Trustworthy AI in Heterogeneous Settings: Federated Learning of Explainable Classifiers

Anonymous Authors

***Abstract*—Trustworthy Artificial Intelligence (AI) has gained significant relevance worldwide. Federated Learning (FL) and eXplainable Artificial Intelligence (XAI) are two among the most relevant paradigms for accomplishing the requirements of trustworthy AI-based applications. On the one hand, FL guarantees data privacy throughout a collaborative learning of an AI model from decentralized data. On the other hand, XAI models ensure transparency, accountability, and trust in AI-based systems by providing understandable explanations for their predictions and decisions. To the best of our knowledge, only few works have explored the combination of FL with inherently explainable models, especially for classification task. In this work, we investigate FL of explainable classifiers, namely Fuzzy Rule-based Classifiers. In the proposed FL scheme, each participant creates its own set of classification rules from its own local training data, resorting to a simple procedure that generates a rule for each training instance. Local rules are sent to a central server which is in charge of aggregating them by removing duplicates and solving conflicts. The aggregated set of rules is then forwarded to the single participants for inference purposes. In our experimental analysis we consider two real-world case studies focusing on heterogeneous settings, namely non-IID (Independent and Identically Distributed) scenarios. Our FL scheme offers significant advantages in terms of classification performance to the participants in the federation, preserving data privacy.**

***Index Terms*—Federated Learning, Explainable Artificial Intelligence, Fuzzy Rule-based Classifier**

I. Introduction

In recent years, Artificial Intelligence (AI) solutions have become increasingly pervasive, extending also to sensitive domains such as healthcare and autonomous driving. Due to the potential impact on human lives that these solutions may have, it is essential to ensure their trustworthiness. The European Commission, for instance, formalized the "Ethic Guidelines for Trustworthy AI" [1] and the "AI Act" [2] to establish guidelines for enabling trustworthy AI. Transparency of AI systems and data privacy stand out as pivotal prerequisites for trustworthiness.

Transparency is defined as the capability to explain AI-based systems to a designated audience: it allows users to understand the rationale behind the results of AI tools and is becoming increasingly relevant for legal accountability [3].

The field of Explainable AI (XAI) focuses on describing the underlying architecture of AI systems and the logic behind their decision-making processes [4]. XAI approaches are categorized into two main groups: post-hoc techniques and algorithms for generating inherently explainable models. Post-hoc techniques aim to enhance the explainability of opaque models like Neural Networks (NNs). Inherently explainable models are designed to be interpretable once trained, providing insights on the reason behind their decisions.

In application domains that involve sensitive data, such as healthcare and finance, moving information from local data owners to a centralized entity to train an AI model could not be feasible due to privacy constraints. To ensure data privacy in learning AI models from decentralized data, Google introduced in 2016 the Federated Learning (FL) paradigm [5]. FL is designed to collaboratively train a global AI model across multiple distributed nodes without exposing their local raw data. In FL each node trains/updates a model exploiting its private data and then shares the updated model typically with a server which aggregates the received models into a global one, generally in a round-based iterative fashion.

Recently, the concept of Fed-XAI, which combines FL with XAI paradigms, has emerged. The synergy between these paradigms is crucial to achieve trustworthiness in AI systems as it enables the simultaneous pursuit of transparency and privacy preservation requirements. In [6] and [7] authors summarize the few contributions in which XAI models have been trained in a federated fashion, along with some real-world applications of Fed-XAI. Most efforts in this domain revolve around FL of interpretable by-design models, such as Fuzzy Rule-based Systems (FRBS), and are specifically tailored for regression tasks [8]–[10]. To the best of our knowledge, only one work reports an approach for FL of rule-based systems for classification tasks [11]: authors devised a strategy for FL of Fuzzy Rule-Based Classifiers (FRBCs) mainly focusing on imbalanced datasets. The study of post-hoc techniques in Fed-XAI is also in its infancy: SHAP [12] is one of the most popular post-hoc methods used for deriving feature importance scores, but its adaptation to the federated setting is not straightforward [13], [14]. Most of the works mentioned above only discuss the independent and identically distributed (IID) scenario which, however, is rather uncommon in the federated setting: in many real-world applications, in fact, it is likely that the local data of the different clients follow different distributions resulting in a non-IID scenario [15].

In this paper, we investigate an approach for FL of FRBCs as explainable by-design classifiers, with a special focus on their performance on data heterogeneous (that is, non-IID) scenarios. First, we extended a recently released open source Fed-XAI framework [16], currently restricted to FL of Takagi-Sugeno-Kang (TSK) FRBS for regression tasks, with the ability to address classification tasks. To this aim, we purposely designed an FL scheme which exploits insights of the well-know Chi algorithm [17] for generating classification

rules from a labelled training set. Then, we carried out an experimental analysis involving two real-world case studies under different data distribution settings. We compared the results achieved with the FRBC generated in a federated fashion with those obtained by learning an FRBC on each node by applying the classical Chi algorithm on local data. Moreover, we compared the proposed FL scheme with the centralized baseline. For this purpose, the classical Chi algorithm is applied on the union of local data, under the assumption (actually unfeasible in the FL setting due to privacy constraints) that local raw data can be moved from nodes where are stored for centralized processing. Finally, results obtained with centralized Chi algorithm are also compared with two state-of-the-art opaque models, namely a Multi-Layer-Perceptron (MLP) and a Random Forest (RF).

The rest of the paper is organized as follows: Section II provides a brief background on Fed-XAI and FRBCs. In Section III we describe the proposed FL scheme for FRBCs. Section IV describes the case studies used in our experimental analysis, and Section V presents and discusses the results of this analysis. Finally, Section VI concludes with appropriate remarks.

II. Background

In this section we first illustrate the main concepts of FL and Fed-XAI paradigms. Then we provide some background on traditional FRBC and on the Chi algorithm.

