Consistent Post-Hoc Explainability in Federated Learning through Federated Fuzzy Clustering

Anonymous Authors

***Abstract*—Ensuring trustworthiness of AI systems by enforcing, for instance, data privacy and model explainability, has become urgent in our society. Recently, the Federated Learning (FL) paradigm has been proposed to preserve data privacy during collaborative model learning. Unfortunately, FL poses critical challenges in the application of post-hoc explanation methods which are used to explain opaque models such as neural networks. In this paper we present an approach for enhancing the explainability of opaque models generated according to the FL paradigm. We focus on one of the most popular methods, namely SHapley Additive exPlanations method (SHAP). Given an input instance, SHAP can explain why an opaque model generated that specific output prediction from the input values. To provide the explanation SHAP needs access to a background dataset, typically consisting of representative training instances. In FL setting, however, the training data are scattered over multiple participants and cannot be shared due to privacy constraints. On the other side, the background dataset should be representative of the overall training set. To this aim, we propose to adopt a federated Fuzzy C-Means clustering for the generation of a common background dataset made up of cluster centers. The resulting background dataset is representative of the actual distribution of the data and can be made available to all participants without violating privacy, thus ensuring accuracy and consistency of the explanations. A thorough experimental analysis shows the validity of the proposed approach also in comparison with baseline and alternative approaches.**

***Index Terms*—Federated Learning, Explainable Artificial Intelligence, SHAP, Fuzzy Clustering, Fuzzy C-Means**

I. Introduction

Explainability and data privacy are two cornerstones for the trustworthiness of Machine Learning (ML) and Artificial Intelligence (AI) systems [1]. Such increasing awareness about trust has recently prompted academia and industry to devise models and techniques capable of meeting these requirements.

The data privacy concern has motivated the development of new paradigms for training ML models in a decentralized setting, including Federated Learning (FL) [2]. FL allows multiple parties to collaboratively train an ML model removing the need to centralize data for training. In essence, a shared global model is learnt in an iterative, round-based, procedure through the aggregation of model updates computed locally by remote data owners. The FL approach is viable over both horizontally (different instances, same features) and vertically (different features, same instances) partitioned data. Typically, models learned in a federated fashion are those optimized through Stochastic Gradient Descent or its variants: this makes FL immediately suitable for Neural Networks (NNs), which are generally deemed as "opaque" models or "black boxes" for their characteristic of being hardly interpretable.

The design of techniques to explain opaque models, as well as the investigation of inherently interpretable models, is at the core of Explainable AI (XAI) [3], [4]. The scope of an explanation can be traced back to two concepts widely employed in the specialized literature, namely global and local interpretability: *global* interpretability relates to the structural properties of a model, whereas *local* interpretability is associated with the inference process and focuses on how the output is produced for any single input instance. Obtaining an explanation for a decision made by an opaque model involves a process of reverse engineering and is typically achieved through the adoption of the so-called *post-hoc* techniques.

This study is positioned at the intersection between FL and XAI: it has the objective of designing an approach for the adoption of a post-hoc technique to explain a model learned according to the FL paradigm. A Multi Layer Perceptron Neural Network (MLP-NN) is considered as an opaque model and the popular SHAP (SHapley Additive exPlanations) [5] method is considered as post-hoc technique.

SHAP is one of the most popular post-hoc techniques used to explain ML model predictions in terms of feature importance by estimating the so-called Shapley values [6]. Based on game theory, SHAP and its variants can be interpreted as methods that exploit knowledge of the training set for evaluating the explanation of any prediction. Specifically, besides the model and the instance to be explained, SHAP necessitates a "background dataset", that provides a reference distribution for the estimate of the contributions of individual features. As the choice of the background dataset can impact the outcome of the explainability process, the training set is commonly employed for this purpose. The high computational complexity of SHAP, however, precludes the use of the entire training set when it is fairly large and the adoption of numerosity reduction techniques becomes essential [7]. Another major challenge arises in the federated setting: the training set is not available in its entirety to any party as it is scattered over multiple physical locations. On the one hand, the privacy requirement prevents raw data from being moved and therefore they cannot be used as background dataset. On the other hand, if any participant relies on its own training set as background dataset, the property of *consistency* of explanations may be lost. In the FL context, consistency is met if the explanations of the same data instance for a global model are the same for different participants. Evidently, SHAP is prone to misalignment of client-side explanations especially when the local datasets follows a non-i.i.d. (non independent and identically distributed) partitioning.

This work describes an approach to overcome these challenges by achieving consistent and accurate explainability using SHAP in the federated setting. In essence, a federated clustering procedure is executed over scattered participants local data as a data summarization technique: the resulting cluster representatives are exploited as background dataset for the execution of the SHAP method. Several recently proposed techniques enable privacy preserving clustering in the FL setting [8]–[10]. In this work we resort to the Federated Fuzzy C-Means (FCM) implementation proposed in [10]. The background dataset (i.e., the centers of the clusters generated by FCM) can be shared with any participant without violating the privacy, so the consistency requirement is fulfilled. Furthermore, the centers are representative of the entire data distribution, ensuring accurate explanations. In this context, an explanation is accurate if it matches the one that would be obtained in the traditional centralized setting (i.e., when SHAP is applied using the union of the local datasets). An extensive experimental analysis on both regression and classification tasks demonstrates the soundness of the proposed approach with respect to several baseline and alternative approaches.

The rest of the paper is organized as follows: Section II describes the background related to the SHAP method. In section III we provide a brief overview of recent works that lie at the intersection between FL and XAI by pursuing post-hoc explainability. Section IV outlines the problem statement and the proposed approach for Federated SHAP based on federated fuzzy clustering. In Section V we describes the experimental setup, including the other approaches used in the comparative analysis and details about models and datasets. Section VI report and discuss the experimental results. Finally, in Section VII we draw some conclusions.

II. Background: the SHAP method for post-hoc explainability

SHAP is one of the most popular post-hoc methods used to explain ML model predictions both in classification and regression tasks. Its steadily increasing popularity is due to several factors: it holds a solid mathematical foundation derived from game theory, it can be applied to several kinds of data (e.g., tabular, images, and textual data), and it is a local method, i.e. it can explain individual predictions.

