Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models*

José Luis Corcuera Bárcena, Mattia Daole, Pietro Ducange, Francesco Marcelloni, Alessandro Renda, Fabrizio Ruffini and Alessio Schiavo

Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, 56122 Pisa, Italy

Abstract

The current era is characterized by an increasing pervasiveness of applications and services based on data processing and often built on Artificial Intelligence (AI) and, in particular, Machine Learning (ML) algorithms. In fact, extracting insights from data is so common in daily life of individuals, companies, and public entities and so relevant for the market players, to become an important matter of interest for institutional organizations. The theme is so relevant that ad hoc regulations have been proposed. One important aspect is given by the capability of the applications to tackle the data privacy issue. Additionally, depending on the specific application field, paramount importance is given to the possibility for the humans to understand why a certain AI/ML-based application is providing that specific output. In this paper, we discuss the concept of Federated Learning of eXplainable AI (XAI) models, in short FED-XAI, purposely designed to address these two requirements simultaneously. AI/ML models are trained with the simultaneous goals of preserving the data privacy (Federated Learning (FL) side) and ensuring a certain level of explainability of the system (XAI side). We first introduce the motivations at the foundation of FL and XAI, along with their basic concepts; then, we discuss the current status of this field of study, providing a brief survey regarding approaches, models, and results. Finally, we highlight the main future challenges.

Keywords

Data Mining, Artificial Intelligence, Machine Learning, Federated Learning, Explainable AI, Decision Tree, Linguistic fuzzy models, FED-XAI

1. Introduction

Artificial Intelligence (AI) and, in particular, Machine Learning (ML) algorithms are becoming de-facto standard pillars for creating innovative services and applications. Nowadays, they are commonly found in many aspects of daily processes, both in public and in private sectors. Actually, their pervasiveness is so deep that institutions have been prompted to address the necessity of specific regulations, also taking ethical principles into account.

In 2017, the G7 partners issued a Declaration, highlighting the importance of the "vision of human-centric AI which drives innovation and growth in the digital economy" [1]. One year

*XAI.it 2022: 3rd Italian Workshop on Explainable Artificial Intelligence, co-located with AI*IA 2022; November 28 - December 2, 2022, Udine, (Italy)*

joseluis.corcuera@phd.unipi.it (J. Corcuera Bárcena); m.daole@studenti.unipi.it (M. Daole); pietro.ducange@unipi.it (P. Ducange); francesco.marcelloni@unipi.it (F. Marcelloni); alessandro.renda@unipi.it (A. Renda); fabrizio.ruffini@ing.unipi.it (F. Ruffini); a.schiavo2@studenti.unipi.it (A. Schiavo)

0000-0002-9984-1904 (J. Corcuera Bárcena); 0000-0003-4510-1350 (P. Ducange); 0000-0002-5895-876X (F. Marcelloni); 0000-0002-0482-5048 (A. Renda); 0000-0001-6328-4360 (F. Ruffini)

later, in 2018, the European Commission tasked a group of independent experts to produce a document called the "Ethic Guidelines for Trustworthy AI" [2], followed in 2021 by a proposal for a "Regulation of the European parliament and of the council laying down harmonized rules on artificial intelligence" [3]. In [2], authors describe a number of requirements that an AI system should meet in order to be considered "trustworthy": among others, "privacy and data governance" and model "transparency" are often regarded by all the stakeholders, from service provider to end-users, as pivotal steps towards trustworthiness.

The ***privacy*** aspect is often considered of utmost importance by data-owning organizations, which are often reluctant to share their data with other parties. This can be due to different internal practices between organizations, or even between different parts of the same organization; also, data are perceived as a precious asset by companies and often exploited as an "unfair advantage" over competitors. Finally, user data often involve sensitive information that needs to be treated carefully to avoid privacy issues.

In these scenarios, with private raw data spread over multiple physical locations, traditional ML approaches are not always feasible, as they require the availability of the overall dataset stored in one centralized server. On the other hand, when data are naturally and necessarily distributed in *isolated silos*, every single data owner may not have sufficient data to properly train an AI model. These considerations lead to the necessity to adopt novel paradigms and propose alternative methodologies. Federated Learning (FL), proposed in the literature a few years ago [4], can be a valid solution to cope with the data privacy issue: the key idea of FL is to learn local AI models from local data, and then aggregate the (locally-computed) models or updates to generate a global aggregated model. Thus, data-privacy is preserved, since raw data are not exchanged between the different clients responsible for the local models, but the overall information is collaboratively used to learn the aggregated model.

The aspect of ***explainability*** is at the heart of so-called trustworthy AI: for example, as mentioned in [2] "[...] *AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned*". Likewise, GDPR recital 71 [5] states that: "[...] *In any case, such 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*". As a consequence, industry and academia are placing increasing attention on eXplainable AI (XAI).