Federated Learning and Fed-XAI

FL is a recent paradigm enabling collaborative training of ML and AI models among multiple participants while preserving data privacy, as only model updates or aggregated statistics are shared [18], [19]. Data partitioning in FL is categorized into horizontal and vertical settings, depending on whether instances or features are partitioned among clients, respectively. Additionally, FL varies in scale: cross-silo FL involves few participants like organizations which are characterised by abundant data and computing resources, while cross-device FL involves many participants like smartphones with limited data and resources. Most FL methods use the federated averaging (FedAvg) protocol [15], an iterative process where participants update a global model using Stochastic Gradient Descent or its variants. This approach suits Deep Learning and NN models but is not immediately applicable to models like Decision Trees (DTs) and Rule-base Systems (RBSs), which are typically not learned through the optimization of a differentiable global objective function.

The concept of Fed-XAI aims to enhance users' trust in AI systems by simultaneously addressing the requirements of privacy preservation (through FL) and explainability (using XAI models and techniques). Research in this area is growing, focusing on both post-hoc methods [20]–[23] and explainable by-design models [8]–[10], [24]. The adoption of post-hoc explanation in FL context is anything but trivial. Several recent works have focused on SHAP, as one of the most popular post-hoc methods [13], [14]: the main issue is that SHAP requires access to training instances (often referred to as the reference dataset) even at explanation time, i.e. when it comes to explaining the output of a model given an input instance. Evidently, in the FL scenario the whole training set is not available at a generic node, which can at most only rely on its local training data. This means that, given the same input instance, FL model and predicted output, different nodes may obtain different explanations. Although tackling these issues presents an interesting research challenge, it is at the same time worth investigating FL of inherently explainable models, which do not require the adoption of post-hoc techniques.

As regards explainable by-design models, one of the main challenges in Fed-XAI is the aggregation of the local models: the FL of models such as DTs and RBSs cannot rely on the traditional FedAvg protocol but rather requires ad-hoc aggregation strategies.

The works discussed in [8] and [10] propose FL schemes for learning TSK-FRBS for regression problems. Both approaches encompass two main stages: i) learning the fuzzy partitions of each input feature and the antecedents of the rules, and ii) learning the consequents for each rule. For the first stage, [8] uses a federated version of the Fuzzy C-Means algorithm to identify global clusters, while [10] employs a local clustering process followed by an aggregation stage on a central server, where similar clusters are merged. After identifying clusters, both studies exploit a typical method for determining antecedent parameters, particularly the membership functions, which involves a Gaussian fitting of the convex envelope of the projected membership values for each cluster. For the second stage, both [10] and [8] apply a federated version of gradient-based learning methods. Another federated algorithm for generating TSK-FRBS for regression problems has been discussed in [9]. In this approach, each client generates a TSK-FRBS from local data and shares it with the server. The server then aggregates the TSK-FRBSs by combining the rules received from the clients and resolving any possible conflict. A conflict occurs when rules from different TSK-FRBSs cover the same specific area of the attribute space (i.e., they have identical antecedents) but have different consequent parameters. Conflict resolution involves creating a single rule from each set of conflicting rules. This rule maintains the common antecedent and has as consequent the average of consequent coefficients of the conflicting rules. This approach ensures a higher explainability than the ones in [10] and [8], because it exploits pre-defined strong fuzzy partitions and adopts the maximum matching strategy in the inference.

The work discussed in [11] introduces a federated approach for incrementally learning the rules of an FRBC. A dedicated weighting scheme is proposed for addressing data imbalance. The approach encompasses two rounds of federation: the first is used for generating the rules and the second computes the rule weights.

Fuzzy Rule Based Classifier

An FBRC basically includes a rule base (RB), a database (DB) containing the definition of the fuzzy sets used in the RB,

and an inference engine. RB and DB comprise the knowledge base of the rule-based system.

Let $X = \{X_1, \ldots, X_F\}$ be the set of input attributes and Y be the output variable. Let U_f, with $f = 1, ..., F$, be the universe of the f^{th} input attribute X_f. Let $P_f = \{A_{f,1}, \ldots, A_{f,j}, \ldots, A_{f,T_f}\}$ be a partition of attribute X_f consisting of T_f fuzzy sets. The output variable Y is a categorical variable with values in the set $\Gamma = \{C_1, \ldots, C_K\}$ of K possible classes. Let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ be a training set composed of N input–output pairs, with $\mathbf{x}_t = [x_{t,1} \ldots, x_{t,F}] \in \mathbb{R}^F$, $t = 1, \ldots, N$ and $y_t \in \Gamma$. Fuzzy sets may be characterized by different membership functions. A popular choice consists in using triangular fuzzy sets: each fuzzy set $A_{f,j}$ is identified by the tuples $(a_{f,j}, b_{f,j}, c_{f,j})$, where $a_{f,j}$ and $c_{f,j}$ correspond to the left and right extremes of the support, and $b_{f,j}$ to the core. In our experiments, we use strong uniform fuzzy partitions with triangular fuzzy sets [25]. The generic m-th rule R_m of an RB is expressed as follows:

$$R_m : \textbf{IF } X_1 \textbf{ is } A_{1,j_{m,1}} \textbf{ AND} \ldots \textbf{AND } X_F \textbf{ is } A_{F,j_{m,F}} \textbf{ THEN } Y \textbf{ is } C_{j_m} \textit{ with } RW_m \quad (1)$$

where C_{j_m} is the class label associated with the rule, and RW_m is the rule weight, i.e., a certainty degree of the classification in the class C_{j_m} for an instance belonging to the subspace delimited by the antecedent part of R_m. RW_m generally is computed as the *certainty factor* (CF_m) [26]. For a generic rule, CF_m is defined as

$$CF_m = \frac{\sum_{\mathbf{x}_t \in C_{j_m}} w_m(\mathbf{x_t})}{\sum_{t=1}^{N} w_m(\mathbf{x_t})} = \frac{Num_m}{Den_m} \quad (2)$$

