Federated TSK Models for Predicting Quality of Experience in B5G/6G Networks

José Luis Corcuera Bárcena, Pietro Ducange, Francesco Marcelloni, Alessandro Renda, Fabrizio Ruffini, Alessio Schiavo
Department of Information Engineering, University of Pisa, Largo Lucio Lazzarino 1, 56122 Pisa, Italy
Email: {joseluis.corcuera, alessio.schiavo}@phd.unipi.it,
{pietro.ducange, alessandro.renda, francesco.marcelloni}@unipi.it,
fabrizio.ruffini@ing.unipi.it

***Abstract*—Real-time applications based on streaming data collected from remote devices, such as smartphones and vehicles, are commonly developed using Artificial Intelligence (AI). Such applications must fulfill different requirements: on one hand, they must ensure good performance and must deliver results in a timely manner; on the other hand, with the objective of being compliant with the AI-specific regulations, they shall preserve data privacy and guarantee a certain level of explainability. In this paper, we describe an AI-based application to predict the Quality of Experience (QoE) for videos acquired by moving vehicles from Beyond 5G and 6G (B5G/6G) network data. To this aim, we exploit a Takagi-Sugeno-Kang (TSK) fuzzy model learned by employing a federated approach, thus meeting, simultaneously, the requests for explainability and data privacy preservation. A thorough experimental analysis, involving also the comparison with an opaque baseline (i.e., a neural network model), is presented and shows that the TSK model can be regarded as a viable solution which guarantees on the one side an optimal trade-off between interpretability and accuracy, and on the other side preserves the data privacy.**

***Index Terms*—Federated Learning, Explainable AI, FED-XAI, Linguistic fuzzy models, QoE**

I. Introduction and motivations

Nowadays, Artificial Intelligence (AI) empowered solutions are gaining significant momentum favoured by the ubiquitous presence of data sources, including mobile, IoT and wearable devices. Looking ahead, the next generations of wireless networks, beyond 5G (B5G) and 6G, are expected to further amplify the pervasiveness of AI. At the same time, they will leverage AI in order to improve the offered services and accommodate a much higher number of connected devices [1].

In this framework, end-user trust in AI systems is considered one of the cornerstones in the design of B5G/6G networks, as attested by the research efforts in flagship initiatives such as HEXA-X[1] Horizon Project, in Europe. According to the recommendation of the European Commission, the pursuit of *trustworthiness* requires that AI systems are i) lawful - complying with all applicable laws and regulations, ii) ethical - adhering to ethical principles and values and iii) robust - both from a technical perspective and taking its social environment into account [2]. Furthermore, the Artificial Intelligence Act [3] has been recently published to define a common ground of rules that AI-based systems must abide by to uphold a set of fundamental rights of the citizens.

One of the key requirements towards trustworthy AI is related to the data privacy and governance. Among the existing approaches aimed at ensuring data privacy while building AI systems based on Machine Learning (ML) models, the Federated Learning (FL) paradigm has recently gained increasing attention [4]. In a nutshell, FL allows multiple parties (or clients) to collaboratively train an ML model without disclosing their private data: raw data are not transferred from local devices to a server for "traditional" centralized processing; instead, model training takes place on the end-user side and only model parameters or updates are shared with the central server for generating a global aggregated model. Intuitively, the knowledge extracted from scattered data is embedded into the federated model.

Another relevant component towards trustworthy AI is transparency, for which "*AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned*" [2]. This is particularly relevant in contexts such as healthcare or autonomous vehicles, where the stakeholders are interested in understanding how a decision is made by the system (e.g., for accountability reasons). The terms explainability and transparency are sometimes used as synonymous. Here, we follow the terminology proposed in [5]: given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand, whereas transparency refers to the characteristic of a model to be inherently understandable for a human. As regards explaining how an AI model works, two approaches are viable: adoption of transparent by-design models and exploitation of post-hoc techniques [5]. Decision trees and rule-based systems pertain to the former approach, as the inference process (based on simple tests on input attributes) somehow resembles the human reasoning and hence it is easy to understand.

This work has been partly funded by the European Commission through the project Hexa-X (Grant Agreement no. 101015956) under the H2020 programme, by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI" and the PNRR "Tuscany Health Ecosystem" (THE) (Ecosistemi dell'Innovazione) - Spoke 6 - Precision Medicine & Personalized Healthcare (CUP I53C22000780001) under the NextGeneration EU programme, and by the Italian Ministry of University and Research (MUR) in the framework of the FoReLab and CrossLab projects (Departments of Excellence)

[1]https://hexa-x.eu/, accessed Feb. 2023

Post-hoc explainability techniques target models generally considered as opaque, such as Neural Networks (NNs) and ensemble methods. The methodologies used in this context include text explanations, visualizations, local explanations, explanations by example, explanations by simplification, and feature relevance [5]. It is worth noticing that modelling capability (measured, e.g., in terms of inference accuracy) and explainability are generally conflicting objectives, especially in scenarios where the phenomenon being modelled is complex: thus it becomes crucial to investigate the trade-off between them. The investigation of XAI models in the federated setting is aimed at further catalyzing the trustworthiness of AI systems and is at the core of the Federated Learning of eXplainable AI (Fed-XAI) models [6].