The SHAP model evaluates the importance of the features by estimating the so-called Shapley values, a concept from the coalitional game theory introduced in 1953 by L. Shapely [6]. The analogy between the game theory and the Shapley values lies in the "rewards" of the "players" in a "game": as the players of a coalitional game contribute in different ways to the total game payout (whereby Shapley values corresponds to different rewards), the F features of a dataset contribute differently to the individual prediction of a model. Specifically, the prediction can be expressed as follows:

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

where f is the predictive model, $\mathbf{x}_i$ is a generic F-dimensional input instance, ϕ_0 is a reference value computed as the average of the predictions values in a background dataset, and ϕ_j are the Shapley values. Thus, when the Shapley value ϕ_j is positive (negative), the j-th feature has a positive (negative) impact on the individual model prediction. Furthermore, the least and the most impactful features are those with the lowest and highest absolute Shapley values, respectively.

The computation of the Shapley values involves testing all the possible combinations of the features (named *coalitions*, following the analogy with the players in game theory) by perturbing the input instance $\mathbf{x}_i$ with values coming from a *background* dataset, also referred to as *reference dataset*. The high computational complexity of the exact calculation of the Shapley values has recently prompted the design of several approaches for estimating them in an efficient way. Among them, KernelSHAP, introduced in the Lundberg and Lee seminal work [5], is widely used as it is model-agnostic, that is, it can be flexibly applied to explain any model.

Three entities are required to generate an explanation with KernelSHAP: an input instance $\mathbf{x}_i$, the predictive model f, and the background dataset. The background dataset serves as a reference when $\mathbf{x}_i$ is perturbed: the features not included in a coalition are set to the values of the corresponding features of instances randomly sampled from the background dataset. Choosing a background dataset that is representative of the actual data distribution is important for an accurate estimate of the Shapley values [11]. Theoretically, the background dataset should coincide with the set of data used for learning the f model. However, KernelSHAP still entails a high computational complexity and using the whole training set is often impractical in real-world applications: thus, it is a common practice to reduce the numerosity of the background dataset (e.g., through sampling) to speed up the estimation of the Shapley values with KernelSHAP [7], [11].

III. Related Works

The use of Shapley values within the context of FL mainly concerns two substantially different aspects that merit clarification: on the one hand, techniques based on Shapley values have been adopted to provide a fair evaluation of the contributions of participants and, consequently, a robust weighting scheme for the FL process, as proposed for instance in [12]. On the other hand, Shapley values are adopted for explainability purposes with regard to predictions of a model learned in the federated setting. The present work is in line with the latter objective. Thus, in the following we report the most recent advances on the topic.

Explainability in FL has been pursued using both ex-ante approaches [8], [13]–[15] and post-hoc techniques [16], [17]. SHAP method evidently falls into the latter category, but it is not the sole approach employed for this purpose. Chen et al. [18], for instance, have proposed a framework for explainability in vertical FL based on a federated counterfactual explanation method. The counterfactual approach aims to explain a single prediction by assessing the smallest

alteration needed in the input instance to prompt the classification into a predefined class. Since counterfactuals are derived based on local private data, the property of consistency of explanations among clients is not met. Fiosina [19] has introduced a method for consistent explainability in horizontal FL based on averaging integrated gradients. The approach yields a global feature importance score but does not address the problem of local explainability. Bogdanova et al. [16] have introduced a novel approach, named DC-SHAP, aimed at achieving consistent explainability for both horizontally and vertically partitioned data within the Data Collaboration (DC) paradigm. Consistency in the horizontal setting is obtained by employing a set of auxiliary synthetic data shared among the users as background dataset, thus mitigating the feature attribution discrepancies among the users. The DC paradigm, however, differs from mainstream FL because it envisages that clients share data (and not models) with the server, after the application of an irreversible transformation. Subsequently, the server combines these intermediate representations into a unified dataset for centralized training of an ML model, which is eventually sent back to the participants.

An adaptation of SHAP has been proposed in [20] as post-hoc method for feature importance explanation in the healthcare domain, based on a hierarchical framework: a first-level FL process allows to collaboratively train a predictive model for patients in the same hospital, whereas a second-level FL process aggregates predictive models coming from different hospitals and generate a final model. The consistency problem is addressed by generating a unique background dataset as follows: synthetic instances are sampled from a Gaussian distribution, whose parameters are estimated for each feature in a hierarchical way, by combining the parameters estimated at the first and at the second level. Results show that there is a discrepancy, albeit not severe, with the centralized feature importance score (i.e., when the background is derived from the whole training set). Furthermore, it is arguable that the assumption that marginal data distribution follows a Gaussian distribution is rarely met in practical applications.

Recently, authors in [21] have proposed an approach to apply SHAP in horizontal FL. In a nutshell, the *federated* explanation for a prediction made by the FL model is obtained by averaging the explanations of the participants. Consistency is evidently met, as any client can rely on the average explanation, and an experimental analysis in an i.i.d. setting highlights that the federated explanations are also consistent with the centralized ones. Nevertheless, the approach is based on the assumption that test data are accessible to all participants: in real-world applications where privacy must be preserved even during inference time and explanations must be obtained with low latency this option is not viable.

IV. Federated SHAP based on Federated Fuzzy Clustering

In this section we first outline the scenario considered in this work and then we describe the proposed approach for obtaining SHAP explanations in the FL setting.

A. Problem Statement

Figure 1 schematizes the setup investigated in this work.

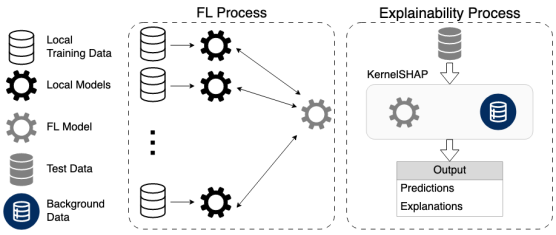


Fig. 1. High level view of the setup investigated in this work.

A star communication topology is considered: multiple parties, with horizontally partitioned data, collaborate in training an FL model under the orchestration of a central server. We assume that such model is opaque (i.e., it does not feature inherent interpretability) thus requiring the adoption of post-hoc explainability techniques. Furthermore, we consider the non-i.i.d. setting: local datasets may follow distributions that are different one from each other and from the overall data distribution. It is worth underlining that such assumptions can be considered fair as they match the most frequent and realistic situations encountered in the FL setting.

Let $P^1, P^2, \ldots, P^M$ be M parties, also referred to as clients, and $(\mathbf{X}^1, \mathbf{Y}^1), (\mathbf{X}^2, \mathbf{Y}^2), \ldots, (\mathbf{X}^M, \mathbf{Y}^M)$, their own training data. $\mathbf{X}^m$ indicates an $N_m \times F$ matrix of N_m records described by F input features, and $\mathbf{Y}^m$ indicates the vector of the N_m associated target values.

