



# Federated Learning of XAI Models in Healthcare: A Case Study on Parkinson's Disease

Pietro Ducange<sup>1</sup> · Francesco Marcelloni<sup>1</sup> · Alessandro Renda<sup>1</sup> · Fabrizio Ruffini<sup>1</sup>

Received: 15 December 2023 / Accepted: 23 July 2024 / Published online: 28 August 2024  
© The Author(s) 2024

## Abstract

Artificial intelligence (AI) systems are increasingly used in healthcare applications, although some challenges have not been completely overcome to make them fully trustworthy and compliant with modern regulations and societal needs. First of all, sensitive health data, essential to train AI systems, are typically stored and managed in several separate medical centers and cannot be shared due to privacy constraints, thus hindering the use of all available information in learning models. Further, transparency and explainability of such systems are becoming increasingly urgent, especially at a time when “opaque” or “black-box” models are commonly used. Recently, technological and algorithmic solutions to these challenges have been investigated: on the one hand, federated learning (FL) has been proposed as a paradigm for collaborative model training among multiple parties without any disclosure of private raw data; on the other hand, research on eXplainable AI (XAI) aims to enhance the explainability of AI systems, either through interpretable by-design approaches or post-hoc explanation techniques. In this paper, we focus on a healthcare case study, namely predicting the progression of Parkinson's disease, and assume that raw data originate from different medical centers and data collection for centralized training is precluded due to privacy limitations. We aim to investigate how FL of XAI models can allow achieving a good level of accuracy and trustworthiness. Cognitive and biologically inspired approaches are adopted in our analysis: FL of an interpretable by-design fuzzy rule-based system and FL of a neural network explained using a federated version of the SHAP post-hoc explanation technique. We analyze accuracy, interpretability, and explainability of the two approaches, also varying the degree of heterogeneity across several data distribution scenarios. Although the neural network is generally more accurate, the results show that the fuzzy rule-based system achieves competitive performance in the federated setting and presents desirable properties in terms of interpretability and transparency.

**Keywords** Federated learning · Explainable artificial intelligence · Fed-XAI · Granular fuzzy models · Parkinson's disease

## Introduction and Motivations

The extensive reliance on artificial intelligence (AI) and machine learning (ML) tools in the healthcare sector poses significant challenges, especially concerning the concept of

*trust*. Any AI system must meet the requirements of robustness, fairness, and transparency throughout its whole life cycle. Furthermore, sensitive health-related data hold an intrinsic value and become a lucrative target for cyber attacks.

The concept of *trustworthy AI* has recently been considered also by government entities: European Union, for example, is at the forefront for AI regulation as witnessed by the proposal of the “AI ACT”<sup>1</sup> (2021), which is often referred to as the first law on AI and is conceived for introducing a common regulatory and legal framework for AI. The European Commission had previously promoted the definition of the “Ethics guidelines for trustworthy AI” [1], which identifies lawfulness, ethics, and robustness as key pillars for trustworthiness and describes the requirements for an AI

---

✉ Alessandro Renda  
alessandro.renda@unipi.it

✉ Fabrizio Ruffini  
fabrizio.ruffini@unipi.it

Pietro Ducange  
pietro.ducange@unipi.it

Francesco Marcelloni  
francesco.marcelloni@unipi.it

<sup>1</sup> Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, Pisa 56122, Italy

<sup>1</sup> <https://artificialintelligenceact.eu/the-act/>, visited May 2024.

system to be deemed trustworthy. The ethical aspects are pivotal in the healthcare domain given the sensitive nature of patient data, the disclosure of which poses serious risks. For example, discrimination based on such data can occur in the insurance field: insurance companies could decide to charge different fees depending on the individual health status. Likewise, securing the non-discrimination in financial services is nowadays perceived as an important matter, as witnessed for example by the regulations on the “right to be forgotten” for cancer survivors [2]. Finally, special attention should be paid to specific domains (e.g., mental health), due to the stigmatized nature of some types of illness.

In the pursuit of trustworthiness, data privacy and transparency emerge as pivotal enablers, especially in the healthcare domain. While data privacy is considered an invaluable right, it somehow clashes with the creation of accurate ML models that, to date, require large amounts of data in their training phase. The common scenario is in fact that many different entities (be they individuals, medical centers, or hospitals) have few or limited amounts of data and are often reluctant to share their assets and sensitive information with other parties. The processes of data mining and knowledge extraction are therefore hampered by the unfeasibility of data collection for centralized processing. The requirement of *transparency* encompasses the traceability of the learning process, beginning from the data gathering phase, and the ability to comprehend the structure and the functioning of the ML model itself. The latter challenge is the central focus of a branch of AI named Explainable AI (XAI) [3–6]. The right for explanation is explicitly mentioned both in the “Ethics guidelines for trustworthy AI” [1], “[...] AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned”, and in the recital 71 of the General Data Protection Regulation (GDPR) [2] “[...] the processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.” Indeed, the ability to understand the inner working of an AI system represents a cornerstone of trust and holds particular significance in high-stakes applications in the healthcare domain.

In this work, we embrace the challenge of enhancing trustworthiness of AI systems in medicine investigating technical enablers for the requirements of data privacy and explainability.

Data access limitations, driven by privacy requirements and by the need to prevent ethical risks associated with the disclosure of sensitive data, have prompted the development of new paradigms for training ML models, including federated learning (FL) [7, 8]. FL enables multiple parties to collaboratively train an ML model without any disclosure of private raw data. Essentially, a shared global model is learned

through proper aggregation of locally computed updates from remote data owners, thus removing the need of centralizing data for training purposes.

The requirement of explainability is typically addressed through two main categories of approaches [3, 4]: the exploitation of interpretable by-design models and the adoption of post-hoc explainability techniques. Interpretability and transparency refer to inherent properties of a model and consist in the ability to understand *how* decisions have been taken and what is the structure of the model itself, respectively. Post-hoc explanation techniques, instead, typically address the goal of explaining *why* a model provides a decision. It follows that an interpretable model results to be also explainable. The property of interpretability is generally attributed to models such as decision trees (DTs) and rule-based systems (RBSs): in fact, they consist of (or can be traced back to) collection of “IF *antecedent* THEN *consequent*” rules. As a consequence, the inference process turns out to be highly understandable. Understandability can be defined as “*the characteristic of a model to make a human understand its function (that is, how the model works) without any need for explaining its internal structure or the algorithmic means by which the model processes data internally*” [3]. It is worth emphasizing that the concept of understandability is strongly related to the targeted audience in terms of their a-priori knowledge and cognitive skills. For example, a rule-based inference requires familiarity with logic and, possibly, an adequate training depending on the specific implementation adopted in the antecedent and consequent parts of the rules.

Post-hoc techniques are applied on models which are referred to as opaque or “black boxes,” such as Neural Networks (NNs) and ensemble models, to explain their outcomes. The vast majority of existing post-hoc methods can be roughly ascribed to the following categories, which emulate different nuances of human reasoning: feature importance explanations, rule-based explanations, prototypes explanations, contrastive/counterfactual explanations, and textual or visual explanations [3, 9]. Notably, since the field of research is constantly evolving, the list should not be considered exhaustive; in addition, different post-hoc strategies can be specifically tailored based on the different kinds of data (e.g., images or text) and based on the specific models involved [9]. In the context of XAI, a further distinction is also made between *local* and *global* explanations: the former refers to the inference process and focuses on how/why the decision is taken for any single input instance. The latter refers to structural properties of the models (thus pertaining to the concept of transparency) or to aggregated information computed over the entire dataset.

The awareness of the importance of explainability and privacy preservation has greatly increased in recent years. While FL inherently tackles the challenge of preserving data privacy

in decentralized ML, it typically lacks integrated solutions for the issue of explainability [10]. Actually, FL was originally conceived for models optimized through stochastic gradient descent (SGD) (e.g., Deep Neural Networks (DNNs)), and the application of post-hoc techniques is not straightforward in the federated setting.

This work lies at the intersection between FL and XAI and contributes to a research area named Fed-XAI (acronym for *federated learning of XAI models*) [11–13]. We explore the adoption of Fed-XAI approaches within the healthcare domain for predicting the progression of Parkinson’s disease (PD), formulated as a regression task. We consider a plausible scenario where sensitive raw data originate from different medical centers, making centralized learning unfeasible. In particular, the task is to predict one of the most commonly used indicators for the severity of PD symptoms, namely the Unified PD Rating Scale (UPDRS, firstly introduced in 1987 [14]), by exploiting real-world voice recordings. The analysis extends a recent work [15] and encompasses two approaches for Fed-XAI in order to explore the trade-off between accuracy and trustworthiness. The first approach adopts an interpretable by-design model, learned in a federated fashion. The second one employs an opaque model, where both training and post-hoc explanation are compliant with the federated setting. On one hand, we can assess the generalization capability of the models in the regression problem by exploiting a real, publicly available dataset; on the other hand, trustworthiness is meant here as the concurrent attainment of explainability in all its nuances, from both local and global perspectives, and privacy preservation through the adoption of FL.

As for the interpretable by-design model, we employ the Takagi-Sugeno-Kang Fuzzy Rule-Based system (TSK-FRBS) [16] which is considered as a transparent and interpretable model: its inference method mimics a cognitive process typical of human reasoning in the form of *if-then rules*. The partitioning of numerical variables into fuzzy sets, which is one of the defining aspects of TSK-FRBS, has proven to enable competitive levels of performance for classification and regression tasks [17] and has a twofold implication. First, fuzziness in rule-based systems enhances semantic interpretability through linguistic representation of numerical variables. Second, a fuzzy set can be interpreted as a formal representation of an information granule, intended as a generic and conceptually meaningful entity [18, 19]: in this context, a fuzzy set allows any number in the real unit interval to represent the membership degree of a feature value to the information granule. As a consequence, the adopted TSK-FRBS fits into the paradigm of granular computing and makes use of information granules in the explainable decision-making pipeline.

As for the opaque model, we employ a well-known biologically inspired model, namely Multi Layer Perceptron Neural

Network (MLP-NN). FL is performed exploiting the popular federated averaging (FedAvg) aggregation strategy [7]. Furthermore, the SHAP explainer [20], purposely adapted to comply with the federated setting based on a recently proposed approach [21], is used for explaining the output of the MLP-NN by attributing the contribution (i.e., importance) of each feature to each prediction.

The main contributions of this work can be summarized as follows:

- We simulate a scenario in which several medical institutions cooperate in creating a PD progression prediction model pursuing the requirements of explainability and privacy preservation;
- To achieve this goal, we implement and exploit two Fed-XAI approaches, based on TSK-FRBS, and MLP-NN plus SHAP, which represent state-of-art techniques for by-design interpretability and post-hoc explainability, respectively, in the federated setting;
- We discuss the accuracy of the approaches under several data distribution scenarios, considering the independent and identically distributed (i.i.d.) case and three different non-i.i.d. cases;
- For each scenario, we compare the FL scheme with two baselines, namely centralized learning and local learning, to verify the suitability of the federated approach;
- We discuss the explainability of MLP-NN and the interpretability of TSK-FRBS, both from a local and a global perspective;
- We discuss about the consistency of explanations provided by the two approaches in the federated setting. Here, consistency is achieved when different participants in the FL process obtain the same explanation given the same input information.

The rest of the paper is organized as follows: in Section 2, we provide a brief overview of recent works that adopt XAI and FL tools in healthcare and more specifically in the context of PD. Furthermore, we describe recent advances in the field of Fed-XAI. Section 3 describes the background related to FL, detailing the approaches for FL of TSK-FRBS and FL of MLP-NN. Furthermore, SHAP is introduced as post-hoc explainability technique, and a recent approach for exploiting SHAP in the federated setting is presented. Section 4 describes the PD progression prediction case study, providing details about the experimental setup: we outline the different data distribution scenarios, the evaluation strategies, and the configuration of the two approaches. In Section 5, we report and discuss the experimental results. The considerations regarding interpretability and explainability are given in Section 6. Finally, in Section 7, we draw some conclusions.

## Related Works

The adoption of AI techniques in healthcare has been widely investigated. In this section, we first discuss the most relevant works concerning the adoption of FL and XAI in this application domain. Then, we discuss existing works related to XAI in Parkinson's disease studies. Finally, we discuss recent algorithmic efforts for combining FL paradigm and XAI approaches.

### Federated Learning and XAI in Healthcare Scenarios

The opportunities and the practical utility of FL in the healthcare domain have been recently acknowledged in the specialized literature [22–24], with applications mainly in the fields of medical imaging [25] and precision medicine [26]. FL is presented as a solution to protect sensitive data for privacy concerns and ethical constraints [27] and also in relation to cyber attacks [28]. At the same time, the interest in XAI is increasingly widespread, especially in the attempt to “open” the so-called black-boxes [29, 30], which have enabled unprecedented performance in the field of deep learning (DL) in medicine. The surveys on XAI for healthcare applications usually delve into the problem of how to present the AI results and their explanations to physicians, medical staff, patients, and caregivers: the explanations should be a tool to understand the outcomes of an AI system, but also a way to allow interaction and enhance stakeholders' trust in AI (human-centered AI). The XAI goal is usually achieved through the adoption of post-hoc methods for opaque models, often concerning image data analysis (e.g., X-rays and CT scans).

Authors of a recently published survey [10] provide a review of clinical cases where post-hoc methods and interpretable by-design models are applied to more than 20 different medical case studies, spanning from COVID-19 diagnosis and early detection to diagnostic for breast cancer. Different data types are exploited, depending on the application: images (e.g., EEG, MRI) are often involved, and SHAP is among the most used post-hoc methods. Example case studies include prediction of depressive symptoms from texts with adoption of a post-hoc method for the estimation of word importance [31] and Alzheimer classification using Random Forest and SHAP [32].

In the same survey [10], the practical utility of FL in healthcare applications is discussed, especially considering DL approaches, horizontal data partitioning, and FedAvg optimization strategy. Example case studies include the detection of COVID-19 from decentralized medical data, with Convolutional Neural Networks applied on anterior and posterior chest X-rays [33]. Authors in [34] exploit tabular electronic health records (demographics, past medical history, vital signs, lab tests results) from five hospitals to pre-

dict mortality in patients diagnosed with COVID-19 within a week of hospital admission. However, it is worth noticing that the applications of FL and XAI are treated separately, emphasizing the substantial lack of works that simultaneously address the requirements of privacy through FL and transparency through XAI in the healthcare domain.

### XAI in Parkinson's Disease Studies

PD is diagnosed in about 10 million people worldwide [35]: after the Alzheimer, it is one of the most prevalent neurodegenerative diseases. Given its socioeconomic relevance, several AI methods have been proposed for supporting diagnosis and monitoring [35, 36]. The most commonly used data types exploited for PD studies include images and speech signals [37, 38].

A few works discuss the topic of explainability in the context of PD studies supported by AI techniques: for instance, authors in [39] apply the LIME [40] post-hoc method on a DNN used to classify healthy from not-healthy subjects using images from SPECT scanning. Authors in [41] provide explanations for different ML model outcomes using three post-hoc methods, namely LIME, SHAP, and SHAPASH (a tool for making ML models more understandable and interpretable for general audience), on a multiclass classification task. Since the aspect of data privacy holds high relevance in this context, few recent works elaborate upon the exploitation of the FL paradigm for PD-related applications [42–44].

