Towards Bridging the FL Performance-Explainability Trade-Off: A Trustworthy 6G RAN Slicing Use-Case

Swastika Roy, *Student Member, IEEE*, Hatim Chergui, *Senior Member, IEEE*
and Christos Verikoukis, *Senior Member, IEEE*

***Abstract*—In the context of sixth-generation (6G) networks, where diverse network slices coexist, the adoption of AI-driven zero-touch management and orchestration (MANO) becomes crucial. However, ensuring the trustworthiness of AI black-boxes in real deployments is challenging. Explainable AI (XAI) tools can play a vital role in establishing transparency among the stakeholders in the slicing ecosystem. But there is a trade-off between AI performance and explainability, posing a dilemma for trustworthy 6G network slicing because the stakeholders require both highly performing AI models for efficient resource allocation and explainable decision-making to ensure fairness, accountability, and compliance. To balance this trade off and inspired by the closed loop automation and XAI methodologies, this paper presents a novel explanation-guided *in-hoc* federated learning (FL) approach where a constrained resource allocation model and an *explainer* exchange—in a closed loop (CL) fashion—soft attributions of the features as well as inference predictions to achieve a transparent 6G network slicing resource management in a RAN-Edge setup under non-independent identically distributed (non-IID) datasets. In particular, we quantitatively validate the faithfulness of the explanations via the so-called attribution-based *confidence metric* that is included as a constraint to guide the overall training process in the run-time FL optimization task. In this respect, Integrated-Gradient (IG) as well as Input $\times$ Gradient and SHAP are used to generate the attributions for our proposed in-hoc scheme, wherefore simulation results under different methods confirm its success in tackling the performance-explainability trade-off and its superiority over the unconstrained Integrated-Gradient *post-hoc* FL baseline.**

***Index Terms*—6G, closed-loop, federated learning, game theory, in-hoc, post-hoc, proxy-Lagrangian, resource allocation, XAI, ZSM**

I. Introduction

6G networks aim to support numerous simultaneous and diverse slices for various vertical use cases, resulting in increased complexity. Extensive research efforts have been dedicated to AI-based autonomous management and orchestration within the zero-touch network and service management (ZSM) framework, standardized by the ETSI, which aims to effectively handle the end-to-end network and services to accommodate the various services offered by 6G networks [1]. On the other hand, FL is a decentralized approach to machine learning that allows for the training of models on distributed data without the need to transfer the data to a central server. The ZSM framework can incorporate FL as a key component for managing zero-touch distributed network slices while preserving the confidentiality of sensitive data [2]. Furthermore, for the widespread adoption of AI in telecommunication networks, addressing the lack of transparency, trust, and explainability in black-box AI models becomes crucial [3]. The deployment of network automation requires a thorough understanding of AI models' decision-making and behavior. This drives the search for AI approaches that offer understandable and explainable outcomes. In this regard, explainable AI (XAI) methods have emerged as a means to scrutinize the decisions made by black-box AI models. These methods build white-box models and generate feature contributions to explain AI decisions, promoting fairness, accuracy, and transparency. The ability to provide confidence and trust in AI models is vital for businesses and organizations deploying AI-enabled systems [4], [3]. Nonetheless, there exists a clear trade-off between performance and explainability which is an ongoing challenge in the field of machine learning. Simpler models like linear regression or decision trees are more interpretable but may have limited predictive performance with complex data. On the other hand, complex models like deep neural networks can achieve higher accuracy but lack interpretability [4]. Additionally, minimizing the model loss conflicts with its explainability. In real-world deployments, it is essential to tackle the trade-off between performance and explainability towards ensuring trustworthy network orchestration. To achieve this goals, we anticipate that *in-hoc* XAI approaches, that leverage the explanations during model's training and optimization, is a path to achieve that trade-off. We therefore introduce an explanation-guided *in-hoc* FL strategy for achieving transparent zero-touch service management of 6G network slices at the RAN-Edge, specifically in a non-independent identically distributed (non-IID) data.

S. Roy is with Iquadrat Informatica and Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, H. Chergui is with i2CAT Foundation, Barcelona, Spain and C. Verikoukis is with University of Patras and ISI/ATHENA, Greece. [e-mail: s.roy@iquadrat.com, chergui@ieee.org, cveri@isi.gr]