The output of the FL process is a collaboratively learned opaque model, which can be used for inference purpose on previously unseen data. The objective is to endow the model with the capability of providing local explanations: for any single test instance, both the prediction and an associated explanation are provided. In this work we consider the KernelSHAP variant of SHAP, as it is a model-agnostic method widely adopted for explaining NN-based models [7], [11].

The explainability process is applied on a unique test set that can be envisaged to be accessible for an entity that does not have training data. For example, the test set can reside on the central server or on a novel client that did not participate in the FL process and just exploits the model for inference purpose. As mentioned in Section II, KernelSHAP requires three elements: besides the test set and the model to be explained, a representative background dataset is also needed. In the traditional case (*centralized* rather than *federated* learning), a background dataset is simply generated by sampling or summarizing the training set, for example via clustering. In the FL setting, however, the training set is spread over multiple participants and cannot be shared due to privacy reason.

The main challenge of adopting SHAP in the FL setting consists in generating a background dataset that enables accurate and consistent explanations, while preserving the privacy of data owners. Our proposed approach for FederatedSHAP is presented in the following subsection.

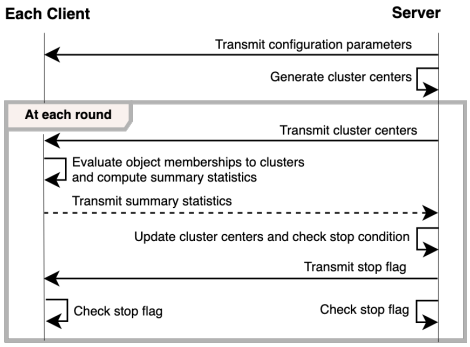


Fig. 2. Sequence diagram of the federated FCM algorithm proposed in [10].

It is worth underlining that a plausible scenario involves private test data being available on any client participating in the FL process. As a consequence, each client could simply rely on its own training data for the generation of the background dataset: this would not violate privacy, since the data would never leave the source, but would possibly undermine the requirement of consistency in the sense that different clients may obtain different explanations for an identical test instance even if employing the same model (the FL one). This may be particularly evident in the non-i.i.d. case since the local data distributions are different. The aspect of consistency is further discussed with empirical evidence from experimental analysis in Section VI-C.

B. Proposed approach

Obtaining a background dataset that is both small and representative of the entire data distribution is crucial for an efficient and effective adoption of the KernelSHAP method. Our proposed approach consists in generating the background dataset by using representatives of clusters obtained by an FL clustering algorithm. We adopt the FL procedure for the execution of the well-known FCM clustering algorithm proposed in [10], which partitions the space into K clusters. We would like to point out that the choice of the clustering algorithm is not critical for our objective. In essence, the clients' private data are not only exploited for training the FL model, but also undergo a federated clustering procedure. At the end, the cluster centers obtained through federated clustering are exploited as background dataset for the execution of the KernelSHAP method. It is worth pointing out that, in this case, the actual objective of clustering is not that of grouping instances but rather finding a compact and representative summarization of the scattered dataset. We observe that the choice of the number K of clusters is independent of the clustering tendency in the data but rather is related to the efficiency in the execution of the KernelSHAP method and its effectiveness.

The federated FCM procedure is schematized in Fig. 2.

At the beginning the server sends to each client the configuration parameters (i.e., the fuzziness factor) and randomly initializes and sends the cluster centers to the clients. The number K of clusters is fixed by the user. The procedure is consistent with the rationale of the traditional FCM, i.e. alternately updating the cluster assignment of objects and the cluster centers until convergence. The cluster assignment evaluation takes place on the client side based on the centers received from the server. The cluster centers are updated on the server side. Notably, the summary statistics shared by the clients to the server do not reveal the raw data, thus ensuring that privacy is preserved. The procedure stops when the distance between the cluster centers over two consecutive rounds is lower than a given threshold. Authors in [10] demonstrate that, given the same random initialization, the Federated FCM algorithm obtains identical results compared to the traditional FCM applied to the union of the local datasets. At the end of the execution of the Federated FCM, the server transmits the cluster centers to all the entities which have joined the federation: these centers form the background dataset.

V. Experimental Analysis

This section describes the experimental analysis from a twofold perspective: first, we formally describe our approach based on Federated FCM and other alternative approaches for generating the background dataset; then, we provide details about datasets and models considered in the experiments.

A. Approaches for background dataset generation for SHAP

Our proposed approach can be summarized as follows:

Federated FCM The background dataset is obtained as

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

where K represents the number of clusters. In other words, the background dataset consists of the K centers obtained through the Federated FCM clustering algorithm executed collaboratively by the M clients to partition the set of data locally stored in the clients. As the results of the FCM algorithm depend on the initial position of the cluster centers, we repeat the procedure 10 times with different seeds for random center initialization.

The approach is compared with baseline and alternative approaches described in the following.

Centralized The background dataset is obtained as the union of the datasets locally stored in the clients.

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

This represents a baseline approach which is typically encountered in the traditional (centralized) setup when the whole training set can be used as background. However, it is worth underlining that this approach is unfeasible in the federated setting because it requires to share private raw data, thus violating privacy. In addition, it is also impractical from a computational point of view as the estimation of the Shapley values with KernelShap is time consuming and grows in complexity with the size of the adopted background dataset. Thus, with the objective of reducing the computational cost, we conceived two variants, namely *FCM-50* and *FCM-100* by summarizing

the *full* background dataset using the traditional FCM algorithm with K=50 and K=100 centers, respectively:

$$\textit{FCM-50} \leftarrow \textit{FCM}_{50}\left(\bigcup_{m=1}^{M}\mathbf{X}^m\right) \tag{4}$$

$$\textit{FCM-100} \leftarrow \textit{FCM}_{100}\left(\bigcup_{m=1}^{M}\mathbf{X}^m\right) \tag{5}$$

Since traditional FCM is applied on the union of training data, even these variants are not viable in the federated scenario where privacy is a mandatory requirement.

Random The background dataset is obtained by randomly sampling K instances from a uniform distribution over the input space:

$$\textit{Random} \leftarrow \textit{Sample}_K\left(\mathcal{U}\left(\mathbf{a},\,\mathbf{b}\right)\right) \tag{6}$$

where $\mathbf{a}$ and $\mathbf{b}$ are the vector of lower and upper bounds of the input features, respectively. Notably, random sampling is repeated 10 times with different seeds.