The acronym *FED-XAI* stands for *Fed*erated learning of *XAI* models and is conceived to provide a leap forward toward trustworthy AI. The objective of Fed-XAI consists in devising methodological and technological solutions as follows: on one hand, to leverage the FL approach for privacy preservation during collaboratively training of ML/AI models. On the other hand, to ensure an adequate degree of explainability of the AI-based systems themselves. Actually, since early works in the FL literature [6],[7], most solutions revolve around the original proposal of Federated Averaging (FedAvg), as a protocol for executing Stochastic Gradient Descent (SGD) in a federated manner. In particular, in [6] the authors showed that deep neural network (DNN) models can be collaboratively trained for tackling image classification and language modeling tasks. While DNNs have achieved unprecedented levels of performance in various application domains, they are generally considered *opaque* models due to their huge number of parameters and non-linear modeling: as such, they do not feature inherent interpretability. Although much less attention has been devoted to FL of XAI models, the interest in this area is steeply increasing. The main goal of this paper is to discuss the current status, the main open challenges and future directions of Fed-XAI.

In the rest of the paper, we first describe some preliminaries about FL and XAI (Section 2). Then, in Section 3, we review the state of the art of Fed-XAI approaches. Section 4 describes the main open challenges towards Fed-XAI and highlights some possible future directions. Finally, in Section 5 we draw some conclusions.

2. Preliminaries

In this section we briefly introduce and discuss some basic concepts of the FL paradigm and XAI models, useful to understand the main characteristics, advantages and disadvantages of the different works compared in the subsequent sections.

2.1. Federated Learning: basic concepts

Several surveys are available on FL [8, 9]. In this section, we present some of the relevant basic concepts. In FL multiple parties (or clients) collaboratively train an ML model. In mainstream FL, the learning procedure is orchestrated by a central server and FL algorithms mainly aim to collaboratively optimize a global differentiable objective function, e.g., through adequate variants of stochastic gradient descent (SGD) such as FedAvg and Federated SGD (FedSGD). FedAvg iterates, in a round-based procedure, over the following steps: (i) the server sends out the global model to a random subset of the data owners; (ii) each data owner updates the model using its local data through one or multiple steps of SGD, and sends it back to the server; (iii) the server takes the average of the locally updated models, weighted according to the number of examples, to obtain a new global model. FedAvg has been used for federated learning of models such as DNN [6] and SVM [10] in many real world applications. A popular case study is the query suggestion improvement on Google Keyboard[1]. FedSGD differs from FedAVG mainly in that clients transmit gradients (rather than model parameters) to the central server, which is then responsible for aggregating them and updating model parameters.

Based on how the data are partitioned in the different local devices, FL schemes are typically categorized in horizontal, vertical, and hybrid learning schemes. In *horizontal FL* the datasets of different parties share the same feature space but may have different sample dimension. Since clients share the same feature space, the local models are usually trained using the same model architecture. In *vertical FL* the datasets of different parties differ in the feature space, as commonly observed in cooperation scenarios between different entities (e.g., taxation and census). In *hybrid FL* the local datasets of the different parties may share or not the same feature space.

Another categorization of FL scenarios is based on the number of involved participants. *Cross-silo* FL refers to a settings with a low number of participants (e.g., from two to few tens) with relatively large amount of data and computational power. In *cross-device* FL, the number of parties is typically much higher, compared to the cross-silo setting, but they feature low

[1] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html, accessed November 2022

computational capability and are generally poorly reliable from a connectivity perspective. This scenario is also more prone to one of the greatest challenges of FL: not only the number of participants can grow fast, but their data may also have different distributions, i.e. non-i.i.d. (independent and identically distributed) setting , and volumes.

2.1.1. Federated Learning Frameworks

In the following, we show a list of commonly used frameworks in which some FL schemes have been implemented. The website reference of each framework, along with the respective developing organization, is reported in Table 1.

TensorFlow Federated Framework (TFF) is an open-source framework for Deep Learning (DL) on decentralized data. It implements a limited number of aggregation strategies and currently supports only a simulation version on a single node [11] (as such, it cannot be deployed on realistic federated setting with multiple nodes). Finally, TFF supports only horizontal partitioning of data and is not ML-framework agnostic.

Federated AI Technology Enabler Framework (FATE) is an open-source project. It was initiated by Webank's AI Department to provide a secure framework supporting a federated Artificial Intelligence system. It is based on the TensorFlow library and the following ML models can be selected for being involved in an FL scheme: DNN, logistic regression and a gradient-boosting decision tree. It supports various implementations of Secure Multi Party Computation protocols and encryption methods and exploits gRPC (recursive acronym for "gRPC Remote Procedure Calls") for interactions. Last versions of the framework allow also to deploy the FL process in a real distributed architecture and also support virtualization based on containers. Similar to TFF, FATE is not ML-framework agnostic.