A. Related Work

This section discusses the state-of-the-art works on explainable AI (XAI) in the telecommunications domain. AI transparency is now more critical than ever with the advent of 6G networks, especially for zero-touch network management [4], [5]. The "human-centric" character of 6G highlights how important it is for XAI to win over humans and keep them in the loop [5]. The authors in [6] specify the challenges of developing XAI methods and show causal inference in 6G technology. The importance of explainability in 5G for essential services, device-to-device architectural security, and causal inference in 6G technology has also been emphasized in [7], [4]. Moreover, The significance of explainability in upcoming 6G networks has been acknowledged by researchers [8], with applications to resolving handover and resource allocation issues. A study [9] compares XAI techniques and suggests that SHAP is the best for identifying the cause of SLA violations. However, the study lacks clear explanations for the discrepancies among the results, casting doubt on the model's reliability. Also, no quantitative metrics are used to evaluate the fidelity of the explanations. Thus, to resolve this gap, in [10], an extensive range of XAI metrics is introduced for qualitative assessment, but no integration-based XAI methods are considered, and there is also no comparative analysis of different XAI techniques. In [4], it is emphasized that the interpretability and transparency of AI/ML models are crucial for the full automation of a ZSM framework in 6G networks. Recently, [11] proposes a neuro-symbolic XAI twin framework to enhance reasoning and trustworthiness in zero-touch IoE service management. However, they do not incorporate a explainable-guided AI approach or address distributed learning with XAI in the ZSM framework like us. Additionally, in [12], the concept of Federated Learning (FL) of XAI is introduced, focusing on its importance in 5G/6G networks. However, their work lacks quantitative validation of explanation faithfulness. In [13], the authors utilize the SHAP method from XAI to apply a trust-aware deep reinforcement learning-based device selection technique. This technique is employed in an FL training process for an autonomous driving scenario. On the other hand, several AI-based research works [14], [2] have focused on solving resource allocation issues in 6G network slices. However, most of these works have not addressed the challenge of providing explainability in their proposed solutions. While some research areas like NLP and Healthcare have started recognizing the trade-off between model performance and explainability [15], the field of telecommunication is still in its early stages concerning this concept. However, in a fortunate development, the authors [4] have emphasized the importance of *in-model* explainable AI (XAI) methods for the beyond 5G/6G field. They discuss the challenges and significance of achieving a better trade-off between explainability and AI model performance. They advocate for developing inherently explainable models and establishing quantifiable metrics to evaluate the effectiveness of XAI. *In-model* explainability aims to create transparent models from the start, reducing reliance on post hoc XAI techniques. This approach allows stakeholders to understand the decision-making process or any decision while balancing performance and explainability. In addition to the aforementioned research areas, a new research approach called Explanation Guided Learning (EGL) has emerged, which aims to exploit the explanations of machine learning models during the training to guide the model towards fulfilling both trustworthiness and performance. Specifically, EGL incorporates additional supervision signals or XAI-driven prior knowledge into the model's reasoning process, which helps improving the confidence of model's predictions [16], which might improve the balance between explainability and AI model performance.

In essence, the state-of-the-art work indicates that as 5G began to appear in 2020, discussions about the potential of 6G technology emerged. In addition to achieving higher bandwidth, enhanced reliability, energy efficiency, and lower latency, researchers see 6G as an advanced network that is "human-centric" and powered by AI. On the other hand, incorporating AI may result in intricate and opaque decision-making procedures. To tackle this issue, XAI has emerged as a promising solution in telecommunications. However, there is a need for more comprehensive comparative analyses and quantifiable metrics for the evaluation of various XAI methods, along with a deeper exploration of emerging approaches like EGL in the context of telecommunications. Quantitative validation and insights into real-world deployment challenges are essential to bridge the gap between research and practical implementation. Addressing these gaps will be instrumental in achieving the desired balance between AI model performance and explainability, particularly in the evolving landscape of 6G networks.

B. Contributions

The main contributions of this paper are as follows:

- We introduce a novel explanation-guided *in-hoc* federated learning approach, where a constrained slice-level CPU resource allocation model and an *explainer* exchange—in a closed loop way—explanations in the form of soft feature attributions as well as predictions to achieve a transparent and explainability-aware 6G network orchestration in a non-IID setup,
- We adopt the integrated gradients (IG) XAI method to generate explanations in terms of feature attributions, and map them to a soft probability space,
- These soft attributions are then used to quantitatively evaluating the model's *confidence metric* which is included as a constraint in the FL optimization task.
- We formulate the corresponding XAI-constrained optimization problem under the *proxy-Lagrangian* framework

Table I
NOTATIONS

Notation	Description
$S_\mu(\cdot)$	Logistic function with steepness μ
$F(\cdot)$	CDF
$\bar{F}(\cdot)$	Complementary CDF
L	Number of local epochs
T	Number of FL rounds
$\mathcal{D}_{k,n}$	Dataset at CL (k,n)
$D_{k,n}$	Dataset size
$\ell(\cdot)$	Loss function
$\mathbf{W}_{k,n}^{(t)}$	Local weights of CL (k,n) at round t
$\mathbf{x}_{k,n}$	Input features
$\hat{x}_{k,n}^{(i,j)}$	Mutated features
$\hat{y}_{k,n}^{(i)}$	Allocated resource by CL (k,n)
α_n	Resource lower-bound for slice n
β_n	Resource upper-bound for slice n
ν_n	Threshold of Confidence bound for slice n
$\mathcal{D}_{k,n}$	Samples whose prediction fulfills the SLA
$\hat{z}_{k,n}^{(i)}$	Predictions after mutation
$\pi_{k,n}^{(i,j)}$	Probability Distribution
$C_{k,n}$	Confidence Metric
$\alpha_{k,n}^{(i,j)}$	Weighted attribution of features
$\lambda_{(\cdot)}$	Lagrange multipliers
R_λ	Lagrange multiplier radius
$\mathcal{L}_{(\cdot)}$	Lagrangian with respect to $(\cdot)$

and solve it via a non-zero sum two-player game strategy, while comparing with a vanilla unconstrained post-hoc IG FL baseline.

