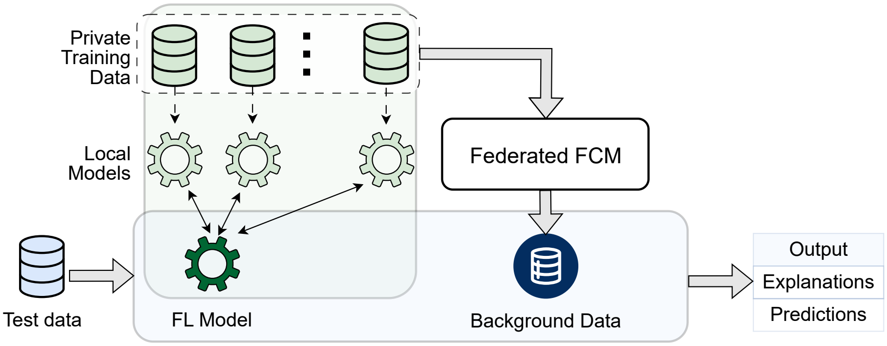


Fig. 2: *FedSHAP-FCM* scheme for tabular datasets.

In our approach, clustering is not employed to discover semantically coherent or well-separated groups, as often pursued in unsupervised learning, but rather as a means of numerosity reduction. The goal is to reduce the number of instances in the dataset while preserving its global structure and statistical representativeness. This is particularly suitable for SHAP, which benefits from having a manageable set of background instances that approximate the training data distribution.

Furthermore, the FedFCM algorithm ensures that data privacy is preserved throughout the process. As detailed in [10], clients share only aggregated statistics with the server, which does not have access to raw data and cannot reconstruct individual samples from the received information. This makes the method well-suited for privacy-sensitive federated environments. It is worth noting that the specific implementation of the clustering algorithm is not critical to our objective: alternative recently proposed approaches [37, 38] could be employed without affecting the core methodology. Similarly, the integration of additional privacy-enhancing techniques, such as cryptographic protocols and secure multiparty computation, remain conceptually orthogonal to the methodology, and can be accommodated without modifying its core logic.

3.3 FedSHAP-GAN for image data

For image datasets, the strategy based on FedFCM is not applicable. Partitioning-based clustering algorithms are known to be ineffective when applied to high-dimensional data such as images, where centroids may fail to capture the underlying structure of the data distribution.

A naïve alternative could be to construct background images by applying perturbations (e.g., noise, blurring, or pixelation) to the original images. However, unlike tabular data, where features are often independent, image data is inherently structured: visual information depends on the spatial arrangement and correlation of pixels. As a result, even perturbed images may preserve recognizable patterns, posing a risk of privacy leakage. Increasing the perturbation magnitude may mitigate this risk but would inevitably degrade the quality of explanations due to the resulting loss of semantic information.

To overcome these limitations, we propose *FedSHAP-GAN*, a novel approach to generate privacy-preserving and representative background datasets using GANs trained in a federated fashion (FedGAN).

The method, illustrated in Figure 3, consists of two main steps:

1. The M participating clients collaboratively train a GAN using an FL scheme. At each round, the server sends the current model to all clients, which update the generator and discriminator locally on their private data. Only model updates are sent back to the server, which aggregates them to refine the global model.
2. After training converges, the final version of the global generator is used to produce S synthetic images from random noise inputs. These images approximate the overall data distribution and serve as the SHAP background dataset.

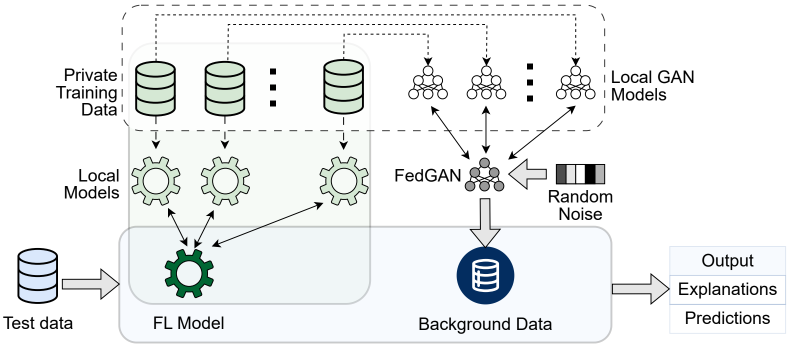


Fig. 3: FedSHAP-GAN scheme for image datasets.

We formally denote this process as:

$$\textit{FedSHAP-GAN}_S \leftarrow \text{Fed-GAN}_S\left((\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^M)\right) \tag{2}$$

This strategy offers two key advantages: it ensures that the background data distribution reflects the entire federated training set, and it fully respects privacy constraints, since no raw data is shared among clients.

It is worth noting that privacy-enhancing techniques, such as cryptographic methods, secure aggregation, and differential privacy, can be applied within the FedGAN framework to further reduce the risk of information leakage. However, these techniques are complementary to the proposed methodology and can be integrated without affecting its core functionality. A detailed analysis of their impact on model performance and computational overhead is beyond the scope of this paper.

Finally, we acknowledge that training a GAN in a federated setting can be computationally and communication-intensive. To assess the trade-off between training effort and the quality of the generated explanations, Appendix E provides a comparative analysis of different training configurations, using the MNIST dataset as a representative example, with 33%, 50%, and 100% of the local training data. For clarity, the main results presented in this paper refer to the setting with full local data availability.

4 Experimental Setup

Our Federated SHAP approaches are compared to several baselines for the generation of the background dataset. We first describe such baseline approaches and highlight their limitations in terms of consistency, accuracy, and privacy. Then, we detail the datasets exploited as test cases and outline the configuration of the approaches. Finally, we discuss the computational complexity of the background generation strategies.

4.1 Baseline approaches for the generation of a background dataset

The four baseline strategies for background generation are referred to as *Centralized*, *Random*, *LDP* (i.e., Local Differential Privacy), and *Local.* Unlike Federated SHAP, they do not simultaneously meet all three requirements of accuracy, consistency, and privacy preservation, nor do they have the same level of applicability across different types of data.

Centralized baseline

The *centralized* strategy creates a background dataset as the union of the local training sets.

$$\textit{Centralized} \leftarrow \bigcup_{m=1}^{M} \mathbf{X}^m \tag{3}$$

This is the approach commonly exploited in a centralized setup. Since all available training data are exploited, the requirements of accuracy and consistency are met. However, in an FL setting, this strategy is unfeasible because it involves the sharing of private raw data, thus violating privacy. Obviously, the size of the background dataset is that of the union of the local training sets.

Random baseline

The background dataset is created by randomly sampling K synthetic instances from a uniform distribution over the input space, which is assumed to be known a-priori. Notice that this assumption is reasonable in many real-world applications.

$$\textit{Random}_K \leftarrow \textit{Sample}_K\left(\mathcal{U}\left(\mathbf{a},\, \mathbf{b}\right)\right) \tag{4}$$

where **a** and **b** are the vectors of lower and upper bounds of the input features, respectively. In the case of image datasets, each pixel corresponds to a feature, and its range is determined based on the values observed in the training set. The strategy does not involve the sharing of private raw data and can be safely adopted in an FL setting. Different clients can refer to the same synthetic background dataset, and this ensures consistency of explanations. However, a background dataset composed of randomly generated instances may fail to capture the underlying distribution of the original data, potentially compromising the accuracy and reliability of the resulting explanations.

Local Differential Privacy (LDP) baseline

Differential privacy [39] is extensively used in the context of FL to provide formal privacy guarantees throughout collaborative model training. In our Fed-XAI system, however, the privacy issue is not limited to the FL of the AI model but also concerns the generation of the background dataset for the application of SHAP in a federated scenario. For this purpose, we introduce a novel baseline based on Local Differential Privacy (*LDP*), which is considered a state-of-the-art approach for privacy-preserving data collection and distribution [40]. Authors in [11] have recently explored the use of data masking techniques, including LDP, in the context of TreeSHAP, which is a variant of SHAP specifically designed for tree-based ML models. However, the scenario examined by the authors pertains to a centralized setting and therefore falls outside the scope of FL. Furthermore, their approach involves deriving both the model and the explanations from perturbed data.

Our LDP strategy shares a similarity with the approach proposed in [11], as we also use LDP in the form of additive noise for data masking, enabling post-hoc explainability. In our *LDP* strategy, the background dataset is obtained as follows: first, Laplace noise is added to local training data; then, the background dataset is generated as the union of the perturbed local training sets.

$$LDP(\epsilon, s) \leftarrow \bigcup_{m=1}^{M} \left(\mathbf{X}^m + Laplace(\mu = 0, \lambda = s/\epsilon)\right) \tag{5}$$

The shape of the Laplace distribution depends on the scale parameter $\lambda = s/\epsilon$, which in turn depends on the sensitivity s and the privacy factor ϵ. The sensitivity of the data is typically set as the maximum possible change for an instance in the dataset, whereas the privacy factor controls the amount of noise: a high ϵ corresponds to low noise, and vice versa. Thus, the *LDP* strategy ensures consistency (because only one background dataset is used), and it offers a trade-off between accuracy and privacy depending on the value of the privacy factor ϵ.

As underlined in Section 3, the *LDP* approach is not suitable for image data. In fact, due to the inherent spatial correlations between pixels, simple perturbations such as noise, blurring, or pixellation do not adequately preserve privacy, as the global structure and coarse-grained visual patterns may still be identifiable.

Local baseline

To avoid any privacy violation at explanation time, each entity could use its own local data as a background dataset to produce the explanation for a novel instance. In our experimental analysis, we suppose that the entities involved in making predictions and explanations are the clients involved in the federated learning; the local background dataset for each of the m clients is simply the local training set:

$$Local^m \leftarrow \mathbf{X}^m \tag{6}$$

This strategy does not violate the privacy constraint. However, this baseline does not ensure the consistency of the explanations, since the explanations obtained by different clients on identical input instances can differ due to their different background datasets. The differences can be particularly relevant in non-IID scenarios. As an obvious consequence, explanations obtained with this strategy may not be accurate.

4.2 Datasets and data distribution scenarios

To evaluate the proposed approach across different tasks and data modalities, we selected a diverse set of publicly available datasets, including binary and multiclass classification problems as well as regression tasks. Specifically, we considered three datasets for binary classification (Phoneme, Magic, Rice [41]), two for multiclass image classification (MNIST [42] and CIFAR-10 [43]), and three for regression tasks (PowerPlant, Concrete, Abalone [41]).

Table 1 summarizes the datasets in terms of *data type* (**Type**), namely tabular (TB) or image (IM), *task type* (**Task**), namely binary classification (BC), multiclass image classification (MC), or regression (RG), *number of training* ($\mathbf{N}_{train}$) and *test* ($\mathbf{N}_{test}$) *instances* used in our experiments and *number of features* (**F**). A more detailed description of the dataset metadata is provided in Appendix B.

Table 1: Summary of the datasets used in the experiments.

Dataset	**Type**	**Task**	**Source**	$\mathbf{N}_{train}$	$\mathbf{N}_{test}$	**F**
Phoneme (PH)	TB	BC	[44]	4863	541	5
Magic (MA)	TB	BC	[41]	17118	1902	7
Rice (RI)	TB	BC	[41]	3429	381	7
PowerPlant (PP)	TB	RG	[41]	8611	957	4
Concrete (CC)	TB	RG	[41]	927	103	8
Abalone (AB)	TB	RG	[41]	3759	418	7
MNIST (MN)	IM	MC	[42]	60000	10000	784
CIFAR-10 (CF)	IM	MC	[43]	50000	10000	3072

Each dataset (except MNIST and CIFAR-10) was randomly split into training and test sets with a 90%-10% ratio, preserving the original distribution of the target variable. For MNIST and CIFAR-10, we adopted the canonical train/test split commonly used in the literature. Additionally, to reduce the computational burden in the explanation phase, we selected a stratified random subset of 1,000 test instances from MNIST and from CIFAR-10 for SHAP evaluation.

To simulate a federated learning scenario, each training set was further divided into 10 disjoint subsets, each assigned to a simulated client. The partitioning followed a non-IID scheme to reflect realistic heterogeneity in FL settings [45]. In particular, we introduced the following types of heterogeneity:

- Quantity skewness: clients are assigned training sets of varying sizes;
- Label distribution skewness: the distribution of the target variable differs across clients;
- Feature distribution skewness: clients are sorted by increasing values of a selected feature.