Open Federated Learning (OpenFL) is an open-source software platform for Federated Learning developed by Intel Labs in the framework of a collaboration with the University of Pennsylvania [12]. It executes a federated training process following a centralized approach: a server component (Aggregator) receives the model's parameters from clients (Collaborators), and aggregates them to compute the global model. The interactions between entities is possible through gRPC over TLS connection. OpenFL natively supports Deep Learning libraries (eg: Pytorch or TensorFlow), allowing the customization of several modules to adapt them for dealing also with user-defined needs. Indeed, OpenFL can be considered ready for being a model agnostic framework. Similar to FATE, OpenFL supports the actual deploy of the FL scheme on a distributed architecture and support virtualization based on containers.

PySift is an open-source Python project for secure and private DL. Since this framework has the largest community of contributors (over 250), rapid development is expected. At the moment, PySyft is not ML-framework agnostic and only works in simulation mode.

IBM Federated Learning (IBM FL) is a production-oriented FL framework: it has been developed to be easy to use and to have a short deployment time in a real distributed environment [13]. It is a modular Python library. The framework support FL of both DL models and "traditional" ML models, such as Linear classifiers (via SGD), Decision Trees and Naive Bayes. An interface is provided to define a standard API to train, save, evaluate, and update a model, as well as generate model updates. IBM Federated Learning can support multiple connection types, including the Flask web framework, gRPC, and WebSockets.

Flower has been presented in [14] as an open-source framework for FL. It allows for the definition of custom FL aggregation strategies and adopts a communication layer based on gRPC. Similar to OpenFL, Flower can be considered as ML-framework agnostic. As a drawback, it lacks security mechanisms to protect data exchanged between FL entities.

Federated Learning Simulator (FLSim) is a recent project by Facebook Research (first released on late 2021) [15]. It is based on PyTorch and currently it supports standard FedAvg and other federated learning methods such as FedAdam, FedProx, FedAvgM, FedBuff, FedLARS, and FedLAMB.

Table 1
Overview of the most popular FL frameworks.

Framework	Developers	URL
TFF	Google Inc	https://www.tensorflow.org/federated
FATE	Webank	https://fate.fedai.org
OpenFL	Intel Labs - University of Pennsylvania	https://github.com/intel/openfl
PySift	Openmined	https://github.com/OpenMined/PySyft
IBM FL	IBM	https://github.com/IBM/federated-learning-lib
Flower	Adap GmbH - several universities	https://flower.dev
FLSim	Facebook Research	https://github.com/facebookresearch/FLSim

2.2. XAI Properties and Models

In the context of XAI, several terms are used interchangeably with the word explainability, but they can have different specific nuances of meaning. Recently, [16] and [17] have tried to summarize the most commonly used terminology: **understandability or intelligibility**: the two terms are associated with a functional understanding of the model in ML, without the need for explaining its inner procedure and representation. **Comprehensibility** is associated with the ability of a learning algorithm to represent its learned knowledge such that the model may be inspected and understood by humans. Thus, it is tightly related to the complexity of a model. **Interpretability** definition is the ability to explain how decisions have been taken or to provide the meaning in a way that is understandable by a human. **Explainability** is defined in [16] as follows: given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand. Finally, **transparency** refers to the characteristic of a model to be inherently understandable for a human.

Although the distinction between the various terms is quite fuzzy, it is possible to identify two major approaches for XAI: models that are interpretable by design (transparent models) and those that can be explained using external XAI techniques (post-hoc explainability).

The **post-hoc explainability** techniques are related to the ways human commonly explain systems and processes by themselves. Existing approaches consist in text explanations, visualizations, local explanations, explanations by example, explanations by simplification and feature relevance. Machine learning models that do not meet any of the requirements imposed to be defined as “transparent” require the use of these post-hoc techniques to explain their decisions. Both some shallow models, e.g. tree ensembles, random forests, support vector machines, and deep models, e.g. Deep Neural Networks, Convolutional Neural Networks, Recurrent

Neural Networks, belong to this category. The taxonomy of the post-hoc explainability techniques presented in [16] encompasses a coarse distinction between model-agnostic techniques, namely those that can be plugged to any model to implement explainability, and model-specific techniques, namely those tailored to explain specific ML models.

The **transparency** property is generally accorded to models such as Rule Based Systems (RBSs), Decision Trees (DTs), Linear/Logistic Regression, k-Nearest Neighbors, Generalized Additive Models and Bayesian Models [16].

It should be underlined that the performance of a model and its transparency are typically conflicting objectives; thus, as shown in Fig. 1, accuracy-oriented solutions are often considered as hard to interpret, while interpretable models may be lacking in performance.