In this paper, we address a time series forecasting problem in an automotive case study resorting to a Fed-XAI approach. The ability to forecast Quality of Experience (QoE) metrics will be crucial in several applications enabled by the future B5G/6G networks; here we consider the QoE perceived on instances of vehicular User Equipment (UE), connected to the wireless network, experiencing a video stream. The perceived quality of the video may be critical for enabling services such as *see-through* or *tele-operated driving* [7], which serve as, or impact on, the driving assistance system. Specifically, we compare two ML models for regression, both learnt in a federated fashion: the first model consists of a Takagi-Sugeno-Kang Fuzzy Rule-Based System (TSK-FRBS) [8], which is generally considered as "light-grey" model. We employ a federated approach for learning TSK-FRBSs with enhanced interpretability, which has been recently proposed by us in [9] under the umbrella of Fed-XAI. The second model, generally considered as opaque, consists of a Multi Layer Perceptron Neural Network (MLP) and its federation procedure leverages the well-established federated averaging method [4]. The comparison between the results of TSK-FRBS and the MLP model used as baseline highlights that the former can achieve accuracy comparable with the latter preserving the explainability characteristic.

This work has the following contributions:

- we exploit a TSK-FRBS learnt in a federated fashion for a QoE forecasting case study, in the context of B5G/6G networks;
- we discuss the effectiveness of our approach through a comparative analysis with an approach based on neural networks, as an example of opaque model;
- we present how the inherent interpretability of TSK-FRBS can be leveraged to explain the decision making process;
- we simulate a real scenario where some of the participants in the federation cannot contribute to the learning process and we discuss how the number of the participants can affect the performance of the federated TSK-FRBS.

The rest of the paper is organized as follows: in Section II we report the basic concepts of Fed-XAI, in Section III we describe our specific application for QoE forecasting, discussing the context and the dataset. In Section IV, we illustrate the experimental setup, whereas in Section V we report and discuss the results. Finally, in Section VI we draw some conclusions.

II. Fed-XAI: background

Since the Federated Learning (FL) paradigm has been introduced a few years ago, several surveys have covered the topic from many facets [6], [10]–[14]. From an algorithmic point of view, mainstream FL approaches revolve around the round-based federated averaging (FedAvg) protocol: at each round the server broadcasts the global model to the participants (or a subset thereof); each involved participant updates the model by performing one or more epochs of Stochastic Gradient Descent (SGD) on its local dataset and transmits the updated model to the server; finally, the server computes the average of the locally updated models, weighted according to the cardinality of the local datasets, to obtain a new global model. It turns out that FedAvg is suitable for handling collaborative learning of models optimized through SGD, e.g. NN models, but it requires proper adaptation for models that are not typically learned through the optimization of a differentiable global objective function. In this regard, Fed-XAI has been conceived as a branch of FL focused on the explainability of the AI systems and it supports the quest for trustworthiness by simultaneously addressing the requirements of privacy preservation and model interpretability. Since a few years, several studies have been exploiting a Fed-XAI approach. Some of them consider post-hoc explainability methods [15]–[19], whereas others involve transparent models, including TSK-FRBSs [9], [20], [21], or Decision Trees [22], [23].

In this work we exploit the federated learning of TSK-FRBSs proposed in [9]. In particular, we adopt first-order TSK-FRBSs characterised by having a linear model of the input variables as consequent of the rules. In the following we first recall some notions underlying TSK-FRBSs and then describe the federated approach to learn such a model.

TSK-FRBS employs if-then rules to perform regression tasks in ML. The rule base is learned from a training set, typically following a two-stage approach: first, in the structure identification stage, the number of rules and the antecedent part of the rules are determined either with grid-partitioning of the input space or exploiting fuzzy clustering methods. Then, in the model parameter identification stage, local linear models are fitted on the subspaces determined in the first stage. The generic k^{th} rule of a first-order TSK-FRBS is expressed in the following form:

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

where F is the total number of input variables in the dataset, $A_{i,j_{k,i}}$ identifies the j^{th} fuzzy set of the fuzzy partition over the i^{th} input variable X_i, and $\gamma_{k,i}$ (with $i = 0, \ldots, F$) are the

coefficients of the linear model, which is used to evaluate the associated output y_k.

In the inference stage, 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 output is evaluated either as the average of the outputs associated with the activated rules (weighted by their firing strengths) or coincides with that of the highest firing strength rule (maximum matching policy).