The resulting client-level distributions are illustrated in Figures B.1 and B.2 in Appendix B.

All features were scaled in the [0, 1] range via min-max normalization. This preprocessing step is compatible with federated settings, under the reasonable assumption that feature ranges (or their estimates) are known to the server in advance.

4.3 ML models and FL algorithms configuration

We apply SHAP to explain the prediction of different opaque models, suitably designed for different tasks and different data types. Since our main purpose is to assess the explainability process in FL setting, we did not carry out a thorough optimization of the models hyperparameters; we simply ensured that the selected configuration achieved reasonable predictive performance. The main characteristic of the models are presented in the following, while additional details are reported in Appendix D.

4.3.1 Models for tabular datasets

For tabular datasets, we employed a Multi-Layer Perceptron (MLP) as a representative example of an opaque model. The network architecture comprises two hidden layers, each with 128 neurons and ReLU activation, followed by an output layer. The latter uses a sigmoid activation for classification tasks and remains linear for regression. Overall, the model includes 17,793 trainable parameters.

Training is performed using the Adam optimizer with a learning rate of 0.01. Binary Cross-Entropy and Mean Squared Error (MSE) are used as loss functions for classification and regression tasks, respectively.

In the FL setup, we adopt the classical Federated Averaging (FedAVG) algorithm for model aggregation [2]. Training is performed with a minibatch size of 64, 5 local epochs per client, and a total of 20 federation rounds for classification tasks and 60 for regression tasks.

4.3.2 Model for image datasets

For the multiclass classification task with the image datasets, we adopted a CNN architecture and federated training setup inspired by the work of Khuu et al. [46].

In the case of MNIST dataset, the CNN consists of two convolutional layers followed by max pooling, a flattening stage, a dense layer with 128 units including dropout, and an output layer with 10 units using softmax activation function for classification. The total number of trainable parameters is 421,642. The loss function used is CrossEntropyLoss, while model optimization was performed using the Adam algorithm with a learning rate of 0.001. The FedAvg algorithm was employed as an aggregation strategy in FL. Local updates were carried out on batches of size 64 for 3 epochs, followed by weight aggregation to obtain the global model. The validation set constituted 20% of the total training data. The training process employed an automated validation mechanism to mitigate overfitting, based on a global validation loss computed as the average of the local validation losses.

In the case of CIFAR-10 dataset, we adopted the architecture of the ResNet model, often used to perform experiments on that dataset (as in [47]). In particular, we used the ResNet-20 variant of ResNet [48] designed for low-resolution images (32×32 pixels in color), such as those in the CIFAR-10 dataset. The network architecture comprises an initial convolutional layer followed by three groups of residual blocks. Each residual block contains two 3×3 convolutional layers, each followed by batch normalization and ReLU activation, with a shortcut connection. The three groups of blocks operate at progressively decreasing spatial resolutions: the first group processes feature maps at 32×32, the second at 16×16 (with downsampling performed in the first block via a stride of 2), and the third at 8×8 (also with downsampling applied in the first block). After completing the convolutional blocks, the network applies a global Average Pooling that reduces each activation map to a single value, followed by a fully connected layer that produces the final distribution over the 10 CIFAR-10 classes via a softmax activation function.

The total number of trainable parameters is 0.27 millions. The loss function used is CrossEntropyLoss, while model optimization was performed using the Adam algorithm with a learning rate of 0.001. The FedAvg algorithm was employed as an aggregation strategy in FL. Local updates were carried out on batches of size 32 for 10 epochs, followed by weight aggregation to obtain the global model. The validation set constituted 20% of the total training data. The training process employed an automated validation mechanism to mitigate overfitting, based on a global validation loss computed as the average of the local validation losses.

4.4 Configuration of the approaches for background generation

The most influential parameter in *FedSHAP-FCM* is the number of clusters K provided as input to the partitioning clustering algorithm, as it also represents the size of the resulting background dataset. We set $K = 50$ for all datasets to obtain a small yet representative background dataset. Although larger background datasets typically yield more robust approximations of Shapley values [49], we intentionally use a smaller background dataset, smaller even than those used in baseline approaches, to show that, even under this constraint, Federated SHAP can achieve competitive performance while preserving computational efficiency.

For the *LDP* strategy, we investigate the trade-off between privacy and accuracy of the explanations by evaluating the achieved results for two values of the ϵ parameter.

Specifically, we created two background datasets considering all the instances of the local training set: one with a high privacy factor $\epsilon = 1$ (*LDPe1*, high noise), and one with a low privacy factor $\epsilon = 10$ (*LDPe10*, low noise). In appendix G we have also reported the results of experiments conducted with an intermediate $\epsilon = 5$. They are not reported here for the sake of brevity. The sensitivity parameter is set to $s = 1$, in accordance with the fact that each feature lies within the range [0, 1] as a result of MinMax normalization.

As regards the experiments with the *FedSHAP-GAN* with MNIST dataset, we adopted the GAN architecture proposed in [34], which was applied in a similar case study on MNIST dataset in a FL scenario. As regards the aggregation strategy and the loss function, we took inspiration from the recent work on FL for GAN, discussed in [36], in which the MNIST dataset was used by the authors in the experimental analysis. Specifically, we employed the classical FedAVG algorithm with full parameter averaging. For the loss function, we adopted the standard minimax formulation of GANs: the discriminator is trained to maximize its ability to distinguish real samples from generated ones, while the generator is optimized to minimize the probability of the discriminator correctly identifying its outputs as fake. In our implementation, this objective is realized using the binary cross-entropy loss. Both the generator and discriminator are trained using the Adam optimizer with fixed hyperparameters.

Details of the GAN generator module are provided in Table D.1 and D.2 in Appendix D. The *FedSHAP-GAN* approach requires several configuration parameters. The first group of parameters pertains to the FL process of the GAN model. Each training session lasts 200 (600) rounds for MNIST (CIFAR), with early stopping implemented to automatically halt the training when necessary. Weight aggregation occurs after each local training epoch. The batch size is set to 64 (128). The clients use the Adam optimizer for the local model training, with a learning rate of 0.0002 (0.0002) for the generator and 0.0001 (0.0002) for the discriminator. The exponential decay rate for the first-moment estimate is 0.5 (0.5), and for the second-moment estimate is 0.999 (0.999). The generator was fed a noise vector sampled from a Gaussian distribution with zero mean and unit standard deviation. The size of the noise vector is 100 (100) and corresponds to the generator's latent space dimensionality.

An example of generated synthetic images is shown in Fig. 4 for the MNIST dataset and in Fig. 5 for the CIFAR-10 dataset, and is an indication of the ability of the network to generate realistic images.

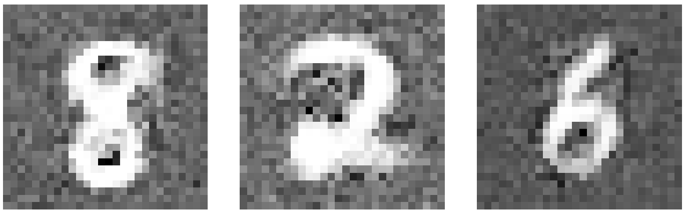


Fig. 4: Examples of MNIST synthetic images generated by the FedGAN.

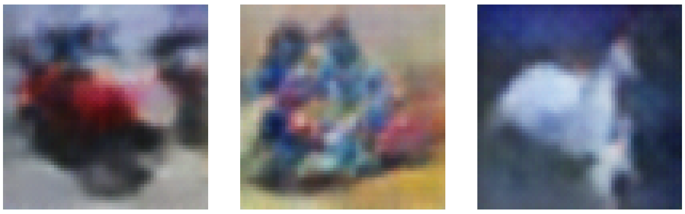


Fig. 5: Examples of CIFAR-10 synthetic images generated by the FedGAN.

Scenario	**Dataset**	**Generated images**	**FID**
Centralized GAN	MNIST	6000	158.38 ± 0.39
FedGAN	MNIST	6000	204.61 ± 0.83
Centralized GAN	CIFAR-10	5000	362.74 ± 0.60
FedGAN	CIFAR-10	5000	447.82 ± 0.58

Table 2: Fréchet Inception Distance (FID) values calculated with respect to the original images (60,000 images for MNIST and 50,000 images for CIFAR-10).

Another influential parameter of *FedSHAP-GAN* is the size of the background dataset, i.e., the number S of synthetic images generated by the FedGAN. We set $S = 6000$ and $S = 5000$, namely the 10% of the entire training set, for MNIST and CIFAR-10, respectively. We have experimentally verified that such a background dataset still allows us to generate accurate explanations within a reasonable time (i.e., a few seconds).

Since the quality of the generated images can impact the final results of the proposed methodology, we calculated the Fréchet Inception Distance (FID), introduced in [50], of the images generated with different settings with respect to the original dataset to quantitatively evaluate the similarity. Table 2 reports the FID values for the different scenarios, repeated for ten different random seeds to ensure variability. The FID values calculated by comparing images generated using our proposed *FedSHAP-GAN* with those generated by a centralized version of the GAN (where all images are available to a central server) are of the same order of magnitude. As expected, the centralized approach yields better results. In addition, we also evaluated the influence of the data distribution in the local training datasets comparing our Non-IID scenario with an IID scenario (which is typically unrealistic in FL settings). The experimental results show that image background generated through the GAN trained under the non-realistic IID conditions yield only marginally improved explanation accuracy compared to the performance observed under realistic Non-IID experimental conditions. For the sake of brevity, we report the results in appendix F.

As regards *Centralized* and *Random* baselines, we recall that the size of the background dataset is equal to the number of the training instances, as shown in Table 1. Finally, for the Local baseline, the size of each local background dataset is equal to the size of the local training sets, as shown in Figures B.1 and B.2 in the Appendix B.

Notably, several strategies for the generation of the background dataset incorporate stochastic elements: *Random* involves sampling from uniform distributions. *FedSHAP-FCM*$_k$ relies on a clustering algorithm that depends on the random initialization of the centroids; *FedSHAP-GAN*$_S$ entails multiple sources of randomness, including latent noise instances from which image generation originates; *LDP* introduces random perturbations drawn from a Laplace distribution. Therefore, we performed 10 repetitions of each of these methods with different seeds, in order to capture the variability introduced by their stochastic components and assess its impact on the explanations generated by SHAP.

4.4.1 Computational complexity of the approaches

The effort required to calculate the Shapley values using the SHAP method is a well-known challenge. For this reason, in the explanation phase we used for both the tabular and the image datasets a high-performance server equipped with two 56-core Intel® Xeon® Platinum 8480+ CPUs, 2 TB of RAM, and an NVIDIA A100-SXM4 GPU with 80 GB of dedicated VRAM.

The computational complexity of KernelSHAP, used for tabular datasets, grows exponentially with the number of features and linearly with the size of the background dataset. Evidently, the latter is the only one we can tune in our experiments to control the complexity at the explanation stage. At the explanation stage, *FedSHAP-FCM* is the most efficient approach, as it operates with the smallest background dataset compared to the other methods. We deliberately extend the background dataset of the baseline approaches to match the size of the training sets: this comes at a cost, in terms of efficiency, but allows them to leverage a larger set of samples, which typically leads to more stable and accurate approximations of the Shapley values. Thus, in this way, the baseline may provide explanations with the highest possible level of accuracy.

GradientExplainer, the SHAP variant used for the image dataset, also has a time complexity that grows linearly with the number of background instances. For the *Centralized* approach, we used all available instances to establish the baseline for the accuracy evaluation. Conversely, for the *Random* and *FedSHAP-GAN* approaches, we reduced the background size to 10% of the available instances, significantly lowering the computational cost compared to the *Centralized* approach, while ensuring that both methods operate under the same computational constraints. The evaluation of explainability results on 1,000 MNIST and CIFAR-10 test images is the most computationally demanding part of our experiments. For MNIST, computing the explanation for a single instance using the full centralized background of 60,000 images required 75 seconds. Employing a background dataset of 6,000 images (as in *FedSHAP-GAN*) reduced the computation time to 7 seconds. Regarding CIFAR-10, we observed similar execution times: computing the explanation for a single instance using the full centralized background of 50,000 images required 98 seconds, which was reduced to 10 seconds when employing a background dataset of 5,000 images.

Another aspect to be considered is the computational effort required to generate the background dataset. Unlike baseline approaches, which simply rely on collecting raw data or computing random perturbations, the proposed Federated SHAP approaches introduce an additional overhead for the background generation itself.