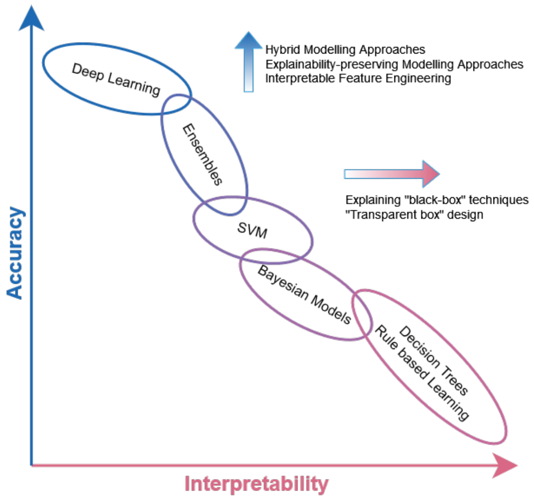
Accuracy
Interpretability
Deep Learning
Ensembles
SVM
Bayesian Models
Decision Trees
Rule based Learning
Hybrid Modelling Approaches
Explainability-preserving Modelling Approaches
Interpretable Feature Engineering
Explaining "black-box" techniques
"Transparent box" design

Figure 1: Trade-off between model interpretability and performance for various classes of models

RBSs and DTs adopt an inference process akin to human reasoning and are intuitively interpretable [18]; in fact, a list of *if-then* rules can be extracted from an induced DT following the paths from the root to each leaf. However, the actual level of interpretability of such models may vary according to different factors. From this perspective, authors in [18] highlighted the distinction between *global* and *local* explanations: the former relates to model transparency and refers to structural properties of the classifiers, such as tree size, number of nodes/leaves in DTs and number of rules in RBSs. The latter is associated with the inference process and analyzes the decision-making process related to the individual instances. Focusing, for the sake of brevity, on the interpretability of DTs, it is possible to identify two factors that impact the interpretability of a model: (i) depending on the problem at hand, pursuing highly accurate models by increasing their complexity may lead to induced trees that are hard to interpret because of their large number of nodes/leaves [17]. This factor relates to global interpretability. (ii) The second factor relates to the semantic interpretability of the rules extracted from the DTs, which should be expressed using linguistic terms whose meaning is easily comprehensible [19]. It relates to both global and local interpretability.

Fuzzy Decision Trees (FDTs) and Fuzzy Rule-Based Systems (FRBSs) leverage fuzzy logic as a tool to derive more readily interpretable rules (formulated verbally over imprecise domains) [20]. In this context, linguistic fuzzy models provide a natural linguistic representation of numeric variables and typically outperform their crisp counterparts in scenarios with some degree of noise and/or uncertainty. However, in the context of FDTs and FRBSs, an input instance may generally activate multiple branches with different activation degrees (strenght of activation): the local interpretability is thus lower if the inference strategy evaluates an output based on all activated paths (typically referred to as the 'weighted average' strategy), whereas it is higher if only the path with the highest activation degree is considered (typically referred to as the 'maximum matching' strategy).

3. Current Status: a review of Fed-XAI approaches

Combining FL paradigm and XAI approaches is attracting increasing attention recently.

Concerning post-hoc explainability, we mention the works presented in [21], [22] and [23]. These three contributions consider the feature importance analysis as the key point for the explainability of their models. Indeed, in [21], authors investigate the feature importance issue in a vertical FL scenario. Since the feature importance of each participant in the FL process may reveal some aspect of the private local data, Shapley values [24] have exploited in the analysis. These values are calculated as the average marginal contribution of a feature value across all possible participants in the FL process. Preliminary experiments have been carried out on simple benchmark datasets and using a KNN algorithm. In [22] and [23] the author adopts a Federated DL model to predict a taxi trip duration within Brunswick region under a horizontal FL setting. The general principle consists in training local models on local data samples and exchanging parameters (i.e., the weights of a DNN) to build a global model according to the FedAvg algorithm. The author shows that the model generated using her FL approach achieves an accuracy level comparable to the one obtained by a centralized model trained on the overall dataset. Feature relevance, based on Integrated Gradients (IGs) [25], has been adopted as post-hoc explainability technique. IGs allow a user to understand if a specific value of a feature has a positive or a negative impact in taking a specific decision. Chen et al. [26] have recently proposed an explainable vertical FL framework, endowing DL models (three layers neural networks) with post-hoc interpretation via a credible federated counterfactual explanation method. In a nutshell, counterfactual explanation is a local explainability technique which aims to explain a prediction by evaluating a minimal change in an instance that would cause the model to classify it in a predefined class.

In the framework of FL of interpretable-by-design models, Takagi–Sugeno-Kang Fuzzy Rule-Based Models (TSK-FRBS) [27] have been mostly considered as XAI models to be learnt in a federated fashion [28, 29, 30]. We recall that TSK-FRBS adopts linguistic if-then rules; an example of the generic k^{th} rule is reported in the following:

$$\begin{aligned} R_k \, : \, \textbf{IF} \quad & X_1 \ is \ A_{1,j_{k,1}} \quad \textbf{AND} \quad \dots \quad \textbf{AND} \quad X_F \ is \ A_{F,j_{k,F}} \\ & \textbf{THEN} \quad y_k = \gamma_{k,0} + \sum_{i=1}^{F} \gamma_{k,i} \cdot x_i \end{aligned} \tag{1}$$

where F is the total number of attributes, $A_{i,j_{k,i}}$ identifies the j^{th} fuzzy set of the fuzzy partition over the i^{th} attribute considered in the k^{th} rule, and $\gamma_{k,i}$ are the coefficient of the linear model, with $i = 0, \ldots, F$.