The term $w_m(\mathbf{x_t})$ represents the matching degree, or strength of activation, for the rule R_m and an input instance $\mathbf{x_t} \in \mathbb{R}^F$. Formally:

$$w_m(\mathbf{x_t}) = \prod_{f=1}^{F} A_{f,j_{m,f}}(x_{t,f}) \quad (3)$$

For the sake of simplicity, in the formula, we have considered the product as t-norm for the logical conjunction. The matching degree captures the compatibility degree between the rule antecedent and the feature values of the input instance. In the computation of the certainty factor, the numerator in Eq. 2 represents the sum of matching degrees for training instances of class C_{j_m} within the fuzzy region defined by the antecedent. The denominator is the sum of matching degrees for all the training instances within this fuzzy subspace, regardless of their associated class.

An RB with M rules can be used for inference purpose, i.e., for determining the class of any given input instance $\hat{\mathbf{x}}$. First, the *association degree* $h_m(\hat{\mathbf{x}})$ of each rule is computed as:

$$h_m(\hat{\mathbf{x}}) = w_m(\hat{\mathbf{x}}) \cdot RW_m \quad (4)$$

Then, a *reasoning method* is applied to determine the predicted class. Maximum matching represents a commonly used reasoning method: an input instance is classified into the class corresponding to the rule with the maximum association degree. In case of tie, the class of the most specific rule or of the rule with the highest RW is typically assigned to the instance.

A popular algorithm for the generation of rules in an FRBC is the Chi algorithm [17]. As discussed in [27], due to its simplicity, this algorithm has also been successfully adopted in different distributed versions for big data classification. The field of distributed ML has similarities with horizontal FL, but also a substantial difference: in FL multiple parties have their own data and are reluctant to share them due to privacy constraints, whereas in distributed ML privacy is not a concern and multiple nodes are employed to enhance processing power and memory for handling big datasets.

In practice, the Chi algorithm relies on a pre-defined DB describing the fuzzy partitions of each attribute and generates a rule for each training instance. The antecedent of a rule is generated considering, for each attribute, the fuzzy set that has been activated by the training instance with the highest membership degree; the consequent is directly specified by the label associated with the training instance itself. Duplicate rules (i.e., those having the same antecedents and the same consequents) are removed and appropriate strategies have been defined for handling conflicting rules (i.e., those having the same antecedents and different consequents), and for assigning a weight to each rule.

III. Federated Fuzzy Rule Based Classifier

The proposed FL algorithm is designed for generating the set of rules of an FRBC in a collaborative and privacy preserving way and is schematized in Fig. 1. Specifically, first, each participant in the federation independently carries out the local training of the model, namely generates a set of rules with its own data. Then, the set of rules is sent to a central server which is in charge of aggregating them. Finally, the aggregated RB is sent back to all the participants, which can use it for inference purposes.

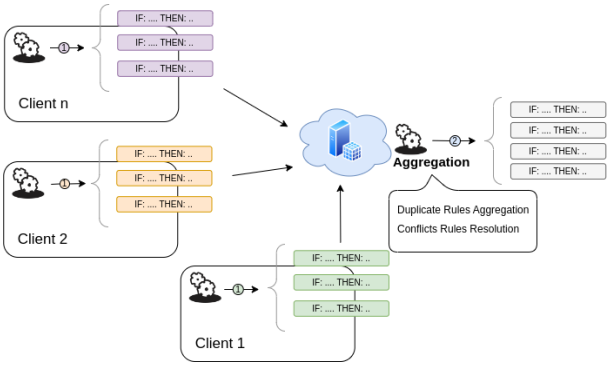


Fig. 1. Overview of the FL algorithm for FRBC generation.

The proposed approach for FL of FRBC stems from the Chi algorithm and exploits a one-shot procedure (i.e., a single round of federation), analogously to what has been proposed

for federated TSK-FRBS [9]. In the following, we describe in detail the steps of our proposed approach:

- The central server configures the learning process by sending a set of hyperparameters to each data owner. Such set includes: the domain of definition of the attributes for data normalization and the number of fuzzy sets T_f ($f = 1, \ldots, F$) for fuzzy partitioning of input attributes.
- Once it has received the hyperparameters, each local node LN^i starts the rule generation stage. It generates a classification rule R_m^i for each training instance, as in the classical Chi algorithm. At this stage duplicate rules are discarded whereas conflicting rules are maintained. Moreover, rather than computing the rule weights, for each rule R_m^i the local node computes and stores the values of the numerator Num_m^i and the denominator Den_m^i of the certainty factor (see Eq. 2)
- Each node LN^i sends the local set of rules along with their associated values of Num_m^i and Den_m^i to the central server.
- The central server creates a temporary RB composed by the juxtaposition of the rules collected from the different local nodes. Notably, the temporary RB may contain duplicate rules (originating from different nodes) and conflicting rules.
- The server creates a final global RB. If a rule has not duplicates or conflicts, it is retained in the RB and its weight is computed as in equation 2. Each set DR_m of duplicate rules is combined into a single global rule R_m whose weight RW_m is determined as follows:

$$RW_m = \frac{\sum_{R_m^i \in DR_m} Num_m^i}{\sum_{R_m^i \in DR_m} Den_m^i}. \tag{5}$$

  For each set of conflicting rules only the one with the highest weight is retained in the final global RB.

The resulting FRBC represents the federated model, which is eventually sent back to the clients for local inference.