FedSHAP-FCM entails the execution of the FedFCM clustering algorithm, whereas *FedSHAP-GAN* requires training a GAN in the federated setting and exploiting it for generating images. The overhead is incurred just once, during the background dataset generation stage, and does not affect the efficiency at explanation time. In our experiments, we measured that FedFCM runs in a few seconds, even on the most demanding dataset (Magic). For the sake of completeness, in appendix C we report the average and standard deviations of the execution times for all the datasets over 10 experiments. Training the GAN takes about 10 (40) minutes for MNIST (CIFAR-10), while generating 6,000 (5,000) images takes less than 10 (13) seconds for MNIST (CIFAR-10). These results suggest that, despite the additional computation required for background dataset generation in Federated SHAP approaches, the overhead remains largely manageable in practice.

5 Experimental Results

This section presents the experimental results, covering three complementary aspects. First, we assess the predictive performance of the opaque models trained in the FL setting, comparing them to their counterparts trained in a centralized configuration. This comparison serves to validate the FL paradigm as a viable learning strategy in our context. Second, we evaluate the accuracy of the explanations produced by the proposed Federated SHAP methods, benchmarking them against alternative background generation strategies introduced in Section 4.1. Finally, we examine the consistency of the explanations across clients, highlighting how the different strategies affect the coherence of local attributions in a federated setting.

5.1 Model performance evaluation

Before evaluating the quality of the explanations provided by our proposed Federated SHAP method, we assess the predictive performance of the underlying models trained in the FL setting.

For each dataset, we report test set performance using standard metrics: accuracy and F1-score for classification tasks, and R^2 and RMSE for regression tasks. These results reflect the predictive reliability of the models upon which the explanations are applied.

To provide a reference baseline, we also include the performance of models trained in a Centralized Learning (CL) setting, in which all client data are aggregated and training occurs on a single, unified dataset. Although this setting breaches privacy constraints and is not feasible in federated contexts, it provides a valuable upper bound for assessing the competitiveness of FL-based training.

Table 3 summarizes the performance results across all datasets. As expected, CL slightly outperforms FL in most cases, due to the availability of the entire dataset and the absence of client heterogeneity. Nonetheless, the federated models achieve comparable performance across both classification and regression tasks, supporting the feasibility of applying post-hoc explainability techniques within a privacy-preserving training framework.

Table 3: Performance metrics on the test set.

	FL	CL	FL	CL
	Accuracy		**F1-score (class 1)**	
Phoneme	0.79	0.85	0.70	0.78
Magic	0.86	0.89	0.79	0.83
Rice	0.89	0.91	0.87	0.89
	Accuracy		**F1-score (weighted)**	
MNIST	0.99	0.99	0.99	0.99
CIFAR-10	0.91	0.92	0.91	0.92
	RMSE		**R^2**	
PowerPlant	5.05	3.86	0.91	0.94
Concrete	10.56	4.01	0.61	0.94
Abalone	2.09	2.01	0.60	0.63

As expected, the models trained according to the CL paradigm slightly outperform their FL counterparts. This is reasonable since, in the CL setting, the full training set is assumed to be available to a single entity for the learning phase. We assume that the small gap in performance does not undermine the considerations about model explainability reported hereafter.

5.2 Explainability analysis

The post-hoc SHAP method explains AI model decisions by assigning Shapley values to features, quantifying their impact on each individual prediction. It is worth noting that in the following analysis, we consider the explanations provided with respect to the model generated in the FL scenario. To evaluate the soundness of the different background generation strategies, we calculate for each strategy s the Shapley values for all the instances in the test set. In this way, we obtain a matrix $\boldsymbol{\Phi}_s$ with shape $N_{test} \times F$, where F is the number of features. To assess the accuracy of a strategy, we calculate the discrepancy with respect to the $\boldsymbol{\Phi}_{Centralized}$ matrix, which is obtained using the full training set as background dataset. As discussed in [9], the discrepancy is quantitatively assessed by calculating the Frobenius norm of the pairwise difference between the generic $\boldsymbol{\Phi}_s$ and the $\boldsymbol{\Phi}_{Centralized}$ matrices. The Frobenius norm of a matrix $\boldsymbol{A}$ with r rows and c columns is defined as the square root of the sum of the absolute squares of its elements:

$$\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{\sum_{i}^{r}\sum_{j}^{c} |a_{ij}|^2} \tag{7}$$

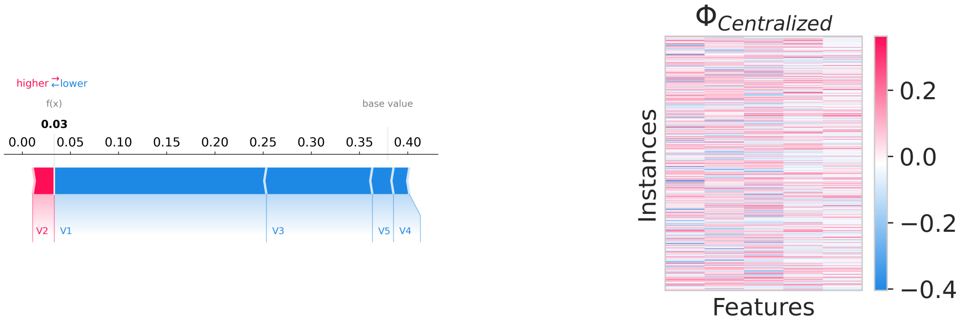
where a_{ij} is the element of $\boldsymbol{A}$ in the ith row and jth column, with $i = 1, \ldots, r$ and $j = 1, \ldots, c$.

The accuracy of a strategy is thus assessed as the discrepancy given by $\|\mathbf{\Phi}_s - \mathbf{\Phi}_{Centralized}\|_F$: obviously, the higher the discrepancy, the lower the accuracy.

Hereafter, we report a separate discussion for the tabular and image data, because of the differences in the strategy of the background generation and in the visualization of the explanations.

5.2.1 Explanations for tabular datasets

To illustrate the process of obtaining numerical results with KernelSHAP on tabular datasets, we take the PH dataset as a representative case. We used the `KernelExplainer` class from the SHAP Python library[2]. Figure 6 provides a visual representation of the results.



(a) Shapley values visualized using the SHAP force plot for an instance of the PH test set (#282), estimated via the *Centralized* approach.

(b) $\mathbf{\Phi}_{Centralized}$ matrix: Shapley values for the PH test set, estimated via the *Centralized* approach.

Fig. 6: Comparison of Shapley value visualizations for the PH test set using the *Centralized* approach.

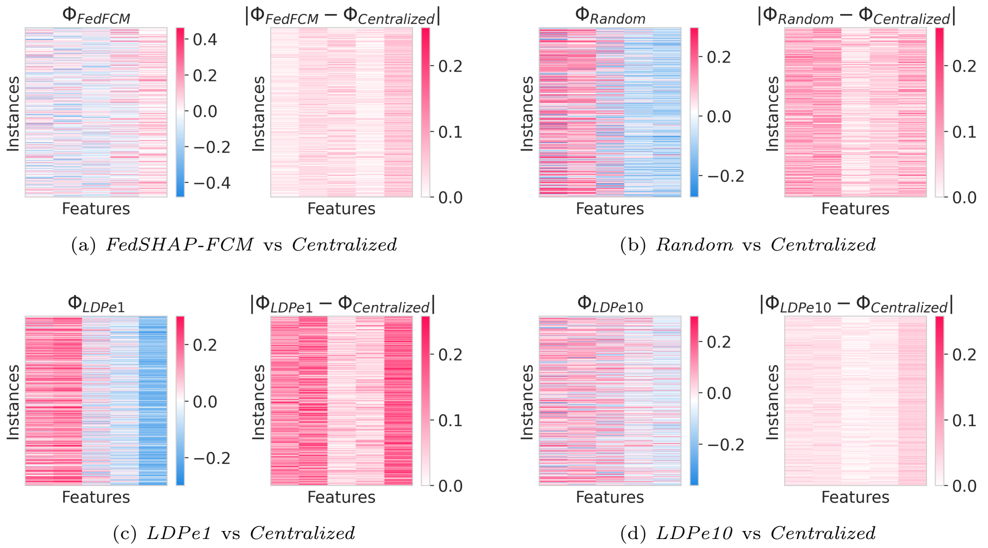
Figure 6a reports an example of data visualization for the Shapley values for an individual decision randomly extracted from the test set (#id 282). For this example, the features $V1, V3, V4$, and $V5$ have a negative impact on the final decision, while the feature $V2$ contributes positively. In detail, the model classifies the instance as class=0, with the Shapley values equal to $\phi_{V1} = -0.22$, $\phi_{V2} = +0.02$, $\phi_{V3} = -0.11$, $\phi_{V4} = -0.02$ and $\phi_{V5} = -0.02$ and finally $\phi_0 = 0.38$.

The same procedure can be applied to all test instances to obtain the $\mathbf{\Phi}$ matrix that summarizes global insights of the explanations. Fig. 6b shows the $\mathbf{\Phi}_{Centralized}$ matrix.

Fig. 7 shows the $\mathbf{\Phi}$ matrices used to calculate the accuracy of all strategies. Figs. 7a, 7b, 7c, 7d show the results for one of the ten different trials obtained by changing the random seed for the *FedSHAP-FCM*, *Random*, *LDP* with $\epsilon = 1$, and *LDP* with $\epsilon = 10$, respectively. In the case of the *Local* strategy, the loss of accuracy is a consequence of inconsistency of explanations. If the explanations obtained from the various clients are

[2] https://shap.readthedocs.io/en/latest/generated/shap.KernelExplainer.html, visited April 2025

different from each other, they cannot coincide with the explanation obtained in the centralized case. For this reason, we do not report the numerical results regarding the local approach here, but rather discuss it in the next section regarding the consistency of explanations.



(a) *FedSHAP-FCM* vs *Centralized*

(b) *Random* vs *Centralized*

(c) *LDPe1* vs *Centralized*

(d) *LDPe10* vs *Centralized*

Fig. 7: Explainability analysis on PH dataset: Shapley values obtained with *FedSHAP-FCM*, *Random* and *LDP* strategies and comparison with *Centralized*.

Taking *FedSHAP-FCM* and *Random* strategies as an example, the evaluation of the Frobenius norm on the resulting matrices leads to the following results: $\|\mathbf{\Phi}_{FedSHAP-FCM} - \mathbf{\Phi}_{Centralized}\|_{\mathrm{F}} = 2.23$ and $\|\mathbf{\Phi}_{Random} - \mathbf{\Phi}_{Centralized}\|_{\mathrm{F}} = 4.31$. This suggests that the approach based on FedFCM is more accurate than the *Random* one, in this particular experiment. Subsection 5.3 presents the full numerical results, also taking into account the 10 repetitions performed for each strategy with different random seeds.

The evaluation of numerical results, however, cannot disregard the understanding of the impact that such discrepancies have on the explanations provided to the end user or the involved stakeholders. To gain a first intuitive insight about this aspect, we analyze the explanations, i.e., the Shapley values, provided by each approach for a single instance. The instance considered represents the worst-case scenario for Federated SHAP, i.e., the one where the distance between the Shapley values obtained by *FedSHAP-FCM* and those obtained by *Centralized* is the greatest. Explanations are visually depicted in Fig. 8.

We can observe how the explanations provided by *FedSHAP-FCM* and *Centralized* are substantially in agreement, even in the considered worst case: for both strategies, the most impactful feature is $V2$, and in general, the signs of the Shapley values

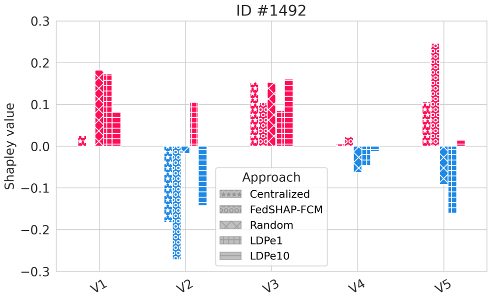


Fig. 8: Shapley values obtained with *Centralized*, *FedSHAP-FCM*, *Random*, *LDPe1*, and *LDPe10* approaches on the instance #1492 of PH dataset, for which the distance between the Shapley values obtained with *FedSHAP-FCM* and with *Centralized* is maximum.

are the same. Vice versa, this is not always true for the *Random* approach, where both signs and rank of feature importance are not always in agreement. A similar consideration applies to the *LDP* strategies, where as the privacy factor ϵ increases, the noise decreases and the explanations become more accurate. This analysis also confirms that the properties of the background dataset are critical to provide accurate and trustworthy explanations.