In this work, we consider the Parkinson Telemonitoring dataset, which has been analyzed in several recent works for both classification [45, 46] and regression [47–49] tasks. None of the works mentioned above, however, considers the aspects of privacy and explainability simultaneously.

The primary goal of our analysis is to understand the potentialities of the Fed-XAI paradigm in a PD-related application. In the following, we provide a brief overview of the most relevant approaches for Fed-XAI proposed in the literature, relaxing the constraint on the application domain.

### Federated Learning of XAI Models

Explainability in FL has been pursued either using post-hoc [20, 50–55] or ex-ante [13, 56–59] approaches. A thorough review of such approaches has been provided in several recent works [11, 12, 60]. Here, we describe the most recent advances on the topic.

Bogdanova et al. [60] have proposed a novel approach (named DC-SHAP) for consistent explainability over both horizontally (different instances, same features) and vertically (different features, same instances) partitioned data for the Data Collaboration (DC) paradigm. This paradigm consists of two stages: first, participants obtain intermediate representations of data through irreversible transformations

and transmit them to a central server (unlike FL, which typically shares models rather than data). Then, the server combines such intermediate representations into a single dataset, trains an ML model, and distributes it back to the participants. Unlike other ex-ante [58] and post-hoc [50, 61] explainability approaches tailored for the decentralized setting, DC-SHAP ensures consistency of explanations: In this context, the property of consistency is met if the explanations of the same data instance for a global model are the same for different participants. As underlined by the authors in [60], model-agnostic post-hoc explainability methods are prone to misalignment of client-side explanations, since they rely on probing the global model with various inputs generated from the local data distribution (typically referred to as *background* or *reference dataset*). In their proposal for horizontally partitioned data, they use a set of auxiliary synthetic data shared among the participants to solve the issue of different background datasets and show how this allows the mitigation of feature attribution discrepancies among the participants. The approach proposed in [51] is conceived to obtain a consistent *global* feature attribution score for horizontal FL. A model-specific post-hoc explainability method, namely Integrated Gradients (IGs) [62], has been adopted for computing feature relevance. The integrated gradients get averaged and thus unified among the clients; however, local explainability is not addressed.

The issue of Fed-XAI for PD has been recently discussed in [63], with the aim of identifying digital bio-markers for the progress of the disease. Three assumptions constitute the privacy model, considering a scenario with multiple hospitals, each with its own patients: (i) input records and corresponding labels are isolated; (ii) the raw inputs are isolated between patients; (iii) the target labels are isolated between hospitals.

A hierarchical framework is adopted to build the FL model: local FL processes allow to collaboratively train a model among patients in the same hospital, whereas a global FL process aggregates models from each hospital for generating the complete model. An adaptation of SHAP is then adopted as post-hoc method for feature importance explanation. To address the issue of misalignment of client-side explanations, background datasets are generated sampling from a Gaussian distribution: the parameters of such Gaussian distribution (mean and variance) are estimated for each feature in a hierarchical way, by combining the parameters estimated intra- and inter-hospitals. It is shown that the average feature importance computed in the federated fashion is qualitatively similar to but quantitatively different from that obtained in the *centralized* fashion, where the union of the participants' training sets can be used as background dataset. Although the proposed method theoretically allows for it, the aspect of local explainability/interpretability is not however discussed.

Authors in [21] have proposed an approach for obtaining SHAP explanations [20] in horizontal FL. Specifically, the explanation of an instance prediction made by the federated ML model is obtained by aggregating the explanation of the participants. Such an approach ensures consistency of explanations and is shown to be a faithful approximation of the SHAP explanations obtained in a centralized setting. However, the approach requires that test instances are available to all participants, which may be undesirable or unfeasible in real-world applications where privacy must be guaranteed also at inference time.

In the framework of FL of interpretable-by-design models, TSK-FRBSs [13, 57, 59] and DTs [56, 58] have been considered as XAI models to be learnt in a federated fashion. Approaches proposed in [57, 59] for federated TSK-FRBS rely on a clustering procedure for the structure identification stage and on a federated adaptation of classical gradient-based learning schemes for adjusting the parameters of the consequent part of the rules. In this work, we consider the approach introduced in [13], which leads to more interpretable TSK-FRBSs compared to the ones considered in [57] and [59]. Additional details on such an approach are reported in Section 3.1.

As for DTs, the IBM FL framework [56] supports, among others, a federated adaptation of the ID3 algorithm for horizontally partitioned data. Specifically, an orchestrating server grows a single decision tree by exploiting client contributions based on their local data, in an iterative, round-based, procedure. Similarly, the approach proposed in [58] allows multiple clients to collaborate in the generation of a global DT by transmitting encrypted statistics, but it refers to the vertical data partitioning scenario. Finally, Polato et al. [64] have proposed a federated version of the AdaBoost algorithm, posing minimal constraints on the learning settings of the clients, enabling a federation of DTs, and without relying on gradient-based methods.

## Background

The categorization of FL approaches is typically based on the data partitioning scheme and the scale of federation. *Data partitioning* can be broadly categorized into horizontal and vertical settings. In the *horizontal* setting, training instances from different participants refer to the same set of features, whereas in the *vertical* setting, the feature set itself is partitioned among participants. The *scale of federation* refers to the number of participants and is typically classified into *cross-silo* FL, involving a low number of participants with ample data and computational power, and *cross-device* FL, where a large number of participants, often represented by

smartphones or personal equipment, may feature a relatively small amount of data and computational power.

The PD progress prediction case study discussed in this work pertains to a cross-silo horizontal FL setting. This section reports background information for the two approaches adopted to address this task, which can be ascribed to the Fed-XAI research field: federated TSK-FRBS and federated MLP-NN with post-hoc explainability.

### Federated Learning of TSK-FRBS

Let  $\mathbf{X} = \{X_1, X_2, \dots, X_M\}$  and  $Y$  be the set of  $M$  input variables and the output variable, respectively. A generic input instance is in the form  $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,M}]^T$  and has an associated output value  $y_i$ . Let  $U_j$  be the universe of discourse of variable  $X_j$  and  $P_j = \{A_{j,1}, A_{j,2}, \dots, A_{j,T_j}\}$  be a fuzzy partition over  $U_j$  with  $T_j$  fuzzy sets, each labeled with a linguistic term. The term  $A_{j,t}$  indicates the  $t^{\text{th}}$  fuzzy set of the fuzzy partition over the  $j^{\text{th}}$  input variable  $X_j$ . A TSK-FRBS consists of a collection of fuzzy *if-then* rules, where the antecedent part of each rule is a conjunction of fuzzy propositions and the consequent part implements a regression model. In case of the commonly used first-order regression model, the generic  $r^{\text{th}}$  rule is expressed as follows:

$$R_r : \mathbf{IF} \ X_1 \text{ is } A_{1,t_{r,1}} \ \mathbf{AND} \ \dots \ \mathbf{AND} \ X_M \text{ is } A_{M,t_{r,M}} \\ \mathbf{THEN} \ y_r = \gamma_{r,0} + \sum_{j=1}^M \gamma_{r,j} \cdot x_j \quad (1)$$

where  $\gamma_{r,j}$  (with  $j = 0, \dots, M$ ) are the coefficients of the linear model that evaluates the output prediction  $y_r$ .

The parameters of the rules are determined through a data-driven approach. The *if* part (antecedent) is generated either using grid-partitioning or fuzzy clustering over the input space. Once the antecedent is determined, the *then* part (consequent) estimation consists of local linear models obtained, for instance, through the least squares method.

At inference time, TSK-FRBS exploits the rule base as follows. Given an input instance, first the strength of activation of each rule is computed as

$$w_r(\mathbf{x}_i) = \prod_{j=1}^M \mu_{j,t_{r,j}}(x_{i,j}) \quad \text{for } r = 1, \dots, N_{\text{rule}} \quad (2)$$

where  $\mu_{j,t_{r,j}}(x_{i,j})$  is the membership degree of  $x_{i,j}$  to the fuzzy set  $A_{j,t_{r,j}}$ . Then, the final output can be evaluated with either *weighted average* or *maximum matching* policy. In the former case, the TSK-FRBS output is computed as the average of the outputs of all the activated rules weighted by their strengths of activation. In the latter case, the output

corresponds to the output of the rule with the maximum strength of activation.

The maximum matching policy enhances the interpretability of TSK-FRBS, since a single rule explains a predicted output for an input instance. Furthermore, the fuzzy linguistic representation of numerical variables fosters the semantic interpretability of the model itself, whose operation, based on the evaluation of rules, turns out to be highly intuitive.

From an algorithmic perspective, FL of TSK-FRBS, as well as of other families of highly interpretable models, requires ad-hoc strategies. In this work, we rely on the approach for building TSK-FRBSs in a federated fashion recently proposed in [13]. We consider horizontally partitioned data: every participant produces a local TSK-FRBS and sends it to the server. Subsequently, the server consolidates the received rule bases by juxtaposing the rules received from the participants and by resolving potential conflicts. A conflict occurs when rules from different local TSK-FRBSs have the same antecedent, thus identifying the same specific region of the input space, but they have different consequents. In this case, the federated TSK-FRBS summarizes conflicting rules in a single rule with the same antecedent as the conflicting rules and with the consequent obtained by computing the weighted average of the regression model coefficients in the consequents of the conflicting rules. Such average takes into account the weight associated with each rule, which is estimated on the local training set as the harmonic mean of its support (how many instances activate the rule), and confidence (average quality of the prediction of the rule).

With the aim of ensuring the consistency of the rules among participants and increasing the system interpretability, the input variables are partitioned by using a strong uniform fuzzy partition with triangular fuzzy sets. An example of strong uniform fuzzy partition with five triangular fuzzy sets is shown in Fig. 1. Here, each fuzzy set is associated with a meaningful label that is used to express linguistically the rules.

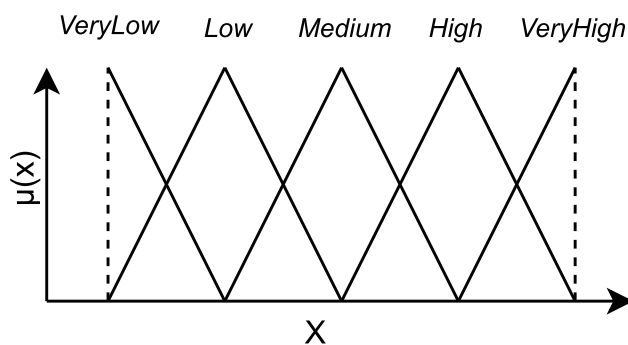


Fig. 1 An example of strong uniform fuzzy partition with five triangular fuzzy sets

It should be noted that building a fuzzy system requires a careful design especially regarding the choice of its hyperparameters (e.g., number, shape, and position of fuzzy sets for each partition), also considering their impact on interpretability [65]. Uniform fuzzy partitions with triangular fuzzy sets are generally deemed highly interpretable since they satisfy the criteria of coverage, completeness, distinguishability, and complementarity [66]. However, in practical applications, a meaningful partition should be agreed with the human users who are expected to interact with the AI system and to interpret the given rules. An interesting future development of this work would consist in examining the choice of the uniform partitioning together with domain experts (e.g., physicians).

### Federated Learning of MLP-NN

Models optimized through Stochastic Gradient Descent (SGD), such as NNs, can be learned in a federated fashion by exploiting an aggregation strategy based on, or derived from, the popular federated averaging (FedAvg) procedure. FedAvg is an iterative, round-based procedure, in which each round encompasses the following steps: the server sends the global model to the participants; each participant locally updates the model through SGD on its local training set and sends the updated model to the server; the server obtains an updated global model by computing the weighted average of the locally updated models, where weights are based on the local training set cardinality. Several extensions of FedAvg have been proposed in the literature, mostly aimed at addressing FL in heterogeneous settings [67, 68]. In this paper, we focus on classical FedAvg and deal in more detail with the issue of explainability of MLP-NN learned in a federated fashion. To this end, in the following, we first describe a popular post-hoc method for explainability, namely SHAP (SHapley Additive exPlanations) method [20], and then an approach for the adoption of SHAP in the federated setting.

#### Post-hoc Explainability: The SHAP Method

One of the most popular post-hoc strategies used to explain a model prediction is to assess the importance of each feature in producing the output. In general, given a model  $f$ , an input instance  $\mathbf{x}_i$ , and a predicted output  $\hat{y}_i = f(\mathbf{x}_i)$ , the explainer assigns to each input component  $x_{i,j}$  a value that reflects how much that particular feature is important for the prediction. These values are interpreted in terms of sign and magnitude: if the sign is positive (negative), then that feature contributes positively (negatively) to the prediction output; as per the magnitude, the larger it is, the higher the impact of the corresponding feature on the output.

In this work, we adopt the SHAP method [20] which is one of the most widely used approaches to assess the feature importance for both regression and classification tasks.

SHAP provides *local* explainability, that is, it explains individual predictions. *Global* explainability insights can be obtained by aggregating the individual explanations over a set of data.

SHAP computes the importance of the individual features using the optimal Shapley values introduced by L. Shapley in 1953 [69] in game theory. In SHAP, the connection between game theory and explainability is that a prediction for an individual instance  $\mathbf{x}_i$  can be explained by conceiving the features  $X_i$  as the “players” of a “game” where the prediction  $\hat{y}_i$  is the game “payout.” Intuitively, the different  $M$  players of the game (features) receive different rewards, called Shapley values  $\phi_j$ , depending on their contribution to the total prediction, i.e.,  $\hat{y}_i = \phi_0 + \sum_{j=1}^M \phi_j$  where  $\phi_0$  is a reference value (baseline) computed as the average of output values. In this game-explanation analogy, the player who contributes with the larger  $\phi_j$  to the total prediction is the most important feature in the explanation.

Since the computation of the Shapley values involves testing all the possible combinations of the features (*coalitions* of the players in the game theory) by perturbing the instance  $\mathbf{x}_i$ , the time increases exponentially with the number of features [70]. Thus, various approaches have been proposed to estimate them efficiently, including SHAP. There are several kinds of SHAP methods, corresponding to different ways of approximating the Shapley values. In this work, we consider the widely adopted KernelExplainer variant of SHAP (KernelSHAP), as it is model-agnostic [21]. Indeed, other methods, such as TreeExplainer, result to be more efficient but are model-specific. Algorithm 1 describes the KernelSHAP procedure.

Notably, KernelSHAP requires a *background dataset* that serves as a reference: whenever a feature is excluded from a coalition, its value is replaced using an instance randomly sampled from such dataset. The choice of a representative background dataset is crucial for obtaining accurate estimates of the Shapley values. For this reason, the training set is typically adopted for this purpose. However, it is not the unique possible choice: a different, generally smaller, dataset can be used at the condition of being representative of the data distribution of the training set. In the literature, representative objects such as medoids or centroids of clusters generated by applying a clustering algorithm on the training set have allowed faster estimations of the Shapley values. Additional details on SHAP and Shapley values can be found in [70].

#### Federated SHAP

Let  $H$  be the number of clients involved in the federation,  $\mathbf{x}_i$  an input instance, and  $f(\mathbf{x}_i)$  the prediction to be explained. In this work, the model  $f(\cdot)$  is an MLP-NN learnt in a federated fashion. Following the setup proposed in [21], the goal

**Algorithm 1** KernelSHAP algorithm, from [70].

**Require:**  $R$ : reference dataset, where each instance is defined by  $M$  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  $j = 1, \dots, M$ .