The first-order TSK model used in this work has been introduced in one of our previous works [9], and is generated as follows. A strong uniform fuzzy partition is defined over each input variable, limiting to three the number of fuzzy sets used in each partition. The choice of using uniform partitions with a limited number of fuzzy sets enhances interpretability from a twofold perspective. On one hand, the geometric characteristics of the partitions make them more interpretable as they satisfy the criteria of coverage, completeness, distinguishability and complementarity [24], which are generally not met when using classical clustering-based, data-driven approaches. On the other hand, the three fuzzy sets of each partition can be mapped to as many highly intuitive linguistic terms (e.g. low, medium, and high) thus increasing semantic interpretability. Furthermore, the inference process is based on maximum matching: in this way the final output can be easily explained by analysing the single rule which determines the output.

The FL approach for building TSK-FRBSs is not iterative, but it generates the global model in one-shot. First, the local TSK-FRBSs are generated by each client and sent to the central server. Then, the server aggregates the received rule bases. The aggregation procedure consists in juxtaposing rules collected from clients, and resolving possible conflicts, i.e., rules from different models having antecedents referring to identical regions of the attribute space and different consequents. Conflict resolution consists in the creation of a single rule from each set of conflicting rules. Such rule retains the common conflictual antecedent as antecedent, and the average of the conflicting rules consequents coefficients as new consequent.

Finally, it is worth emphasising that the FL approach can be considered as an alternative to two baseline learning settings, schematically depicted in Figure 1.

In FL schematized in Fig. 1a, clients collaborate in obtaining a single federated model without compromising their privacy. In Local Learning (LL) shown in Fig. 1b, each client individually learns a model from its local data only. In this setting, raw data privacy is preserved but there is no collaboration among clients. Hence, the assessment of the performance of a FL approach entails measuring the gain with respect to the local learning setting. In Centralized Learning (CL) schematised in Fig. 1c, data from all clients are first gathered and stored on a single device and then are used to learn the model. This setting represents the utmost form of collaborative training, but implies the violation of users privacy, as raw data need to be transmitted.

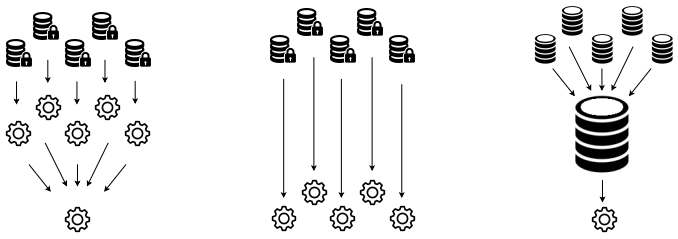

Fig. 1. Schematized representation of three learning settings: (a) federated learning (FL), (b) local learning (LL), (c) centralized learning (CL).

In our previous work [9] we have experimented the approach on several benchmark datasets showing that the FL scheme generally achieves better results compared to models generated locally, yet being sometimes outperformed by the centralized approach (we point out that this approach is not however viable due to privacy limitation).

III. Fed-XAI for video streaming QoE prediction

The development of new generation networks B5G/6G is currently ongoing and will enable several services, which will be likely based on AI as enabling technology. In this work, we consider an automotive application where connected vehicles are User Equipments (UEs) of a mobile network. Each UE is connected to its respective base station (BS) and receives a live video acquired by the camera of the vehicle in front of it, possibly enabling advanced driving assistance support systems, such as safety distance evaluation. An essential requirement to offer such services is that the video should be continuously displayed with high quality. One indicator, typically used in the context of telecommunications, is the Quality of Experience (QoE): a measure of end-user satisfaction in enjoying a service [25].

A prediction of real-time QoE to monitor the system, early-detect problems and/or predict incoming failures can be achieved using AI-based approaches [25]–[27]. However, there are some challenges to consider: i) the availability of representative datasets (network operators are reluctant to share data, in particular for B5G/6G scenarios); ii) systems need to comply with stringent requirements, both technical, ethical and legislative. In particular, the prediction must be delivered in a timely way (if the system is too slow in providing outputs, the prediction would be outdated and, thus, useless). Next, the system must be able to leverage knowledge in data from different UEs without directly accessing it in order not to violate users' privacy. We describe a method to deal with these challenges using: i) an ad-hoc dataset created via a simulation carried out with Simu5G [28] (an open-source *model library*

for the OMNeT++ simulation framework[2]), and ii) exploiting a Fed-XAI model designed to be trustworthy and, at the same time, achieve good accuracy. The application is conceived in a FL-as-a-Service (FLaaS) fashion, providing the B5G/6G network with flexible mechanisms that allow end-users to exploit the FL service. This scenario was already introduced in one of our previous works [29], but the learning of the model for QoE prediction was not addressed by employing a federated approach as it would be natural in this domain. In the following, we describe the generation of the dataset we used in the experimental analysis and its preprocessing.