This approach assumes the knowledge of realistic values for the lower and upper bounds of each feature, which is reasonable in many real-world applications. Since synthetic data are generated, the approach does not require disclosure of private raw data and can be safely adopted in the federated setting. However, being not data-driven, it likely results in a background dataset that is not representative of the actual distribution of data.

Local Let us assume that the entity interested in the explanation is the m-th generic client which has participated in the training. The background dataset is obtained as follows:

$$\textit{Local}^m \leftarrow \textit{FCM}_K\left(\mathbf{X}^m\right) \tag{7}$$

For any client, the background dataset consists of the K centers obtained through the execution of the FCM clustering algorithm over its own local training set. The local approach ensures privacy preservation but poses the problem of the consistency of explanations: the explanations obtained from the M clients may differ from each other even if the instance to be explained is identical.

Table I provides a comparison of the four approaches with regard to the properties of the resulting background datasets.

TABLE I
PROPERTIES OF THE BACKGROUND DATASET BASED ON THE APPROACH EMPLOYED FOR ITS GENERATION.

	Ensure consistency of explanations	Represent the actual data distribution	Preserve privacy
FedFCM	✓	✓	✓
Centralized	✓	✓	✗
Random	✓	✗	✓
Local	✗	*Only local data*	✓

In our experimental analysis the parameter K is set to 50, which is found as a reasonable value in the R[1] and Python[2] implementations of KernelSHAP. A comprehensive investigation of the impact of such parameter is an interesting future development of this work. A preliminary insight is hereby provided within the centralized approach, where the size of the background dataset varies in $\{50, 100, |\bigcup_{m=1}^{M}\mathbf{X}^m|\}$.

B. Datasets and data distribution scenarios

We consider a cross-silo FL setting, in which raw data are scattered over ten different clients following a horizontal partitioning scheme. The objective of the analysis is to discuss the appropriateness of the proposed approach based on Federated FCM in terms of accuracy and consistency of explanations.

It is worth underlining that the approach is versatile and applicable to both regression and classification tasks: in fact, KernelSHAP is readily suited for generating explanations for both families of tasks and the clustering procedure used for data summarization is clearly decoupled from the supervised learning stage. Thus, we employed two classification datasets, Magic and Rice, and two regression datasets, California and Abalone, all characterized by numerical features.

As per the setup discussed in Section IV-A, each dataset is divided into a training set, which is then further partitioned across clients, and a unique test set, which follows the overall data distribution, with a 90%-10% split percentage. Table II summarizes the characteristics of the four datasets.

TABLE II
DATASETS DESCRIPTION.

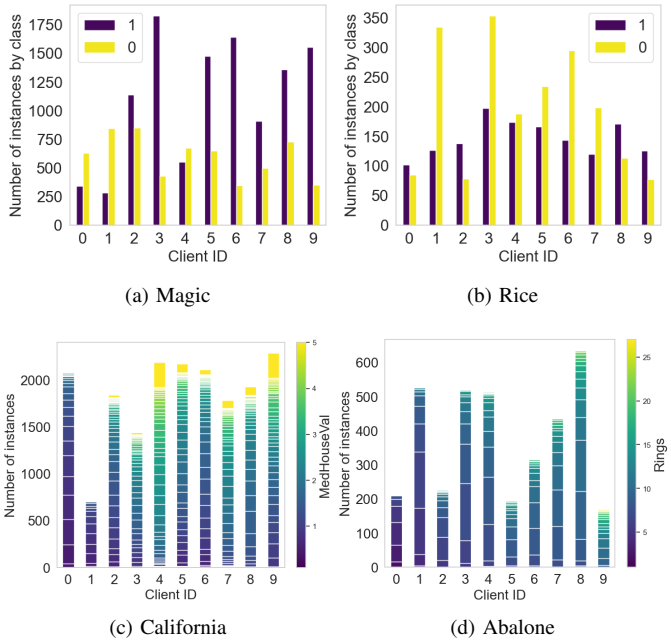
Dataset	**Source**	**Task**	**N**	$\mathbf{N}_{train}$	$\mathbf{N}_{test}$	F
Magic (MA)	[22]	C	19020	17118	1902	7
Rice (RI)	[22]	C	3810	3429	381	7
California (CA)	[23]	R	20640	18576	2064	8
Abalone (AB)	[22]	R	4177	3759	418	7

The Magic dataset (Major Atmospheric Gamma Imaging Cherenkov Telescopes) is generated by a Monte Carlo program which simulates the registration of high energy gamma particles in an atmospheric Cherenkov telescope. The classification task consists in discriminating between background and gamma signal events. The Rice dataset is generated from pictures of two rice species. From each image, several morphological features are extracted and a binary classification task is enabled. The California dataset contains housing information collected in California from the 1990 Census. The regression task consists in the prediction of the median house value using as input spatially aggregated information, such as the housing median age, over the USA California state. The Abalone dataset contains data on the abalone marine snails. The regression task consists in exploiting several shell physical measurements to determine the number of rings, which is

[1]https://cran.r-project.org/web/packages/kernelshap, visited 2024/01
[2]https://shap.readthedocs.io/en/latest/index.html, visited 2024/01

related to the age of the snail and is typically used as a proxy both by farmers and customers to determine their price.

A non-i.i.d. scenario is induced when partitioning the datasets among the various clients participating in the FL process. Specifically, we force both a quantity skewness and a label distribution skewness [24]: the former indicates that different clients may hold different amounts of local training data; the latter indicates that the marginal distributions of the target variable may vary across clients. For example, the California training set was divided into ten contiguous geographical areas aggregated starting from the 52 Californian administrative boundaries, whereas the Abalone dataset was divided into ten clients with incremental average number of rings. Figure 3 contains a visual representation of the distribution of the training sets among clients. Finally, a MinMax normalization is applied to all datasets to clip the features range in the unit interval. This is feasible in the federated setting under the reasonable assumption that the range of each feature, or an estimate thereof, is known to the server.



(a) Magic (b) Rice

(c) California (d) Abalone

Fig. 3. Number of instances per client. Color indicates the marginal distribution of the target. (a,b) Classification tasks – (c,d) Regression tasks.