1: Sample  $K$  coalitions  $z'_k \in \{0, 1\}^M$ , with  $k \in \{1, \dots, K\}$ , 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”.

2: **for**  $k = 1, \dots, K$  **do**

3: Compute  $z_k = h_x(z'_k)$ , where  $h_x : \{0, 1\}^M \rightarrow \mathbb{R}^M$  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  $R$ .

4: Compute the prediction of the model  $f(h_x(z'_k))$

5: Compute the weight for each  $z'_k$  with the following SHAP kernel:

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|} |z'| (M-|z'|)}$$

where  $|z'|$  is the number of non-zero elements in  $z'$ .

6: **end for**

7: Estimate the Shapley values  $\phi_j$  by optimizing the loss function

$$L(f, g, \pi_x) = \sum_{k=1}^K [f(h_x(z'_k)) - g(z'_k)]^2 \pi_x(z'_k) \quad (3)$$

with

$$g(z') = \phi_0 + \sum_{j=1}^M \phi_j z'_j \quad (4)$$

8: **return** the Shapley values  $\phi_j$

is to achieve a federated explanation for  $f(\mathbf{x}_i)$  considering the input instance simultaneously available to all the clients. It is worth underlining that this may be undesirable or unfeasible if the instance  $\mathbf{x}_i$  is subject to the privacy constraint. Notably, in some particular scenarios, for example, if the patients require multiple medical consultations, the instance  $\mathbf{x}_i$  could, under specific agreements, be shared to the other clients. Another scenario is represented by the presence of a benchmark dataset available to research entities, with the objective of comparing the goodness of the explainability produced by different methods.

As discussed in Section 3.2.1, the application of the SHAP method requires a background dataset. Typically, this is the dataset used to train the prediction model. However, in the federated setting, local training sets belong to different enti-

ties and cannot be shared due to privacy issues. Therefore, the server has no reference data to be used as background. Federated SHAP proposed in [21] represents a possible strategy to overcome this issue and to achieve federated explanations by exploiting the additive property of the Shapley values.

In the federated SHAP procedure, first of all, each client estimates the Shapley values using the local dataset as background dataset; then the values are transmitted to the server that evaluates their average. In this way, the data privacy is preserved, since the raw data are never shared. Furthermore, it is shown that the average can be considered a good approximation of the Shapley values calculated if the union of the local training sets was available to the server.

Schematically, the overall procedure for FL with post-hoc explainability technique entails the following steps: (i) each participant  $h$  (with  $h = 1, \dots, H$ ) contributes to the creation of an FL model. At the end of the federated procedure, the federated model is made available to each participant; (ii) given an unseen instance  $\mathbf{x}_i$ , each participant  $h$  computes the Shapley values  $\phi_j^{(i,h)}$  with  $j = 1, \dots, M$ , evaluating KernelSHAP locally, by exploiting the private training set as background dataset; (iii) the Shapley values  $\phi_j^{(i,h)}$  estimated at participant level are transmitted to the server, which performs simple averaging to obtain the federated estimation of the Shapley values  $\phi_j^{(i)}$  for explaining the prediction  $f(\mathbf{x}_i)$ .

## Case Study: Federated Learning for Parkinson’s Disease Progress Prediction

In this section, we describe the case study for the evaluation of the Fed-XAI approaches. First, details about the PD telemonitoring dataset are provided. Then, we describe the experimental setup in terms of data partitioning scenarios and evaluation strategies. Finally, we give the configurations of the ML models adopted in the two Fed-XAI approaches based on MLP-NN and TSK-FRBS, respectively.

### The Parkinson Telemonitoring Dataset

The Parkinson Telemonitoring dataset is a well-known regression dataset available within the UCI Machine Learning Repository [71]. The dataset is composed of 5875 instances of biomedical voice measurements from 42 patients with early-stage PD. Data are acquired remotely during a 6-month trial. Each instance corresponds to one voice recording, characterized by 22 features as reported in Table 1. The regression task consists in predicting the total Unified PD Rating Scale score (*total\_UPDRS*) associated with a given voice recording. Differently from *motor\_UPDRS*, which is related only to the motor symptoms, *total\_UPDRS* is related to the overall set of symptoms.

**Table 1** Parkinson Telemonitoring dataset description

Feature name	Brief description
Subject#	Patient identifier
Age	Subject age
Sex	Subject gender: ‘0’ - male, ‘1’ - female
Test_time	Time since recruitment into the trial
Jitter[% , Abs, RAP, PPQ5, DDP]	Measures of variation in fundamental frequency
Shimmer, Shimmer [dB, APQ3, APQ5, APQ11, DDA]	Measures of variation in amplitude
NHR, HNR	Two measures of ratio of noise to tonal components in the voice
RPDE	A nonlinear dynamical complexity measure
DFA	Signal fractal scaling exponent
PPE	A nonlinear measure of fundamental frequency variation
Motor_UPDRS	Clinician’s score, linearly interpolated
Total_UPDRS	Clinician’s score, linearly interpolated

## Federated Learning Scenarios

In this paper, extending the preliminary analysis performed in [15], we consider the challenging setting in which the raw dataset is not available on a single node for centralized processing, as in traditional ML, but it is instead scattered over multiple physical locations, e.g., hospitals or healthcare institutions. Specifically, we simulate several scenarios featuring 10 hospitals (cross-silo FL setting), in order to evaluate the performance of two Fed-XAI approaches under different horizontal data partitioning schemes that could be encountered in real-world situations.

In the following, we formally define the four scenarios considered in our experimental analysis. Let  $P_h(\mathbf{x}, y)$  be the local distribution of input data  $\mathbf{x}$  and associated target values  $y$  (*total\_UPDRS*) for the hospital  $h$ , and  $P(\mathbf{x}, y)$  the overall data distribution.

**Scenario IID** It is a simple independent and identically distributed (*i.i.d.*) setting; formally,

$$P_h(\mathbf{x}, y) \sim P(\mathbf{x}, y) \quad \forall h \in \{1, \dots, 10\} \quad (5)$$

The training data of the ten hospitals follow the same distribution, with about 500 instances each.

**Scenario NIID-Q** (acronym for **non-i.i.d.** quantity skew). It is a *non-i.i.d.* setting with *quantity skew* [72]: different hospitals can hold different amounts of training data, which follow the same overall distribution.

**Scenario NIID-F** (acronym for **non-i.i.d.** feature skew). It is a *non-i.i.d.* setting with *feature distribution skew* [72] based on the *age* feature; formally,

$$P_g(\mathbf{x}, y) \neq P_h(\mathbf{x}, y) \quad \forall g, h \in \{1, \dots, 10\}, g \neq h \quad (6)$$

Each hospital contains training data from only a specific range of ages (e.g., 56 to 57, 58 to 59, . . . , more than 75 years old). In this scenario, we aim to have training sets with as similar amount of data as possible.

**Scenario NIID-FQ** (acronym for **non-i.i.d.** feature and quantity skew). It is a *non-i.i.d.* setting with both *quantity skew* and *feature distribution skew* based on the *age* feature. Each hospital contains training data from only a specific range of ages; furthermore, different hospitals can hold different amounts of data.

The four scenarios concern different partitioning schemes for *training* data. As for the *testing* data, we consider the case of an external publicly available test set, valid for all the scenarios. The test set follows the overall data distribution (i.e., representative of all age groups) and has 588 instances. The distribution of the training data in the four scenarios is summarized in Table 2 and in Fig. 2.

As for the NIID-F and NIID-FQ scenarios, it is worth underlining that other features besides age may be affected by bias or skewness. However, this contingency still meets the definition of feature distribution skew. Therefore, the four scenarios enable a thorough and extensive evaluation of the performance of the two Fed-XAI approaches based, respectively, on MLP-NN and TSK-FRBS.

## Evaluation Settings

Typically, the performance evaluation of a model in the federated setting is performed not only in absolute terms, but also comparatively against two baseline settings [15, 73, 74]: local learning and centralized learning. Figure 3 provides a schematic overview of the three learning settings.

Federated learning (FL), local learning (LL), and centralized learning (CL) can be summarized for the dataset under consideration as follows:

- FL: the hospitals collaborate in obtaining a single federated model without sharing their raw data. The privacy of sensitive data is preserved.
- LL: each hospital locally learns a model from its private training data. As a consequence, the privacy of sensitive data is preserved, as in the FL case, but there is no collaboration among different hospitals.
- CL: training data from all hospitals are collected in a single central repository in the server and exploited to learn

**Table 2** Local training set cardinality ( $N_{(Train)}$ ) and related age range for each hospital and each scenario

Hospital	IID		NIID-Q		NIID-F		NIID-FQ	
	$ N_{train} $	Age range	$ N_{train} $	Age range	$ N_{train} $	Age range	$ N_{train} $	Age range
0	529	[36-85]	212	[36-85]	561	[36-55]	321	[36-49]
1	529	[36-85]	212	[36-85]	473	[56-57]	1099	[55-58]
2	529	[36-85]	212	[36-85]	655	[58-59]	269	[59-59]
3	529	[36-85]	212	[36-85]	487	[60-62]	275	[60-61]
4	529	[36-85]	424	[36-85]	506	[63-65]	212	[62-62]
5	529	[36-85]	636	[36-85]	380	[66-66]	886	[63-66]
6	529	[36-85]	847	[36-85]	540	[67-68]	255	[67-67]
7	528	[36-85]	844	[36-85]	428	[71-72]	713	[68-72]
8	528	[36-85]	844	[36-85]	591	[73-74]	347	[73-73]
9	528	[36-85]	844	[36-85]	666	[75-85]	910	[74-85]

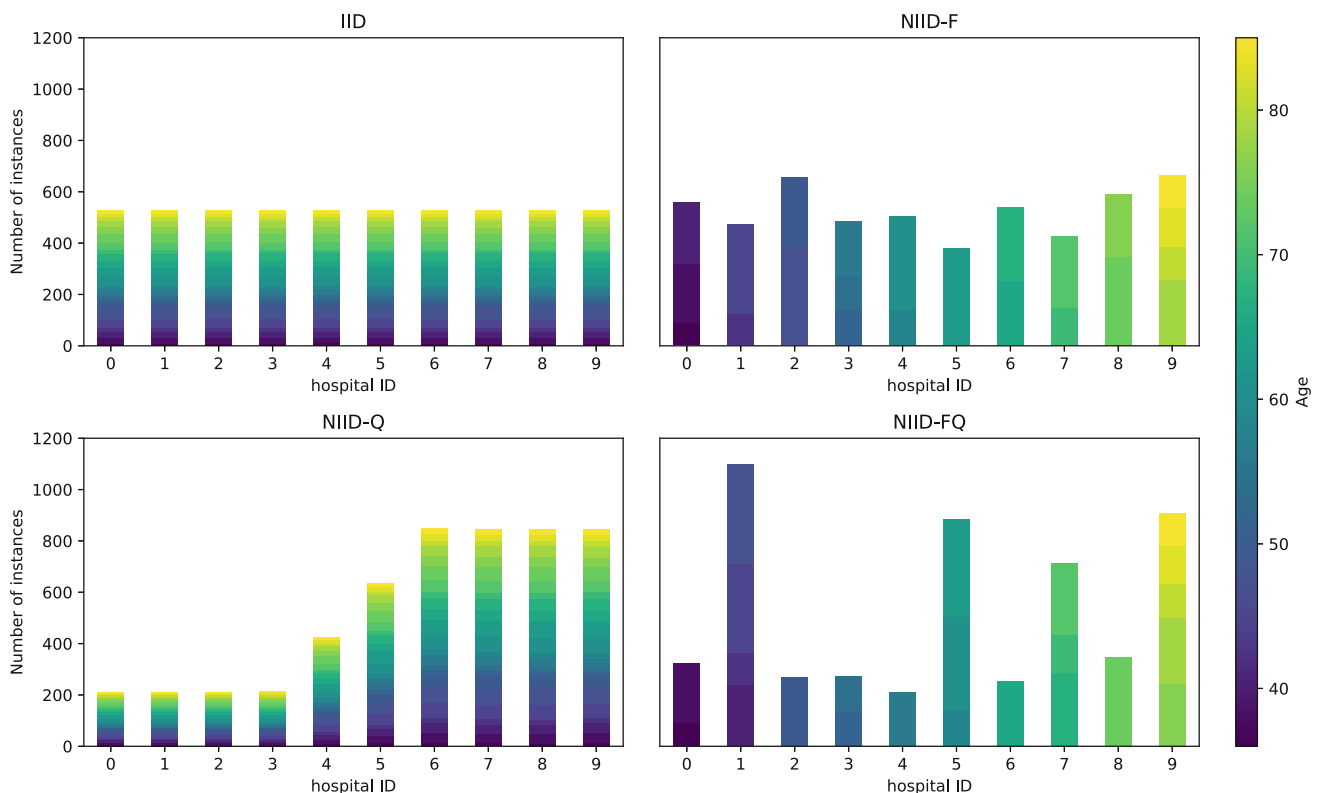
a global model. CL implies indeed maximum collaboration among hospitals, but violates privacy, as private sensitive data are moved from their owner to the server.

A model learned in the FL setting is expected to be more accurate than the ones learned in the LL setting. On the other hand, a model learned in the CL setting can outperform the other models (both LL and FL), in terms of accuracy, because it can rely on the union of the training datasets. The CL

approach, however, is not viable in real applications where privacy protection is a mandatory constraint.

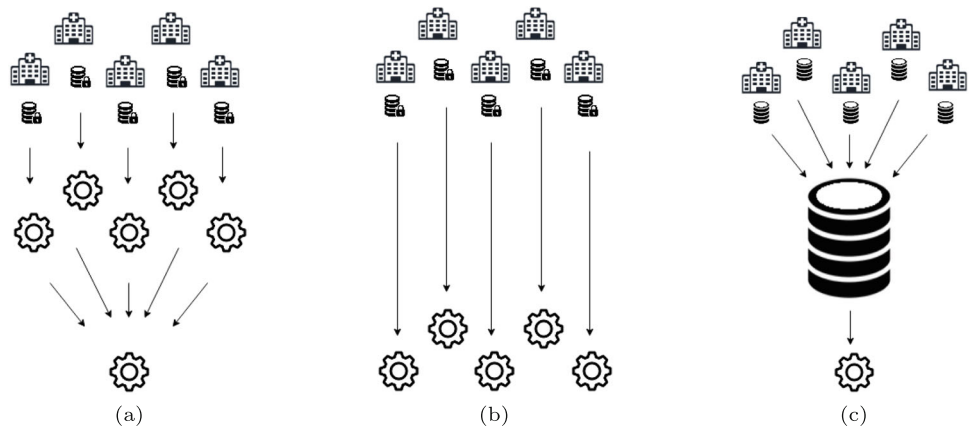
### Regression Problem and Fed-XAI Models

The PD progress prediction is formulated as a regression problem where the target variable is the *Total\_UPDRS* score. In our experiments, we replicated the preprocessing steps adopted in [15], namely, (i) a robust scaling using 0.025 and



**Fig. 2** Barplot of the training data partitioning scheme over ten hospitals in the four scenarios. Data for different ages are represented in different colors

**Fig. 3** Schematized representation of the three learning settings: **a** federated learning (FL), **b** local learning (LL), **c** centralized learning (CL)



0.975 quantiles is applied to the input features to remove outliers and clip the distribution in the range [0,1], and (ii) the output variable is normalized in the range [0,1].