A. Dataset description

Raw data are publicly available[3] and consist of a set of Quality of Service (QoS) and QoE metrics obtained by simulating scenarios in which 15 vehicles (UEs) acquire video-streams while moving. In the simulations, video-streams flow from a video server towards the UEs and each UE collects data for 120 seconds at discrete time-tagged moments. One experiment is defined as one independent replica of one simulation lasting 120 seconds for one UE. The dataset used in this work consists of 24 independent runs. Several time-tagged metrics reported in Table I are associated with each simulation run. More details on the dataset simulation and content are reported in [29].

TABLE I
DESCRIPTION OF THE METRICS INCLUDED IN THE DATASET: CQI = CHANNEL QUALITY INDICATOR; SINR: SIGNAL TO INFERENCE PLUS NOISE RATIO; RTP: REAL-TIME TRANSPORT PROTOCOL DL = DOWNLINK. ALL METRICS ARE RECORDED BY THE UES, EXCEPT THE AVGSERVEDBLOCKSDL ONE, WHICH IS INSTEAD RECORDED BY THE BSS

Name	Description
Context	
UE position	(x, y, z) coordinates of the UE in the floorplan
UE speed	Speed of the UE in m/s
QoS metrics	
avgServedBlocksDl	Number of Resource Blocks occupied in downlink
averageCqiDl	CQI values reported in DL
rcvdSinrDl	SINR value measured at packet reception
servingCell	ID of the new serving cell after the handover
frameSize	Size of the displayed frame (Byte)
rtpPacketSize	Size of the RTP packet (Byte)
end2EndDelay	Time between transmission and reception of an RTP packet
interArrivalTimeRtp	Interarrival time between two RTP packets
rtpLoss	RTP packets of frame lost
QoE metrics	
framesDisplayed	Frame percentage arrived at the time of its display
playoutBufferLength	Frame buffer size
firstFrameElapsedTime	Three values: 1) timestamp of the UE request 2) timestamp of the sender 3) time between the request and the first frame displayed

[2]OMNeT++ Website: https://omnetpp.org, accessed May 2022

[3]http://docenti.ing.unipi.it/g.nardini/ai6g_qoe_dataset.html, last visited Feb. 2023

B. QoE prediction task as a regression problem

We formulate the QoE prediction problem as a regression task. The target of the regression is to estimate the average value of the *framesDisplayed* metric at a specific time horizon H in the near future. The target value is estimated considering as input to our regression model a set of statistics calculated on the historical values of the metrics shown in Table I, within a time window W.

The following steps have been carried out to transform the raw dataset into a regression dataset suitable for the downstream adoption of AI approaches:

- before calculating the statistics, any missing value in the metrics has been filled with pre-defined values;
- in our experiments, we fixed H and W equal to 1 and 3 seconds, respectively. We extracted and considered as input features for our regression model the following statistics for each metric: mean, median, max, min, variance, standard deviation, kurtosis, skewness, Q1, Q3, and the number of actual samples used for computing the statistics;
- we clipped extracted features in the range [0, 1] after robust scaling using quantiles (0.025, 0.975).
- the analysis carried out in [29], where a Decision Tree was applied to predict the QoE in a non-Federated fashion, suggested us that a reduced set of the extracted features can be effective for the target estimation. Specifically, we selected the following features:
  - framesDisplayed_Q3,
  - framesDisplayed_mean,
  - playoutBufferLength_mean,
  - interArrivalTimeRtp_max,
  - framesDisplayed_median,
  - playoutBufferLength_counter,
  - distanceBS_variance,
  - distanceBS_stdev,
  - interArrivalTimeRtp_counter,
  - interArrivalTimeRtp_skew,
  - framesDisplayed_kurtosis,
  - end2endDelay_mean,
  - rcvdSinrDl_Q3,
  - end2endDelay_counter,
  - distanceBS_skew.

More details on how we built the original regression dataset from the raw data con be found in [29].

IV. EXPERIMENTAL SETUP

The experimental analysis follows the footprints of the work described in [29], where a Decision Tree was applied to predict the QoE in a non-Federated fashion. Unlike the previous work, we use the TSK-FRBS model and we compare it with a MLP model, as opaque reference model, in terms of predictive accuracy. We trained both the models according to the FL, LL and CL settings.

The first 20 out of 24 runs of each UE are used as training sets for the models, and the last 4 runs are employed as

test sets to evaluate their generalization capabilities. From now on, we will refer to the last four runs as Run 1, Run 2, Run 3, and Run 4. We already described the TSK-FRBS model and its parameters in Section II. As for the MLP, the architecture consists of a linear stack of two fully connected layers having 64 neurons each, with ReLu (Rectified Linear Unit) as activation function, followed by a single unit linear layer, which is a typical setup for scalar regression. Mean Squared Error (MSE) is used as Loss function and Adam as optimizer. Although finding the optimal hyperparameters was out of the scope of this work, we tried out different architectures, monitoring training and validation loss on our data (using a basic hold-out validation as evalutation method). We chose the architecture standing right at the border between underfitting and overfitting.