C. Details on classification and regression models

We considered an MLP-NN as a representative of an "opaque" model to be learned in an FL fashion. The MLP-NN has two hidden layers, each with 128 neurons and ReLu activation, followed by an output layer. For the binary classification tasks, a sigmoid activation function is considered along with the BinaryCrossentropy loss. For the regression tasks, the output layer consists of a linear unit and the Mean Squared Error (MSE) loss is considered. Adam is adopted as optimizer [25]. In the FL setting, we exploit the classical FedAvg as aggregation strategy [2], and we set the minibatch size to 64, the number of local epochs to 5, and the overall number of federation rounds to 20 for classification and 80 for regression tasks. It is worth underlining that pursuing the best possible performance metrics is not the main scope of this work: we merely searched for a configuration that allows the FL model to address the supervised learning task while achieving satisfactory performance, without any thorough optimization of the hyperparameters. To ensure that the explainability analysis is valid and meaningful, we ascertain that FL models achieve reasonable performance. We considered accuracy and F1-score as metrics for classification tasks, R^2 and RMSE as metrics for regression tasks.

VI. Experimental Results

Table III reports the performance metrics of the MLP-NN on the test set. The models learned in an FL fashion are compared with models learned via a traditional centralized setting (CL), where the whole training set is assumed to be available on a single server. Clearly, this only serves as a baseline and is not applicable when privacy preservation is mandatory since it requires data sharing.

TABLE III
Performance metrics on the test set. FL and CL indicates the Federated and the Centralized Learning setting, respectively.

	FL	**CL**	**FL**	**CL**
	Accuracy		F1-score	
Magic	0.87	0.89	0.90	0.92
Rice	0.90	0.90	0.88	0.88
	RMSE		R^2	
California	0.80	0.56	0.55	0.78
Abalone	2.09	2.01	0.60	0.63

For all datasets, the results of the FL model highlight that the supervised learning tasks are successfully addressed, even if it is generally outperformed by the CL counterpart. The reasons for this gap can be manifold: first, the CL setting considers the availability of the entire batch of training data; second, no optimization of the FL configuration parameters was performed, including the choice of an aggregation strategy different from FedAvg which may suffer from poor convergence on non-i.i.d. data [26]. In any case, this slight drop in metrics does not undermine the significance of the explainability analysis applied to the FL model, which will be the focus of the following sections. Specifically, we first provide a fine-grained analysis on one of the datasets (Abalone) and then report a thorough discussion of the outcomes derived from the proposed and alternative approaches considered in this work for federated explanations with SHAP on all datasets.

A. Interpreting the results: the case of Abalone dataset

The Shapley values reveal, a-posteriori, the additive importance score of each feature for each prediction performed by the model. Figure 4 illustrates a very common way of representing Shapley values, based on a randomly selected example from the Abalone test set.

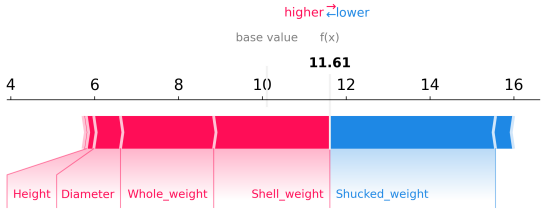


Fig. 4. SHAP force plot. Visual representation of Shapley values for instance #3259 of Abalone test set.

The force plot in Fig. 4 shows the contribution of each feature to the prediction. The base value, computed as the average of predicted values for the examples in the background dataset, is $\phi_0 = 10.11$. Features represented in pink have a positive contribution, whereas those in blue have a negative one. In the example, the largest contributions to the predicted value $\hat{y} = 11.61$ come from *Shell_weight* and *Whole_weight* (which contribute positively: $\phi_{Shell\_weight} = 2.77$, $\phi_{Whole\_weight} = 2.22$) and from *Shucked_weight* (which contributes negatively: $\phi_{Shucked\_weight} = -3.95$).

Let $\mathbf{\Phi}$ be the $N_{test} \times F$ matrix of Shapley values obtained for all the instances in the test set. In the following, the subscript is used to indicate the approach adopted for the estimation of the Shapley values. A compact way of representing the $\mathbf{\Phi}$ matrix is shown in Fig. 5.

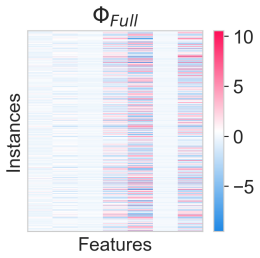
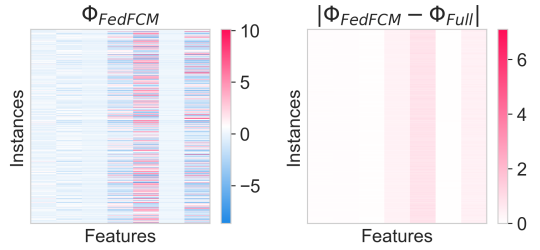


Fig. 5. Shapley values for the Abalone test set, estimated with the *Centralized Full* approach.

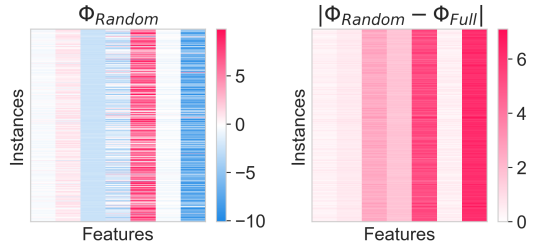
The heatmap displays the $\mathbf{\Phi}_{Full}$ matrix, whereby Shapley values have been estimated using the full training set (union of local training sets) as background dataset. Figure 5 provides an early glimpse of the feature importance on the test set. However, further insight can be gained by comparing the SHAP matrices obtained with different approaches.

Figure 6 shows a comparison of both the *FedFCM* approach (Fig. 6a) and the *Random* approach (Fig. 6b) with the baseline approach, represented by centralized *Full*. For the sake of brevity, only one of the ten repetitions is presented.

For each approach, in addition to the $\mathbf{\Phi}$ matrix (left), also the element-wise absolute value of the difference between $\mathbf{\Phi}$ and $\mathbf{\Phi}_{Full}$ is reported (right). A visual analysis of the heatmaps suggests that the *FedFCM* approach shows a very similar pattern compared to the centralized *Full* approach. This is also reflected in the relatively low values of the difference matrix: using a limited number of instances obtained via Federated



(a) Shapley values obtained with *FedFCM* (left) and comparison with *Full* (right).



(b) Shapley values obtained with *Random* (left) and comparison with *Full* (right).

Fig. 6. Shapley values for the Abalone test set.