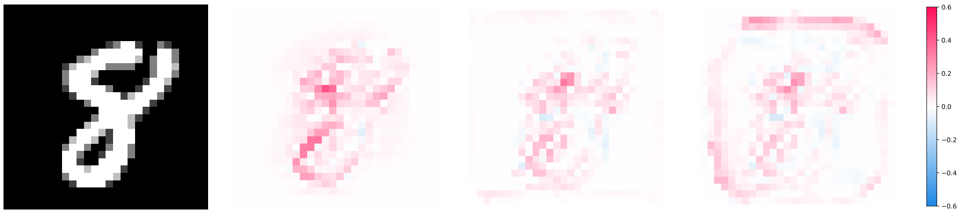
5.2.2 Explanations for the image datasets

For image data, explanations were generated using the `GradientExplainer` component of the SHAP library[3]. Each explanation is represented as a 28 × 28 heatmap for MNIST and a 32 × 32 heatmap for CIFAR-10, where pixel intensities indicate the Shapley value of the corresponding input feature. Red and blue tones indicate positive and negative contributions to the model's decision, respectively, with color saturation proportional to importance magnitude.

Considering MNIST, Fig. 9 highlights the qualitative differences in SHAP explanations produced with different background datasets for the same test instance. As expected, the explanation obtained with the *Centralized* background (Fig. 9c) provides a clear and coherent attribution: positive contributions (in red) align with the characteristic loops of the digit "8", while negative contributions (in blue) appear in less relevant peripheral regions. Notably, the explanation obtained with our proposed *FedSHAP-GAN* strategy (Fig. 9c) closely matches the centralized reference in both structure and polarity. The key discriminative areas are correctly identified, confirming that the synthetic background generated by the federated GAN is effective in approximating the true data distribution. In contrast, the explanation generated using a *Random* background (Fig. 9d) is considerably more diffuse and noisy, with scattered contributions and limited interpretability. This reinforces the importance of using realistic and representative background datasets for SHAP in federated contexts.

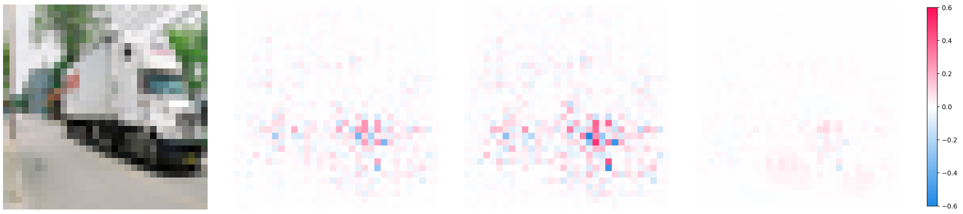
[3] https://shap.readthedocs.io/en/latest/generated/shap.GradientExplainer.html

We observe that, for the specific dataset, both the *FedSHAP-GAN* and *Random* strategies assign non-negligible SHAP values to the border pixels surrounding the digit, forming a visible frame. This artifact is completely absent in the *Centralized* explanation. Importantly, the border effect is significantly less pronounced when using our proposed *FedSHAP-GAN* strategy, while it appears markedly stronger and more spatially uniform in the *Random* case. This contrast suggests that the GAN is able to approximate the structural distribution of the original data to a much greater extent than random background generation, which fails to respect the semantics and layout of the input space. Indeed, GAN-generated images may occasionally introduce subtle artifacts near the image edges, whereas random images, being entirely unstructured, can result in systematic misalignment. These edge mismatches can lead the model to interpret the border pixels as informative, thereby distorting the attribution process. This analysis underscores the critical importance of ensuring that background datasets are both realistic and well-aligned with the input data distribution when applying SHAP in federated settings.



(a) (b) (c) (d)

Fig. 9: (a) Random image (#810) from the MNIST test set; SHAP explanation using: (b) *Centralized* approach, (c) *FedSHAP-GAN* approach, (d) *Random* approach.



(a) (b) (c) (d)

Fig. 10: (a) Random image (#810) from the CIFAR-10 test set; SHAP explanation using: (b) *Centralized* approach, (c) *FedSHAP-GAN* approach, (d) *Random* approach.

To support this qualitative insight with quantitative evidence, we computed the Shapley values for all test instances using each strategy. The resulting explanations were flattened and assembled into a matrix $\mathbf{\Phi}_s$ for each strategy s, having shape $N_{\text{test}} \times$

F (with $F = 784$ in the MNIST case). As for tabular datasets, explanation accuracy is assessed by measuring the Frobenius norm between each $\mathbf{\Phi}_s$ and the reference matrix $\mathbf{\Phi}_{Centralized}$.

Similar considerations can be made for the CIFAR-10 dataset. Fig. 10 illustrates how the different approaches provide different explanations when applied to a randomly selected image (ie, a truck). Notably, the explanation generated by using *FedSHAP-GAN* shows greater similarity to that obtained via the *Centralized* approach, compared to the explanations derived from the *Random* procedure. In particular, the intensity and spatial patterns of the Shapley values in Fig. 10c are much closer to the ones in Fig. 10b, suggesting that the same areas (likely, the truck's edges) are being identified as relevant. On the contrary, the explanation provided by Fig. 10d contains much lower and disperse Shapley values, indicating that the generated image may fail to capture key contribution regions. This suggests that the proposed federated strategy is more effective in preserving the underlying patterns. Also in this case, to provide a quantitative comparison, we computed the Shapley values across the test set for each approach s. Given the three-channel RGB structure of the CIFAR-10 colored images, the Shapley values are calculated independently for each color channel and then flattened as a unique Φ_s matrix. Finally, the Frobenius norms between each Φ_s and the reference $\Phi_{Centralized}$ matrix are used to compare the accuracies of explanations.

5.3 Accuracy analysis of the explanations

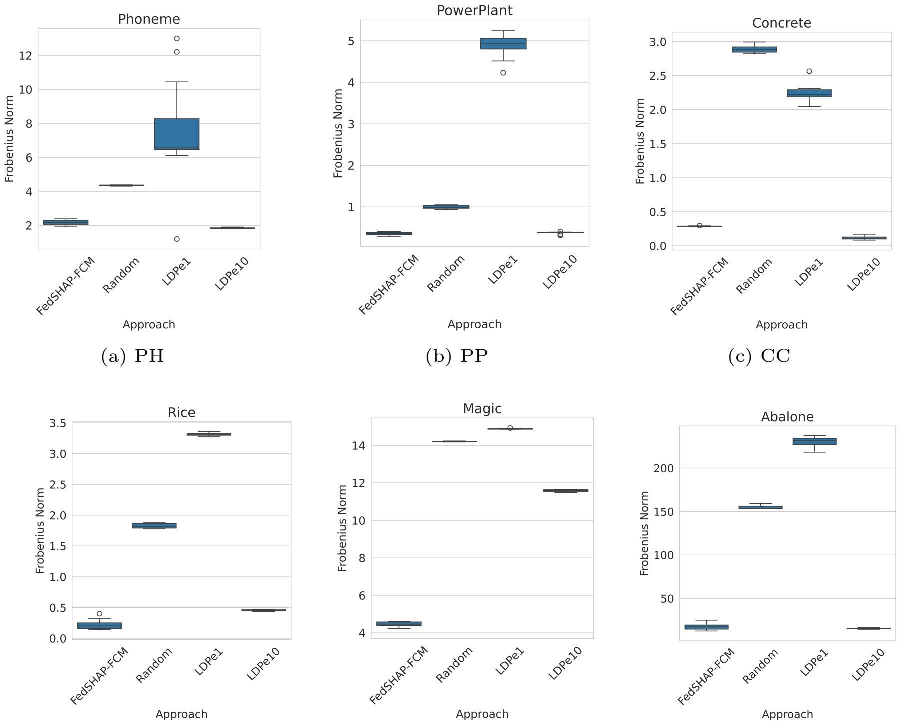
We report the numerical results of the accuracy of the explanations related to the different strategies for the generation of the background datasets, separately for tabular and image case studies.

5.3.1 Accuracy of explanations for the tabular datasets

Figure 11 reports, for each strategy s and for each dataset, the boxplots of the ten values of $\|\mathbf{\Phi}_s - \mathbf{\Phi}_{Centralized}\|_{\mathrm{F}}$, obtained with as many repetitions with different seeds.

The results confirm the suitability of the proposed Federated SHAP strategy for the tabular datasets taken into account. The boxplots illustrate that, in general, the explanations produced by *FedSHAP-FCM* are highly accurate, i.e., more closely aligned with those obtained using the *Centralized* baseline, compared to the other strategies (*Random* and *LDP*).

As expected, the privacy factor plays a key role for the *LDP* strategy. On the one hand, with a high privacy factor ($\epsilon = 1$, *LDPe1*) the additive noise makes the explanations completely inaccurate and often worse than in the *Random* case, where the background instances are synthetic examples randomly sampled from the domain. On the other hand, with a low privacy factor ($\epsilon = 10$, *LDPe10*), the magnitude of the perturbation is reduced, and the explanations are accurate and comparable to those obtained with *FedSHAP-FCM*. However, it is worth highlighting two key differences between *FedSHAP-FCM* and *LDP*. First, the background dataset in *FedSHAP-FCM* consists of only 50 elements, whereas in *LDP*, it matches the size of the training set,



(a) PH (b) PP (c) CC

(d) RI (e) MA (f) AB

Fig. 11: Boxplots of the discrepancy of the *FedSHAP-FCM*, *Random* and *LDP* approach with respect to the *Centralized* approach, in terms of Frobenius norm of the pairwise difference of $\mathbf{\Phi}$ matrices.

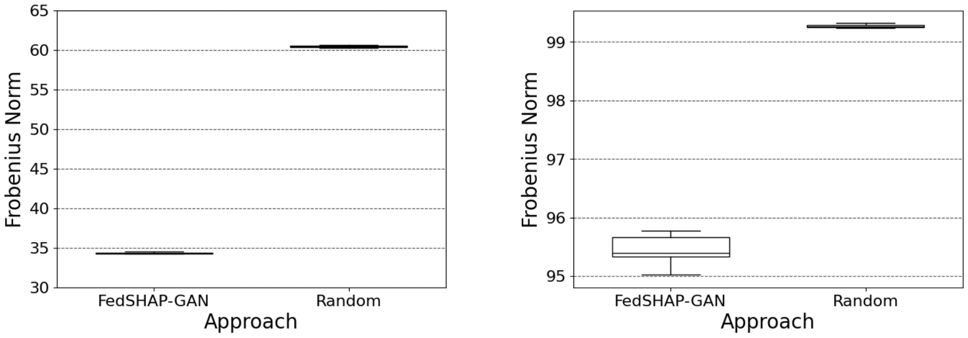
making it one or two orders of magnitude larger, depending on the dataset. This confirms that *FedSHAP-FCM* can provide accurate explanations while maintaining high computational efficiency. Second, *FedSHAP-FCM* utilizes a background composed of centroids rather than real perturbed instances, which can be considered beneficial in terms of privacy protection, especially with a low privacy factor as in *LDPe10*.

The variability induced by the seed is limited and never affects the ranking of the various strategies in terms of accuracy. The highest variability is observed for *LDPe1*, particularly on the PH and PP datasets.

5.3.2 Accuracy of explanations for the image datasets

In the case of image data, we evaluate the accuracy of the explanations obtained using our proposed *FedSHAP-GAN* strategy, which relies on a synthetic background generated via a federated GAN. As a baseline, we consider a *Random* strategy, where the background dataset consists of unstructured, noisy images.

Figure 12 reports, for each strategy, the distribution of the discrepancy scores $\|\mathbf{\Phi}_s - \mathbf{\Phi}_{Centralized}\|_{\mathrm{F}}$ over ten repetitions with different random seeds.



(a) MNIST

(b) CIFAR-10

Fig. 12: Boxplots of the discrepancy of the *FedSHAP-GAN* and *Random* approaches with respect to the centralized *Centralized* approach, measured as the Frobenius norm of the pairwise difference of $\mathbf{\Phi}$ matrices, for MNIST (12a) and CIFAR-10 (12b) case studies.

The results confirm that *FedSHAP-GAN* outperforms the *Random* baseline: its explanations are notably closer to those produced by the centralized model, highlighting the effectiveness of the proposed method in approximating the true data distribution without direct access to real training images.