Notably, the antecedent part of the RB obtained in the federated setting is equal to that obtained in the centralized case. However, the weights of the rules are not necessarily the same: in the federated case, a participant will contribute to the global computation of the certainty factor (i.e., RW) only for the rules generated by exploiting its local training set. Conversely, in the centralized case, each training instance contributes to the computation of the certainty factor for all the rules for which it has a non-zero strength of activation. Obviously, a discrepancy in the rule weights between federated and centralized settings may lead to a different outcome of the conflict resolution process and therefore to a difference in the consequent part of the resulting FRBC models.

As discussed in Section I, the FL scheme discussed in this section has been integrated in a publicly available open source framework which supports the implementation in Python of FL schemes for FRBSs[1].

[1]Source code will be available upon acceptance of this paper

IV. Experimental case studies

In this section, we describe the two case studies considered in our experimental analysis. The first one regards the recognition of high energy gamma particles in the atmosphere and is based on the publicly available MAGIC Gamma Telescope [28] dataset. The second one pertains to a vehicle networking environment and was designed within the framework of *HEXA-X*[2], the EU flagship project for Beyond 5G (B5G) and 6G networks. This scenario regards the prediction of the Quality of Experience (QoE) for video streaming across multiple vehicles.

In the following, details of the two case studies are illustrated.

A. Gamma Signal Detection

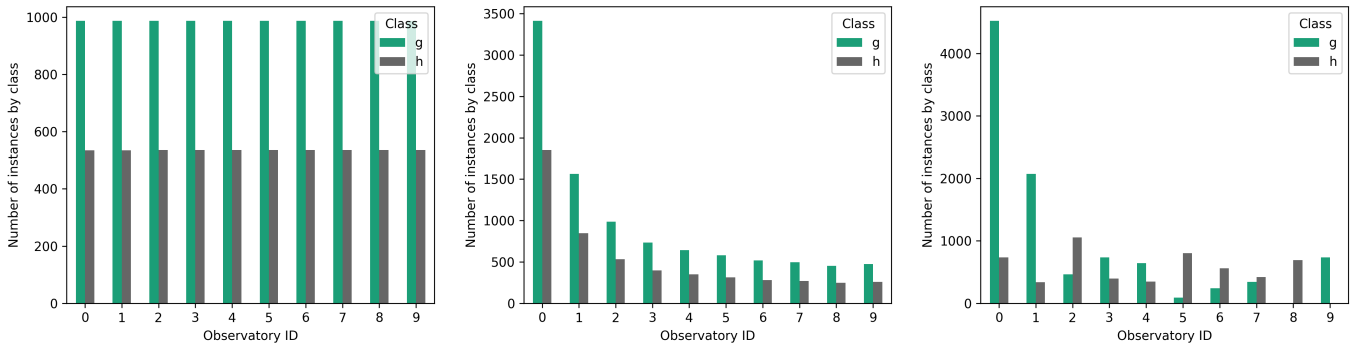
The case study considers a ground-based atmospheric Cherenkov gamma telescope that registers the observed high-energy gamma particles employing an imaging technique. This type of telescope detects high-energy gamma rays by capitalizing on the radiation emitted by charged particles generated within the electromagnetic showers. These showers are initiated by gamma particles and develop in the atmosphere. The telescope collects photons, forming distinct patterns.

In this case study, the primary objective is to differentiate between instances originating from primary gamma rays and those from hadronic showers initiated by cosmic rays in the upper atmosphere. The dataset consists of 19,020 instances, split into 12,332 instances belonging to class *gamma* (g) and 6,688 instances belonging to class *hadron* (h). In our experiments, we simulated a scenario in which 10 observatories, each equipped with a telescope, are involved in a federated study. To this aim, we consider three different distributions of the data of the MAGIC Gamma Telescope dataset across the 10 observatories. Specifically, we defined the following scenarios:

- **IID**: all the observatories have approximately the same class distribution and the same volume (i.e., number of instances). As a consequence, each observatory data is a representative instance of the overall dataset.
- **Q-NIID** (non-IID scenario with *quantity* skew [29]): the class distribution remains consistent, whereas the volume of local dataset varies across observatories.
- **QL-NIID** (non-IID scenario with *quantity* and *label* skew): in this case, both the volume and the distribution of classes vary across observatories. As a consequence, local data distributions are generally different from each other and from the overall data distribution.

Figure 2 shows the details of the different training data distributions for the Gamma Signal Detection scenario. These distributions can simulate three different real-world situations that may occur in a scientific setting where different research centers are not allowed to share their raw data.

[2]https://hexa-x.eu/, visited on January 2024



(a) IID (b) Q-NIID (c) QL-NIID

Fig. 2. Gamma Signal Detection case study: barplot of the different data distribution scenarios.

B. QoE Prediction in B5G/6G Networks

The second case study, shown in Fig. 3, has been introduced and described in [30] and deeply analyzed in [31]. It involves an automotive environment where connected vehicles serve as User Equipments (UEs) within the mobile network. Each UE is connected to its respective base station (BS) and receives a live video feed from the camera of the vehicle ahead. This setup potentially supports advanced driving assistance systems, like safety distance evaluation. A crucial requirement for providing these services is the continuous display of the video in high quality. Thus, an AI-based service which continuously monitors the state of the network and provides an alert in case of decrease of the video quality is envisioned to be deployed in next generation networks [32].

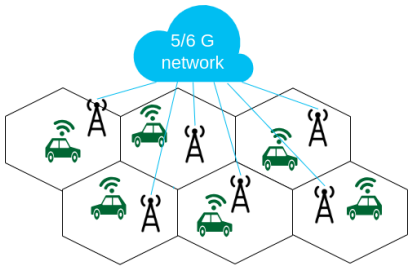


Fig. 3. Schematized representation of the QoE prediction case study.