Unlike in [15], the univariate feature selection procedure is carried out independently for the three learning settings, for a fair comparison of the entire regression pipelines. We select the  $G = 4$  best features in terms of Mutual Information (MI) [75] with the target variable. The estimate of MI and the subsequent feature selection is done individually by each participant in the LL setting, based on the local training sets, and globally in the CL setting, based on the union of the training sets. As for the FL setting, the federated feature selection procedure is schematized in Fig. 4, considering the example of the IID scenario. Each participant computes the MI score for all the features and transmits such information to the server. The server computes the average MI score for each feature and communicates the  $G$  best features to each participant. Thereafter, the FL process starts considering only the selected subset of features. In the example of Fig. 4 concerning the IID setting, the federated feature selection procedure selects the following features: *age*, *test\_time*, *DFA*, and *HNR*. Note that the feature importance scores of each participant may change depending on the data distribution scenario: thus, the selected features may vary and may generally differ from the CL setting.

The choice of the  $G$  value is guided by the following considerations: a reduced number of features generally improves the explainability task, both for post-hoc and interpretable by-design approaches. In addition, TSK-FRBSs struggle to handle high dimensional datasets [76]: the set of candidate rules grows exponentially with the number of features, thus jeopardizing the accuracy and the interpretability of the system. We have verified that  $G = 4$  ensures a good generalization capability for both models and an increase in the number of features does not lead to a significant improvement in performance.

In our experiments, for each data distribution scenario, we trained a TSK-FRBS and an MLP-NN according to FL, LL,

and CL settings. The experimental analysis is approached from a twofold perspective: model *accuracy* and model *explainability*. We assess the accuracy of the predictions obtained by the regression models as in [15] by using two popular metrics, namely Root Mean Squared Error (RMSE) and Pearson correlation coefficient ( $r$ ). They are defined as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \tag{7}$$

$$r = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}} \tag{8}$$

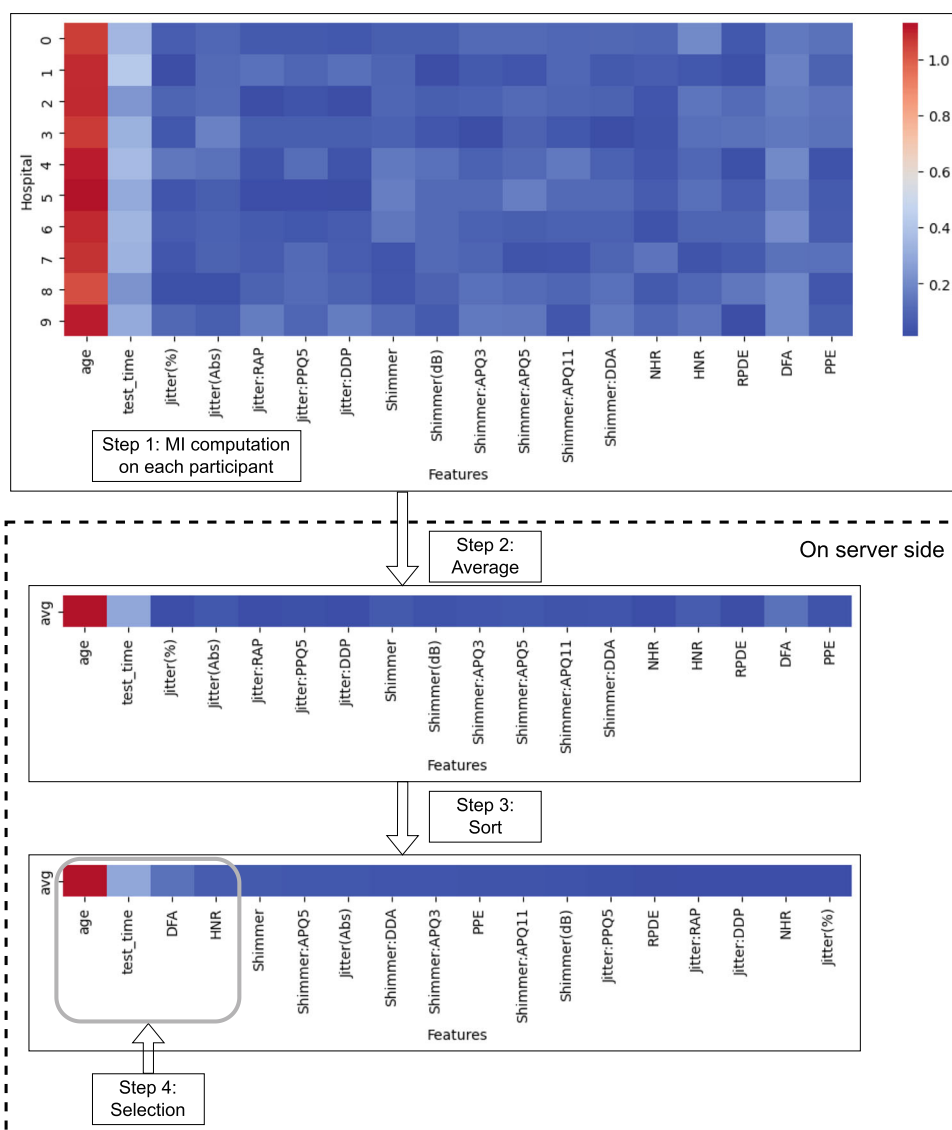
where  $N$  is the number of samples considered for the evaluation, and  $y_i$  and  $\hat{y}_i$  are the ground truth value and the predicted value associated with the  $i$ -th instance, respectively. Finally,  $\bar{y}$  is the mean of ground truth values, and  $\bar{\hat{y}}$  is the mean of the predicted values. Obviously, the goal is to minimize RMSE and maximize  $r$ .

It is worth underlining that the evaluation of an FL system typically covers other aspects besides accuracy such as computation and communication efficiency. These aspects, however, represent often crucial requirements or potential bottlenecks in *cross-device* FL, with many devices featuring limited computational resources [77]. In a *cross-silo* FL scenario, as the one considered in this work, such aspects are generally deemed less critical.

### Interpretable By-design Fed-XAI: TSK-FRBS Configuration

As in [15], we employ a first order TSK-FRBS model (described in Section 3.1). We adopt a strong uniform fuzzy partition on the features with five triangular fuzzy sets, as shown in Fig. 1. The choice of *five* fuzzy set is driven by the indication of the specialized literature and by the pursuit of a reasonable trade-off between model complexity

**Fig. 4** A schematic overview of the federated feature selection procedure. The example concerns the IID scenario



and generalization capability. The number of linguistic terms associated with a linguistic variable should be below the limit of  $7 \pm 2$  [78]. Indeed, it has been shown that this represents a threshold for information processing capability, and thus exceeding it undermines the interpretability of the system [79]. With the aim of describing linguistically a given rule, the five fuzzy sets can be labeled with the following linguistic terms: *VeryLow*, *Low*, *Medium*, *High*, and *VeryHigh*. Furthermore, it should be noted that different features may be partitioned using a different number of fuzzy sets, e.g., by exploiting domain knowledge to enhance understandability. Although this represents an interesting future development, we did not conduct extensive hyperparameters optimization. Rather, we verified in the CL setting that beyond 5 fuzzy sets, there is a substantial increase in model complexity, without significant improvement in terms of performance metrics. The choice of 5 fuzzy sets ensures a high linguistic

interpretability and represents a reasonable trade-off between model complexity and generalization capability.

#### Post-hoc Explainable Fed-XAI: MLP-NN Configuration

The MLP-NN consists of two hidden layers with 128 neurons, each with the ReLu activation function. The Mean Squared Error (MSE) is adopted as loss function and Adam as optimizer. The minibatch size is set to 64. The overall number of epochs is set to 100 in the CL and LL settings. In the FL setting, we set the number of local epochs and the number of rounds as 5 and 20, respectively. We have not performed a thorough optimization of hyperparameters for each learning setting and each data distribution scenario individually; however, we have empirically observed that a further increase in the capacity of the models in terms of number of layers, number of neurons, and training epochs does not

**Table 3** RMSE and  $r$  values on the training and test sets, obtained by the TSK-FRBS and MLP-NN models for the three learning settings and for each data partitioning scenario

	TSK-FRBS			MLP-NN					
	RMSE Train	Test	$r$ Train	RMSE Train	Test	$r$ Train			
IID	LL	7.369 ± 1.612	14.197 ± 7.722	<b>0.736 ± 0.104</b>	0.360 ± 0.088	8.202 ± 0.336	0.655 ± 0.033	<b>8.975 ± 0.239</b>	0.527 ± 0.028
	FL	<b>7.916 ± 0.236</b>	<b>7.896</b>	<b>0.677 ± 0.016</b>	<b>0.661</b>	8.590 ± 0.117	0.601 ± 0.010	8.319	0.605
	CL	7.679 ± 0.157	7.595	0.699 ± 0.012	0.686	<b>5.189 ± 0.208</b>	<b>0.877 ± 0.008</b>	<b>5.370</b>	<b>0.861</b>
NIID-Q	LL	7.822 ± 1.764	26.570 ± 41.601	0.720 ± 0.102	0.346 ± 0.182	8.287 ± 0.342	0.651 ± 0.033	<b>9.193 ± 0.403</b>	<b>0.492 ± 0.075</b>
	FL	<b>8.421 ± 0.331</b>	<b>8.523</b>	<b>0.631 ± 0.033</b>	<b>0.602</b>	8.755 ± 0.301	0.581 ± 0.023	8.602	0.568
	CL	7.695 ± 0.223	7.595	0.699 ± 0.012	0.686	<b>5.544 ± 0.240</b>	<b>0.863 ± 0.014</b>	<b>5.645</b>	0.851
NIID-F	LL	3.355 ± 1.319	113.049 ± 91.015	0.903 ± 0.098	-0.115 ± 0.253	3.710 ± 1.192	0.887 ± 0.093	<b>31.971 ± 24.690</b>	-0.008 ± 0.239
	FL	15.307 ± 7.391	16.848	0.465 ± 0.359	<b>0.461</b>	10.061 ± 3.606	0.380 ± 0.259	<b>10.268</b>	0.205
	CL	7.477 ± 2.158	7.595	0.615 ± 0.209	0.686	<b>5.515 ± 1.053</b>	0.728 ± 0.262	<b>5.702</b>	<b>0.839</b>
NIID-FQ	LL	3.441 ± 2.785	33.815 ± 25.799	0.890 ± 0.111	0.043 ± 0.158	3.763 ± 2.681	0.883 ± 0.106	<b>22.232 ± 21.441</b>	0.075 ± 0.203
	FL	9.734 ± 2.972	<b>11.182</b>	0.435 ± 0.252	<b>0.399</b>	10.489 ± 5.026	0.281 ± 0.471	12.880	-0.166
	CL	6.671 ± 2.083	7.595	0.619 ± 0.222	0.686	5.680 ± 1.049	0.701 ± 0.278	<b>5.496</b>	<b>0.863</b>

For the LL setting, we report the average values ± standard deviation obtained by the models learned locally in each participant

lead to a significant increase in the generalization capability of the MLP-NN.

## Analysis of the Experimental Results

Table 3 presents the RMSEs and  $r$  coefficients obtained by the TSK-FRBS and MLP-NN models for all the learning settings and the data distribution scenarios. As regards the LL setting, we report the average values  $\pm$  standard deviation obtained by the models learned locally in each participant. In the table, we have highlighted in bold the best results for each row, considering the comparison between TSK-FRBS and MLP-NN. Notably, in certain cases, the result is obtained from a distribution of values and is expressed in terms of mean and standard deviation: this occurs in the LL setting when ten local models are evaluated on the test sets and, regardless of the learning setting, when performance metrics are evaluated on ten training sets from as many hospitals. In such cases, we highlighted in bold the best result only if there exists a statistical difference in metrics values between TSK-FRBS and MLP-NN. The statistical significance has been assessed through a pairwise Wilcoxon signed-rank test [80] with confidence level  $\alpha = 0.05$ .

In general, federated models outperform the local counterparts, both in terms of RMSE and  $r$ . The benefit of FL over the LL setting is particularly evident for the TSK-FRBS and especially in the non-i.i.d. settings.

Federated TSK-FRBS and federated MLP-NN achieve comparable performances. The most noticeable difference occurs in the NIID-F setting, in which the two metrics provide diverse insights: on one hand, RMSE indicates that the deviation of predictions from true values is lower for the MLP-NN (10.268) compared to the TSK-FRBS (16.848); on the other hand, predictions and true values are more correlated for TSK-FRBS ( $r = 0.461$ ) than for MLP-NN ( $r = 0.205$ ).

The non-i.i.d. setting with quantity skew (NIID-Q) does not harm particularly the performance of the models: RMSE and  $r$  values are comparable to those of the IID setting, for

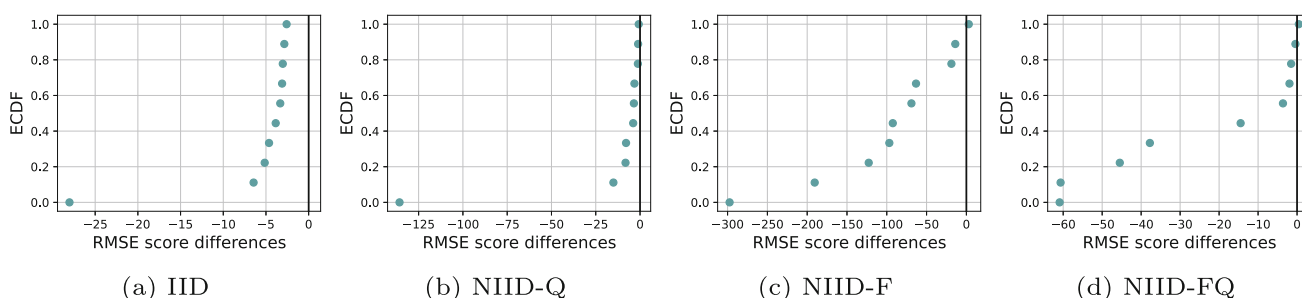
both TSK-FRBS and MLP-NN. In the case of TSK-FRBS, the simple average of the performances measured on the training sets shows poor performance for the LL setting, where some models likely suffer from low data availability. The aggregation strategy based on the rule weight (defined as a combination of confidence and support) ensures that this unfavorable situation is mitigated in the FL setting.

Scenarios with feature distribution skew (NIID-F and NIID-FQ) turn out to be the most challenging for both models. The generalization capability of models built in the LL setting is rather poor, due to exposure to data from a limited age range during training: both TSK-FRBS and MLP-NN perfectly model the training data (resulting in low RMSE values and an  $r$  coefficient around 0.9 on training sets) but fail in properly predicting the *total\_UPDRS* score of the test instances (resulting in high RMSE and low  $r$ ). Discrepancy in RMSE and  $r$  values between training and test sets is noticeable in the LL setting, whereas it is limited or negligible in the FL setting.