As regards the FL parameters for the MLP, let K be the total number of UEs; let C be the fraction of UEs participating in the training stage; let E be the number of local epochs each client executes over its data; let B be the local mini-batch size and R the number of federation rounds. In the experiments, we set $K = 15$, $C = 1$, $E = 5$, $B = 64$ and $R = 5$.

The quality of prediction of the models is evaluated through the MSE and the coefficient of determination (R^2). For the purpose of performance assessment, we evaluate the metrics as follows: regardless of the learning setting, we consider the actual partition of the dataset across UEs and we evaluate the models generated by CL, FL and LL on the local training and test sets.

V. Experimental results

In this section, we report and discuss the results. We consider four perspectives: prediction accuracy, interpretability of the used models, communication cost and sensitivity to the number of clients involved in the federation. All the four perspectives are relevant in our use case and we need to find a good trade-off among them.

A. Prediction accuracy

Table II shows the average values of MSE and R^2 obtained on the training and test sets for the different approaches.

TABLE II
AVERAGE VALUES OF MSE AND R^2 ON THE TRAINING AND TEST SETS, OBTAINED WITH THE THREE LEARNING SETTINGS, FOR EACH OF THE MODELS USED.

	FL		**LL**		**CL**	
	Train	Test	Train	Test	Train	Test
MSE (TSK)	0.052	0.066	0.030	0.094	0.045	0.057
MSE (MLP)	0.056	0.060	0.047	0.062	0.047	0.055
R^2 (TSK)	0.614	0.559	0.799	0.376	0.692	0.617
R^2 (MLP)	0.560	0.590	0.675	0.576	0.678	0.628

We can observe how the FL approach outperforms the LL one both for MLP and TSK-FRBS, but, as expected, does not achieve the performance of the CL approach. The values obtained by the FL models in Table II show that the difference between the baseline MLP and the TSK-FRBS on the test set is around 10%. In the specific application, such difference is deemed as acceptable, also considering that interpretability is a major requirement. In Table II we have also shown the CL setting results with the aim of presenting a complete overview. For the sake of simplicity, and since CL represents an unfeasible scenario in practice (recalling the discussed data privacy constraints), from here on we will mostly omit the CL results.

The average values of MSE and R^2 give a general overview of the behaviour of the models. With the aim of carrying out a more effective and robust evaluation and comparison, we execute a statistical test. Namely, we adopted the pairwise Wilcoxon signed-rank test of the distributions of 60 values of both MSEs and R^2 metrics: such values consists of the metrics computed on the local test sets by single run (4 Runs each) of the 15 UEs involved in the experimentation. Table III shows the results of the statistical test. These results confirm that the FL approach outperforms the LL one, both considering MSE and R^2. We also observe that the federated MLP (FL-MLP) approach achieves better results than the federated TSK-FRBS (FL-TSK).

TABLE III
RESULTS OF THE PAIRWISE WILCOXON SIGNED-RANK TEST APPLIED TO THE DISTRIBUTIONS OF MSES AND R^2 OBTAINED ON THE TEST SET FOR A SIGNIFICANCE LEVEL OF $\alpha = 0.05$. R^+ (R^-) IS THE SUM OF RANKS OF THE DIFFERENCES IN WHICH THE FIRST (SECOND) REPORTED FL MODEL OUTPERFORMED THE SECOND (FIRST) ONE.

MSE				
Comparison	$\mathbf{R^+}$	$\mathbf{R^-}$	p**-value**	**Hypothesis**
FL-TSK vs LL-TSK	1563	267	2e-06	Rejected
FL-MLP vs FL-TSK	1247.5	582.5	0.014	Rejected
$\mathbf{R^2}$				
Comparison	$\mathbf{R^+}$	$\mathbf{R^-}$	p**-value**	**Hypothesis**
FL-TSK vs LL-TSK	1575	255	1e-06	Rejected
FL-MLP vs FL-TSK	1243	587	0.016	Rejected

Figures 2 and 3 report the average MSEs of the two models on the test set (average computed over Runs 1, 2, 3, and 4) of each UE. The FL setting allows an improvement in performance for all clients with respect to LL models. As regards the TSK model, Clients 6 and 10 particularly benefit from the adoption of the FL model which shows good performance while LL models severely overfit. This confirms that the FL models generalize better than the LL ones thanks to the additional knowledge contained in data of the other clients.