- We present a comparative analysis of additional XAI methods that are used to generate attributions for the *in-hoc* scheme and assess their confidence metrics via distribution plots.
- We showcase the impact of network conditions (channel quality indicator (CQI), OTT traffics and MIMO full-rank usage) on the output CPU allocation.

C. Notations

We summarize the notations used throughout the paper in Table I.

II. System Model and Problem Statement

A. System Model

As depicted in Fig. 1, a 6G RAN-Edge topology under a per-slice central unit (CU)/distributed unit (DU) functional split is considered, wherein DUs are co-located with the transmission/reception point (TRP), while each CU k is a virtual network function (VNF) running on top of a commodity hardware at the Edge domain and implements a closed loop (CL) $k\,(k = 1, \ldots, K)$ consisting of a key performance indicators (KPIs) collection as well as AI-enabled slice resource allocation functions. This CL concept adheres to the ZSM framework

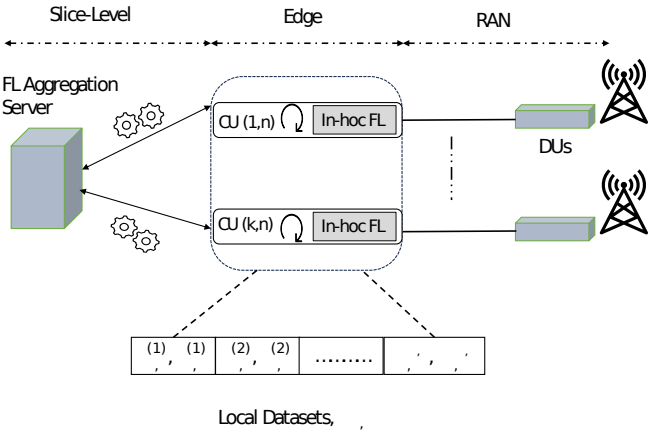


Figure 1. Decentralized closed loops (CLs) architecture

ETSI [1], which considers autonomous feedback loops between data monitoring, AI-driven data analytics, and decision-making functions to achieve particular network management tasks. In this regard, the architecture entails synthesizing data of several over-the-top (OTT) applications, which serves to build local datasets for each slice $n\,(n = 1, \ldots, N)$, i.e., $\mathcal{D}_{k,n} = \{\mathbf{x}_{k,n}^{(i)}, y_{k,n}^{(i)}\}_{i=1}^{D_{k,n}}$, where $\mathbf{x}_{k,n}^{(i)}$ stands for the input features vector, which includes OTT traffics, channel quality indicator (CQI) and multiple-input multiple-output (MIMO) full-rank usage, while $y_{k,n}^{(i)}$ represents the supervised output, which is the CPU load as shown in Table II. The considered slices include several OTTs each, namely,

- **eMBB:** Netflix, Youtube and Facebook Video,
- **Social Media:** Facebook, Whatsapp and Instagram,
- **Browsing:** Apple, HTTP and QUIC,

where the corresponding accumulated datasets are non-IID due to the different traffic patterns induced by the heterogeneous users' distribution and channel conditions. Note that the intuitive link between the input features and the output resides in the fact that when the radio conditions are enhanced, the transmission and computing queues are relaxed, delivering thereby a higher number of packets which also results in increasing the CPU load. Particularly, our system model incorporates MIMO technology, recognizing its pivotal role in enhancing data transmission and bolstering reliability through spatial diversity. Moreover, it introduces complexity into resource allocation decisions, particularly between CUs and DUs. MIMO's adaptability to changing network conditions significantly impacts CPU load requirements, making it a critical factor in resource optimization. We combine MIMO-related insights with other relevant features to drive the optimization of resource allocation and enhance overall network performance. Next, to enhance model accuracy and preserve data privacy

as well as reduce transport overhead, the architecture adopts a federated learning (FL) approach, where local closed loops (CLs) participate in training using their respective synthesized datasets. The FL process is guided by XAI ensuring transparency and interpretability during model training. The knowledge gained from each local CL is then aggregated through an E2E slice-level server, collectively improving the global AI model's explainability and AI model performance.