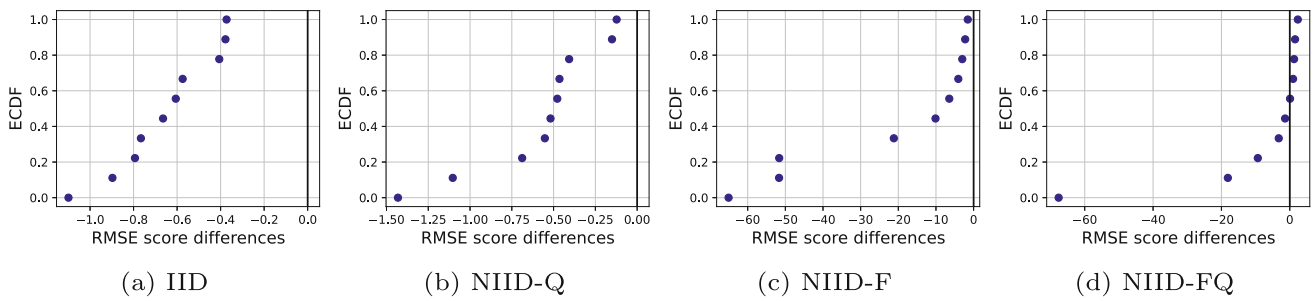
It is worth highlighting that the performances in the FL setting are always worse than those obtained in the CL setting, both for TSK-FRBS and MLP-NN. In general, the centralized MLP-NN is able to achieve the best performance with a slight improvement in terms of RMSE and  $r$  over the centralized TSK-FRBS. The superior performance of the centralized model can be attributed to the utilization of all data for conventional training. However, it is deemed unfeasible when privacy preservation represents a critical requirement.

The results in Table 3 provide an aggregate view of the LL setting. A better understanding of the outcomes can be gained by analyzing the specific performance obtained in each hospital: such detailed results are illustrated in Figs. 5 and 6 through the empirical cumulative distribution function (ECDF) for the RMSE metric.

For both models, the ECDF is reported for the values of the difference, for each hospital, of the RMSE achieved in the FL setting and the 10 locally computed values of RMSE obtained in the LL setting: each plot, therefore, is made up of 10 points. The plot can be interpreted as follows: if a point lies



**Fig. 5** TSK-FRBS: empirical cumulative distribution function (ECDF) of the differences of RMSE scores between FL and LL for the four data partitioning schemes



**Fig. 6** MLP-NN: empirical cumulative distribution function (ECDF) of the differences of RMSE scores between FL and LL for the four data partitioning schemes

in the negative half-plane (negative RMSE difference), then the RMSE value of the FL model is lower (and therefore the FL model is better) than the RMSE value of the LL model. The fine-grained analysis shown in Figs. 5 and 6 confirms that the FL setting generally outperforms the LL setting.

Finally, we report on the overall complexity of the models, which will be further discussed in Section 6. In the case of the MLP-NN, the network architecture is fixed: the complexity, intended as the number of parameters, does not change with the learning setting. In the case of TSK-FRBS, the complexity can be assessed in terms of the number of rules in the rule base. Figure 7 shows the complexity for each data distribution scenario and each learning setting.

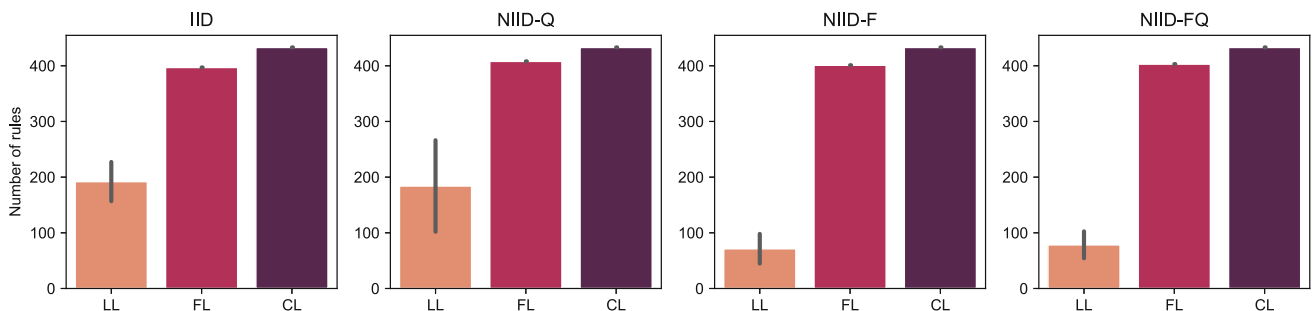
The number of rules of the federated TSK-FRBS never exceeds, by construction, that of the centralized TSK-FRBS, which in any case is limited (433 rules). As expected, the federated TSK-FRBS is more complex than locally learned ones. In the IID and NIID-Q scenarios (in which each hospital has data representative of the entire distribution), the 10 local models have approximately half the number of rules of the federated ones, indicating that common antecedents are often found in the rule aggregation phase. The complexity of locally learned TSK-FRBSs in the presence of feature distribution skew (NIID-F and NIID-FQ) is significantly lower than the one in IID setting, and the gain in accuracy provided by the federated models comes at a cost in terms of number of rules, which is approximately five times higher.

### Explainability Analysis

The extensive adoption of AI systems in the healthcare field depends not only on the achievement of adequate levels of accuracy, but also on how much they are perceived to be trustable. In particular, the ability to explain how the outcomes have been produced by the models is more and more required and represents the main driver of XAI. This section discusses the aspect of transparency of AI systems, focusing on how it is defined for the two Fed-XAI approaches analyzed in this paper. First, we analyze the *explainability* of the MLP-NN model, in which SHAP is used as post-hoc method. Then, we discuss the *interpretability by-design* of the TSK-FRBS. We consider models built in a federated fashion according to the IID data partitioning scheme: the discussion of the outcomes is limited to such case, but the pipeline for explainability analysis can be easily replicated for any data distribution scenario. Furthermore, we recall that all the input variables and the output variable are normalized in the unit interval [0, 1]: the considerations in this section refer to the predicted values before inverse transformation.

### Post-hoc Explainability of MLP-NN

In this section, we discuss the explainability of the MLP-NN after the application of the agnostic post-hoc method SHAP. We recall that, given an input instance, the Shapley



**Fig. 7** Number of rules of the TSK-FRBS for each learning setting and each data partitioning scheme. Error bar represents the standard deviation

value associated with each feature represents the *contribution* given by such feature to the predicted value. In this sense, for each prediction, SHAP explains why the model produces a particular output.

We adopt the Federated SHAP approach proposed in [21] and introduced in Section 3.2.2. KernelSHAP is employed in each hospital to estimate the Shapley values considering the full local training set as background dataset.

It is worth underlining a first crucial aspect concerning the explainability of MLP-NN: unlike interpretable-by-design Fed-XAI approaches, the post-hoc method affects the overall efficiency of the systems, both from a computation and a communication point of view. As for the former aspect, the estimation of the Shapley values with KernelShap is time consuming and the runtime increases with the number of features and the size of the adopted background dataset [70]. As for the latter aspect, Federated SHAP introduces a communication overhead, as Shapley values need to be transmitted by the participants for central aggregation. Conversely, the interpretable-by-design TSK-FRBS has no computation and communication overhead for generating the explanations.

### MLP-NN: Global Insights

Globally, an MLP-NN is generally considered “opaque,” due to the presence of several layers of non-linear information processing. In our case, the structure consists of two hidden layers with 128 neurons, resulting in 17,281 trainable parameters. The high number of parameters and the relations among these parameters make very hard to provide a global explanation of the model. Thus, indirect methods based, for instance, on *feature importance* are typically used to provide global explainability information [3].

As Shapley values represent additive feature importance scores for each particular prediction, the overall feature importance can be assessed by computing the average of the absolute Shapely values across the data [70]. Of course, the larger the average absolute value of the contribution given by a feature, the greater the importance of that feature. The assessment of the feature importance of a model is typically independent of the test data. In the case of the MLP-NN, it can be estimated as follows. First, each hospital  $h$  evaluates the importance  $I$  of the feature  $i$  on its training data:

$$I_{j,h}^{(MLP)} = \frac{1}{N_h} \cdot \sum_{i=1}^{N_h} |\phi_j^{(i,h)}| \quad \forall j \in \{1, \dots, M\}. \quad (9)$$

where  $\phi_j^{(i,h)}$  is the Shapely value for a feature  $j$  and a training instance  $i$  at hospital  $h$ , and  $N_h$  is the size of the training set at hospital  $h$ . Then, each client can transmit locally computed features importance to the server, and the overall features

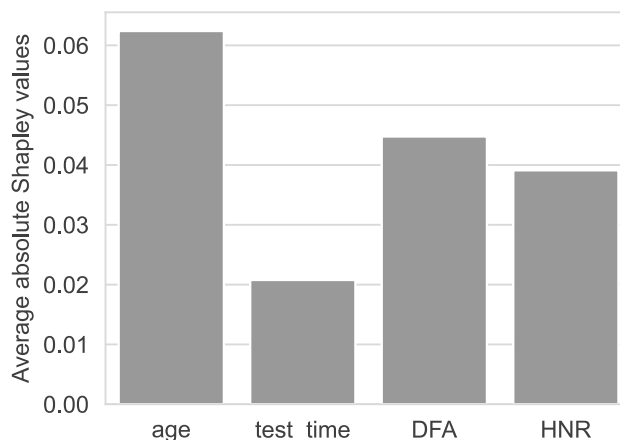


Fig. 8 Feature importance scores for the MLP-NN

importance for the federated model can be computed by the server as follows:

$$I_j^{(MLP)} = \sum_{h=1}^H \frac{N_h}{N} \cdot I_{j,h}^{(MLP)} \quad \forall j \in \{1, \dots, M\}. \quad (10)$$

where  $H$  is the number of hospitals and  $N = \sum_{h=1}^H N_h$ .

Figure 8 shows the global feature importance scores for the MLP-NN, as per (10): in the IID setting, the most relevant feature is *age*, while *test\_time* seems to be less relevant than vocal features, namely *DFA* and *HNR*, which in turn are of similar relevance.

### MLP-NN: Local Insights

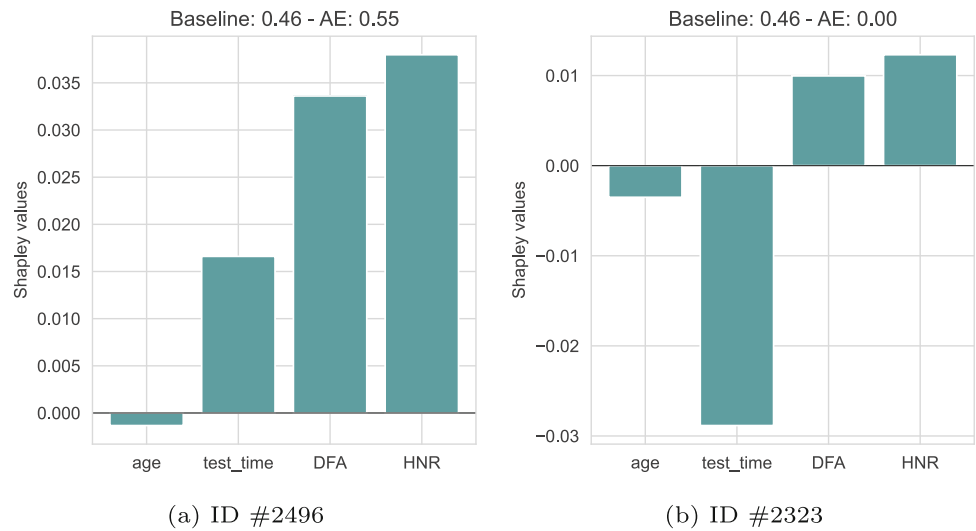
Figure 9 reports the SHAP values for two instances of the test set; they correspond to two cases where both models (MLP-NN and TSK-FRBS) obtain high and low errors, respectively. The absolute error (AE) made by the MLP-NN is around 0.55 for instance #2496 and around 0 for instance #2323.

The Shapley values reveal, a-posteriori, the relevance of the corresponding features in the prediction performed by the model: as expected, they are different when considering different instances. In the former case (Fig. 9a), *age* has little negative influence, while the other features have a large and positive impact on the output value. In the latter case (Fig. 9b), the most influential feature is *test\_time*, which has a negative impact on the output. Notably, the SHAP values for individual features are evaluated, for both instances, with respect to the same baseline value  $\phi_0 = 0.46$ .

### Interpretability By-design of TSK-FRBS

TSK-FRBSs are often considered as “light gray box” models [81]: their operation is highly interpretable, since they consist

**Fig. 9** MLP-NN local explainability: Shapley values for two instances in the test set. The absolute error for each instance is reported along with the baseline value



of a collection of linguistic, fuzzy, *if-then* rules. However, in the first-order TSK-FRBSs used in this paper, the adoption of a linear model in the consequent part, which certainly improves the accuracy with respect to the zero-order TSK-FRBS, makes the interpretation of a single rule less intuitive than the zero-order counterpart.

A substantial difference with respect to the MLP-NN model analyzed in Section 6.1 is that interpretability information is given without additional overhead in terms of computation and communication (as it is the case for the calculation of Shapley values on the MLP-NN model).

The following analysis aims to characterize both global and local interpretability of TSK-FRBS learnt in a federated fashion.

**TSK-FRBS: Global Insights**

The global interpretability of TSK-FRBSs can be quantitatively assessed by measuring the complexity of the system in terms of the number of rules and/or parameters. Less complex models, i.e., those with fewer rules, can be generally considered more interpretable [66]. Figure 7 reports the number of rules for each learning setting and each data partitioning scenario. As underlined in Section 5 for the IID case, the number of rules in the FL setting is rather limited (i.e., 397) and just double that in the LL case (despite the presence of 10 participants).

The model is therefore comprehensively described by the rule base, which can be represented in the intelligible form reported in the following.

$R_1$  : **IF** *age is VeryLow & test\_time is VeryLow & DFA is VeryLow & HNR is High*  
**THEN** :  $total\_UPDRS = -0.357 + 2.48 \cdot age - 0.51 \cdot test\_time - 0.59 \cdot DFA + 0.75 \cdot HNR$

...

$R_{397}$  : **IF** *age is VeryHigh & test\_time is VeryHigh & DFA is VeryHigh & HNR is Medium*  
**THEN** :  $total\_UPDRS = 2.46 - 1.31 \cdot age - 1.10 \cdot test\_time + 0.11 \cdot DFA + 0.09 \cdot HNR$

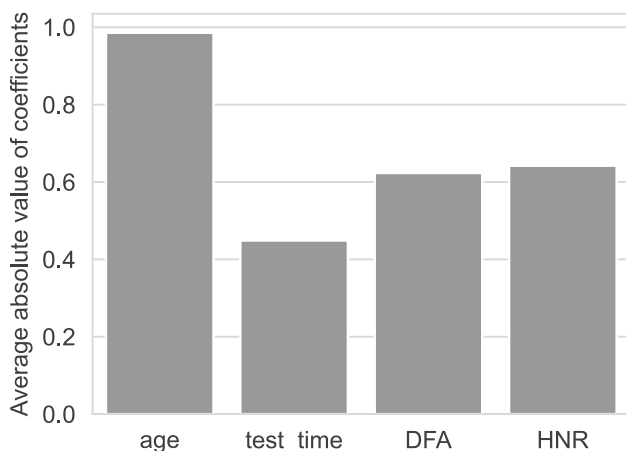
The number of parameters of a rule in a TSK-FRBS can be estimated as follows, considering the presence of four input variables: it is given by the sum of four parameters for the antecedent part (one for each input variable, to identify a fuzzy set of the a-priori partitioning) and five parameters for the linear model of the consequent part. Ultimately, the considered TSK-FRBS has 3573 parameters overall.