While some discrepancy with the *Centralized* reference remains, likely due to the fully synthetic nature of the background, the explanations produced by FedSHAP-GAN are consistently more accurate than the ones generated by the *Random* strategy and stable across trials. This reinforces the validity of using generative models to support post-hoc explainability in privacy-preserving federated contexts.

To assess whether increasing the size of the synthetic background improves explanation quality, we conducted an additional experiment using the *FedSHAP-GAN* strategy with a background composed of 60,000 synthetic images for the MNIST test case, matching the size of the original training set. The average distance between the *Centralized* and the *FedSHAP-GAN* configuration with 6,000 images is equal to 34.36, and it closely matches the value of 34.25 observed for the setup of *FedSHAP-GAN* with 60,000 images. This confirms that using a background of moderate size, such as 6,000 synthetic images, is sufficient to generate reliable and faithful SHAP explanations, while significantly reducing memory usage and computational cost.

5.4 Analysis of the consistency of explanations

The proposed *FedSHAP-FCM* and *FedSHAP-GAN* strategies inherently meet the consistency requirement, because the background used to calculate the Shapley values is common and available to all clients. In other words, two distinct clients would obtain exactly the same explanation given two identical test instances. In fact, the cluster centers obtained with FedFCM and the synthetic images generated by the FedGAN can be shared with any client without violating data privacy.

In this section, we analyze the issue of consistency for the *Local* approach, where no common background is available, and each client applies the SHAP procedure using its own local training data as the background dataset. We will show that consistency is not ensured, especially for non-IID situations. In the following, we outline the issue of consistency of the local approaches considering individual examples of explanations for both tabular and image datasets.

5.4.1 Consistency Analysis for the Tabular Dataset

To assess the consistency of the explanations across clients, we analyze the Shapley values obtained for a single instance of the PH dataset (instance #0) using the *Local* approach, in which each client generates its own explanations based solely on its local training data as background dataset. For comparison, we also report the explanation obtained for the same instance using the proposed *FedSHAP-FCM* strategy.

Figure 13 shows the Shapley values for the five input features, with each color representing the output from a different client in the *Local* case, and the blue bars corresponding to the explanation obtained using the *FedSHAP-FCM* background dataset. As evident from the figure, the *Local* explanations exhibit a high degree of variability across clients. In some cases (e.g., features $V1$ and $V2$), different clients assign opposite signs to the same feature, leading to conflicting interpretations. This variability is expected due to the non-IID nature of local training data, which causes each client to operate over a different feature-label distribution. The resulting explanations, though technically valid for each local dataset, are inconsistent when viewed from a global perspective. In contrast, the explanation obtained with *FedSHAP-FCM* provides a unified and stable interpretation, as it is based on a federated summary of the full training distribution.

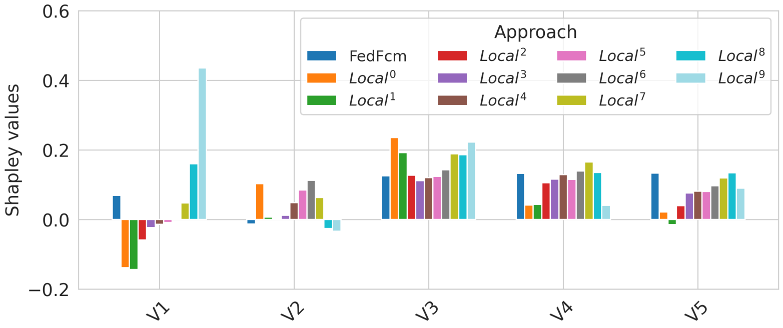


Fig. 13: Shapley values for instance #0 of the PH dataset: comparison between client-specific explanations obtained with the *Local* approach and the explanation provided by *FedSHAP-FCM*.

As already discussed in Section 5.2.1, the *Local* approach is also unable to guarantee the accuracy of the explanations, since each client uses a limited and biased subset of the training data as background dataset. This not only reduces the representativeness of the background distribution but also amplifies the inconsistency of the attributions, potentially leading to misleading explanations when deployed in practice.

5.4.2 Consistency Analysis for the Image Dataset

To investigate the consistency of explanations in the image domain, we consider as a representative of the image data the MNIST dataset and we simulate a realistic scenario involving three new clients (A, B, and C) that did not participate in the federated training of the image classification model. These clients possess only a small amount of local data, with significantly unbalanced and non-IID label distributions.

Figure 14 shows the data distribution for the three clients. Client A has a relatively balanced dataset that includes all classes, whereas Clients B and C lack coverage for several classes, making the scenario representative of practical deployment conditida confrontare ons in FL systems.

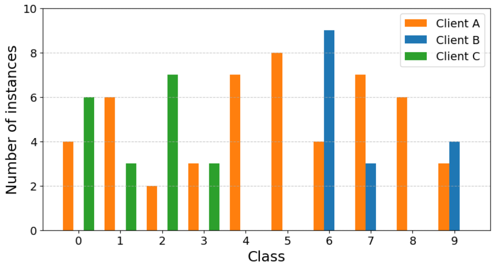


Fig. 14: Data distribution of three new joining clients.

In this setting, explanations are generated using three strategies: *Local* (each client uses its own local dataset as background), *FedSHAP-GAN* (the background includes the centroids of the clusters generated by FedFCM), and *Centralized* (the background dataset is obtained joining the local training sets of the clients which participated in the FL). Figures 15 and 16 report, respectively, the SHAP explanations for two sample instances from the MNIST test set: #689 (classified as digit *6*) and #808 (classified as digit *8*).

For instance #689, we observe that the SHAP explanations produced by the three clients under the *Local* strategy differ substantially. As shown in subfigures 15d–15f, the salient regions of the image vary widely between clients, with Client A producing the most plausible explanation, likely due to its more complete data distribution (cf. Figure 14). Clients B and C, which lack samples of class 6 in their local datasets, produce explanations with low alignment to the true decision logic.

Conversely, the explanation obtained using *FedSHAP-GAN* (Fig. 15c) is qualitatively close to the centralized one (Fig. 15b), capturing similar regions with comparable attribution strength.

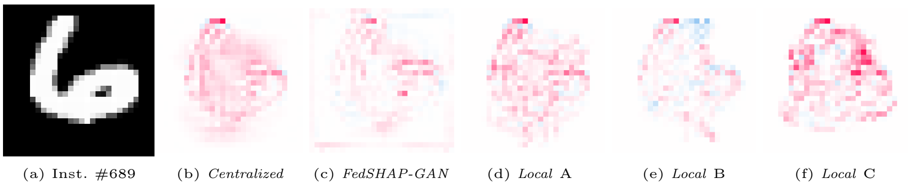


Fig. 15: Comparison of explanations for instance #689: (a) original image, (b) explanation with *Centralized*, (c) with *FedSHAP-GAN*, (d–f) explanations by Clients A, B, and C using the *Local* approach.

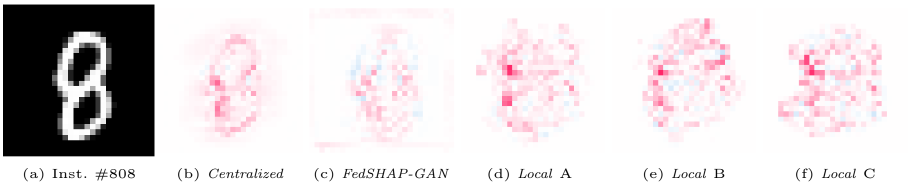


Fig. 16: Comparison of explanations for instance #808: (a) original image, (b) explanation with *Centralized*, (c) with *FedSHAP-GAN*, (d–f) explanations by Clients A, B, and C using the *Local* approach.

A similar pattern is visible, for instance #808 (Figure 16). In this case, only Client A has examples of class 8 in its background dataset; the explanations produced by Clients B and C appear noisy and structurally inconsistent. Again, the explanation provided by *FedSHAP-GAN* remains aligned with the centralized one.

These results highlight two key observations: (i) the *Local* strategy leads to inconsistent and sometimes misleading explanations, particularly when the local dataset is small and unbalanced, and (ii) our proposed *FedSHAP-GAN* method provides stable, centralized-like explanations even for clients that were not involved in the training process.

As discussed in Section 5.2.1, the *Local* strategy also fails to guarantee explanation accuracy. In this case, the issue is further exacerbated by the spatial nature of image data, where poorly representative background distributions can result in noisy or even contradictory attributions. *FedSHAP-GAN*, by contrast, offers a robust alternative capable of supporting consistent and accurate post-hoc explanations in federated settings.

6 Conclusions

In this work, we addressed the challenge of making opaque AI models explainable in privacy-preserving, distributed machine learning scenarios. Specifically, we proposed a novel framework, *Federated SHAP*, aimed at generating post-hoc explanations that are both accurate and consistent, while complying with the constraints of the Federated Learning paradigm.

Two strategies were developed and evaluated. For tabular data, we introduced *FedSHAP-FCM*, which leverages a federated version of the Fuzzy C-Means clustering algorithm to summarize the training data of each client into a shared, privacy-preserving background dataset. For image data, we proposed *FedSHAP-GAN*, where a Generative Adversarial Network is trained in a federated fashion to synthesize realistic reference images that approximate the training distribution without revealing sensitive information.

Our extensive experimental evaluation covered both classification and regression tasks across a variety of benchmark datasets. The results confirm that *FedSHAP-FCM* achieves a strong balance between privacy, explanation accuracy, and consistency. It consistently outperformed baseline methods such as *Local*, where each client uses its own training data as background at explanation time, resulting in inconsistent and biased explanations; *Random*, which builds a shared background from uniformly sampled synthetic data, leading to uninformative attributions; and *LDP*, where clients perturb their local data with noise to preserve privacy before contributing to a shared background, introducing an inherent accuracy-privacy trade-off. Notably, *FedSHAP-FCM* was able to match or exceed the explanation accuracy of *LDP*, even when using a smaller and more efficient background dataset.

In the case of image data, *FedSHAP-GAN* provided more faithful and structured explanations than those obtained with random synthetic backgrounds, especially in terms of spatial coherence. While a residual gap remains compared to the centralized baseline, the method offers a practical and effective solution for federated explainability when visual features and high-dimensional inputs are involved. Additional analyses also showed that using a moderately sized background is sufficient to achieve good performance, making the method efficient and scalable.

Moreover, we demonstrated that the explanations produced by *Federated SHAP* remain consistent and meaningful even for new clients joining the federation after training, and with highly heterogeneous local data distributions. This is a critical feature in real-world federated systems, where clients often operate with limited and non-representative data.

Future work will extend this approach to other data types such as textual data, explore its integration with alternative post-hoc methods, and involve human-in-the-loop evaluation to further validate the interpretability and trustworthiness of the explanations.

Acknowledgments. This work has been partly funded by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - *FAIR - Future Artificial Intelligence Research* - Spoke 1 *Human-centered AI* under the NextGeneration EU program, and from the PNRR *Tuscany Health Ecosystem* (THE) (*Ecosistemi dell'Innovazione*) - Spoke 6 - *Precision Medicine & Personalized Healthcare* (CUP I53C22000780001) under the NextGeneration EU program. Additionally, this research has been supported by the Italian Ministry of University and Research (MUR) in the framework of the *FoReLab* and *CrossLab* projects (*Departments of Excellence*).