Given an input pattern $\mathbf{x} = [x_1, x_2, \ldots, x_F]^T$, first the strength of activation of each rule is computed as follows:

$$w_k(\mathbf{x}) = \prod_{f=1}^{F} \mu_{f,j_{k,f}}(x_f) \quad \text{for} \quad k = 1, \ldots, K \tag{2}$$

where $\mu_{f,j_{k,f}}(x_f)$ is the membership degree of x_f to the fuzzy set $A_{f,j_{k,f}}$. Then, the inference process generates an output as the weighted average of the outputs obtained from the K activated rules. Formally:

$$\hat{y}(\mathbf{x}) = \sum_{k=1}^{K} \left(\frac{w_k(\mathbf{x})}{\sum_{h=1}^{K} w_h(\mathbf{x})} \right) \cdot y_k(\mathbf{x}) \tag{3}$$

In [28] and [29], two main stages are considered in the FL scheme, namely (i) learning the fuzzy partitions of each input feature and the antecedent of the rules, and (ii) learning the rule consequent of each rule. As for the first stage, the authors of [28] consider a local clustering procedure followed by an aggregation stage carried out on the centralized server. The aggregation is based on merging similar clusters. As regards the work discussed in [29], a federated version of the Fuzzy C-Means algorithms has been adopted for the identification of the global clusters. Once the clusters have been identified, a typical approach for evaluating the antecedent parameters, and specifically the membership functions, consists in evaluating a Gaussian fitting of the convex envelop of the projected membership values for each cluster. Finally, for the generation of the consequent of each rule, both works consider the application of a federated version of the gradient-based learning schemes. It is worth noting that in [29] a more intensive experimental analysis has been carried out, in which several benchmark datasets have been considered and several FL scheme configurations have been experimented. Results have shown that in most of the cases the FL-based approach achieves results comparable to centralized one.

In our recent work presented in [30], we have proposed an FL scheme for learning more explainable TSK-FRBSs than the classical ones considered in [28] and [29]. An overview of the approach is shown in Fig. 2.

First, we have adopted fuzzy uniform partitions with a limited number of fuzzy sets (up to 5) rather than partitions generated by using the classical clustering-based, data-driven, approach. Indeed, the latter can lead to the generation of fuzzy partitions composed of several, possibly highly overlapping, fuzzy sets, whereas the proposed approach guarantees the highest semantic interpretability. Then, we have suggested the adoption of an inference strategy based on the maximum voting rather than the classical weighted averaging method. The proposed FL approach is not iterative but it generates the global model in one-shot: first the local TSK fuzzy rules are generated by each client and sent to the central server. Then, the server aggregates the received rules. The aggregation procedure consists in juxtaposing rules collected from clients, and resolving possible conflicts, which emerge when rules from different models, having antecedents referring to identical or overlapping regions of the attribute space, have different consequents. New consequents of the aggregated rules are calculated as the weighted average

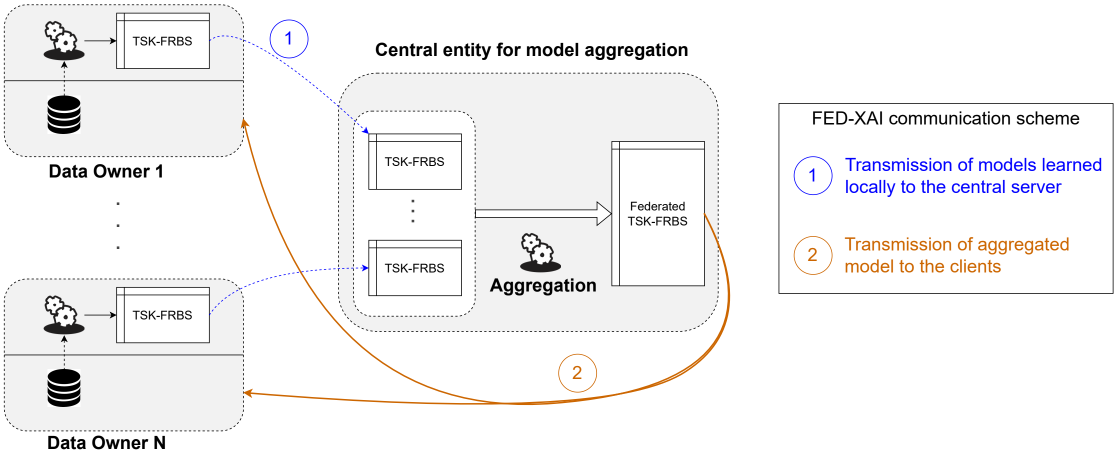
TSK-FRBS
Data Owner 1
TSK-FRBS
Data Owner N
1
Central entity for model aggregation
TSK-FRBS
TSK-FRBS
Aggregation
Federated TSK-FRBS
2
FED-XAI communication scheme
1
Transmission of models learned locally to the central server
2
Transmission of aggregated model to the clients