In our experimental analysis, we adopt the publicly available QoE prediction dataset [30]. It encompasses 15 UEs within a geographical space served by seven BSs to which the vehicles are connected. For each UE different metrics (regarding contextual information, Quality of Service and QoE) are gathered from 24 simulations, each lasting approximately 120 seconds. Further details are available in [30]. The training set comprises the data from the initial 20 simulations, while the last 4 simulations are designated for the test set. Notably, each UE has its own dedicated test set. We created a binary classification dataset by following the details provided in [30]. Indeed, we define the QoE prediction problem as a classification task. The target is to estimate the future level (*Good* or *Poor*) of the video quality at a specific time horizon of 3 seconds in the future. The target value is estimated considering as input to our classification model a set of statistics calculated on the historical values of the metrics within a time window of 10 seconds. Specifically, following the preprocessing steps proposed in [30], we considered three network metrics, namely the Signal to Interference plus Noise Ratio (SINR) value measured at packet reception, the number of resource blocks occupied and the distance from the serving cell. For each metric, we extracted 10 statistics (mean, median, max, min, variance, standard deviation, kurtosis, skewness, Q1 and Q3). Thus, in total we consider 30 input features. The value of the target was defined as Good if the value of the metric *framesDisplayed* (the fraction of frames arrived at the time the video was displayed) is greater than 0.8, and Poor otherwise.

In Figure 4 we show the two training data distribution scenarios adopted for the second case study.

Specifically, we consider:

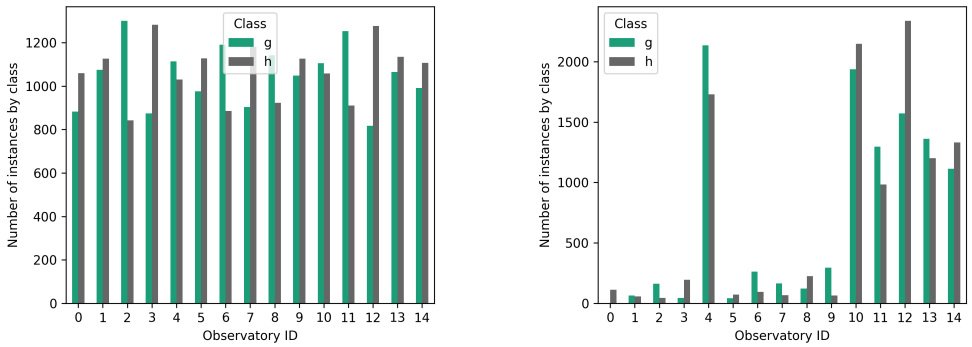
- **IID**: we maintain the original data distribution, with the instances of 20 video sessions for each UE. It is worth noticing that due to the nature of the dataset the actual class distribution may slightly vary across participants.
- **QL-NIID**: it is obtained by randomly shifting video sessions among the different UEs, forcing half of them to include a few sessions.

Unlike the previous case study, the Q-IID scenario is not discussed: in fact, due to the nature of the simulated video sessions, altering the data volume across participants inevitably results also in a label skewness.

It is worth to recall that each UE has its own test set, which is always the same in both cases, and is composed by the instances of 4 video sessions.

C. Experimental Setup

For each dataset and each data distribution scenario, we evaluate the results achieved with FRBCs designed considering three different learning schemes: FL, Local Learning (LL) and Centralized Learning (CL). In FL, local nodes collaborate in obtaining a single federated model without compromising



(a) IID (b) QL-NIID

Fig. 4. QoE prediction case study: barplot of the different data distributions scenarios.

the data privacy, as described in Section III. In LL each node individually learns a model from its local data only. In this setting, raw data privacy is preserved but there is no collaboration among nodes. Hence, the assessment of the performance of an FL approach entails measuring the gain with respect to the LL setting. Finally, in CL data from all nodes are first gathered and stored on the central server and then are used to learn the model. This setting represents the utmost form of collaborative training, but implies the violation of data privacy, as raw data need to be transferred to the server. The CL scheme is considered as a baseline for assessing whether the FL scheme achieves acceptable results in terms of classification performance.

Notably, the same testing data are considered for each learning scheme and for each data distribution scenario. In this way, we ensure a fair evaluation of the results in the comparative experimental analysis. As for the QoE case study, in the original dataset each UE has its own test set. As for the Gamma Signal Detection case study, we carried out a 5-fold cross-validation (CV) analysis. For each iteration of the CV the test set of each local node follows the same distribution as the original data (IID), whereas the training data varies according to the distribution scenario considered.

To further enforce the comparative analysis, we consider two additional baseline opaque models in our CL experiments, namely an MLP and an RF. Although it may be valuable to assess the performance of such models also in the federated setting, we recall that our focus is on both model performances and explainability. For the latter purpose, however, such models would require the design of post-hoc explainability techniques compliant with the federated setting (as discussed in Section II) which are outside the scope of this work.

The training of the FRBC model is influenced by a single hyperparameter, namely the number of fuzzy sets T_f used to partition each input attribute. We carried out our experimental analysis considering the same value of T_f for each input attribute and setting T_f equal to 3, 5 and 7. Here, for the sake of brevity, we show only the results achieved considering T_f equal to 5 and 3 for the Gamma Signal Detection and QoE Prediction case studies, respectively. Indeed, these settings provide the best results in terms of trade-off between the classification performance and the complexity of the FRBCs.

V. Experimental Results

In this section, we report the results achieved on the two considered case studies.

A. Gamma Signal Detection Results

As discussed in Section IV-A, data are distributed across 10 local nodes, i.e., 10 observatories, following 3 different distribution scenarios.

Table I shows an aggregated view of the classification performance, in terms of F1-score per class, achieved considering the three learning schemes for FRBC and the CL scheme for MLP and RF. The table reports the average F1-scores over the 10 local nodes along with their standard deviation. For each node results were averaged over the five iterations of the CV.