In the case of TSK-FRBS, a measure of feature importance can be obtained by averaging the absolute values of the coefficients of the linear models in the rule base. Formally, the importance  $I$  of the feature  $j$  is evaluated as follows:

$$I_j^{(TSK)} = \frac{1}{N_{rule}} \cdot \sum_{r=1}^{N_{rule}} |\gamma_{r,j}| \quad \forall j \in \{1, \dots, M\}. \quad (11)$$

where  $\gamma_{r,j}$  is the coefficient for the variable  $j$  in the rule  $r$ .

Figure 10 suggests that feature importance computed for the TSK-FRBS model is consistent with the one computed for the MLP-NN by the SHAP method in the IID setting. *Age* and *test\_time* are identified as the most and the least relevant features, respectively. Furthermore, the importance of *HNR*

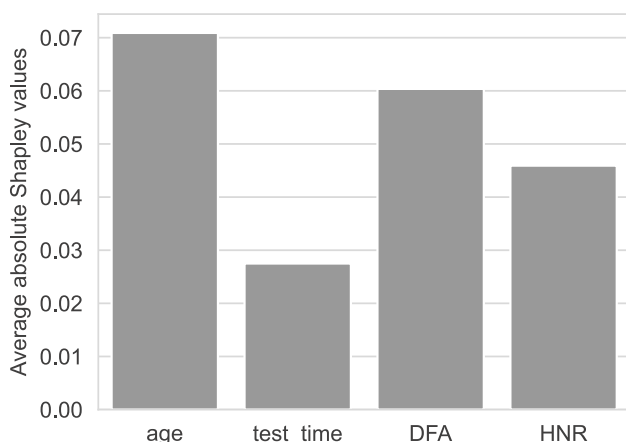


**Fig. 10** Feature importance scores for the TSK-FRBS

and *DFA* is similar, consistently with what is observed for the MLP-NN.

A direct comparison of the importance values between Fig. 8 (MLP-NN) and 10 (TSK-FRBS) is not meaningful. Indeed, in the case of MLP-NN, Fig. 8 shows the average absolute contribution for each feature with respect to the baseline value. In case of TSK-FRBS, Fig. 10 represents the average absolute value of the coefficients of the linear models used in the case of TSK-FRBS.

It is worth underlining that model-agnostic nature of SHAP can be exploited to compute post-hoc explanations also on the TSK-FRBS model. Thus, we can directly compare the feature importance obtained by averaging the absolute values of the coefficients of the linear models in the rule base and reported in Fig. 10, with that obtained by averaging the absolute Shapley values across the data for TSK-FRBS (9) and (10). The latter approach results in the scores reported in Fig. 11.



**Fig. 11** Feature importance scores for the TSK-FRBS evaluated in terms of Shapley values

It is interesting to note that *age* and *test\_time* are still identified as the most and least important features, indicating a summary agreement among the results obtained with different important attribution methods in the IID scenario. There is a discrepancy, on the other hand, between the relative values of *DFA* and *HNR*, possibly because the two approaches estimate importance values with different criteria. Furthermore, it is worth noting that feature importance scores computed for TSK-FRBS using SHAP are consistent with those computed for MLP-NN and shown in Fig. 8, also in terms of range of values.

The identification of *age* as the most important feature regardless of the model and the attribution method adopted is not surprising. First, in the IID scenario, each hospital can have data for all age ranges and therefore the feature entails a high variability. Second, such evidence is reflected in the specialized literature, which indicates that age is the best predictor of the progression of Parkinson’s disease and the most important risk factor for the development of the disease [82].

### TSK-FRBS: Local Insights

The interpretability of the TSK-FRBS derives from its structure and the type of inference strategy we use. Indeed, for any given input instance, the predicted output depends on a single rule: the antecedent part isolates a region (a hypercube) of the input space, where its consequent part defines a local linear model. The coefficients of this model indicate how the input features contribute to form the TSK-FRBS output in that region: a positive (negative) coefficient for a given feature indicates that the output increases (decreases) with that feature. Notably, all instances belonging to that region will refer to the same linear model.

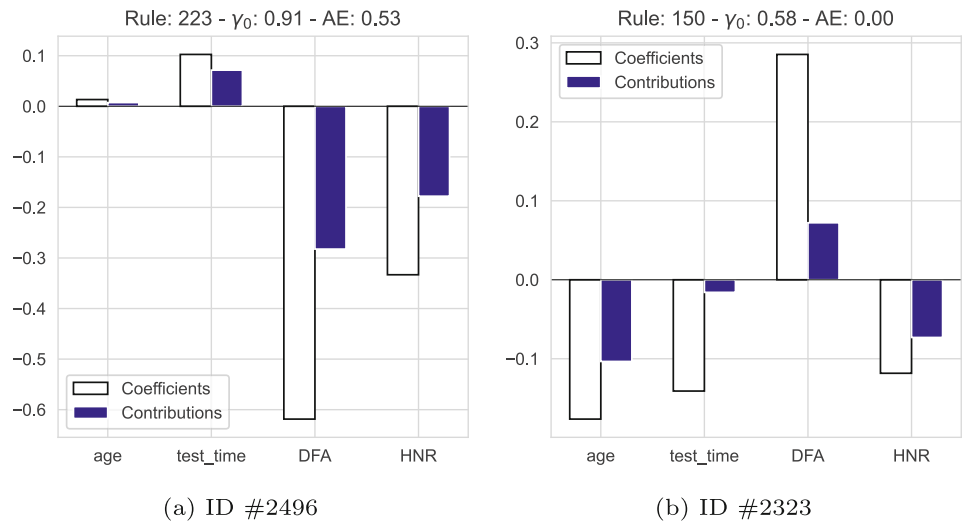
Given an instance  $\mathbf{x}_i$  and a rule  $r$ , which is the one with the maximum strength of activation for  $\mathbf{x}_i$ , the *actual contribution* of each feature to the prediction  $\hat{y}_i$  is obtained as  $\gamma_{r,j} \cdot x_{i,j}$ . In other words, contributions are obtained as the element-wise product between the coefficients of the linear model and the feature values of an instance. Figure 12 reports both the feature contributions and the coefficients of the linear model of the TSK-FRBS for the same two instances of the test set analyzed in the case of the MLP-NN (instances #2496 and #2323).

The two instances activate different rules: as a consequence, the contributions are significantly different. Furthermore, in general, the contributions are reduced compared to the coefficients, since each feature is normalized in the unit interval.

### Local Explanations: Comparison Between TSK-FRBS and MLP-NN

A relevant outcome can be drawn by comparing the barplots of Figs. 9 and 12: although the absolute errors are simi-

**Fig. 12** TSK-FRBS local explainability: coefficients of the linear model and actual feature contributions for two instances in the test set. The absolute error for each instance is reported above the relevant plot along with the ID of the rule considered for the prediction and the term  $\gamma_0$  of the linear model



lar (we have verified that the predicted values are similar as well), the two models “reason” differently and assign different—sometimes diametrically opposed—contributions to the features, also because the term  $\gamma_0$  is different from the baseline value of SHAP.

To better examine this aspect, we focus on a set of instances and analyze the explanations provided by the two models. Specifically, we consider the instance resulting in a high AE for both models (ID #2496) and all the instances of the test set that activate the same rule (namely,  $R_{223}$ ) of the TSK-FRBS. In this way, we isolate four instances (ID #827, ID #2266, ID #2496, ID #5146) which are inevitably close to each other in the input space.

The local interpretability of the TSK-FRBS is straightforward: predictions are obtained by applying the following rule:

$R_{223}$  : **IF** *age* is *Medium* & *test\_time* is *High* & *DFA* is *Medium* & *HNR* is *Medium*  
**THEN** :  $total\_UPDRS = 0.91 + 0.01 \cdot age + 0.10 \cdot test\_time - 0.62 \cdot DFA - 0.33 \cdot HNR$

In the case of MLP-NN, the prediction is explained through the Shapley values: Fig. 13 shows the *contributions* for the MLP-NN considering the four instances of the test set that activate rule  $R_{233}$  of the TSK-FRBS system.

It is evident that the Shapley values for the MLP-NN vary greatly even though the instances are fairly close in the input space: as an example, *age* has a negative contribution for ID #827 and a positive one for ID #5146. For this reason, it is equally evident that a correspondence cannot be found between the explanations offered by SHAP for the MLP-NN and the interpretation of the linear model of the TSK-FRBS. For example, SHAP always assigns a positive contribution to *DFA*, while the relevant coefficient is negative for TSK-FRBS.

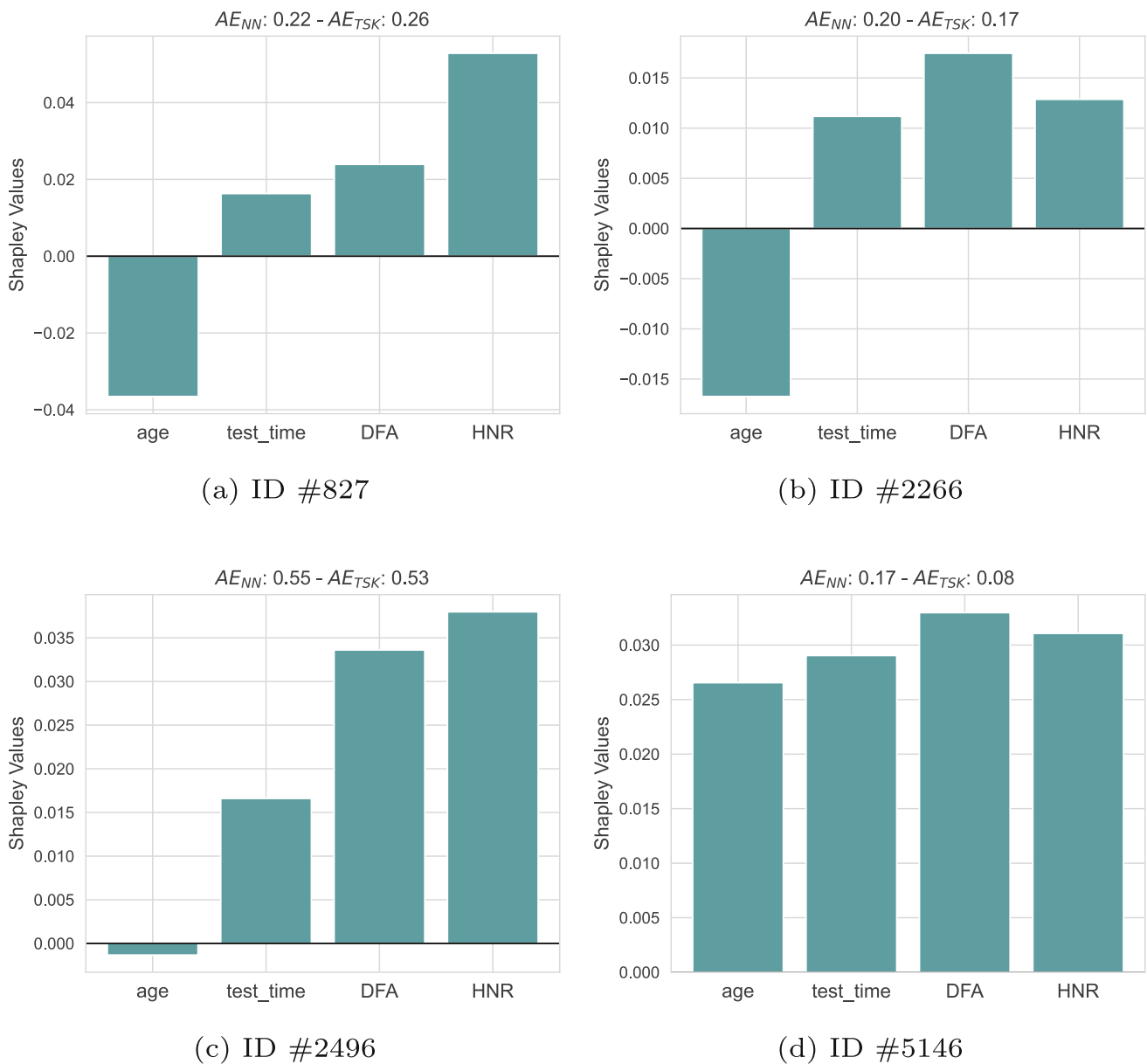
As noted above, the divergence in explanations between TSK-FRBS and MLP-NN in the IID scenario does not correspond to a divergence in output values. We have verified that the predicted outputs are similar (and indeed the reported AE values are similar): different models, which achieve similar results, lead to different explanations from a local point of view. This is not to be considered odd: our analysis entails different feature importance methods (inherent and post-hoc) and different models (TSK-FRBS and MLP-NN, respectively). Actually, it has been empirically shown that the feature importance score may suffer from numerical instability (when model, instance and attribution method are the same), solution diversity (if different models are considered, but on the same instance with the same attribution methods), or disagreement problem (if different attribution methods are considered, but on the same instance and the same model) [83]. These scenarios are due to the so-called Rashomon effect [84], whereby for a given dataset there may exist many models with equally good performance but with different solution strategies.

### Consistency of Explanations

As mentioned in Section 2.3, the property of consistency in the FL setting, introduced in [60], is met if different participants receive the same explanation of an output obtained with the federated model given the same data instance.

Evidently, for all the different operative scenarios discussed in this paper, the explanations for the federated TSK-FRBS are consistent: any local explanation obtained in a given hospital depends only on the input instance and on the activated rule of the federated model.

Conversely, the approach proposed in [21] and adopted in this work as post-hoc technique for the MLP-NN explainability does ensure the consistency of the explanations only



**Fig. 13** Shapley values calculated for the MLP-NN model on the instances in the test set that activate rule  $R_{233}$  of the TSK-FRBS. The absolute error (AE) for each instance and for each model is reported above the relevant plot

in the situation where the test instances are shareable to all the clients. The local explanation obtained in a given hospital, in fact, depends not only on the input instance and the federated MLP-NN, but also on the background dataset used for estimating the Shapley values. Since each hospital has its own private dataset, the Shapley values for the same input instance may differ, in general, from one hospital to another. The Federated SHAP approach allows obtaining an explanation for each test instance by averaging the local explanations from different hospitals. On one hand, this ensures that a unique and unambiguous explanation is obtained. On the other hand, this requires that any test instance is shared with

other hospitals at inference time, which may be problematic due to privacy and/or latency issues.

In the following, we quantitatively assess the misalignment of client-side explanations obtained with SHAP for the MLP-NN, before applying the averaging operation that characterizes Federated SHAP. We consider an input test instance (ID# 2496, discussed also in the previous examples) and suppose that it is available at every hospital. Specifically, we evaluate how the prediction for such instance would be explained on different clients, in case that sharing Shapley values for averaging is precluded for privacy reason.