B. Interpretability Analysis of the Federated TSK-FRBS for QoE Prediction

Complexity, often calculated as the total number of rules, is usually adopted as a measure of the global interpretability of a TSK-FRBS: the lower the complexity, the higher the interpretability. Table IV shows the number of rules for the TSK-FRBSs obtained by the FL, LL and CL approaches. In the case of the LL setting, we report the average and standard

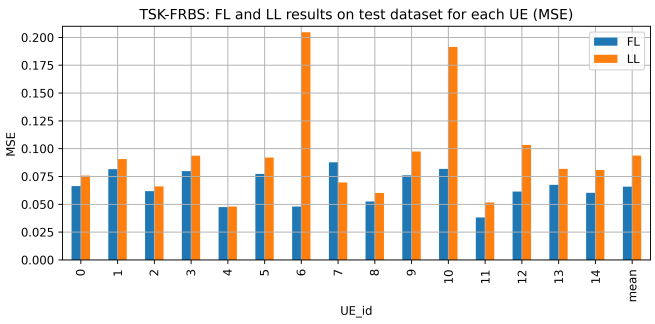

Fig. 2. Average MSEs on test set for the FL and LL settings of the TSK-FRBS model for each UE

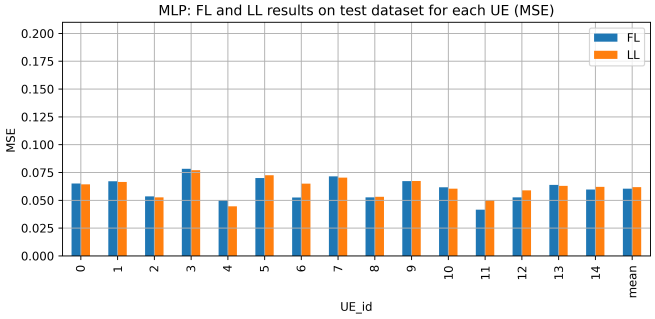

Fig. 3. Average MSEs on test set for the FL and LL settings of the MLP model for each UE

deviation of the number of rules of the local TSK-FRBSs generated by the fifteen UEs. As expected, the number of rules of the FL and CL approaches is equal. It is worth to notice that the average number of rules of the local TSK-FRBSs is around one third of the number of rules of the TSK-FRBSs generated with the FL and CL settings. Indeed, in the LL approach each UE uses only the fraction of the data available locally to train the model, consisting of about 2300 samples on average. In CL all the data (about 35400 samples) are used in the training phase, and in FL rules are aggregated as discussed in Section II. However, the prediction capability of models generated using FL and CL is statistically higher than the one of the single local models. Thus, from a global interpretability point of view, we can state that the federated TSK-FRBS represents a good trade-off solution, especially because its prediction capability (considering the R^2) is, on average, almost two times higher that the one of local models.

TABLE IV
MODEL COMPLEXITY: NUMBER OF RULES OF THE TSK-FRBSS.

	FL	LL	CL
TSK-FRBS	997	289.1 $\pm$ 21.6	997

Another important aspect to take into consideration is the local intepretability of the TSK-FRBS, discussed in Section II. Indeed, this model adopts only one rule for making a prediction, namely the one with the highest firing strength. Thus, to some extent our TSK-FRBS is able to explain why and how an output has been generated. As an example, in the following we show a rule (Eq. 3) of the TSK-FRBS, generated in FL fashion for QoE prediction task:

$$
\begin{aligned}
R_k : \textbf{IF } & framesDisplayed\_Q3 \ is \ High \\
\textbf{AND } & framesDisplayed\_mean \ is \ Medium \\
\textbf{AND } & playoutBufferLength\_mean \ is \ Medium \\
\textbf{AND } & interArrivalTimeRtp\_max \ is \ High \\
\textbf{AND } & framesDisplayed\_median \ is \ Low \\
\textbf{AND } & playoutBufferLength\_counter \ is \ Medium \\
\textbf{AND } & distanceBS\_variance \ is \ Low \\
\textbf{AND } & distanceBS\_stdev \ is \ Low \\
\textbf{AND } & interArrivalTimeRtp\_counter \ is \ High \\
\textbf{AND } & interArrivalTimeRtp\_skew \ is \ High \\
\textbf{AND } & framesDisplayed\_kurtosis \ is \ Low \\
\textbf{AND } & end2endDelay\_mean \ is \ High \\
\textbf{AND } & rcvdSinrDl\_Q3 \ is \ Medium \\
\textbf{AND } & end2endDelay\_counter \ is \ High \\
\textbf{AND } & distanceBS\_skew \ is \ Medium \\
\textbf{THEN } : & QoE = -0.210 \\
& + 0.246 \cdot framesDisplayed\_Q3 \\
& + 0.465 \cdot framesDisplayed\_mean \\
& + 0.636 \cdot playoutBufferLength\_mean \\
& - 0.291 \cdot interArrivalTimeRtp\_max \\
& + 0 \cdot framesDisplayed\_median \\
& + 0.293 \cdot playoutBufferLength\_counter \\
& + 0.001 \cdot distanceBS\_variance \\
& + 0.019 \cdot distanceBS\_stdev \\
& + 0.223 \cdot interArrivalTimeRtp\_counter \\
& - 0.21 \cdot interArrivalTimeRtp\_skew \\
& + 0 \cdot framesDisplayed\_kurtosis \\
& - 0.257 \cdot end2endDelay\_mean \\
& + 0.454 \cdot rcvdSinrDl\_Q3 \\
& + 0.223 \cdot end2endDelay\_counter \\
& - 0.031 \cdot distanceBS\_skew
\end{aligned}
\tag{3}
$$

It is worth recalling that a high semantic interpretability of the antecedent part of the rules is ensured by using uniform strong fuzzy partitions with only 3 fuzzy sets labeled as *Low*, *Medium* and *High*.