The FL approach improves the performance of its LL counterpart. Although in the IID scenario the two learning schemes allow achieving comparable performances, the federated FRBC significantly outperforms the local ones in the non-IID settings, especially under both quantity and label skewness (QL-NIID). The federated models are not particularly affected by the non-IID settings, whereas those learned locally lose some generalization capability, as highlighted by the decrease in F1-scores. The only exception for the FL setting is represented by the average F1-scores on class h, which in the QL-NIID case is a couple of percentage points lower than in the IID case. Furthermore, the FL scheme obtains performance levels that are nearly on par with those achieved using the CL scheme. The possible misalignment of rule weights between the FL scheme and the CL scheme, discussed in Section III, has a rather limited impact on performance in this case study.

Opaque models (RF and MLP) outperform the FRBC, especially on the minority class h. Nevertheless, the rule-based model achieves classification performance that can still be considered acceptable and represents a different trade-off, compared to opaque models, between accuracy and interpretability: on one hand, in fact, the reason behind the

TABLE I
GAMMA SIGNAL DETECTION SCENARIO: F1-SCORES ACHIEVED IN THE DIFFERENT EXPERIMENTS. AVERAGE VALUES ± STANDARD DEVIATION.

	FRBC IID		**FRBC Q-NIID**		**FRBC QL-NIID**		**FRBC**	**MLP**	**RF**
class	LL	FL	LL	FL	LL	FL	CL	CL	CL
g	0857 ± 0.009	0.867 ± 0.007	0.850 ± 0.009	0.866 ± 0.007	0.685 ± 0.284	0.859 ± 0.006	0.867 ± 0.007	0.909 ± 0.005	0.910 ± 0.001
h	0.676 ± 0.021	0.690 ± 0.018	0.662 ± 0.022	0.688 ± 0.017	0.553 ± 0.210	0.666 ± 0.016	0.690 ± 0.017	0.818 ± 0.008	0.817 ± 0.005

decision obtained from a rule is easily understood by a human. On the other hand, the gain in classification accuracy of opaque models comes at the cost of sacrificing inherent interpretability.

The average values of F1-score shown in Table I, however, provide only a rough indication of the performances, but do not accurately reflect the actual situation experienced on each local node. To provide a fine grained picture of the performance of LL, FL and CL schemes, we report in Figure 5, the empirical cumulative distribution functions (ECDFs) for the F1-score on the g class (which represents the event of interest viz. the gamma radiation).

Specifically, in the figure we report the ECDF for the values of the difference, for each local node, of the F1-score between the FL scheme and the LL scheme (Δ_{FL-LL}, dark blue circles) and between the FL scheme and the CL scheme (Δ_{FL-CL}, cyan diamonds) for the three considered distributions IID (Fig. 5a), Q-NIID (Fig. 5b) and QL-NIDD (Fig. 5c). Each curve has 10 points, coherently with the number of observatories considered in our experiments. Whenever a point lies in the positive half-plane (positive F1-score difference) it indicates that the F1-score of the FL scheme is higher (and therefore better) compared to the other one (either CL or LL). We observe that the difference between FL and LL is stable within the positive half-plane, regardless of the data distribution scenario. As expected, the most significant improvement is obtained for the QL-NIID distribution. Additionally, the points representing the difference between the FL and the CL schemes are located around the black vertical line (indicating a difference equal to zero), attesting the similarity in their performance.

In Table II we show the average number of rules in the RBs for the different data distribution scenarios considered in our experiments.

TABLE II
GAMMA SIGNAL DETECTION CASE STUDY: NUMBER OF RULES FOR THE DIFFERENT DATA DISTRIBUTION SCENARIOS AND THE DIFFERENT LEARNING SCHEMES. AVERAGE VALUES OVER 5-FOLD CV ARE REPORTED ALONG WITH THE STANDARD DEVIATION.

	IID	**Q-NIID**	**QL-NIID**
LL	445.5 ± 5.8	425.5 ± 211.0	423.0 ± 133.4
FL	1789.0 ± 16.6	1788.8 ± 16.9	1788.4 ± 17.2
CL	1789.0 ± 16.6	1788.8 ± 16.9	1788.4 ± 17.2

The algorithm used for rule generation implies that the number of rules is positively correlated, in general, with the number of training instances. This motivates the high standard deviation observed for the LL scheme in the non-IID scenarios. As far as the FL scheme is concerned, the number of rules is almost constant across the data distribution scenarios: although the central server receives a different number of rules from each local node the aggregation algorithm generates approximately the same number of rules. Furthermore, as expected, the values obtained in FL perfectly coincide with those of the CL setting. As stated in Section III, this is due to the fact that both schemes use, directly or indirectly, all the training instances.

As regards the explainability of the FRBCs generated with the three different learning schemes, we highlight that models generated with the LL schemes are more compact, and thus feature higher global explainability than the ones generated via FL. As highlighted before, this also entails a poor classification capability for the LL setting, especially when considering non-IID scenarios. Models learned with FL and CL schemes are more complex and indeed harder to interpret from a global point of view. The global explainability of such models could be enhanced by exploiting methods for reducing the RB complexity, e.g., through an a-posteriors rule selection algorithm [33].