Figure 2: Overview of our approach for Federated Learning of TSK-FRBSs. Inspired by [31].

of the coefficients of the original rules, where the weight of each rule depends on its support and confidence values. We have experimented the proposed approach on several benchmark datasets and we have demonstrated that the FL scheme achieves better results than models generated locally. Moreover, we have shown that the results obtained by the FL scheme are comparable to those achieved by three different versions of centralized approaches for learning TSK-FRBSs. In the comparison, we have also included the classical version of the learning algorithm, which adopts the clustering-based approach for the generation of rule antecedents, and the weighted average strategy for making inference.

As regards decision trees, two recent FL approaches have been recently proposed in the literature[13, 32].

The IBM Federated Learning framework [13] supports, as previously described, also a Federated Decision Tree based on the ID3 algorithm. In the proposed approach, a single decision tree is generated at centralized server while the clients just provide some counting information based on their local data. At each round, the server composes the list of candidate values of the input features to be split and the list of class labels to query the count information received by the client. Then the server computes the information gain and split accordingly. Classical stopping conditions used for decision trees are adopted, such as reaching the maximum depth of the tree or the absence in all nodes of values for splitting.

In [32] authors discuss a vertical federated learning for Tree-based models, where a privacy preserving approach, based on a Partially Homomorphic Encryption, is adopted. Specifically, all clients contribute to build the structure of the decision tree by providing iterative encrypted statistics to a super client which is in charge of selecting the splitting points of the most relevant attribute. Throughout the whole process, no intermediate information is disclosed to any client. Authors compared the results achieved by their privacy-based FL approaches with their non-private counterparts. Slight losses in accuracy have been highlighted when considering privacy-based FL approaches.

In addition, a recent work described in [33] proposes the application of a federated version of the AdaBoost algorithm. Interestingly, the approach poses minimal constraints on the learning settings of the clients, thus enabling a federation of models such as DTs and SVM, without relying on gradient-based methods.

Recently, in [31] we have envisioned that FED-XAI may represent a relevant enabling technology in advanced 5G towards 6G systems and have discussed its applicability to an automated vehicle networking use case. Specifically, we have presented a framework to evaluate the FED-XAI approach involving online training based on real data from live cellular networks. In our vision, FED-XAI deployed on next generation wireless networks, such as 6G, is expected to bring benefits as a methodology for achieving seamless availability of decentralized, lightweight and communication efficient intelligence. It is worth noticing that the work in [31] summarizes the results achieved during the first year of activities of the HEXA-X Flagship EU project on 6G[2]. Moreover, under the framework of HEXA-X, FED-XAI has been recently awarded as key innovation[3] by the EU Innovation Radar.

4. Open Challenges

With the goal of applying FED-XAI in different contexts, thus ensuring high-levels of performance and interpretability, several challenges are open and need further attention. In our understanding, the major challenges of FED-XAI are related to: (i) how to ensure strong **privacy constraints** (e.g., to avoid data leakage possibilities), (ii) how to **merge** XAI local models (e.g., manage conflicts between different rules created in different clients in rule-based systems or aggregating DNN weights limiting data transmission), (iii) how to cope with **massive data streaming** scenarios in which concept drift issues are often experienced (as an example, in [34] this problem was addressed, in a classical centralized learning scheme, using a Hoeffding decision tree, able to adapt its structure as new streams of data arrive).

Additionally, there are challenges related to **datasets** to be used as benchmarks in the context of FED-XAI. Indeed, most of the experimental analyses carried out using FL schemes consider datasets composed by images or text from different domains [35, 36]. These datasets are suitable for being analyzed adopting DL methods and, thus, have been also adapted for being used in FL schemes of DL models. However, when dealing with XAI models such as decision trees or rule-based models, image and text datasets cannot be directly used without a preliminary feature extraction stage (in the case of DL models, this stage is usually embedded in the models themselves). In addition, the extracted features must be “interpretable,” that is, they must be metrics understandable by the user of an XAI model-based system. Otherwise, the construction of XAI models would be useless. Moreover, a standardization of the experimental setup for analyzing both i.i.d. and non-i.i.d. data distribution among the clients should be defined for ensuring the repeatability of the experiments for fair comparisons.