References

[1] High Level Expert Group on AI: Ethics Guidelines for Trustworthy AI, Technical Report. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (2019)

[2] McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-Efficient Learning of Deep Networks from Decentralized Data. PMLR (2017). https://proceedings.mlr.press/v54/mcmahan17a.html

[3] Guidotti, R., Monreale, A., Pedreschi, D., Giannotti, F.: In: Sayed-Mouchaweh, M. (ed.) Principles of Explainable Artificial Intelligence, pp. 9–31. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-76409-8_2

[4] Ducange, P., Marcelloni, F., Renda, A., Ruffini, F.: Fundamentals on explainable and interpretable artificial intelligence models. Trustworthy AI in Medical Imaging, 279–296 (2025)

[5] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions **30** (2017)

[6] Shapley, L.S.: In: Kuhn, H.W., Tucker, A.W. (eds.) 17. A Value for n-Person Games, pp. 307–318. Princeton University Press, Princeton (1953). https://doi.org/10.1515/9781400881970-018 . Last checked: 2023-11-21

[7] Bogdanova, A., Imakura, A., Sakurai, T.: Dc-shap method for consistent explainability in privacy-preserving distributed machine learning. Human-Centric Intelligent Systems **3**(3), 197–210 (2023) https://doi.org/10.1007/s44230-023-00032-4

[8] López-Blanco, R., Alonso, R.S., González-Arrieta, A., Chamoso, P., Prieto, J.: Federated learning of explainable artificial intelligence (fed-xai): A review. In: Ossowski, S., Sitek, P., Analide, C., Marreiros, G., Chamoso, P., Rodríguez, S. (eds.) Distributed Computing and Artificial Intelligence, 20th Int'l Conf., pp. 318–326. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-38333-5_32

[9] Ducange, P., Marcelloni, F., Renda, A., Ruffini, F.: Consistent post-hoc explainability in federated learning through federated fuzzy clustering. In: 2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–10 (2024). https://doi.org/10.1109/FUZZ-IEEE60900.2024.10611761

[10] Bárcena, J.L.C., Marcelloni, F., Renda, A., Bechini, A., Ducange, P.: Federated *c*-means and fuzzy *c*-means clustering algorithms for horizontally and vertically partitioned data. IEEE Transactions on Artificial Intelligence **5**(12), 6426–6441 (2024) https://doi.org/10.1109/TAI.2024.3426408

[11] Bozorgpanah, A., Torra, V.: Explainable machine learning models with privacy. Progress in Artificial Intelligence **13**(1), 31–50 (2024)

[12] Li, W., Chen, J., Wang, Z., Shen, Z., Ma, C., Cui, X.: Ifl-gan: Improved federated learning generative adversarial network with maximum mean discrepancy model aggregation. IEEE Transactions on Neural Networks and Learning Systems **34**(12), 10502–10515 (2023) https://doi.org/10.1109/TNNLS.2022.3167482

[13] Shenaj, D., Rizzoli, G., Zanuttigh, P.: Federated learning in computer vision. Ieee Access **11**, 94863–94884 (2023)

[14] Bárcena, J.L.C., Daole, M., Ducange, P., Marcelloni, F., Renda, A., Ruffini, F., Schiavo, A.: Fed-xai: Federated learning of explainable artificial intelligence models, vol. 3277, pp. 104–117 (2022)

[15] Lopez-Ramos, L.M., Leiser, F., Rastogi, A., Hicks, S., Strümke, I., Madai, V.I., Budig, T., Sunyaev, A., Hilbert, A.: Interplay between Federated Learning and Explainable Artificial Intelligence: a Scoping Review (2024)

[16] Corcuera Bárcena, J.L., Ducange, P., Ercolani, A., Marcelloni, F., Renda, A.: An approach to federated learning of explainable fuzzy regression models. In: 2022 IEEE Int'l Conf. on Fuzzy Systems (FUZZ-IEEE), pp. 1–8 (2022). https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882881

[17] Wilbik, A., Grefen, P.: Towards a federated fuzzy learning system. In: 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–6 (2021). https://doi.org/10.1109/FUZZ45933.2021.9494392

[18] Bárcena, J.L.C., Ducange, P., Marcelloni, F., Renda, A.: Increasing trust in ai through privacy preservation and model explainability: Federated learning of fuzzy regression trees. Information Fusion **113**, 102598 (2025)

[19] Briola, E., Nikolaidis, C.C., Perifanis, V., Pavlidis, N., Efraimidis, P.: A federated explainable ai model for breast cancer classification. In: Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference. EICC '24, pp. 194–201. Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3655693.3660255

[20] Sandeepa, C., Siniarski, B., Wang, S., Liyanage, M.: Sherpa: Explainable robust algorithms for privacy-preserved federated learning in future networks to defend against data poisoning attacks. In: 2024 IEEE Symposium on Security and Privacy (SP), pp. 4772–4790 (2024). https://doi.org/10.1109/SP54263.2024.00271

[21] Corbucci, L., Guidotti, R., Monreale, A.: Explaining black-boxes in federated learning. In: Longo, L. (ed.) Explainable Artificial Intelligence, pp. 151–163. Springer, Cham (2023)

[22] Asiri, A., Wang, W., Wu, F., Vo, H., Yu, S.: Fedxai for detecting ddos on iot edge networks in federated learning. In: Proceedings of the 2024 34th International Telecommunication Networks and Applications Conference (ITNAC 2024) (2024).

https://doi.org/10.1109/ITNAC62915.2024.10815590

[23] Ben Saad, S., Brik, B., Ksentini, A.: A trust and explainable federated deep learning framework in zero touch b5g networks. In: GLOBECOM 2022 - 2022 IEEE Global Communications Conference, pp. 1037–1042 (2022). https://doi.org/10.1109/GLOBECOM48099.2022.10001371

[24] Kalakoti, R., Bahsi, H., Nomm, S.: Explainable federated learning for botnet detection in iot networks, pp. 22–29 (2024). https://doi.org/10.1109/CSR61664.2024.10679348

[25] Sarker, M.A.A., Shanmugam, B., Azam, S., Thennadil, S.: Enhancing smart grid load forecasting: An attention-based deep learning model integrated with federated learning and xai for security and interpretability. Intelligent Systems with Applications **23**, 200422 (2024) https://doi.org/10.1016/j.iswa.2024.200422

[26] Abtahi, A., Aminifar, A., Aminifar, A.: Privacy-preserving federated interpretability. In: Proceedings of the 2024 IEEE International Conference on Big Data, BigData 2024, pp. 7592–7601 (2024). https://doi.org/10.1109/BigData62323.2024.10825590

[27] Fatema, K., Dey, S.K., Anannya, M., Khan, R.T., Rashid, M., Chunhua, S., Mazumder, R.: Federated xai ids: An explainable and safeguarding privacy approach to detect intrusion combining federated learning and shap (2025)

[28] Guan, H., Yap, P.-T., Bozoki, A., Liu, M.: Federated learning for medical image analysis: A survey. Pattern Recognition, 110424 (2024)

[29] Gecer, M., Garbinato, B.: Federated learning for mobility applications. ACM Computing Surveys **56**(5), 1–28 (2024)

[30] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 1135–1144. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778 . https://doi.org/10.1145/2939672.2939778

[31] Gipiškis, R., Tsai, C.-W., Kurasova, O.: Explainable ai (xai) in image segmentation in medicine, industry, and beyond: A survey. ICT Express (2024)

[32] Daole, M., Ducange, P., Marcelloni, F., Miglionico, G.C., Renda, A., Schiavo, A.: Xaimed: A diagnostic support tool for explaining ai decisions on medical images. Proceedings of the 1st International Conference on Explainable AI for Neural and Symbolic Methods - EXPLAINS, 27–37 (2024) https://doi.org/10.5220/0012942000003886

[33] Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B., Yang, M.-H.: Gan inversion:

A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3121–3138 (2022)

[34] Hardy, C., Le Merrer, E., Sericola, B.: Md-gan: Multi-discriminator generative adversarial networks for distributed datasets. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), pp. 866–877 (2019). https://doi.org/10.1109/IPDPS.2019.00095

[35] Li, W., Chen, J., Wang, Z., Shen, Z., Ma, C., Cui, X.: Ifl-gan: Improved federated learning generative adversarial network with maximum mean discrepancy model aggregation. IEEE Transactions on Neural Networks and Learning Systems **34**(12), 10502–10515 (2023) https://doi.org/10.1109/TNNLS.2022.3167482

[36] Zhang, J., Zhao, L., Yu, K., Min, G., Al-Dubai, A.Y., Zomaya, A.Y.: A novel federated learning scheme for generative adversarial networks. IEEE Transactions on Mobile Computing **23**(5), 3633–3649 (2023)

[37] Stallmann, M., Wilbik, A.: On a framework for federated cluster analysis. Applied Sciences **12**(20) (2022) https://doi.org/10.3390/app122010455

[38] Pedrycz, W.: Federated fcm: Clustering under privacy requirements. IEEE T. Fuzzy Syst., 1–1 (2021) https://doi.org/10.1109/TFUZZ.2021.3105193

[39] Dwork, C.: Differential privacy: A survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) Theory and Applications of Models of Computation, pp. 1–19. Springer, Berlin, Heidelberg (2008)

[40] Mahawaga Arachchige, P.C., Liu, D., Camtepe, S., Nepal, S., Grobler, M., Bertok, P., Khalil, I.: "local differential privacy for federated learning". In: Atluri, V., Di Pietro, R., Jensen, C.D., Meng, W. (eds.) Computer Security – ESORICS 2022, pp. 195–216. Springer, Cham (2022)

[41] Kelly, M., Longjohn, R., Nottingham, K.: The UCI Machine Learning Repository. https://archive.ics.uci.edu

[42] LeCun, Y., Cortes, C., Burges, C.J.: Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist **2** (2010)

[43] Krizhevsky, A.: Learning multiple layers of features from tiny images. (2009). https://api.semanticscholar.org/CorpusID:18268744

[44] Vanschoren, J., Rijn, J.N., Bischl, B., Torgo, L.: Openml: networked science in machine learning. ACM SIGKDD Explorations Newsletter **15**(2), 49–60 (2014) https://doi.org/10.1145/2641190.2641198

[45] Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., *et al.*: Advances and open

problems in federated learning. Foundations and trends® in machine learning **14**(1–2), 1–210 (2021)

[46] Khuu, D.-P., Sober, M., Kaaser, D., Fischer, M., Schulte, S.: Data poisoning detection in federated learning. In: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, pp. 1549–1558 (2024)

[47] Bolchini, C., Bosio, A., Cassano, L., Miele, A., Pappalardo, S., Passarello, D., Ruospo, A., Sanchez, E., Reorda, M.S., Turco, V.: Benchmark suite for resilience assessment of deep learning models. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 1–1 (2025) https://doi.org/10.1109/TCAD.2025.3578297

[48] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[49] Yuan, H., Liu, M., Kang, L., Miao, C., Wu, Y.: An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models (2023)

[50] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a nash equilibrium. CoRR **abs/1706.08500** (2017) 1706.08500

[51] Biecek, P., Burzykowski, T.: Explanatory model analysis. Chapman and Hall/CRC, New York (2021)

[52] Molnar, C.: Interpretable machine learning: a guide for making black box models explainable (2023). https://christophm.github.io/interpretable-ml-book

[53] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning (ICML'17), 3319–3328 (2017)

[54] Mo, B.-Y., Nuannimnoi, S., Baskoro, A., Khan, A., Ariesta Dwi Pratiwi, J., Huang, C.-Y.: Clusteredshap: Faster gradientexplainer based on k-means clustering and selections of gradients in explaining 12-lead ecg classification model. In: Proceedings of the 13th International Conference on Advances in Information Technology. IAIT '23. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3628454.3631199

[55] Corcuera Bárcena, J.L., Marcelloni, F., Renda, A., Bechini, A., Ducange, P.: A federated fuzzy c-means clustering algorithm. In: Int'l WS on Fuzzy Logic and Applications 2021, vol. 3074, pp. 1–9 (2021)

Appendix A Background: SHAP post-hoc method

Among the many post hoc explanation methods available, SHAP [5] is arguably the most widely used. This predominance is due to different factors: i) its mathematical foundation, based on game theory, is solid and renowned, ii) it can be used both for regression and classification tasks, iii) it can be applied to tabular, images, as well as textual datasets, iv) the interpretation of the explanations is quite straightforward; and v) its model-agnostic variants can be applied to any kind of AI model.

In detail, SHAP assigns the importance of the features by estimating the Shapley values, a concept introduced by L. S. Shapley in [6] for cooperative games. In the context of game theory, the Shapley values are the different rewards assigned to different players of a coalitional game, depending on the different ways in which each player contributes to the total game payout. In the context of XAI, the Shapley values correspond to the different contributions of the individual input features (the players) to the AI-model output (the game payout): as a consequence, Shapley values serve as a measure of feature importance. Formally, given a model f and an input instance $\mathbf{x}_i$ with F input features, the output $f(\mathbf{x}_i)$ is expressed as the sum of the Shapley values:

$$f(\mathbf{x}_i) = \phi_0 + \sum_{j=1}^{F} \phi_j \tag{A1}$$

where ϕ_j (with $j = 1...F$) are the F Shapley values and ϕ_0 is a reference value given by the average of the $f()$ results over a reference dataset.

The interpretation of the Shapley values is intuitive: the larger the absolute Shapley value, the greater the impact of the corresponding feature on the model decision. Positive (negative) values correspond to positive (negative) contributions to the model output.