**Fig. 14** Shapley values for the MLP-NN for instance ID #2496 for each hospital. IID Scenario

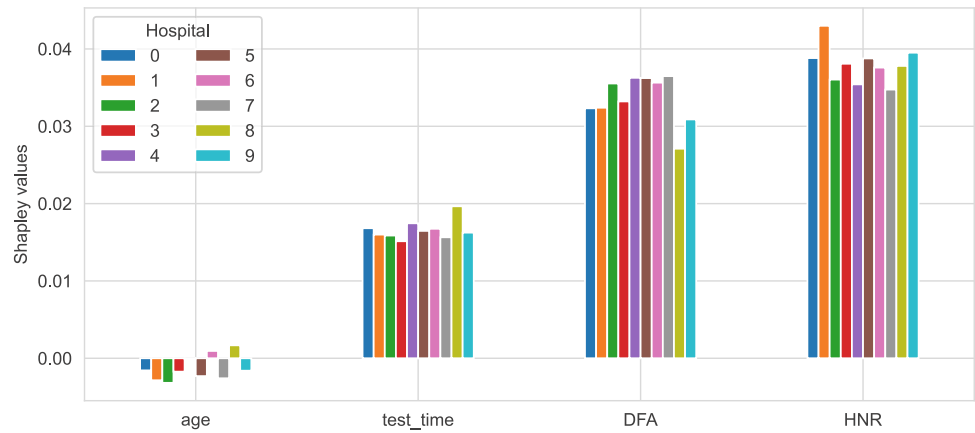


Figure 14 reports the Shapley values for each hospital and each feature in the IID scenario.

The barplot suggests that the explanations are consistent in the IID scenario, albeit showing some slight variability, which is reasonable since the background datasets are identically distributed. Indeed, explanations are in line with the average pattern reported in Fig. 14c.

The variability among client-side explanations turns out to be substantial in non-i.i.d. scenarios. Figure 15 shows the Shapley values for the same instance for each hospital and each feature when considering the NIID-FQ scenario. We recall that such scenario entails both a *quantity* skew and a *feature* skew on the *age* feature. Furthermore, it is worth mentioning that the feature selection process is part of the FL pipeline. As a consequence, the set of selected features depends on the data distribution scenario: this explains the presence of different features compared to the IID case, with *Jitter(Abs)* replacing *HNR*.

Figure 15 suggests that the misalignment of explanations is severe, especially for the contribution assigned to the *age* feature by different hospitals. The relevant Shapley value goes from negative in hospitals with younger patients to positive in hospitals with older patients. Probing the global model with data from heterogeneous distributions results in a differ-

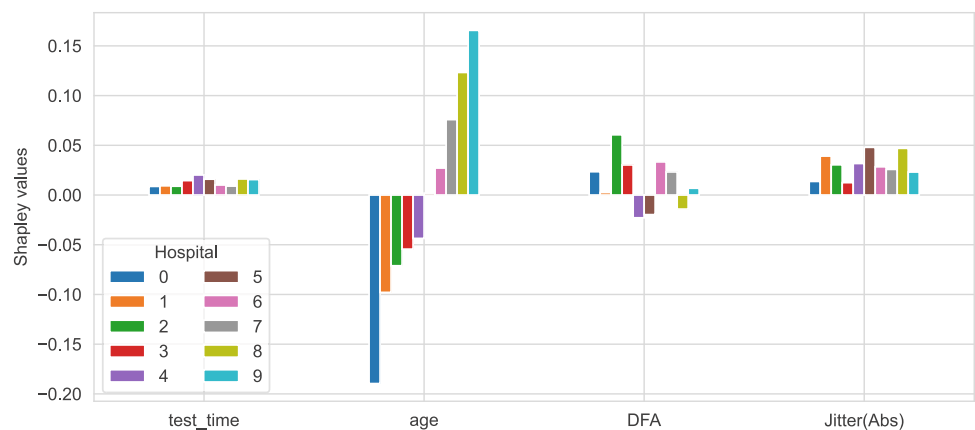
ence also in the importance assigned to the *DFA* feature. In summary, the same instance, analyzed with the same model, is explained in very different ways on different hospitals. Thus, the property of consistency among explanations is not achieved.

The consistency of SHAP explanations in the FL setting can be achieved by avoiding the use of private training data as background: Chen et al. [63], for example, propose to use synthetic background datasets generated sampling from a Gaussian distribution whose parameters are estimated on the server side based on the contributions of all participants. However, such explanations may be different from those obtained using actual training data. Ensuring both consistency and accuracy of explanations, intended as agreement with the centralized case, is one of the interesting future developments of this work.

### Conclusions

In this paper, we have addressed the problem of developing a trustworthy AI system for a healthcare application, with specific focus on a Parkinson’s disease (PD) progression prediction task. For this purpose, we designed two

**Fig. 15** Shapley values for the MLP-NN for instance ID #2496 for each hospital. NIID-FQ scenario



approaches that simultaneously meet the requirements of data privacy preservation and explainability, which are usually deemed crucial for enabling trustworthiness. The first approach adopts a Takagi-Sugeno-Kang Fuzzy Rule-Based system (TSK-FRBS), which is interpretable by-design. TSK-FRBSs make use of fuzzy sets as information granules, thus guaranteeing high semantic interpretability. The second approach employs a Multi Layer Perceptron Neural Network (MLP-NN): as a “black-box” model, it requires the adoption of a post-hoc technique for explainability purposes. In this paper, we have adopted SHAP, which is considered as a state of art feature importance explainability method.

For both approaches, the federated learning (FL) paradigm has been exploited as it inherently enables privacy preservation during global model training procedures in decentralized settings. In detail, we devised an experimental setting assuming that sensitive data originate from ten hospitals and cannot be shared for privacy reasons. In order to cover several real-world situations, four (one i.i.d. and three non-i.i.d.) scenarios with different degrees of heterogeneity are simulated.

The critical analysis of the two approaches has concerned the following aspects: (i) the accuracy of the models, in terms of Root Mean Squared Error (RMSE) and Pearson correlation coefficient  $r$ , depending on the learning setting (federated learning, local learning, centralized learning) and the four data distribution scenarios, and (ii) the explainability of the models at global and local levels.

The key findings can be summarized as follows. From the perspective of performance metrics, results highlight that the federated models are able to outperform the ones learned by using only local data, both in terms of RMSE and  $r$  values, hence highlighting the benefits of the federation. This is particularly evident in the non-i.i.d. settings. Also, results suggest that federated TSK-FRBS and federated MLP-NN achieve comparable performance, within the context of the considered case study.

As regards explainability, we have presented the results of the post-hoc explainability of MLP-NN and of the by-design interpretability of TSK-FRBS, also providing a comparative analysis of the two approaches. It turns out that the two approaches can lead to different local explanations, even if the underlying models achieve similar results in terms of regression metrics. A first major difference between the two approaches lies in their nature: the TSK-FRBS model provides insights about “how” an outcome is obtained, whereas the post-hoc method provides insight about “why” an outcome is provided. Consequently, a qualitative comparison is more reasonable than a quantitative one. We have to take into account, however, that if a model is able to provide a glimpse on how an outcome has been obtained, implicitly it is also making manifest why that outcome has been computed from the inputs.

More precisely, the TSK-FRBS is a collection of linguistic *if-then* rules: its global interpretability is usually assessed in terms of the number of rules and/or parameters: a lower number of rules corresponds to systems considered more interpretable. An equivalent global picture is not immediately available in the case of MLP-NN: the adoption of the SHAP method enables local explainability (i.e., it explains individual predictions), and a global explainability indication can be obtained in terms of feature importance, by aggregating the individual explanations over a set of data. The comparison of the feature importance obtained for the MLP-NN (as the average of the absolute Shapley values) and for the TSK-FRBS (as the average of the absolute values of the coefficients of the linear models in the rule base) in the IID setting reveals a fair agreement: both methods identify *age* as the most important feature and *test\_time* as the least important one.

Local explanations of the two approaches convey inherently different messages. For the MLP-NN, the Shapley values represent the importance of each feature to each prediction with respect to a baseline value, computed as the average output value. For the TSK-FRBS, the predicted output depends, in our setting, just on the mostly activated rule: the antecedent part isolates a region of the input space and is expressed using linguistic terms associated with fuzzy sets (information granules), whereas its consequent part defines a local linear model whose coefficients indicate how the input features contribute to form the output. Focusing on a specific region of the input space, corresponding to the one isolated by the antecedent of a rule in the TSK-FRBS, we have evaluated the explanations provided by the two models for the instances located within that region. As expected, the explanations obtained with SHAP for the MLP-NN are generally different for the different instances and do not match what is expressed by the relevant rule of the TSK-FRBS. In essence, different models, which produce similar outputs and achieve similar overall results, lead to different explanations from a local point of view.

Finally, it is worth underlining that the property of consistency holds for the federated TSK-FRBS: for the same input data, different participants get the same explanation from the common federated model. In the case of the SHAP post-hoc technique, this is not as straightforward: if participants use their local training data as background dataset for the estimation of the Shapley values, the feature importance scores are necessarily different; if the federated SHAP strategy of averaging locally computed Shapley values is adopted, a single value is obtained. This ensures consistency of explanation but requires that the privacy preservation constraint is relaxed at test time. How to achieve consistency of post-hoc explanation and simultaneously to ensure privacy preservation represents an interesting future development of this work.

Overall, the Fed-XAI approach offers remarkable potential and can have important practical implications in high-stake applications such as healthcare: in this domain, in fact, data centralization for ML model training is not only difficult to implement from a technical perspective, but it also raises ethical concerns related to privacy and is subject to limitations imposed by regulatory policies. On one hand, FL removes the need for data centralization while still allowing ML models training to benefit from ample and diverse data, which is crucial to address urgent challenges such as health disparities, under-served populations, and rare diseases [85]. On the other hand, the adoption of XAI tools endow the AI system with the capability of explaining its decisions, which is paramount in these kinds of applications.

We have discussed strengths and weaknesses of the two federated approaches for learning XAI models. In the future, we aim to enrich the comparative analysis with an additional level of assessment by collecting feedbacks on explainability from human experts (e.g., physicians) and by considering other case studies.

**Author Contribution** All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Alessandro Renda and Fabrizio Ruffini. The first draft of the manuscript was written by Francesco Marcelloni, Alessandro Renda, and Fabrizio Ruffini, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** Open access funding provided by Università di Pisa within the CRUI-CARE Agreement. This work has been partly funded by the PON 2014–2021 “Research and Innovation,” DM MUR 1062/2021, Project title: “Progettazione e sperimentazione di algoritmi di federated learning per data stream mining,” PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI” and the PNRR “Tuscany Health Ecosystem” (THE) (Ecosistemi dell’Innovazione) - Spoke 6 - Precision Medicine & Personalized Healthcare (CUP I53C22000780001) under the NextGeneration EU programme, and by the Italian Ministry of University and Research (MUR) in the framework of the FoReLab and CrossLab projects (Departments of Excellence).

**Data Availability** The dataset is publicly available at <https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring>.

## Declarations

**Conflict of Interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the

permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## 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>.
2. : GDPR. Available from: <https://gdpr-info.eu/recitals/no-71/> [cited 06.10.2022].
3. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020;58:82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>.
4. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv*. 2018. <https://doi.org/10.1145/3236009>.
5. Guidotti R, Monreale A, Pedreschi D, Giannotti F. In: Sayed-Mouchaweh M, editor. Principles of explainable artificial intelligence. Cham: Springer International Publishing; 2021. p. 9–31.
6. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable Artificial Intelligence (XAI): what we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*. 2023;99:101805. <https://doi.org/10.1016/j.inffus.2023.101805>.
7. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAY. Communication-efficient learning of deep networks from decentralized data. In: Singh A, Zhu J, editors. Proceedings of the 20th international conference on artificial intelligence and statistics. vol. 54 of Proceedings of Machine Learning Research. PMLR; 2017. p. 1273–1282. Available from: <https://proceedings.mlr.press/v54/mcmahan17a.html>.
8. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol*. 2019. <https://doi.org/10.1145/3298981>.
9. Bodria F, Giannotti F, Guidotti R, Naretto F, Pedreschi D, Rinzivillo S. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*. 2023;37(5):1719–78. <https://doi.org/10.1007/s10618-023-00933-9>.
10. Chaddad A, Lu Q, Li J, Katib Y, Kateb R, Tanougast C, et al. Explainable, domain-adaptive, and federated artificial intelligence in medicine. *IEEE/CAA Journal of Automatica Sinica*. 2023;10(4):859–76. <https://doi.org/10.1109/JAS.2023.123123>.
11. Corcuera Bárcena JL, Daole M, Ducange P, Marcelloni F, Renda A, Ruffini F, et al. Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models. In: XAI.it: 3rd Italian workshop on explainable artificial intelligence, co-located with AI\*IA; 2022. Available from: <https://ceur-ws.org/Vol-3277/paper8.pdf>.
12. López-Blanco R, Alonso RS, González-Arrieta A, Chamoso P, Prieto J. Federated Learning of Explainable Artificial Intelligence (FED-XAI): a review. In: Ossowski S, Sitek P, Analide C, Murreiros G, Chamoso P, Rodríguez S, editors. Distributed computing and artificial intelligence, 20th international conference. Cham: Springer Nature Switzerland; 2023. p. 318–26.
13. Corcuera Bárcena JL, Ducange P, Ercolani A, Marcelloni F, Renda A. An approach to federated learning of explainable fuzzy regression models. In: IEEE International conference on fuzzy systems (FUZZ-IEEE); 2022. p. 1–8.
14. Daneault JF, Carignan B, Sadikot AF, Duval C. Are quantitative and clinical measures of bradykinesia related in advanced Parkinson’s disease? [Article]. *Journal of Neuroscience Meth-*