FCM as background dataset provides a good approximation of what would be obtained if the entire training set were available. This outcome is somehow expected: the Federated FCM coincides with the traditional FCM applied on the entire dataset (as discussed in Section IV) and the application of summarization techniques to reduce the numerosity of the background dataset is a widely established practice [7], [11]. Conversely, the randomly generated synthetic background leads to a matrix of Shapley values $\mathbf{\Phi}_{Random}$ that diverges substantially from the centralized baseline $\mathbf{\Phi}_{Full}$, meaning that such explanations can be qualified as inaccurate. As anticipated in Section IV-B, the rationale lies in the fact that the background dataset is not representative of the actual distribution of the data.

A quantitative assessment of the difference between $\mathbf{\Phi}$ matrices can be obtained by resorting to a matrix norm. Without loss of generality, here we consider the Frobenius norm, which is defined as follows for a generic matrix A with s rows and t columns:

$$\|\boldsymbol{A}\|_{\mathrm{F}} = \sqrt{\sum_{i}^{s}\sum_{j}^{t}|a_{ij}|^2} \tag{8}$$

With reference to the matrix reported in Fig. 6, we obtain $\|\mathbf{\Phi}_{FedFCM} - \mathbf{\Phi}_{Full}\|_{\mathrm{F}} = 19.5$ and $\|\mathbf{\Phi}_{Random} - \mathbf{\Phi}_{Full}\|_{\mathrm{F}} = 176.1$. The distance assessment confirms the superiority of the approach based on Federated FCM, as it is obviously desirable to minimize the Frobenius norm in the comparison with the centralized case, used as a baseline. However, the numerical evaluation does not enable a straightforward understanding of whether – and to what extent – the explanations are actually consistent or not. In other words, we are unable to judge a

priori how the discrepancy $\|\Phi_{FedFCM} - \Phi_{Full}\|_{\mathrm{F}} = 19.5$ is reflected in the actionable explanations provided by the system to any stakeholder.

To shed light on the perceived difference in the explanations, we refer to the specific instance of the test set for which the distance between the Shapley values obtained with *FedFCM* and with *Full* is maximum (ID #3140). Figure 7 reports the Shapley values of the three approaches (*Full*, *FedFCM* and *Random*) for such an instance.

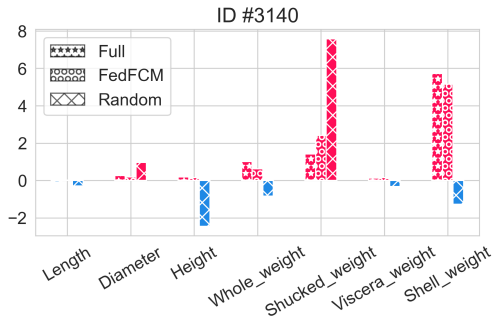


Fig. 7. Shapley values obtained with *Full*, *FedFCM* and *Random* approaches on the instance #3140 of Abalone dataset, for which the distance between the Shapley values obtained with *FedFCM* and with *Full* is maximum.

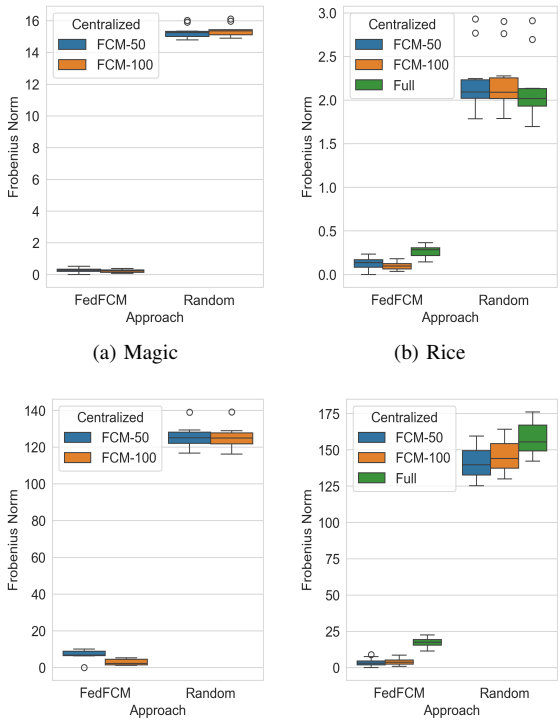
In the example where the *FedFCM* approach deviates more from the *Full* centralized one, their explanations are however substantially consistent. Both identify *Shell_weight* as the most influential feature and, in general, the sign and rank of feature contributions are always preserved. This is not the case for explanations obtained with the *Random* approach: the *Shucked_weight* feature has the major positive contribution and several features (*Height*, *Whole_weight* and *Shell_weight*) have opposite values of importance compared to the other approaches. We recall that the model is the same for all three approaches and therefore the predicted value is also the same.

The analysis therefore confirms that the choice of the background dataset is a critical aspect of the KernelSHAP method and highlights that the approach based on federated clustering allows for accurate explanations in the FL setting.

B. Numerical results on Classification and Regression Datasets

In this section, the numerical results concerning the four datasets described in Section V-B are reported.

First, we measure the discrepancy of both the *FedFCM* and the *Random* approach with the baseline centralized approaches (*Full*, *FCM-100* and *FCM-50*) in terms of Frobenius norm of the pairwise difference of $\mathbf{\Phi}$ matrices. For each comparison, ten values of the norm are obtained from as many trials with different random seeds. Figure 8 shows the results in the form of boxplots. It is worth noticing that the evaluation of *Full* centralized KernelSHAP on the two largest datasets (Magic and California) was not possible: using the entire training set as a background to obtain explanations on the entire test set proves to be prohibitively time consuming even on relatively sophisticated computational units (Apple M1 Pro, 16-GB).



(a) Magic (b) Rice

(c) California (d) Abalone

Fig. 8. Boxplots of the discrepancy of both the *FedFCM* and the *Random* approach with the baseline centralized approaches in terms of Frobenius norm of the pairwise difference of $\mathbf{\Phi}$ matrices.

Results confirm the general validity of the preliminary outcome observed for the Abalone dataset in Section VI-A. Explanations obtained with federated clustering procedure are much closer to the ideal centralized case compared to those obtained with a random background dataset. The absolute norm values are higher in regression tasks than in classification tasks, as only in the latter case the predictions (and indeed also the explanations) are bounded in the unit interval [0,1]. Also, the value is not normalized for the size of the test set, which varies among datasets. Interestingly, different random initialization of centers does not entail significant variability of the Frobenius norm for the *FedFCM* approach.

Table IV reports the Frobenius norm of the difference between $\mathbf{\Phi}$ matrices obtained with the centralized approaches for each dataset.