B. Problem Statement

Unlike post-hoc XAI strategies, the aim of this paper is to design an explanation-guided *in-hoc* FL-based resource allocation scheme for transparent zero-touch 6G network slicing at the RAN-Edge domain while balancing the trade-off between the performance and explainability. To achieve this, the proposed approach incorporates the XAI-based *confidence metric* as constraint in a closed-loop manner during the model learning process. This allows for iterative optimization and adjustment of the model based on the insights gained from XAI considerations. Indeed, this approach is needed to strengthen the intelligence entities of our ZSM-based slice-level resource management [11]. Nevertheless, in our resource allocation scheme, we leverage feature attributions obtained through Integrated Gradients (IG) to understand the factors influencing CPU resource allocation for network slices. These attributions are mapped to a soft probability space, providing a probabilistic view of feature importance. Notably, the model's confidence metric, derived from soft attributions, serves as a constraint in the allocation optimization. This approach yields transparent and explainability-aware resource allocation, aligning decisions with influential factors and the model's confidence. Moreover, we emphasize feature attributions, including those from MIMO technology, which plays a unique role in our system model, as discussed earlier. Insights from MIMO along with other features are integral to understanding factors guiding CPU resource allocation. This serve as the basis for our model's confidence metric, based on soft attributions, ensuring alignment with our goal of transparency and explainability in resource allocation.

III. Proposed Approach

Towards bridging the AI performance-explainability trade-off for trustworthy RAN slicing resource allocation, we detail in this section the proposed architecture shown in Fig. 2.

A. EGL-Driven Trustworthy Resource Allocation

Inspired by the CL and EGL principle, we propose an explanation-aided *in-hoc* federated learning architecture where the local learning is performed iteratively with run-time explanation. The overall working principle of the proposed model is manifested in Fig. 2 by clearly indicating the steps. For each local epoch, the Learner module feeds the posterior symbolic model graph to the Tester block which yields the test features and the corresponding predictions $z_{k,n}^{(i)}$ to the Explainer as shown from steps 1 to 3. Here, the model graph encapsulates the neural network's architecture (i.e., the arrangement of layers and their connectivity) and the learned parameters, which are the weights associated with each connection between neurons or nodes in the network. These weights are learned during training to make the model's predictions more accurate. The latter first generates the features attributions using one of the feature attribution XAI methods. It then converts these attributions to a soft probability distribution, as indicated in step 4, which is translated afterward into a confidence metric by the Confidence Mapper and fed back to the Learner to include it in the local constrained optimization, as pointed out in stages 5 and 6. Indeed, for each local CL (k, n), the predicted amount of resources $\hat{y}_{k,n}^{(i)}$, $(i = 1, \ldots, D_{k,n})$, should minimize the main loss function with respect to the ground truth $y_{k,n}^{(i)}$ and guided by the gained insights from XAI. Hence, as depicted in those mentioned steps 1 to 7, the optimized local weights at round t, $\mathbf{W}_{k,n}^{(t)}$, are sent to the server which generates a global FL model for slice n as,

$$\mathbf{W}_{n}^{(t+1)} = \sum_{k=1}^{K} \frac{D_{k,n}}{D_n} \mathbf{W}_{k,n}^{(t)}, \tag{1}$$

where $D_n = \sum_{k=1}^{K} D_{k,n}$ is the total data samples of all datasets related to slice n. The server then broadcasts the global model to all K CLs that use it to start the next round of turbo local optimization. Specifically, it leverages a two-player game strategy to jointly optimize over the objective and original constraints as well as their smoothed surrogates and detailed in the sequel.

Table II
Dataset Features and Output

Feature	Description
`OTT Traffics per TRP`	Apple, Facebook, Instagram, NetFlix, HTTPS, Whatsapp, and Youtube
`CQI`	Channel quality indicator reflecting the average quality of the radio link of the TRP
`MIMO Full-Rank`	Usage of MIMO full-rank spatial multiplexing in (%)

Output	Description
`CPU Load`	CPU resource consumption (%)

B. Model Testing and Explanation

In this subsection, operational details of Model Tester and Explainer blocks of Fig. 2 are discussed more descriptively. As

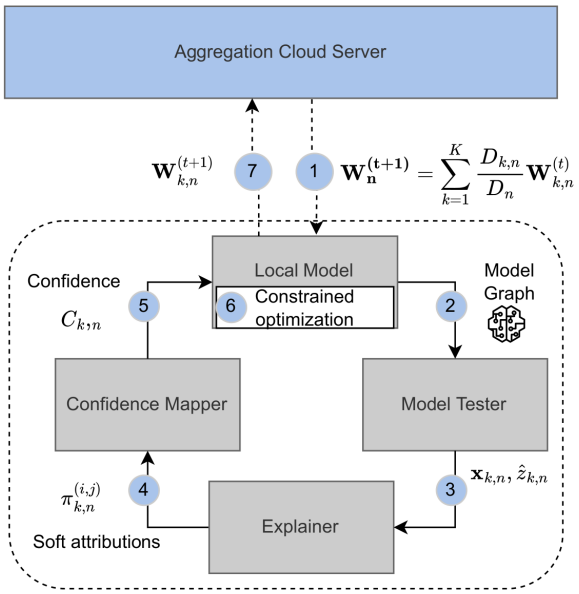


Figure 2. *in-hoc* FL for transparent resource allocation.