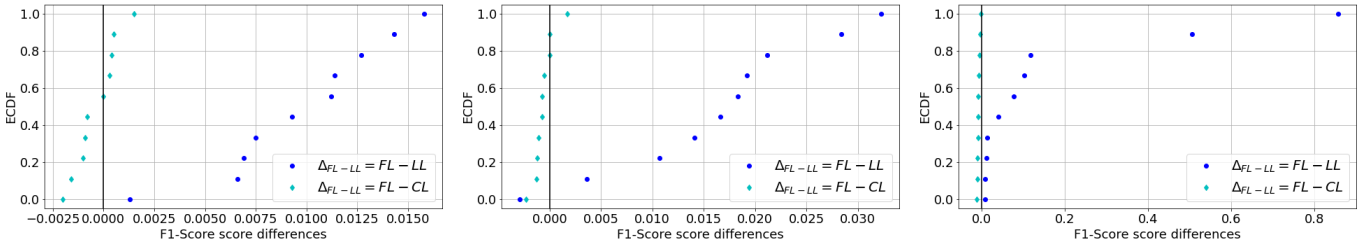
It is worth noticing that model explainability does not only concern its structural properties but also the inference process (local explainability). In our FRBCs the predicted output is determined by a single rule, which allows for straightforward interpretation: the antecedent of the rule defines a specific region within the search space, while its consequent describes the estimated class. Unlike opaque models, local explainability in FRBC is guaranteed "by construction" and does not require the adoption of post-hoc techniques.

B. QoE Prediction Results

The QoE prediction case study involves 15 UEs. We recall that each local node has its own test set which encompasses four video streaming simulations and does not vary among the various data distribution scenarios.

Table III shows the results, in terms of average F1-scores, achieved in the different experiments.

Also in this case study the FL scheme outperforms the LL counterparts. Moreover, the FRBCs generated by the FL scheme achieve performance nearly equivalent to the ones obtained by the CL approach. The QL-NIID scenario leads to a drop in average F1-score for the models learned according to LL scheme compared to the IID scenario, whereas the federated FRBC mantains its performance level. The generalization capability of opaque models is greater compared to FRBCs, but the discrepancy is less evident compared to the other case study.



(a) IID (b) Q-NIID (c) QL-NIID

Fig. 5. Gamma Signal Detection case study: ECDFs of the differences of F1-scores for the g class between FL and LL (Δ_{FL-LL}, dark blue) and between FL and CL (Δ_{FL-CL}, light blue) for the three data distribution scenarios.

TABLE III
QOE PREDICTION IN B5G/6G NETWORKS: F1-SCORES ACHIEVED IN THE DIFFERENT EXPERIMENTS. AVERAGE VALUES ± STANDARD DEVIATION.

	FRBC IID		**FRBC QL-NIID**		**FRBC**	**MLP**	**RF**
class	LL	FL	LL	FL	CL	CL	CL
Good	0.670 ± 0.082	0.708 ± 0.074	0.587 ± 0.216	0.708 ± 0.076	0.708 ± 0.074	0.7256 ± 0.071	0.720 ± 0.069
Poor	0.721 ± 0.091	0.743 ± 0.069	0.677 ± 0.109	0.741± 0.070	0.741 ± 0.069	0.774 ± 0.071	0.781 ± 0.068

To provide additional understanding of the aggregated results shown in Table III, we report in Fig. 6 the ECDFs related to the F1-score for the Poor class.

The ECDFs confirm that FL brings benefits compared to the LL setting to most of the participants involved in the federation compared to the LL scheme, in terms of classification performance. In both data distribution scenarios the ECDFs between FL and CL are close to the vertical black lines, meaning that the difference in performance between FL and CL models over local test sets is generally close to zero.

Table IV shows the number of rules for all the FRBCs under the different scenarios. For the LL scheme, we show the average number of rules calculated considering the different RBs of each local node, along with the standard deviation. The analysis of the values in the table confirms what was observed in the previous case study. The number of rules is generally higher than in the other case study, likely due to the larger volume of the QoE dataset. The number of rules obtained considering the FL and CL schemes is exactly the same. Although the case study encompasses 15 participants, the federated model has a number of rules that is higher than that of the LL models by a factor of 5.9 in the IID setting and 7.3 in the non-IID setting. This is clearly due to the effective management of duplicate and conflicting rules.

TABLE IV
QOE PREDICTION CASE STUDY: NUMBER OF RULES FOR THE DIFFERENT DATA DISTRIBUTION SCENARIOS AND DIFFERENT LEARNING SCHEMES.

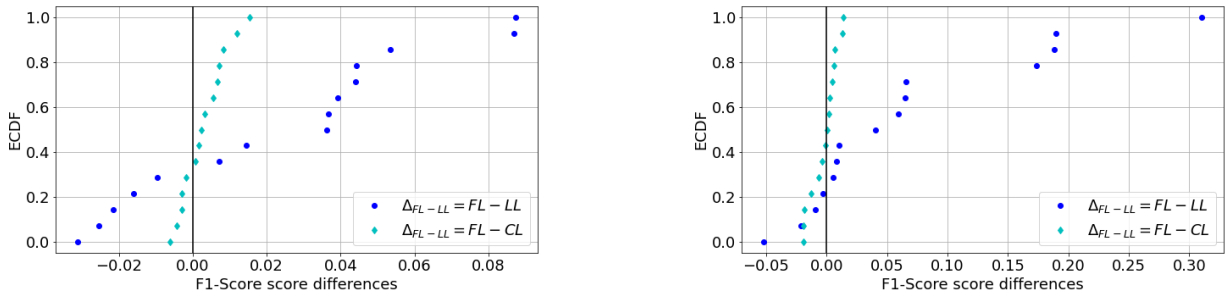
	IID	**QL-NIID**
LL	811.7 ± 37.3	657.4 ± 755.6
FL	4809	4809
CL	4809	4809