TABLE IV
FROBENIUS NORM OF THE DIFFERENCE BETWEEN CENTRALIZED APPROACHES FOR EACH DATASET.

	MA	**RI**	**CA**	**AB**
$\|\mathbf{\Phi}_{Full} - \mathbf{\Phi}_{FCM-100}\|_{\mathrm{F}}$	*NA*	0.2	*NA*	14.4
$\|\mathbf{\Phi}_{Full} - \mathbf{\Phi}_{FCM-50}\|_{\mathrm{F}}$	*NA*	0.2	*NA*	19.5
$\|\mathbf{\Phi}_{FCM-100} - \mathbf{\Phi}_{FCM-50}\|_{\mathrm{F}}$	0.3	0.1	5.2	5.4

It can be noticed that the values of the norm are quite limited and comparable to the discrepancy observed for *FedFCM* approach in Fig. 8. The maximum value observed for the

Abalone dataset (19.5 for *FedFCM* vs *Full*) is coherent with the one reported in Section VI-A, which was shown to have no relevant influence on the explanations even for the most impactful record in the test set. It is worth underlining that increasing the size of the background entails an increased computational cost. Table V shows the runtime for the centralized approaches, where their execution was feasible.

TABLE V
RUNTIME IN SECONDS OF THE CENTRALIZED APPROACHES FOR EACH DATASET.

	MA	**RI**	**CA**	**AB**
$\mathbf{\Phi}_{Full}$	*NA*	2276	*NA*	2735
$\mathbf{\Phi}_{FCM-100}$	2957	94	1097	97
$\mathbf{\Phi}_{FCM-50}$	1666	53	735	58

Although an accurate estimate would require repeated testing, the values of runtime obtained for single trials are indicative of the dependence on the size of the background and of the test set.

C. Consistency analysis of local explanations

The property of consistency in FL setting, as introduced in [16], is achieved when distinct participants receive identical explanations for an output generated by the FL model given identical input instances. The numerical results demonstrate that the estimation of the Shapely values obtained with the *FedFCM* approach are a faithful approximation of those obtained with the *Centralized* approaches. Furthermore, the centers derived with such a privacy preserving procedure constitutes a collaboratively built background dataset that can be shared to any party, thus ensuring consistency of explanations.

As summarized in Table I, *Local* approaches are data-driven and privacy preserving, but unlike the *FedFCM* approach they do not ensure consistency of explanation. The misalignment of client-side explanations obtained with KernelSHAP is due to the fact that the background datasets (derived from local training sets) vary from client to client and therefore is expected to be particularly evident in non-i.i.d. settings.

In this section, we first provide a fine-grained consistency analysis of the explanations for the Abalone dataset and then we discuss the outcomes for the four datasets. For the purpose of this analysis, we assume that the same set of instances (that is, the test set) is available to each client.

Figure 9 reports the Shapley values for the test instance #3259 (the same of Fig. 4, taken as an example) estimated with the *FedFCM* approach (black bar) and the *Local* approach.

The barplot highlights that the local explanations exhibit large variability among the clients, which is ascribed to the fact that the local training sets are not identically distributed as shown in Figure 3d. As a consequence, the same individual prediction from the same model is explained by different relative feature importance. This is particularly evident when comparing explanations for Client 0 and Client 9, whereby the average value of the target variable *Rings* in the training set is

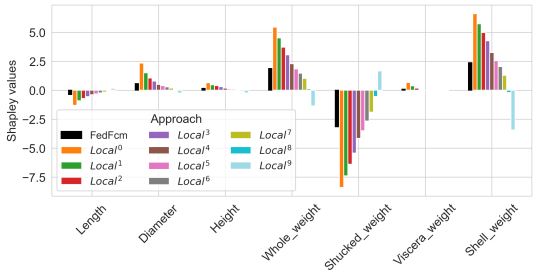
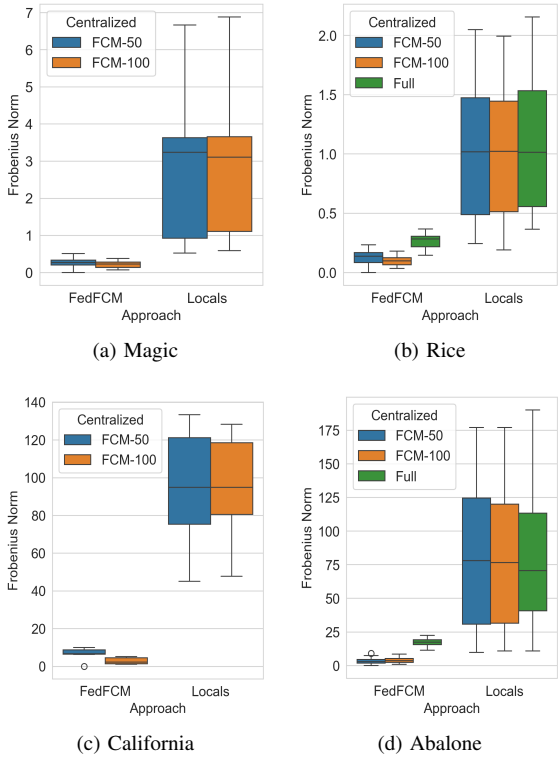


Fig. 9. Shapley values for instance #3259 of Abalone dataset for each client.

lowest and highest, respectively, and resulting Shapley values have always opposite signs.

Finally, we report the discrepancy of each *Local* approach with the baseline centralized approaches (*Full*, *FCM-100* and *FCM-50*) in terms of Frobenius norm of the pairwise difference of the matrices containing the Shapley values (Fig. 10). For the sake of completeness, we also report on the same plot the boxplots for the *FedFCM* approach.



(a) Magic (b) Rice

(c) California (d) Abalone

Fig. 10. Boxplot of the discrepancy of both the *FedFCM* and the *Local* approach with the baseline centralized approaches in terms of Frobenius norm of the pairwise difference of $\mathbf{\Phi}$ matrices.

As expected, the variability of the *Local* differences from the centralized approaches is very pronounced and significantly higher than the ones obtained with *FedFCM*. The lower whiskers of the *Local* boxplots are found at low Frobenius norm values, comparable with those of the *FedFCM* case (except for the California dataset). However, the matrices of

explanations generated with the *Local* strategy are in general more distant from the *Centralized* ones.