The computation of the Shapley values involves calculating the marginal contribution of all the possible coalitions of features. A single marginal contribution of a feature j is the difference between the prediction of the model for an instance $\mathbf{x}_i$ and the prediction of the model for an input $\tilde{\mathbf{x}}_i$ where all the feature values are the same as in $\mathbf{x}_i$ except for the value of the feature j that is replaced by a perturbed value. This is repeated for all possible coalitions of features, where a coalition is defined by the presence or absence of any number of features. This can be achieved by perturbing the instance extracting values from a *background* (also called *reference*) dataset. Since the exact calculation of the Shapley values is demanding in terms of computational effort, several approximation techniques have been proposed over time. Among them, Lundberg et al. introduced several SHAP variants in [5] such as KernelSHAP and GradientExplainer [51, 52]. KernelSHAP is widely used because it is a model-agnostic method (that is, it can be applied to explain any kind of ML model). The related procedure is described in the Algorithm 1.

GradientExplainer is an extension of the integrated gradients method [53], that is, a feature attribution method designed for differentiable models.

We refer to [54] for a discussion on GradientExplainer. Briefly, given the individual instance x_i and a background instance r, different gradient points G can be calculated

Algorithm 1: KernelSHAP algorithm

```
Require: $R$: reference dataset, where each instance is defined by $F$ features;
Require: $f$: predictive model;
Require: $\mathbf{x}_i$: instance for which the prediction $f(\mathbf{x}_i)$ needs to be explained;
Output: Shapley values $\phi_j$, for $j = 1, \ldots, F$.
Sample $K$ coalitions $z'_k \in \{0,1\}^F$, with $k \in \{1, ..., K\}$, from the possible coalitions that can be generated. A value of 1 means that the corresponding feature value is "present" and 0 that it is "absent".
for $k = 1, \ldots, K$ do
    Compute $z_k = h_x(z'_k)$, where $h_x : \{0,1\}^F \to \mathbb{R}^F$ maps a coalition of features into the original feature space. ▷ For example, a coalition $z'_k = (0, 1, 0, 1)$ is transformed into an instance $z_k$ as follows: since the second and fourth features are present in the coalition $z'_k$, the corresponding values of $z_k$ are taken from the input instance $x_i$. Since the first and third features are absent in the coalition $z'_k$, the corresponding values of $z_k$ are taken from an instance randomly sampled from the reference dataset $R$.
    Compute the prediction of the model $f(h_x(z'_k))$
    Compute the weight for each $z'_k$ with the following SHAP kernel:
        $\pi_x(z') = \frac{(F-1)}{\binom{F}{|z'|}|z'|(F-|z'|)}$
    where $|z'|$ is the number of non-zero elements in $z'$.
end for
Estimate the Shapley values $\phi_j$ by optimizing the loss function
    $L(f, g, \pi_x) = \sum_{k=1}^{K} [f(h_x(z'_k)) - g(z'_k)]^2 \pi_x(z'_k)$    (A2)
with
    $g(z') = \phi_0 + \sum_{j=1}^{F} \phi_j z'_j$    (A3)
return the Shapley values $\phi_j$
```

as:

$$G = \pi x_i + (1 - \pi) r \tag{A4}$$

with π a constant in the range $[0, 1]$. The average of the gradients over the background dataset gives an estimate of the behaviour of the model around x, and the Shapley values can be calculated as

$$\phi_i = E\left[\frac{\partial f(G)}{\partial x_i}\right] \tag{A5}$$

In this sense, the expected gradients approximate the Shapley values [5] assuming that a linear approximation around the point is realistic. GradientExplainer is known to be computationally more efficient with respect to KernelSHAP, and it is often used to explain NN models. Other SHAP variants are specially optimized for specific purposes: TreeSHAP, for example, is commonly adopted to explain tree-based models.

Appendix B Datasets description

The Phoneme (PH) dataset originates from a European project (ROARS), aimed at the development of a real-time analytical system for French and Spanish speech recognition. This dataset is available in the OpenML repository [44] and is used for a binary classification task of distinguishing between nasal and oral vowels. In particular, the input features are the amplitudes of the first five harmonics, normalised by the total energy (integrated on all the frequencies).

The Magic (Major Atmospheric Gamma Imaging Cherenkov Telescopes) dataset is a well-known classification dataset available in the UCI Machine Learning Repository dataset collection [41]. It consists of 19,020 instances generated via Monte Carlo simulation to model the detection of high-energy gamma particles by an atmospheric Cherenkov telescope. The classification task consists of discriminating between background and gamma signal events thanks to 10 different attributes.

The Rice dataset is generated from pictures of two rice species, the Osmancik and the Cammeo species. From each image, seven morphological features are extracted, and a binary classification task is enabled. Also, this dataset is available in the UCI Machine Learning Repository dataset collection [41].

The Power Plant (PP) dataset contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the plant was set to work with full load. This dataset is available in the UCI repository [41] and has four numerical continuous features used for the regression task: predict the hourly output power given the hourly average ambient variables temperature (AT), the ambient pressure (AP), the relative humidity (RH) and the exhaust vacuum (V).

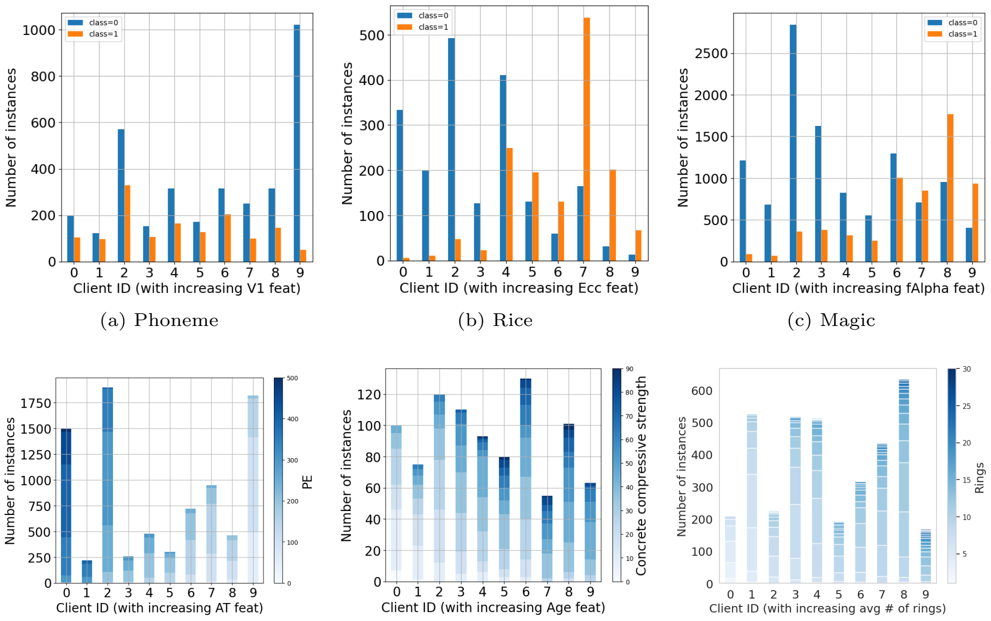
The Concrete (CC) dataset consists of 1030 instances and is used to predict the compressive strength of high-performance concrete as a regression task. It is composed of 8 features and is available in the UCI repository [41].

The MNIST (MN) dataset [42] contains 70,000 handwritten digits (i.e. 10 labels) and is commonly used as a reference dataset for comparing classification models.

The CIFAR-10 (CF) dataset [43] contains 60,000 colour images in 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck) and is commonly used as a reference dataset for comparing classification models.

The abalone snails, considered worldwide as a food delicacy, are the subject of the Abalone (AB) dataset. The regression task has the objective of determining the number of rings of the snails, using as input the physical information of the shell. The number of rings is commonly related to the age, which is used to determine their price.

The dataset is then split into training and test sets, with non-IID scenarios enforced as described in Section 4. The class distributions across the 10 clients' training sets for each of the datasets considered in this study are shown in Figures B.1 and B.2.



(a) Phoneme

(b) Rice

(c) Magic

(d) PowerPlant

(e) Concrete

(f) Abalone

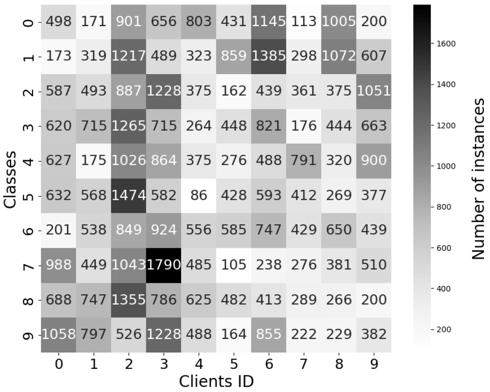
Fig. B.1: Number of instances per client. The color indicates the marginal distribution of the target. (a, b, c) Binary classification tasks – (d, e, f) Regression tasks.

Appendix C Federated Fuzzy C-Means Clustering algorithm

The Federated Fuzzy C-Means (FedFCM) procedure is schematized in Fig. C.1.

First, a center initialization procedure is executed: the central server transmits the initial configurations to the clients, including the number of clusters K (defined a priori by the user) and the value of the fuzziness parameter λ. The server also randomly initializes the cluster centroids and sends them to the clients. As an alternative to random initialization, the authors in [10] suggest a federated version of k-means++ for the careful seeding of the clustering algorithm.

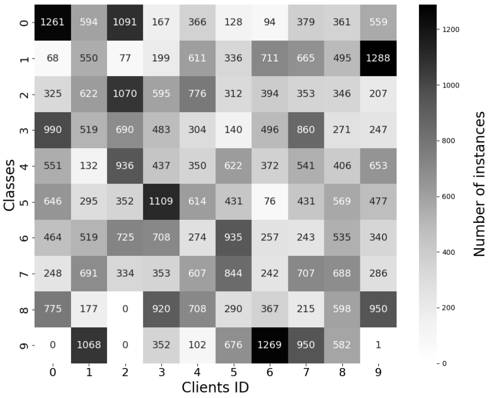
At each iteration of the protocol, the clients locally update the cluster membership degrees of the objects, based on the centroids received. Then, each client calculates



	0	1	2	3	4	5	6	7	8	9
0	498	171	901	656	803	431	1145	113	1005	200
1	173	319	1217	489	323	859	1385	298	1072	607
2	587	493	887	1228	375	162	439	361	375	1051
3	620	715	1265	715	264	448	821	176	444	663
4	627	175	1026	864	375	276	488	791	320	900
5	632	568	1474	582	86	428	593	412	269	377
6	201	538	849	924	556	585	747	429	650	439
7	988	449	1043	1790	485	105	238	276	381	510
8	688	747	1355	786	625	482	413	289	266	200
9	1058	797	526	1228	488	164	855	222	229	382

(a) MNIST

Fig. B.2: Number of instances per client and per class in the MNIST dataset. The color indicates the marginal distribution of the target.



	0	1	2	3	4	5	6	7	8	9
0	1261	594	1091	167	366	128	94	379	361	559
1	68	550	77	199	611	336	711	665	495	1288
2	325	622	1070	595	776	312	394	353	346	207
3	990	519	690	483	304	140	496	860	271	247
4	551	132	936	437	350	622	372	541	406	653
5	646	295	352	1109	614	431	76	431	569	477
6	464	519	725	708	274	935	257	243	535	340
7	248	691	334	353	607	844	242	707	688	286
8	775	177	0	920	708	290	367	215	598	950
9	0	1068	0	352	102	676	1269	950	582	1

(a) CIFAR-10

Fig. B.3: Number of instances per client and per class in the CIFAR-10 dataset. The color indicates the marginal distribution of the target.

aggregate statistics and sends them to the server. Shared statistics consist in (i) the sum of the membership degrees of the local objects to each cluster, raised to the λ power, and (ii) the weighted sum of feature values of objects based on their membership degrees to each cluster. The server uses this information to update the centroids. The process is repeated until the variation of centroids between two successive iterations is less than a predefined threshold.