depicted in stage 2 of Fig. 2, upon the reception of the updated model graph, the Tester uses a batch drawn from the local dataset to reconstruct the test predictions $\mathbf{z}_{k,n}^{(i)}$. All the graph, test dataset and the predictions are fed to the Explainer at stage 3 to generate the attributions, i.e., a quantified impact of each single feature on the predicted output. Let $\mathbf{a}_{k,n}^{(i)} \in \mathbb{R}^Q$ denote the attribution vector of sample i, which can be generated by any attribution-based XAI method. In our solution, for estimating the attributions of the considered features, we mainly leverage the low-complexity Integrated Gradient (IG) scheme [17], which is based on the gradient variation when sampling the neighborhood of a feature. At stage 4, the Explainer then calculates what we call *soft attributions* by mapping the attributions into a probability space as follows,

$$\pi_{k,n}^{(i,j)} = \frac{\exp\left\{\left|\alpha_{k,n}^{(i,j)}\right|\right\}}{\sum_{q=1}^{Q} \exp\left\{\left|\alpha_{k,n}^{(i,q)}\right|\right\}}, \; j = 1, \ldots, Q, \tag{2}$$

where, $\alpha_{k,n}^{(i,j)} = a_{k,n}^{(i,j)} / x_{k,n}^{(i,j)}$ stands for the weighted attribution, since the sensitivity of the model's output with respect to an input feature in the neighborhood of the input is approximately given by the ratio of the attribution for that input feature to the value of that feature [18].

C. Confidence Mapping

To characterize the trustworthiness of the local model, we invoke the explanation-based confidence metric $C_{k,n}$ [18]. Its rationale lies in the fact that slightly shifting the value of high-magnitude features (in the sens of attributions) is an acceptable way to determine the model's conformance. In essence, the explanation (feature attributions) informs the process of assessing confidence. When a model's explanation highlights certain features as highly influential, the confidence in the model's predictions can be gauged by examining how robust those predictions are when these significant features are altered in the input neighborhood. Therefore, the explanation (attributions) directly guides the method for evaluating and understanding the model's confidence in its predictions. In our context, such an approach is vital because there will likely be no change in the SLA group (SLA violation or non-violation) if we sample over low attribution features space. In this respect, the *Confidence Mapper* at stage 5 of Fig. 2 starts by performing a feature mutation, where it selects from the dataset feature $x_{k,n}^{(i,j)}$ with probability $\pi_{k,n}^{(i,j)}$, and changes it to the baseline value, i.e., zero,

$$\hat{x}_{k,n}^{(i,j)} = x_{k,n}^{(i,j)} \times (1 - p), \; p \sim \mathcal{B}\left(1, \pi_{k,n}^{(i,j)}\right), \tag{3}$$

to force the change of the model's prediction into the opposite category. By adding Fig. 3, we have clarified the elucidation of the feature mutation process within the Confidence Mapper box of Fig. 2. This figure breaks the procedure into two steps, with numbered blue markers designating each step in order. In Stage (a), eligible features for mutation are identified, with the Feature Selector making these selections based on prior discussions. In classification tasks, the categories (or subsets) are merely the classes. In contrast, we cast CPU resource allocation as a regression problem since it is well-suited for predicting continuous variables, allowing the FL model to estimate the CPU load as a numerical value. In this case, the subsets are defined according to an SLA threshold, i.e.,

$$\mathcal{D}_{k,n} = \mathcal{U}_{k,n} \cup \bar{\mathcal{U}}_{k,n}, \tag{4}$$

where $\mathcal{U}_{k,n}$ contains the samples whose prediction fulfills the SLA, i.e., their CPU load lies in an interval $[\alpha_n, \beta_n]$. The aforementioned transformation leads to a mutated dataset $\{\hat{x}_{k,n}^{(i,j)}\}$. The Confidence Mapper reports then the fraction of samples in the neighborhood for which the decision of the model,

$$\hat{z}_{k,n}^{(i)} = \mathcal{M}_{k,n}(\mathbf{W}_{k,n}^{(t)}, \hat{\mathbf{x}}_{k,n}^{(i)}), \tag{5}$$

does not move to the other set, that is, conforms to the original decision, as the conservatively estimated confidence measure [18], i.e.,

$$C_{k,n} = \frac{1}{u_{k,n}} \sum_{i=1}^{u_{k,n}} \max\left\{\mathbb{1}_{\mathbb{R}^-}\left(\alpha_n - \hat{z}_{k,n}^{(i)}\right), \mathbb{1}_{\mathbb{R}^-}\left(\hat{z}_{k,n}^{(i)} - \beta_n\right)\right\}, \tag{6}$$

where $\hat{z}_{k,n}^{(i)}$ are the predictions after mutation, while $u_{k,n}$ stands for the size of the considered original category $\mathcal{U}_{k,n}$. This process is visually depicted in stage (b) of Fig. 3. Here, the

Confidence Analysis block, upon receiving the newly mutated dataset $\mathbf{D}_1$, computes the confidence metric and later shares it with the local model, contributing to its optimization process.