VII. Conclusion

In this paper we have proposed an approach for simultaneously addressing two requirements towards trustworthy AI: data privacy preservation during the learning stage and explainability of the resulting model. The Federated Learning (FL) paradigm enables collaborative model learning in a privacy preserving manner. However, most existing FL solutions revolve around Deep Learning and Neural Network (NN) models which are generally considered opaque, i.e. hard to interpret. Several approaches have been proposed for the post-hoc explainability of such models but their adaptation to the FL setting is not straightforward. We focused on the one of the most popular post-hoc methods, namely the SHapley Additive exPlanation (SHAP) method, and designed a novel approach for obtaining accurate and consistent explanations in the federated setting. In our view, an explanation is accurate if it coincides with the one that would have been obtained if the scattered training data could have been put together, thus relaxing the privacy requirement. An explanation is consistent if, given the same input instance, FL model and predicted output, any client obtains the same explanation.

The crux of the problem consists in properly designing a common background dataset, which is required by SHAP for calculating the individual explanations. Our proposal consists in executing a federated clustering procedure over scattered participants local data as a data summarization technique: the resulting cluster representatives constitute the common background dataset for the execution of the SHAP method. We resorted to the KernelSHAP variant of SHAP to explain the individual predictions of a Multi Layer Perceptron Neural Network (MLP-NN) learned in an FL fashion. As a federated clustering procedure we adopted a recently proposed federated version of the popular Fuzzy C-means (FCM) algorithm.

We applied the proposed methodology to four open-access datasets suitably partitioned to simulate an FL setting, covering both classification and regression tasks. The analysis of the results shows how the explanations provided by the proposed Federated SHAP method are fairly accurate, compared to the centralized case. We also compared our approach based on federated clustering with two alternative approaches: the *Random* approach which generates the background dataset by randomly sampling from a uniform distribution over the input space, and the *Local* approach where each client generates its own background dataset by applying the FCM clustering algorithm to its local training set. The experimental results show that the former approach lacks accuracy while the latter lacks both accuracy and consistency. Future work will broaden the analysis over different dimensions: generalizing the proposed approach for handling also categorical variables, and investigating the adoption of other post-hoc explainability methods in the FL setting, possibly involving different data types (e.g., images and texts) and different tasks (e.g., time series prediction).

References

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

[2] B. McMahan, E. Moore *et al.*, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. of the 20th Int'l Conf. on Artificial Intelligence and Statistics*, vol. 54. PMLR, 2017, pp. 1273–1282.

[3] A. Barredo Arrieta, N. Díaz-Rodríguez *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inform. Fusion*, vol. 58, pp. 82–115, 2020.

[4] S. Ali, T. Abuhmed *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Inform. Fusion*, vol. 99, p. 101805, 2023.

[5] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Adv. Neur. In.*, vol. 30, 2017.

[6] L. S. Shapley, *17. A Value for n-Person Games*. Princeton University Press, 1953, pp. 307–318.

[7] P. Biecek and T. Burzykowski, *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021.

[8] A. Wilbik and P. Grefen, "Towards a Federated Fuzzy Learning System," in *2021 IEEE Int'l Conf. on Fuzzy Systems (FUZZ-IEEE)*, 2021, pp. 1–6.

[9] W. Pedrycz, "Federated FCM: Clustering Under Privacy Requirements," *IEEE T. Fuzzy Syst.*, pp. 1–1, 2021.

[10] J. L. Corcuera Bárcena, F. Marcelloni *et al.*, "A federated fuzzy c-means clustering algorithm," in *Int'l WS on Fuzzy Logic and Applications 2021*, vol. 3074, 2021, pp. 1–9.

[11] C. Molnar, "Interpretable machine learning: a guide for making black box models explainable," 2023.

[12] Z. Liu, Y. Chen *et al.*, "GTG-Shapley: Efficient and Accurate Participant Contribution Evaluation in Federated Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 4, may 2022.

[13] J. L. Corcuera Bárcena, P. Ducange *et al.*, "An Approach to Federated Learning of Explainable Fuzzy Regression Models," in *2022 IEEE Int'l Conf. on Fuzzy Systems (FUZZ-IEEE)*, 2022, pp. 1–8.

[14] M. Daole, A. Schiavo *et al.*, "OpenFL-XAI: Federated Learning of Explainable Artificial Intelligence Models in Python," *SoftwareX*, vol. 23, p. 101505, 2023.

[15] X. Zhu, D. Wang *et al.*, "Horizontal Federated Learning of Takagi–Sugeno Fuzzy Rule-Based Models," *IEEE T. Fuzzy Syst.*, vol. 30, no. 9, pp. 3537–3547, 2022.

[16] A. Bogdanova, A. Imakura *et al.*, "DC-SHAP Method for Consistent Explainability in Privacy-Preserving Distributed Machine Learning," *Human-Centric Intelligent Systems*, vol. 3, no. 3, pp. 197–210, 2023.

[17] R. López-Blanco, R. S. Alonso *et al.*, "Federated Learning of Explainable Artificial Intelligence (FED-XAI): A Review," in *Distributed Computing and Artificial Intelligence, 20th Int'l Conf.*, 2023, pp. 318–326.

[18] P. Chen, X. Du *et al.*, "EVFL: An explainable vertical federated learning for data-oriented Artificial Intelligence systems," *J Syst. Architect.*, vol. 126, p. 102474, 2022.

[19] J. Fiosina, "Explainable Federated Learning for Taxi Travel Time Prediction," in *Int'l Conf. on Vehicle Technology and Intelligent Transport Systems, VEHITS - Proceedings*, vol. 2021-April, 2021, p. 670 – 677.

[20] Y. Chen, X. Yang *et al.*, "FedDBM: Federated Digital Biomarker for Detecting Parkinson's Disease Progress," in *2023 IEEE Int'l Conf. on Multimedia and Expo (ICME)*, 2023, pp. 678–683.

[21] L. Corbucci, R. Guidotti *et al.*, "Explaining Black-Boxes in Federated Learning," in *Explainable Artificial Intelligence*, 2023, pp. 151–163.

[22] K. N. Markelle Kelly, Rachel Longjohn, "The UCI machine learning repository." [Online]. Available: https://archive.ics.uci.edu

[23] L. Torgo, "California housing dataset," accessed January 2024. [Online]. Available: https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html

[24] P. Kairouz, H. B. McMahan *et al.*, "Advances and Open Problems in Federated Learning," pp. 1–210, 2021. [Online]. Available: http://dx.doi.org/10.1561/2200000083

[25] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd Int'l Conf. on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[26] T. Li, A. K. Sahu *et al.*, "Federated Optimization in Heterogeneous Networks," in *Proceedings of Machine Learning and Systems*, vol. 2, 2020, pp. 429–450.