From an *architectural* point of view, following the suggestions in [37], FED-XAI applications should be designed for being deployed on edge-computing platforms. Indeed, despite *cloud computing* paradigm, in which data are transmitted from the source to a centralized data center in the network core, **edge-computing** brings computation and data storage close to the

[2] https://hexa-x.eu, accessed November 2022

[3] https://www.innoradar.eu/innovation/45988, accessed November 2022

sources of data, thus ensuring low latency and reducing network congestion. In the context of edge computing, Multi-Access Edge Computing (MEC), that is a type of network architecture that brings to the edge of the network server platforms with high computational and storage capabilities, is the perfect candidate for supporting the deploy of FED-XAI schemes [38]. Indeed, MEC architecture strongly supports virtualization, thus ensuring data privacy and isolation. Moreover, MEC architecture can be the perfect venue for the deploy of FED-XAI schemes and applications implemented using FL frameworks which (i) are agnostic from the tool used for generating local models, (ii) allow the usage of user-defined models and aggregation strategies and (iii) have a native support for virtualization. Finally, it is worth noting that a standardization of FL learning on MEC architecture would be beneficial for all the stakeholders involved.

5. Conclusions

With the increasing pervasiveness of daily-life applications based on big data, governmental entities have started to discuss and define regulations to boost the effectiveness of the new methodologies, especially using AI/ML based approaches, while ensuring the population fundamental rights. In this context, data privacy and the capability to understand the model outputs are two of the fundamental aspects to be ensured. Among other approaches to tackle the data privacy, Federated Learning is considered as an effective methodology because it is based on the concept of creating AI/ML models without sharing raw data between different data owners, but still combining knowledge extracted from all of them. In a nutshell, this is achieved training the models on the local data and then updating a global model without sharing data, but model characteristics. The additional need for the stakeholders to understand the model outputs suggested the genesis of FED-XAI approaches, that is Federated Learning approaches of eXplainable Artificial Intelligence models. In this paper, we have introduced the basic concepts of FL, XAI and FED-XAI, and have reported a brief survey of interesting works using those concepts. The main problems are related to: i) achieve results comparable to a centralized approach where all data are available, ii) to find an optimal trade-off between explainability and accuracy of the results, and iii) to explain the models. Our understanding is that this research field is at its early stages, but with increasing interest. We think that the interest is given by the fact that the methodology is both capable of ensuring data privacy and explainability, while also enabling good level of performance. At the present, some work has to be done for the FED-XAI concept to reach its maturity: between others, standardizing the terminology, defining experimental datasets to make easier the comparison between studies, making the current frameworks flexible and robust, and finally taking the architectural point of view into account. We believe that, given its key elements, the FED-XAI approach will be a common presence in the future AI-based application ecosystem.

Acknowledgments

This work has been partly funded by the European Commission through the H2020 project Hexa-X (Grant Agreement no. 101015956) and by the Italian Ministry of University and Research (MUR) in the framework of the CrossLab Project (Departments of Excellence).

References

[1] G7 ict and industry ministers' declaration making the next production revolution inclusive, open and secure, 2017. URL: http://www.g7italy.it/sites/default/files/documents/G7%20ICT_Industry_Ministers_Declaration_%20Italy-26%20Sept_2017final_0/index.pdf.

[2] High-Level Expert Group on AI, Report, European Commission, 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[3] Proposal for a regulation of the european parliament and of the council, 2021.

[4] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2019) 1–19.

[5] Gdpr, 2018. URL: https://gdpr-info.eu/recitals/no-71/.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y. Arcas, Communication-Efficient Learning of Deep Networks from Decentralized Data, in: A. Singh, J. Zhu (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1273–1282.

[7] J. Konečný, H. B. McMahan, D. Ramage, P. Richtárik, Federated optimization: Distributed machine learning for on-device intelligence, 2016. doi:10.48550/ARXIV.1610.02527.

[8] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, Y. Gao, A survey on federated learning, Knowledge-Based Systems 216 (2021) 106775.

[9] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, G. Srivastava, A survey on security and privacy of federated learning, Future Generation Computer Systems 115 (2021) 619–640.

[10] E. Bakopoulou, B. Tillman, A. Markopoulou, Fedpacket: A federated learning approach to mobile packet classification, IEEE Transactions on Mobile Computing 21 (2022) 3609–3628. doi:10.1109/TMC.2021.3058627.

[11] I. Kholod, E. Yanaki, D. Fomichev, E. Shalugin, E. Novikova, E. Filippov, M. Nordlund, Open-source federated learning frameworks for iot: A comparative review and analysis, Sensors 21 (2021). doi:10.3390/s21010167.

[12] G. A. R. et al., Openfl: An open-source framework for federated learning, 2021. URL: https://arxiv.org/abs/2105.06413. doi:https://doi.org/10.48550/arXiv.2105.06413.