The rule suggests that a situation where framesDisplayed_mean, playoutBufferLength_mean, and rcvdSinrDl_Q3 are Medium, the predicted output is positively affected by those features (higher values of the coefficients associated with these features). However, the rule also suggests that QoE decreases when interArrivalTimeRtp_max is high. This can be interpreted as follows: on one hand, when the current scenario shows no large anomalies from a nominal situation regarding number of frames displayed, length of the playout buffer and SINR, we could expect an incoming high value of QoE. On the other hand, a degradation of the QoE is expected when the inter-arrival time between two packets in the past window is high. Thanks to the explainability provided by each rule, some countermeasures can be taken in the situations where a QoE degradation is predicted. Indeed, we can expect that the

rules activated for each specific prediction can be stored in a logging file or shown on an interactive dashboard. The logging file or the dashboard may be used by an operator for actually deciding if taking any specific countermeasures.

C. Communication cost analysis

In the proposed QoE prediction application, the aggregator must be able to deploy the generated model to the clients in a timely manner to ensure a highly performing service. Since the environment where the vehicles are moving is dynamical, and the signal can degrade due to the presence of obstacles or tunnels, it is important that the transmission time is adequate. One of the main factors influencing the time is the size of the data amount to be exchanged. In the case of the MLP, this size is about 6 MB: for each of the 5 federation rounds, the 15 clients exchange with the server about 5000 parameters (the MLP weights). Then, the server sends the aggregated model back to each client. As regards the TSK-FRBS, the amount of data exchanged between each client and the server is about 70 kB. After that, the aggregated model sums up to 240 kB and can be redeployed locally in the clients. Thus, the total size of exchanged information is about 0.3 MB, an order of magnitude smaller than the MLP scenario.

D. Sensitivity to the number of clients involved in the FL of TSK-FRBSs

We performed a sensitivity analysis on the number of clients involved in the FL of TSK-FRBSs by changing this number from a minimum of 3 to a maximum of 15 clients (default strategy). We aimed to recreate a possible real-world scenario, when, for instance, some clients are not available or have not been authorized to share local models or have lost the connection. We expect that the smaller is the number of clients involved in the federation, the lower is the accuracy. We randomly sampled 3, 6, 9, and 12 clients out of the total 15, repeating the procedure 10 times, and we calculated the results on the test set of Runs 1, 2, 3 and 4. Figure 4 shows the MSEs obtained by the FL approach for the different numbers of clients. We can note that actually the MSE increases with the decrease of the number of clients, thus confirming our expectations. The non-overlapped confidence intervals suggest that the differences between the MSEs in the various scenarios are significant, supporting the intuition that a higher fraction of participating clients is preferable.

The difference between the MSEs obtained by a federation with all the possible clients involved (15) and the minimum number (3) is evident. We observe however that the TSK-FRBSs learned with this least favourable scenario still outperform the ones learned with the LL approach. Indeed, the average MSE on the test set is 0.094 with the LL approach and 0.077 with the federation of 3 clients.

VI. Conclusions

This paper has presented the results obtained by the application of TSK-FRBS models learned with a federated approach to the prediction of Quality of Experience metrics from simulated time series of Beyond 5G/6G networks data. The approach allows simultaneously addressing the requirements of trustworthiness and data privacy. The results showed that the TSK-FRBSs learned through a federated approach overcome the ones learned by using only the local training sets. Since one of the challenges is being able to balance between model performance and explainability, we compared the results achieved by the TSK-FRBS learned with the federated approach with a baseline represented by a federated multi-layer perceptron neural network. The results have highlighted that the baseline overcomes the TSK-FRBS in terms of MSE and R^2 scores, but does not guarantee the same level of interpretability. Concerning the communication cost, the size of the data exchanged by the baseline federated approach is significantly larger than the one of the federated TSK-FRBS. Finally, we carried out a sensitivity analysis on the number of clients participating in the federation, confirming how the accuracy is considerably affected by this number.

The results presented in the paper are encouraging and show that the application of a federated approach to learn explainable AI models in the context of Beyond 5G and 6G networks is viable. As future work, we plan to improve the accuracy of the TSK-FRBS models by exploiting second-order TSK-FRBSs and modifying accordingly the federated learning approach for taking these models into consideration.