VI. CONCLUSION

In this paper we investigated an approach for Federated Learning (FL) of Fuzzy Rule-based Classifiers (FRBCs) within the framework of trustworthy AI. On one hand, FL ensures privacy preservation in decentralized collaborative model learning; on the other hand FRBCs are generally deemed as highly interpretable models, thus meeting the transparency requirement for enhancing users' trust. Our FL scheme for the generation of FRBCs consists of a one-shot procedure and is designed as follows: first, each local data owner generates a set of rules based on its private data and shares them with a central server along with summary information on the activation degree of each rule; then the server exploits the summary information for the computation of the weight of each rule and aggregates the received rules by handling duplicates and conflicts. We assessed the performance of our federated FRBC under two real-world case studies exploring both independent and identically distributed (IID) and non-IID scenarios. Furthermore, the FL scheme is compared with two alternative learning approaches, namely centralized and local learning. Centralized learning removes the privacy constraint and entails the adoption of the classical FRBC generation algorithm on the union of local training sets. Local learning waives the collaboration among participants in the sense that each data owner learns an FRBC based only on its private data. Results highlight that the federated approach allows outperforming the models learned locally and obtaining classification performance close to the centralized FRBC. Although the federated FRBCs are slightly outperformed by state-of-art opaque models learned in a centralized fashion, the proposed FL scheme enables the effective and efficient generation of classification models that feature both satisfactory levels of classification performance and of inherent interpretability. Interesting future developments of this work include the design of strategies for reducing the complexity of the federated model and approaches for handling streaming data.

REFERENCES

[1] High Level Expert Group on AI, "Ethics Guidelines for Trustworthy



(a) IID

(b) QL-NIID

Fig. 6. QoE prediction case study: ECDFs of the differences of F1-scores of the Poor class between FL and LL (Δ_{FL-LL}, dark blue) and between FL and CL (Δ_{FL-CL}, light blue) for the two data distribution scenarios..

AI, Technical Report," 2019, European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[2] "Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts." 2021, European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[3] F. Doshi-Velez, M. Kortz *et al.*, "Accountability of AI Under the Law: The Role of Explanation," *SSRN Electronic Journal*, 11 2017.

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

[5] B. McMahan, E. Moore *et al.*, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. of the 20th Int'l Conf. on Artificial Intelligence and Statistics*, ser. Proc. of Machine Learning Research, vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282.

[6] J. L. Corcuera Bárcena, M. Daole *et al.*, "Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models," in *XAI.it 2022: 3rd Italian Workshop on Explainable Artificial Intelligence, co-located with AI*IA 2022*, 2022.

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

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

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

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

[11] L. J. Dust, M. L. Murcia *et al.*, "Federated Fuzzy Learning with Imbalanced Data," in *2021 20th IEEE Int'l Conf. on Machine Learning and Applications (ICMLA)*, 2021, pp. 1130–1137.

[12] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *ADV NEUR IN*, vol. 30. Curran Associates, Inc., 2017.

[13] L. Corbucci, R. Guidotti, and A. Monreale, "Explaining Black-Boxes in Federated Learning," in *Explainable Artificial Intelligence*, L. Longo, Ed. Cham: Springer Nature Switzerland, 2023, pp. 151–163.

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

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

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

[17] Z. Chi and H. Yan, *Fuzzy algorithms: with applications to image processing and pattern recognition*. World scientific, 1996, vol. 10.

[18] M. Aledhari, R. Razzak *et al.*, "Federated learning: A survey on enabling technologies, protocols, and applications," *IEEE Access*, vol. 8, pp. 140 699–140 725, 2020.

[19] V. Mothukuri, R. M. Parizi *et al.*, "A survey on security and privacy of federated learning," *FUTURE GENER COMP SY*, vol. 115, pp. 619–640, 2021.

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

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

[22] ——, "Interpretable Privacy-Preserving Collaborative Deep Learning for Taxi Trip Duration Forecasting," in *Int'l Conf. on Vehicle Technology and Intelligent Transport Systems, Int'l Conf. on Smart Cities and Green ICT Systems*. Springer, 2022, pp. 392–411.

[23] G. Wang, "Interpret federated learning with shapley values," *arXiv preprint arXiv:1905.04519*, 2019.

[24] Y. Wu, S. Cai *et al.*, "Privacy Preserving Vertical Federated Learning for Tree-Based Models," *Proc. VLDB Endow.*, vol. 13, no. 12, p. 2090–2103, sep 2020.

[25] M. Gacto, R. Alcalá, and F. Herrera, "Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures," *Information Sciences*, vol. 181, no. 20, pp. 4340–4360, 2011, special Issue on Interpretable Fuzzy Systems.

[26] O. Cordón, M. J. del Jesus, and F. Herrera, "A proposal on reasoning methods in fuzzy rule-based classification systems," *INT J APPROX REASON*, vol. 20, no. 1, pp. 21–45, 1999.

[27] P. Ducange, F. Marcelloni, and R. Pecori, "Fuzzy Hoeffding Decision Tree for Data Stream Classification," *INT J COMPUT INT SYS*, vol. 14, pp. 946–964, 2021.

[28] R. Bock, "MAGIC Gamma Telescope," UCI Machine Learning Repository, 2007, DOI: https://doi.org/10.24432/C52C8B.

[29] P. Kairouz, H. B. McMahan *et al.*, "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[30] J. L. C. Bárcena, P. Ducange *et al.*, "Towards Trustworthy AI for QoE prediction in B5G/6G Networks," in *First Int'l Workshop on Artificial Intelligence in Beyond 5G and 6G Wireless Networks (AI6G 2022)*, Padova, Italy, 2022.

[31] J. L. Corcuera Bárcena, P. Ducange *et al.*, "Enabling federated learning of explainable AI models within beyond-5G/6G networks," *Computer Communications*, vol. 210, pp. 356–375, 2023.

[32] A. Renda, P. Ducange *et al.*, "Federated Learning of Explainable AI Models in 6G Systems: Towards Secure and Automated Vehicle Networking," *Information*, vol. 13, no. 8, 2022.

[33] X. Lu and Y. Bai, "A new rule reduction method for fuzzy modeling," *IEEE T FUZZY SYST*, vol. 28, no. 11, pp. 3023–3031, 2019.