[13] H. Ludwig, N. Baracaldo, G. Thomas, Y. Zhou, A. Anwar, S. Rajamoni, Y. Ong, J. Radhakrishnan, A. Verma, M. Sinn, M. Purcell, A. Rawat, T. Minh, N. Holohan, S. Chakraborty, S. Whitherspoon, D. Steuer, L. Wynter, H. Hassan, S. Laguna, M. Yurochkin, M. Agarwal, E. Chuba, A. Abay, Ibm federated learning: an enterprise framework white paper v0.1, 2020. URL: https://arxiv.org/abs/2007.10987. doi:10.48550/ARXIV.2007.10987.

[14] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, P. P. de Gusmão, N. D. Lane, Flower: A friendly federated learning research framework, arXiv preprint arXiv:2007.14390 (2020).

[15] Federated learning simulator (flsim), 2021. URL: https://github.com/facebookresearch/FLSim.

[16] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.

[17] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, F. Marcelloni, Evolutionary fuzzy

systems for explainable artificial intelligence: Why, when, what for, and where to?, IEEE Computational Intelligence Magazine 14 (2019) 69–81. doi:10.1109/MCI.2018.2881645.

[18] J. M. Alonso, P. Ducange, R. Pecori, R. Vilas, Building explanations for fuzzy decision trees with the expliclas software, in: 2020 IEEE Int'l Conf. on Fuzzy Systems (FUZZ-IEEE), IEEE, 2020, pp. 1–8.

[19] M. J. Gacto, R. Alcalá, F. Herrera, Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures, Inf. Sci. 181 (2011) 4340–4360.

[20] A. Bechini, J. L. C. Bárcena, P. Ducange, F. Marcelloni, A. Renda, Increasing accuracy and explainability in fuzzy regression trees: An experimental analysis, in: 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2022, pp. 1–8.

[21] G. Wang, Interpret federated learning with shapley values, arXiv preprint arXiv:1905.04519 (2019).

[22] J. Fiosina, Explainable federated learning for taxi travel time prediction, in: VEHITS, 2021.

[23] J. Fiosina, 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, pp. 392–411.

[24] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[25] D. Janzing, L. Minorics, P. Blöbaum, Feature relevance quantification in explainable ai: A causal problem, in: International Conference on artificial intelligence and statistics, PMLR, 2020, pp. 2907–2916.

[26] P. Chen, X. Du, Z. Lu, J. Wu, P. C. Hung, Evfl: An explainable vertical federated learning for data-oriented artificial intelligence systems, Journal of Systems Architecture 126 (2022) 102474. URL: https://www.sciencedirect.com/science/article/pii/S1383762122000583. doi:https://doi.org/10.1016/j.sysarc.2022.102474.

[27] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control, IEEE T SYST MAN CYB (1985) 116–132.

[28] A. Wilbik, P. Grefen, Towards a federated fuzzy learning system, in: 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2021, pp. 1–6.

[29] X. Zhu, D. Wang, W. Pedrycz, Z. Li, Horizontal federated learning of takagi–sugeno fuzzy rule-based models, IEEE Transactions on Fuzzy Systems 30 (2022) 3537–3547. doi:10.1109/TFUZZ.2021.3118733.

[30] J. L. C. Bárcena, P. Ducange, A. Ercolani, F. Marcelloni, A. Renda, An approach to federated learning of explainable fuzzy regression models, in: 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2022, pp. 1–8.

[31] A. Renda, P. Ducange, F. Marcelloni, D. Sabella, M. C. Filippou, G. Nardini, G. Stea, A. Virdis, D. Micheli, D. Rapone, et al., Federated learning of explainable ai models in 6g systems: Towards secure and automated vehicle networking, Information 13 (2022) 395.

[32] Y. Wu, S. Cai, X. Xiao, G. Chen, B. C. Ooi, Privacy preserving vertical federated learning for tree-based models 13 (2020) 2090–2103.

[33] M. Polato, R. Esposito, M. Aldinucci, Boosting the federation: Cross-silo federated learning

without gradient descent, 2022. doi:10.1109/IJCNN55064.2022.9892284.

[34] P. Ducange, F. Marcelloni, R. Pecori, Fuzzy hoeffding decision tree for data stream classification., Int. J. Comput. Intell. Syst. 14 (2021) 946–964.

[35] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, A. Talwalkar, Leaf: A benchmark for federated settings, arXiv preprint arXiv:1812.01097 (2018).

[36] D. Caldarola, B. Caputo, M. Ciccone, Improving generalization in federated learning by seeking flat minima, arXiv preprint arXiv:2203.11834 (2022).

[37] Q. Xia, W. Ye, Z. Tao, J. Wu, Q. Li, A survey of federated learning for edge computing: Research problems and solutions, High-Confidence Computing 1 (2021) 100008.

[38] C. Feng, Z. Zhao, Y. Wang, T. Q. Quek, M. Peng, On the design of federated learning in the mobile edge computing systems, IEEE Transactions on Communications 69 (2021) 5902–5916.