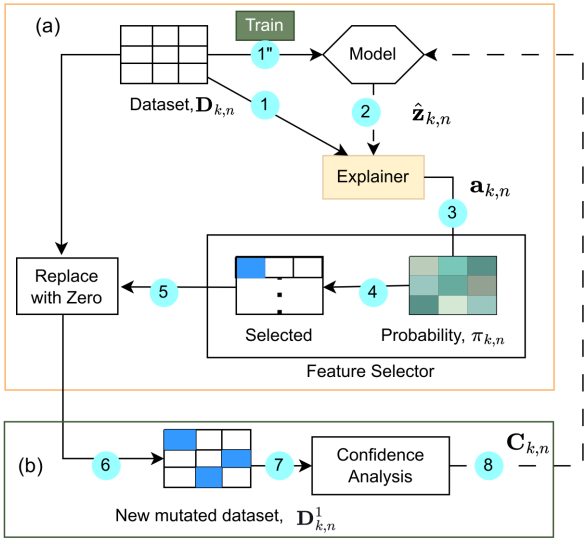


Figure 3. Feature mutation approach of confidence mapper

D. Explainability-Aware Resource Allocation

We cast the slice-level resource allocation as a confidence-constrained regression task. For this purpose, we consider the datasets corresponding to the different slices as summarized in the Table II and section II, where resources at CU-level are dynamically allocated to slices according to their traffic patterns and radio conditions (CQI and MIMO full-rank usage). In this respect, we formulate the constrained optimization using the proxy-Lagrangian framework and solve it via a non-zero sum two-player game strategy. Specifically, the confidence metric should be higher than a predefined threshold ν_n. This translates into solving a statistically constrained local resource allocation problem as shown at stage 6 of Fig. 2, i.e.,

$$\min_{\mathbf{W}_{k,n}^{(t)}} \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \ell\left(y_{k,n}^{(i)}, \hat{y}_{k,n}^{(i)}\left(\mathbf{W}_{k,n}^{(t)}, \mathbf{x}_{k,n}\right)\right), \tag{7a}$$

$$C_{k,n} \geq \nu_n, \tag{7b}$$

which is solved by invoking the so-called *proxy Lagrangian* framework [19]. This consists first on considering two Lagrangians as follows:

$$\begin{aligned}\mathcal{L}_{\mathbf{W}_{k,n}^{(t)}} =& \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \ell\left(y_{k,n}^{(i)}, \hat{y}_{k,n}^{(i)}\left(\mathbf{W}_{k,n}^{(t)}, \mathbf{x}_{k,n}\right)\right) \\ &+ \lambda_1 \Psi_1\left(\mathbf{W}_{k,n}^{(t)}\right),\end{aligned} \tag{8a}$$