A thorough analysis of privacy in FedFCM is presented in our previous work [10]. Although the shared aggregated statistics do not include raw data, a semi-honest server could attempt to reconstruct individual client data by solving a system of equations

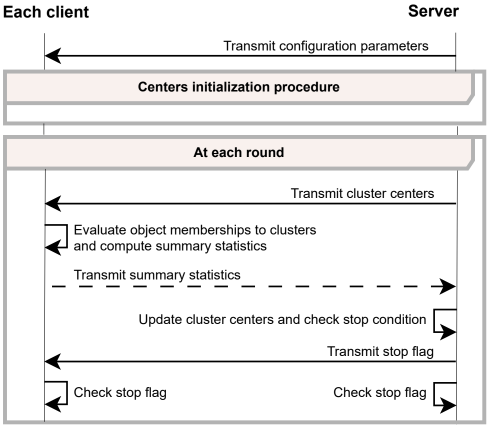

Each client
Server
Transmit configuration parameters
Centers initialization procedure
At each round
Transmit cluster centers
Evaluate object memberships to clusters and compute summary statistics
Transmit summary statistics
Update cluster centers and check stop condition
Transmit stop flag
Check stop flag
Check stop flag

Fig. C.1: Sequence diagram of the FedFCM procedure, as proposed in [10].

derived from these statistics. However, this reconstruction is only feasible under specific conditions, which are rarely encountered in practice. In particular, it requires the server to know the number of local data objects, which must also be relatively small compared to the number of clusters. Even under this unlikely worst-case scenario, the FedFCM procedure can still be applied safely by limiting participation to clients whose local dataset sizes meet the necessary conditions for privacy preservation.

A distinctive feature of FedSHAP-FCM is its equivalence with the centralized version of the algorithm, given the same initialization. In other words, by applying FedFCM to the distribution of data among clients, the same result is obtained as if the traditional FCM algorithm were run on the union of all data in a single centralized base. This property, known as losslessness, was formally demonstrated in [55].

At the end of the execution, the final centroids are shared with all participants in the federation: these represent a global summary of the distributed data and can be used as a common knowledge base (background dataset).

Table C.1 reports the execution times required to generate the background datasets from the various datasets used in this study, assuming execution on a server equipped with a 10-core Intel® i7-1265U CPUs, 16 GiB of RAM.

Appendix D Details on GAN generator

Tables D.1 and D.2 report the architecture of the GAN generator model employed for the generation of the background, with the *FedSHAP-GAN* approach, for the MNIST and for the CIFAR-10 datasets, respectively.

Appendix E *FedSHAP-GAN* performance with different training conditions

To evaluate the robustness of *FedSHAP-GAN*, we trained three variants of the generative model using 33%, 50%, and 100% of the local training sets, denoted as

Table C.1: Execution times (in seconds) for background dataset generation, reported as mean and standard deviation over 10 experiments.

	Mean	**Standard deviation**
Phoneme	0.33	0.05
Magic	2.29	0.34
Rice	0.42	0.09
PowerPlant	0.96	0.22
Concrete	0.07	0.02
Abalone	0.27	0.04

Table D.1: Architecture of the Generator Model of the FedSHAP-GAN for the MNIST dataset.

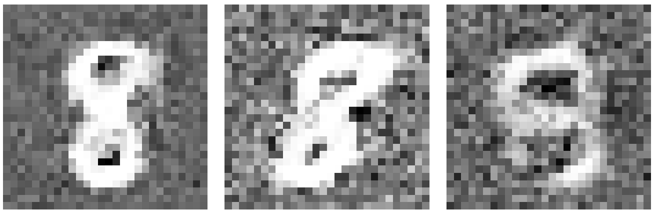
Layer		Output Size	Activation
Name	Type		
Input	Noise Vector	100	-
1	Fully Connected	512	ReLU
2	Fully Connected	512	ReLU
3	Fully Connected	784	Tanh
Output	Reshape	28×28×1	-

Table D.2: Architecture of the Generator Model of the FedSHAP-GAN for the CIFAR-10 dataset.

Layer		Output Size	Activation
Name	Type		
Input	Noise Vector	$100 \times 1 \times 1$	-
1	ConvTranspose2d	$64 \times 8 \times 4 \times 4$	ReLU
2	BatchNorm2d	$64 \times 8 \times 4 \times 4$	-
3	ConvTranspose2d	$64 \times 4 \times 8 \times 8$	ReLU
4	BatchNorm2d	$64 \times 4 \times 8 \times 8$	-
5	ConvTranspose2d	$64 \times 2 \times 16 \times 16$	ReLU
6	BatchNorm2d	$64 \times 2 \times 16 \times 16$	-
7	ConvTranspose2d	$64 \times 1 \times 32 \times 32$	ReLU
8	BatchNorm2d	$64 \times 1 \times 32 \times 32$	-
9	ConvTranspose2d	$3 \times 64 \times 64$	Tanh

FedSHAP-GAN33, *FedSHAP-GAN50*, and *FedSHAP-GAN*, respectively. The training times are 3.3 minutes for the model with 33% data, 4.9 minutes for the model with 50% data, and 9.4 minutes for the model trained with 100% data. The resulting models were used to generate synthetic background datasets for SHAP-based explanations, and their effectiveness was assessed by measuring the discrepancy with the *Centralized* baseline as the Frobenius norm of the difference between the corresponding SHAP matrices.

Figure E.1 provides a qualitative comparison of the synthetic digits produced by each variant.



(a) *FedSHAP-GAN* (b) *FedSHAP-GAN50* (c) *FedSHAP-GAN33*

Fig. E.1: An example of synthetic images generated with the three different GAN models trained on the MNIST dataset.

As expected, the visual quality of the generated images increases with the size of the training set. Samples generated by *FedSHAP-GAN* and *FedSHAP-GAN50* appear realistic and representative of the class, while *FedSHAP-GAN33* shows increased noise and reduced diversity. While GAN33's reduced training time can be useful in resource-constrained scenarios, it comes at the cost of lower sample quality.

Figure E.2 shows the distribution of the Frobenius norm over 10 runs for each background dataset generation strategy.

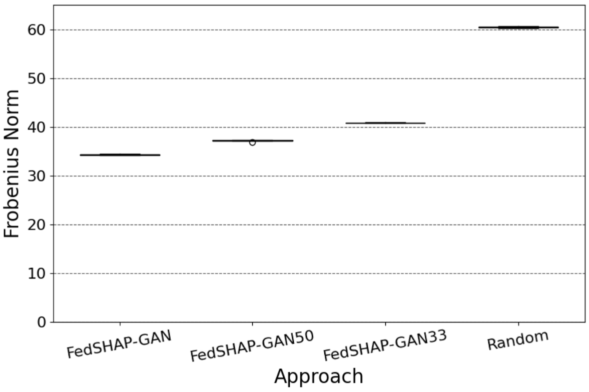


Fig. E.2: Boxplots of the discrepancy of the *FedSHAP-GAN*, *FedSHAP-GAN50*, *FedSHAP-GAN33* and *Random* approaches with respect to the *centralized* approach, in terms of Frobenius norm of the pairwise difference of $\mathbf{\Phi}$ matrices.

Scenario	Generated images	FID
FedSHAP-GAN(IID)	6000	192.68 ± 0.45
FedSHAP-GAN(Non-IID)	6000	204.61 ± 0.83
FedSHAP-GAN50(Non-IID)	6000	242.37 ± 0.67
FedSHAP-GAN33(Non-IID)	6000	281.09 ± 0.52

Table F.1: FID values (average and standard deviation) calculated with respect to the original images (60,000 images for MNIST).

On the other hand, *FedSHAP-GAN* achieves the best explanation performance, but requires the highest training time. These results highlight the trade-off between computational efficiency and explanation accuracy.

The plot suggests that *FedSHAP-GAN* achieves the highest accuracy (i.e., lowest discrepancy from the *Centralized* baseline). *FedSHAP-GAN50* and *FedSHAP-GAN33* exhibit a noticeable and progressively increasing drop in performance. As expected, the Random strategy performs the worst, with significantly higher discrepancies from the centralized approach.

Appendix F Impact of the clients' data distributions on *FedSHAP-GAN* performance

Our experiments on the image datasets were conducted under a scenario in which the data distributions across the ten clients are Non-IID, reflecting the conditions typically encountered in realistic FL settings. To evaluate the impact of training data distributions, we also experimented with a non-realistic scenario where the ten different clients have the same data distribution (i.e., an IID setting) of the full MNIST training dataset. Under these experimental conditions, we trained a GAN in a federated setting (in the following, we denote this GAN as FedSHAP-GAN(IID)) and compared its explanation accuracy against the three variants of FedSHAP-GAN described in Appendix E, which were trained using 33%, 50%, and 100% of the local Non-IID training data. For clarity, we denoted these three variants as FedSHAP-GAN33(Non-IID), FedSHAP-GAN50(Non-IID), and FedSHAP-GAN(Non-IID), respectively. Our comparison methodology consists of two steps. First, we calculated the Fréchet Inception Distance (FID) between the images generated by the different scenarios and the original training dataset, in order to evaluate and compare the quality of the synthetic images. The results, obtained from ten different image background datasets generated using ten different random seeds, are presented in Table F.1. As expected, the best FID value is achieved in the IID scenario, followed closely by FedSHAP-GAN(Non-IID), which utilizes 100% of local training data under non-IID conditions.

Subsequently, we calculated the Frobenius norm of the Shapely values corresponding to the explanations on the test dataset to quantify the impact attributable to the difference in the distribution settings. Figure F.1 reports the boxplots of the discrepancy of the FedSHAP-GAN(IID), FedSHAP-GAN(Non-IID), FedSHAP-GAN50(Non-IID), FedSHAP-GAN33(Non-IID) and Random approaches with respect to the centralized approach in terms of Frobenius norm of the pairwise difference of $\mathbf{\Phi}$ matrices. As observed, the results indicate that although distributional heterogeneity has a measurable effect on the explanation accuracy, the overall impact remains moderate when comparing IID and Non-IID scenarios.

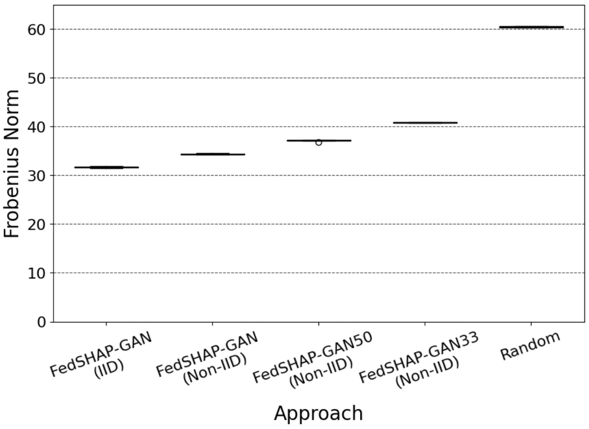


Fig. F.1: Boxplots of the discrepancy of the *FedSHAP-GAN(IID)*, *FedSHAP-GAN(Non-IID)*, *FedSHAP-GAN50(Non-IID)*, *FedSHAP-GAN33(Non-IID)* and *Random* approaches with respect to the *centralized* approach, in terms of Frobenius norm of the pairwise difference of $\mathbf{\Phi}$ matrices.

Appendix G LDP performance with different configurations of epsilon parameter

In our experiments, we considered two values of the privacy factor, ϵ =1, corresponding to a high privacy scenario, and ϵ =10, corresponding to a low privacy scenario. Here we also report the results of the experiment performed using an intermediate value of $\epsilon = 5$. Fig. G.1 reports the results of the Frobenius norms of the pairwise difference of $\mathbf{\Phi}$ matrices for Rice dataset, chosen as a representative of the tabular dataset. The results obtained with $\epsilon = 5$ demonstrate intermediate values relative to those observed under conditions of $\epsilon = 1$ and $\epsilon = 10$, consistent with expectations.

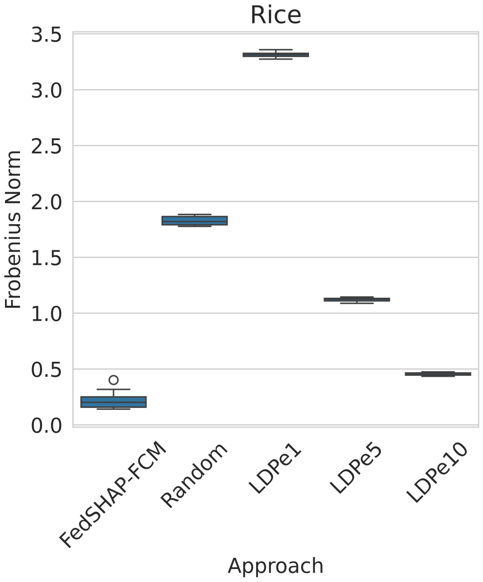


Fig. G.1: Boxplots of the discrepancy of the FedSHAP-FCM, Random and LDP approaches with respect to the Centralized approach, in terms of Frobenius norm of the pairwise difference of $\mathbf{\Phi}$ matrices for Rice dataset.