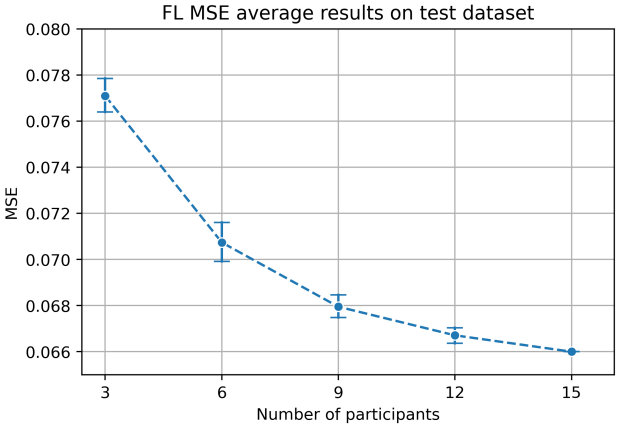

Fig. 4. Average MSEs obtained by TSK-FRBSs generated by FL over 10 randomly repeated experiments extracting n clients without replacement, with n=3,6,9,12. Error bars indicate the 95% confidence interval.

References

[1] F. Miltiadis, L. Vasiliki, M. Jafar, M. Mattia, E. U. Soykan, B. Tamas, R. Nandana, R. Nuwanthika, L. Le Magoarou, P. Pietro *et al.*, "Pervasive artificial intelligence in next generation wireless: The Hexa-X project perspective," in *First International Workshop on Artificial Intelligence in Beyond 5G and 6G Wireless Networks (AI6G 2022)*, 2022.

[2] "Ethics Guidelines for Trustworthy AI, Technical Report," 2019, European Commission. High Level Expert Group on AI. 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 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.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282.

[5] 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, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[6] J. L. C. Bárcena, M. Daole, P. Ducange, F. Marcelloni, A. Renda, F. Ruffini, and A. Schiavo, "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. [Online]. Available: https://ceur-ws.org/Vol-3277/paper8.pdf

[7] 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*, vol. 13, no. 8, p. 395, 2022.

[8] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 1, pp. 116–132, 1985.

[9] J. L. Corcuera Bárcena, P. Ducange, A. Ercolani, F. Marcelloni, and A. Renda, "An Approach to Federated Learning of Explainable Fuzzy Regression Models," in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2022, pp. 1–8.

[10] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

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

[12] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[13] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, and B. He, "A survey on federated learning systems: vision, hype and reality for data privacy and protection," *arXiv preprint arXiv:1907.09693*, 2019.

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

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

[16] J. Fiosina, "Explainable Federated Learning for Taxi Travel Time Prediction," in *VEHITS*, 2021.

[17] ——, "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.

[18] P. Chen, X. Du, Z. Lu, J. Wu, and P. C. Hung, "EVFL: An explainable vertical federated learning for data-oriented Artificial Intelligence systems," *Journal of Systems Architecture*, vol. 126, p. 102474, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1383762122000583

[19] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[20] A. Wilbik and P. Grefen, "Towards a Federated Fuzzy Learning System," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2021, pp. 1–6.

[21] X. Zhu, D. Wang, W. Pedrycz, and Z. Li, "Horizontal Federated Learning of Takagi–Sugeno Fuzzy Rule-Based Models," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 9, pp. 3537–3547, 2022.

[22] 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, and A. Abay, "IBM Federated Learning: an Enterprise Framework White Paper V0.1," 2020. [Online]. Available: https://arxiv.org/abs/2007.10987

[23] Y. Wu, S. Cai, X. Xiao, G. Chen, and B. C. Ooi, "Privacy Preserving Vertical Federated Learning for Tree-Based Models," *Proc. VLDB Endow.*, vol. 13, no. 12, p. 2090–2103, sep 2020. [Online]. Available: https://doi.org/10.14778/3407790.3407811

[24] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025511001034

[25] V. Vasilev, J. Leguay, S. Paris, L. Maggi, and M. Debbah, "Predicting QoE Factors with Machine Learning," in *2018 IEEE Int'l Conf. on Communications (ICC)*, 2018, pp. 1–6.

[26] A. Renda, P. Ducange, G. Gallo, and F. Marcelloni, "XAI Models for Quality of Experience Prediction in Wireless Networks," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2021, pp. 1–6.

[27] H. E. Dinaki, S. Shirmohammadi, E. Janulewicz, and D. Côté, "Forecasting Video QoE With Deep Learning From Multivariate Time-Series," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 512–521, 2021.

[28] G. Nardini, D. Sabella, G. Stea, P. Thakkar, and A. Virdis, "Simu5G–An OMNeT++ Library for End-to-End Performance Evaluation of 5G Networks," *IEEE Access*, vol. 8, pp. 181 176–181 191, 2020.

[29] J. L. Corcuera Bárcena, P. Ducange, F. Marcelloni, G. Nardini, A. Noferi, A. Renda, G. Stea, and A. Virdis, "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)*, 2022.