Algorithm 1: *In-hoc* Explainable Federated Learning

Input: K, m, η_λ, T, L. `# See Table III`
Server initializes $\mathbf{W}_n^{(0)}$ and broadcasts it to the CLs
for $t = 0, \ldots, T-1$ **do**
  parallel for $k = 1, \ldots, K$ **do**
  Initialize $M =$ `num_constraints` and $\mathbf{W}_{k,n,0} = \mathbf{W}_n^{(t)}$
  Initialize $\mathbf{A}^{(0)} \in \mathbb{R}^{(M+1)\times(M+1)}$ with $\mathbf{A}_{m',m}^{(0)} = 1/(M+1)$
  for $l = 0, \ldots, L-1$ **do**
    Receive the graph $\mathcal{M}_{k,n}$ from the local model
    `# Test the local model and calculate the attributions`
    $a_{k,n}^{i,j} =$ `Int. Gradient` $\left(\mathcal{M}_{k,n}\left(\mathbf{W}_{k,n,l}, \mathbf{x}_{k,n}\right)\right)$
    `# Generate soft attributions`
    $\pi_{k,n}^{(i,j)} = \frac{\exp\left\{\left|\alpha_{k,n}^{(i,j)}\right|\right\}}{\sum_{p=1}^{L}\exp\left\{\left|\alpha_{k,n}^{(i,p)}\right|\right\}}$, $j = 1, \ldots, Q$,
    `# Mutate the test dataset`
    Mutate $x_{k,n}^{(i,j)}$ with probability $\pi_{k,n}^{(i,j)}$
    `# Calculate the confidence metric`
    $C_{k,n} = \frac{\sum_{i=1}^{u_{k,n}} \max\left\{\mathbb{1}_{\mathbb{R}-}\left(\alpha_n - \hat{z}_{k,n}^{(i)}\right), \mathbb{1}_{\mathbb{R}-}\left(\hat{z}_{k,n}^{(i)} - \beta_n\right)\right\}}{u_{k,n}}$
    Let $\lambda^{(l)}$ be the top eigenvector of $\mathbf{A}^{(l)}$
    `# Solve problem (7) via oracle optimization`
    Let $\hat{\mathbf{W}}_{k,n,l} = \mathcal{O}_\delta\left(\mathcal{L}_{\mathbf{W}_{k,n,l}}(\cdot, \hat{\lambda}^{(l)})\right)$
    Let $\Delta_\lambda^{(l)}$ be a gradient of $\mathcal{L}_\lambda(\hat{\mathbf{W}}_{k,n,l}, \lambda^{(l)})$ w.r.t. λ
    `# Exponentiated gradient ascent`
    Update $\tilde{\mathbf{A}}^{(l+1)} = \mathbf{A}^{(l)} \odot \cdot \exp\left\{\eta_\lambda \Delta_\lambda^{(l)}(\lambda^{(l)})\right\}$
    `# Colunm-wise normalization`
    $\mathbf{A}_m^{(l+1)} = \tilde{\mathbf{A}}_m^{(l+1)} / \left\|\mathbf{A}_m^{(l+1)}\right\|_1$, $m = 1, \ldots, M+1$
  end
  return $\hat{\mathbf{W}}_{k,n}^{(t)} = \frac{1}{L^\star}\sum_{l=0}^{L-1}\hat{\mathbf{W}}_{k,n,l}$
  Each local CL (k,n) sends $\hat{\mathbf{W}}_{k,n}^{(t)}$ to the server.
  end parallel for
  return $\mathbf{W}_n^{(t+1)} = \sum_{k=1}^{K}\frac{D_{k,n}}{D_n}\hat{\mathbf{W}}_{k,n}^{(t)}$
  and broadcasts the value to all local CLs.
end

$$\mathcal{L}_\lambda = \lambda_1 \Phi_1\left(\mathbf{W}_{k,n}^{(t)}\right) \tag{8b}$$

where Φ_1 and Ψ_1 represent the original constraints and their smooth surrogates, respectively. Specifically, the indicator term in (7b) are replaced with Logistic functions and the soft maximum is used as surrogate of the maximum as,

$$\begin{aligned}\Psi_1\left(\mathbf{W}_{k,n}^{(t)}\right) = \nu_n - \frac{1}{u_{k,n}}\sum_{i=1}^{u_{k,n}} \log\Big\{&\exp\left[S_\mu\left(\hat{z}_{k,n}^{(i)} - \alpha_n\right)\right] \\ &+ \exp\left[S_\mu\left(\beta_n - \hat{z}_{k,n}^{(i)}\right)\right]\Big\} \leq 0,\end{aligned} \tag{9}$$

where L_μ stands for the Logistic function with steepness parameter μ, i.e.,

$$S_\mu(\theta) = \frac{1}{1 + e^{-\mu\theta}}. \tag{10}$$

This optimization task turns out to be a non-zero-sum two-player game in which the $\mathbf{W}_{k,n}^{(t)}$-player aims at minimizing $\mathcal{L}_{\mathbf{W}_{k,n}^{(t)}}$, while the λ-player wishes to maximize $\mathcal{L}_\lambda$ [20, Lemma 8]. While optimizing the first Lagrangian w.r.t. $\mathbf{W}_{k,n}$ requires differentiating the constraint functions $\Psi_1(\mathbf{W}_{k,n}^{(t)})$, to differentiate the second Lagrangian w.r.t. λ we only need to evaluate $\Phi_1\left(\mathbf{W}_{k,n}^{(t)}\right)$. Hence, a surrogate is only necessary for the $\mathbf{W}_{k,n}$-player; the λ-player can continue using the original constraint functions. The local optimization task can be written as,

$$\min_{\mathbf{W}_{k,n}\in\Delta} \quad \max_{\lambda,\, \|\lambda\|\leq R_\lambda} \quad \mathcal{L}_{\mathbf{W}_{k,n}^{(t)}} \tag{11a}$$

$$\max_{\lambda,\, \|\lambda\|\leq R_\lambda} \quad \min_{\mathbf{W}_{k,n}\in\Delta} \quad \mathcal{L}_\lambda, \tag{11b}$$

where thanks to Lagrange multipliers, the λ-player chooses how much to weigh the proxy constraint functions, but does so in such a way as to satisfy the original constraints, and ends up reaching a nearly-optimal nearly-feasible solution [21]. These steps are all summarized in Algorithm 1.