- ods. 2013;219(2):220–3. <https://doi.org/10.1016/j.jneumeth.2013.08.009>.
15. Corcuera Bárcena JL, Ducange P, Marcelloni F, Renda A, Ruffini F. Federated learning of explainable artificial intelligence models for predicting Parkinson's disease progression. In: Longo L, editor. *Explainable artificial intelligence*. Cham: Springer Nature Switzerland; 2023. p. 630–48.
  16. Takagi T, Sugeno M. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*. 1985;SMC-15(1):116–32. <https://doi.org/10.1109/TSMC.1985.6313399>.
  17. Fernandez A, Herrera F, Cordon O, Jose del Jesus M, Marcelloni F. Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to? *Comp Intell Mag*. 2019;14(1):69–81. <https://doi.org/10.1109/MCI.2018.2881645>.
  18. Zadeh LA. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*. 1997;90(2):111–27. [https://doi.org/10.1016/S0165-0114\(97\)00077-8](https://doi.org/10.1016/S0165-0114(97)00077-8). (**Fuzzy Sets: Where Do We Stand? Where Do We Go?**).
  19. Yao JT, Vasilakos AV, Pedrycz W. Granular computing: perspectives and challenges. *IEEE Transactions on Cybernetics*. 2013;43(6):1977–89. <https://doi.org/10.1109/TSMCC.2012.2236648>.
  20. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc.; 2017. Available from: [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf).
  21. Corbucci L, Guidotti R, Monreale A. Explaining black-boxes in federated learning. In: Longo L, editor. *Explainable artificial intelligence*. Cham: Springer Nature Switzerland; 2023. p. 151–63.
  22. Li H, Li C, Wang J, Yang A, Ma Z, Zhang Z, et al. Review on security of federated learning and its application in healthcare. *Future Generation Computer Systems*. 2023;144:271–90. <https://doi.org/10.1016/j.future.2023.02.021>.
  23. Hwang H, Yang S, Kim D, Dua R, Kim JY, Yang E, et al. Towards the practical utility of federated learning in the medical domain. In: Mortazavi BJ, Sarker T, Beam A, Ho JC, editors. *Proceedings of the conference on health, inference, and learning*. vol. 209 of *Proceedings of Machine Learning Research*. PMLR; 2023. p. 163–181. Available from: <https://proceedings.mlr.press/v209/hwang23a.html>.
  24. Rahman A, Hossain MS, Muhammad G, Kundu D, Debnath T, Rahman M, et al. Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues. *Cluster computing*. 2023;26(4):2271–311. <https://doi.org/10.1007/s10586-022-03658-4>.
  25. Sohan MF, Basalamah A. A systematic review on federated learning in medical image analysis. *IEEE Access*. 2023;11:28628–44. <https://doi.org/10.1109/ACCESS.2023.3260027>.
  26. De Falco I, Della Cioppa A, Koutny T, Scafuri U, Tarantino E. Model-free-communication federated learning: framework and application to precision medicine. *Biomedical Signal Processing and Control*. 2024;87:105416. <https://doi.org/10.1016/j.bspc.2023.105416>.
  27. Nguyen DC, Pham QV, Pathirana PN, Ding M, Seneviratne A, Lin Z, et al. Federated learning for smart healthcare: a survey [article]. *ACM Computing Surveys*. 2022. <https://doi.org/10.1145/3501296>. (**Cited by: 209; All Open Access, Green Open Access**).
  28. Ali M, Naeem F, Tariq M, Kaddoum G. Federated learning for privacy preservation in smart healthcare systems: a comprehensive survey [article]. *IEEE Journal of Biomedical and Health Informatics*. 2023;27(2):778–89. <https://doi.org/10.1109/JBHI.2022.3181823>. (**Cited by: 36; All Open Access, Green Open Access**).
  29. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*. 2023. <https://doi.org/10.1007/s12559-023-10179-8>.
  30. Patrício C, JaC Neves, Teixeira LF. Explainable deep learning methods in medical image classification: a survey. *ACM Comput Surv*. 2023. <https://doi.org/10.1145/3625287>.
  31. Uddin MZ, Dysthe KK, Følstad A, Brandtzaeg PB. Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*. 2022;34(1):721–44. <https://doi.org/10.1007/s00521-021-06426-4>.
  32. El-Sappagh S, Alonso JM, Islam SMR, Sultan AM, Kwak KS. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci Rep*. 2021;11(1):2660.
  33. Dong N, Voiculescu I. Federated contrastive learning for decentralized unlabeled medical images. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part III*. Berlin, Heidelberg: Springer-Verlag; 2021. p. 378–87. [https://doi.org/10.1007/978-3-030-87199-4\\_36](https://doi.org/10.1007/978-3-030-87199-4_36).
  34. Vaid A, Jaladanki SK, Xu J, Teng S, Kumar A, Lee S, et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. *JMIR Med Inform*. 2021;9(1):e24207.
  35. Bounsall K, Milne-Ives M, Hall A, Carroll C, Meinert E. Artificial intelligence applications for assessment, monitoring, and management of parkinson disease symptoms: protocol for a systematic review. *JMIR Res Protoc*. 2023;12:e46581. <https://doi.org/10.2196/46581>.
  36. El-Sappagh S, Alonso-Moral JM, Abuhmed T, Ali F, Bugarín-Diz A. Trustworthy artificial intelligence in Alzheimer's disease: state of the art, opportunities, and challenges. *Artif Intell Rev*. 2023;56(10):11149–296. <https://doi.org/10.1007/s10462-023-10415-5>.
  37. Gómez-Vilda P, Rodellar-Biarge V, Nieto-Lluis V, Muñoz-Mulas C, Mazaira-Fernández LM, Martínez-Olalla R, et al. Characterizing neurological disease from voice quality biomechanical analysis. *Cognitive Computation*. 2013;5(4):399–425. <https://doi.org/10.1007/s12559-013-9207-2>.
  38. Arias-Vergara T, Vázquez-Correa JC, Orozco-Arroyave JR. Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech. *Cognitive Computation*. 2017;9(6):731–48. <https://doi.org/10.1007/s12559-017-9497-x>.
  39. Magesh P, Myloth R, Tom R. An explainable machine learning model for early detection of Parkinson's disease using lime on DaTSCAN imagery. *Computers in Biology and Medicine*. 2020;11(126):104041. <https://doi.org/10.1016/j.compbiomed.2020.104041>.
  40. Ribeiro MT, 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. New York, USA: Association for Computing Machinery; 2016. p. 1135–1144.
  41. Junaid M, Ali S, Eid F, El-Sappagh S, Abuhmed T. Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson's disease. *Computer Methods and Programs in Biomedicine*. 2023;234:107495. <https://doi.org/10.1016/j.cmpb.2023.107495>.
  42. Jorge J, Barros PH, Yokoyama R, Guidoni D, Ramos HS, Fonseca N, et al. Applying federated learning in the detection of freezing of gait in Parkinson's disease. In: *IEEE/ACM 15th International conference on Utility and Cloud Computing (UCC)*; 2022. p. 195–200.

43. Sarlas A, Kalafatelis A, Alexandridis G, Kourtis MA, Trakadas P. Exploring federated learning for speech-based Parkinson's disease detection. In: Proceedings of the 18th international conference on availability, reliability and security. ARES '23. New York, USA: Association for Computing Machinery; 2023.
44. Dipro SH, Islam M, Al Nahian A, Sharmita Azad M, Chakrabarty A, Reza T. A federated learning based privacy preserving approach for detecting Parkinson's disease using deep learning. In: 2022 25th International Conference on Computer and Information Technology (ICCIT); 2022. p. 139–144.
45. Grover S, Bhartia S, Akshama, Yadav A, R SK. Predicting severity of Parkinson's disease using deep learning. *Procedia Computer Science*. 2018;132:1788–94. <https://doi.org/10.1016/j.procs.2018.05.154>. (**International Conference on Computational Intelligence and Data Science**).
46. Gunduz H. Deep sets learning-based Parkinson's disease classification using vocal feature. *IEEE Access*. 2019;7:115540–51. <https://doi.org/10.1109/ACCESS.2019.2936564>.
47. Nilashi M, Ibrahim O, Samad S, Ahmadi H, Shahmoradi L, Akbari E. An analytical method for measuring the Parkinson's disease progression: a case on a Parkinson's telemonitoring dataset. *Measurement*. 2019;136:545–57. <https://doi.org/10.1016/j.measurement.2019.01.014>.
48. Shahid AH, Singh MP. A deep learning approach for prediction of Parkinson's disease progression. *Biomedical Engineering Letters*. 2020;10:227–39. <https://doi.org/10.1007/s13534-020-00156-7>.
49. Xue Z, Zhang T, Lin L. Progress prediction of Parkinson's disease based on graph wavelet transform and attention weighted random forest. *Expert Systems with Applications*. 2022;203: 117483. <https://doi.org/10.1016/j.eswa.2022.117483>.
50. Chen P, Du X, Lu Z, Wu J, Hung PCK. EVFL: an explainable vertical federated learning for data-oriented Artificial Intelligence systems. *Journal of Systems Architecture*. 2022;126:102474. <https://doi.org/10.1016/j.sysarc.2022.102474>.
51. Fiosina J. Explainable federated learning for taxi travel time prediction. vol. 2021-April; 2021. p. 670 - 677. Available from: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121125924&partnerID=40&md5=9a076bc80768b97eee10c0883e2557e7>.
52. Fiosina J. Interpretable privacy-preserving collaborative deep learning for taxi trip duration forecasting. In: International conference on vehicle technology and intelligent transport systems, international conference on smart cities and green ICT systems. Springer; 2022. p. 392–411.
53. Wang G. Interpret federated learning with Shapley values. arXiv preprint [arXiv:1905.04519](https://arxiv.org/abs/1905.04519). 2019;.
54. Sidhpura J, Shah P, Veerkhare R, Godbole A. FedSpam: privacy preserving SMS spam prediction. In: Tanveer M, Agarwal S, Ozawa S, Ekbal A, Jatowt A, editors. *Neural information processing*. Singapore: Springer Nature Singapore; 2023. p. 52–63.
55. Ben Saad S, Brik B, Ksentini A. A trust and explainable federated deep learning framework in zero touch 5G networks. In: GLOBECOM - IEEE Global Communications Conference; 2022. p. 1037–1042.
56. Ludwig H, Baracaldo N, Thomas G, Zhou Y, Anwar A, Rajamoni S, et al.: IBM federated learning: an enterprise framework white paper V0.1. [arXiv](https://arxiv.org/abs/2007.10987). Available from: <https://arxiv.org/abs/2007.10987>.
57. Wilbik A, Grefen P. Towards a federated fuzzy learning system. In: IEEE International conference on fuzzy systems (FUZZ-IEEE); 2021. p. 1–6.
58. Wu Y, Cai S, Xiao X, Chen G, Ooi BC. Privacy preserving vertical federated learning for tree-based models. *Proc VLDB Endow*. 2020;13(12):2090–103. <https://doi.org/10.14778/3407790.3407811>.
59. Zhu X, Wang D, Pedrycz W, Li Z. Horizontal federated learning of Takagi-Sugeno fuzzy rule-based models. *IEEE Transactions on Fuzzy Systems*. 2022;30(9):3537–47. <https://doi.org/10.1109/TFUZZ.2021.3118733>.
60. Bogdanova A, Imakura A, Sakurai T. DC-SHAP method for consistent explainability in privacy-preserving distributed machine learning. *Human-Centric Intelligent Systems*. 2023;3(3):197–210. <https://doi.org/10.1007/s44230-023-00032-4>.
61. Zheng S, Cao Y, Yoshikawa M. Secure Shapley value for cross-silo federated learning. *Proc VLDB Endow*. 2023, 16(7), pp. 1657–1670. <https://doi.org/10.14778/3587136.3587141>
62. Janzing D, Minorics L, Blöbaum P. Feature relevance quantification in explainable AI: a causal problem. In: International Conference on artificial intelligence and statistics. PMLR; 2020. p. 2907–2916. Available from: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85122610367&partnerID=40&md5=32e4054213ccb02f05bfa648c864fbb>.
63. Chen Y, Yang X, He Y, Miao C, Chan P. FedDBM: federated digital biomarker for detecting Parkinson's disease progress. In: IEEE International Conference on Multimedia and Expo (ICME); 2023. p. 678–683.
64. Polato M, Esposito R, Aldinucci M. Boosting the federation: cross-silo federated learning without gradient descent. In: International Joint Conference on Neural Networks (IJCNN); 2022. p. 1–10.
65. Alonso Moral JM, Castiello C, Magdalena L, Mencar C. In: *Designing interpretable fuzzy systems*. Cham: Springer International Publishing; 2021. p. 119–168. Available from: [https://doi.org/10.1007/978-3-030-71098-9\\_5](https://doi.org/10.1007/978-3-030-71098-9_5).
66. Gacto MJ, Alcalá R, Herrera F. Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures. *Information Sciences*. 2011;181(20):4340–60. <https://doi.org/10.1016/j.ins.2011.02.021>. (**Special Issue on Interpretable Fuzzy Systems**).
67. Zhu H, Xu J, Liu S, Jin Y. Federated learning on non-IID data: a survey. *Neurocomputing*. 2021;465:371–90. <https://doi.org/10.1016/j.neucom.2021.07.098>.
68. Morafah M, Wang W, Lin B. A practical recipe for federated learning under statistical heterogeneity experimental design. *IEEE Transactions on Artificial Intelligence*. 2023;p. 1–14. <https://doi.org/10.1109/TAI.2023.3297090>.
69. Shapley LS. In: Kuhn HW, Tucker AW, editors. 17. A value for n-person games. Princeton University Press: Princeton; 1953. p. 307–18.
70. Molnar C.: *Interpretable machine learning: a guide for making black box models explainable*. Available from: <https://christophm.github.io/interpretable-ml-book>.
71. Tsanas A, Little MA, McSharry PE, Ramig LO. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*. 2010;57(4):884–93. <https://doi.org/10.1109/TBME.2009.2036000>.
72. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al.: *Advances and open problems in federated learning*. Available from: <http://dx.doi.org/10.1561/22000000083>.
73. Bakopoulou E, Tillman B, Markopoulou A. FedPacket: a federated learning approach to mobile packet classification. *IEEE Transactions on Mobile Computing*. 2022;21(10):3609–28. <https://doi.org/10.1109/TMC.2021.3058627>.
74. Corcuera Bárcena JL, Ducange P, Marcelloni F, Nardini G, Noferi A, Renda A, et al. Enabling federated learning of explainable AI models within beyond-5G/6G networks. *Computer Communications*. 2023;210:356–75. <https://doi.org/10.1016/j.comcom.2023.07.039>.
75. Cover TM. *Elements of information theory*. John Wiley & Sons; 1999.
76. Cózar J, Ossa Ldl, Gámez JA. TSK-0 fuzzy rule-based systems for high-dimensional problems using the apriori principle for rule

- generation. In: International Conference on Rough Sets and Current Trends in Computing. Springer; 2014. p. 270–279.
77. Li Q, Wen Z, Wu Z, Hu S, Wang N, Li Y, et al. A survey on federated learning systems: vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*. 2023;35(4):3347–66. <https://doi.org/10.1109/TKDE.2021.3124599>.
  78. Alonso JM, Magdalena L, Guillaume S. HILK: a new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism. *International Journal of Intelligent Systems*. 2008;23(7):761–94. <https://doi.org/10.1002/int.20288>.
  79. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*. 1956;63(2):81–97. <https://doi.org/10.1037/h0043158>.
  80. Wilcoxon F. Individual comparisons by ranking methods. In: *Breakthroughs in statistics*. Springer; 1992. p. 196–202.
  81. Fuchs C, Spolaor S, Nobile MS, Kaymak U. pyFUME: a Python package for fuzzy model estimation. In: *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*; 2020. p. 1–8.
  82. Obeso JA, Rodriguez-Oroz MC, Goetz CG, Marin C, Kordower JH, Rodriguez M, et al. Missing pieces in the Parkinson's disease puzzle [Review]. *Nature Medicine*. 2010;16(6):653–61. <https://doi.org/10.1038/nm.2165>.
  83. Müller S, Toborek V, Beckh K, Jakobs M, Bauckhage C, Welke P. An empirical evaluation of the Rashomon effect in explainable machine learning. In: Koutra D, Plant C, Gomez Rodriguez M, Baralis E, Bonchi F, editors. *Machine learning and knowledge discovery in databases: Research Track*. Cham: Springer Nature Switzerland; 2023. p. 462–78.
  84. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science*. 2001;16(3):199–231. <https://doi.org/10.1214/ss/1009213726>
  85. Pati S, Baid U, Edwards B, Sheller M, Wang SH, Reina GA, et al. Federated learning enables big data for rare cancer boundary detection. *Nature Communications*. 2022. <https://doi.org/10.1038/s41467-022-33407-5>.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.