At line 1 of Algorithm 1, we declare the baseline parameters. Then, at line 2, the aggregation server initializes the weights and broadcasts them to all participating CLs. Lines 3 to 4 describe the execution of the FL optimization task, as mentioned earlier at each FL round.Noted that, the preference of proxy-Lagrangian optimization over the well know convex-concave procedure [22] is firmly justified. This decision is primarily driven by the intricacy of the non-convex, dataset-dependent XAI constraints, which CCP may not accurately capture. Moreover, it provides computational efficiency, ensuring timely resource allocation decisions, unlike CCP, which often involves computationally expensive iterations for convexification. The versatility and numerical stability of proxy-Lagrangian methods make them well-suited for handling a wide range of SLA constraints and statistical metrics. In summary, proxy-Lagrangian optimization outperforms CCP in terms of efficiency and flexibility in addressing B5G resource allocation challenges.

IV. Results

In this section we evaluate the proposed *in-hoc* FL framework by first justifying the use of feature attributions as a pillar to build the explainability-aware constrained resource allocation model. We then present the FL convergence, confidence score and performance metrics and showcase their underlying trade-offs. We afterward study the correlation between features attributions and draw some important conclusions. Finally we present a time complexity analysis to assess the computational efficiency of the proposed framework.To implement the model Tester and Explainer, we invoke `DeepExplain` framework, which includes state-of-the-art gradient and perturbation-based attribution methods [23]. It provides an attribution score based on the feature's contribution to the model's output, which we integrate with our proposed constrained resource allocation FL framework in a closed-loop iterative way. Besides, our simulation-based experiments were conducted in a Python environment on Ubuntu 20.04. The proposed federated neural network was built with TensorFlow V1, featuring fully connected layers activated by the leaky ReLU function. We used the Adam optimizer with a learning rate η_λ of 0.02 to update the model's parameters.

Table III
Settings

Parameter	Description	Value
N	# Slices	3
K	# CLs	100
$D_{k,n}$	Local dataset size	1000 samples
α_n	SLA lower bound	0%
β_n	SLA upper bound	3%
T	# Max FL rounds	30
L	# Local epochs	100
R_λ	Lagrange multiplier radius	Constrained: 10^{-5}
η_λ	Learning rate	0.02

A. Parameter Settings and Baseline

Here, we define the parameters to tackle the FL optimization task. The details of the considered slices and associated KPIs are mentioned in section II. We use vectors α, β for the cpu resource bounds corresponding to the different slices for a particular resource and ν for the explainability confidence metrics threshold. The parameters settings are presented in Table III. Recently, there has been growing interest in combining FL and XAI approaches. In this work, the considered baseline is the state-of-the-art unconstrained FL with post-hoc explanation [24]. The authors [24] investigate the feature importance challenge in vertical FL scenarios and propose a method, the Sharp Federated algorithm, which utilizes Shapley values to determine feature importance in a post-hoc way.

B. Bridging the Performance-Explainability Trade-off

To study the performance-explainability trade-off of the proposed strategy, we plot both the convergence curves and the confidence metric vs. the FL rounds in Fig. 4 and Fig. 5, respectively, considering our *in-hoc* FL scheme and the baseline unconstrained IG post-hoc. It shows that the confidence metric of the proposed in-hoc approach for the different slices conserves a higher value above 85%, while presenting a similar convergence trend as the post-hoc FL baseline. In contrast, the latter fails to ensure the confidence of the model, since its confidence metric decreases as we gradually approach the convergence. This behavior conveys that the in-hoc strategy addresses the trade-off successfully, and guarantees explainability and trust in the training phase. For analysis completeness,

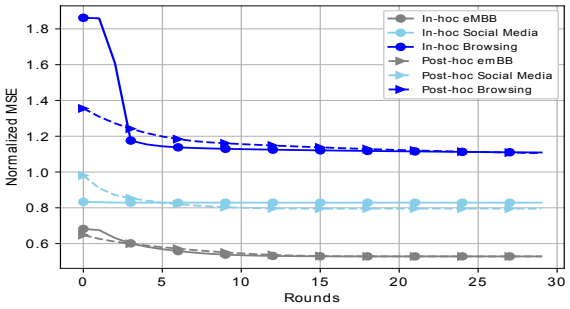


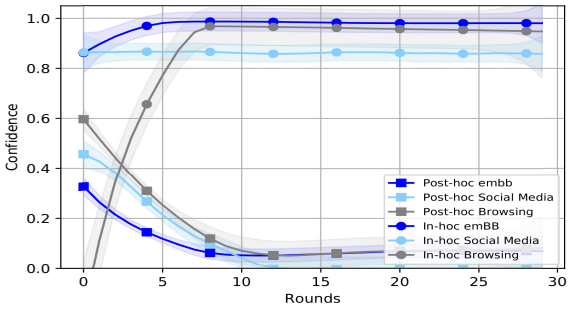
Figure 4. FL training MSE loss vs. FL rounds with $\nu = [0.82, 0.83, 0.85]$

the Explainer block in Fig. 2 is also implemented using both the perturbation-based XAI method SHAP and gradient-based method Input×Gradient [25], where Fig. 5-(b) confirms that our proposed *in-hoc* FL algorithm preserves the same behavior during the testing phase. Moreover, the confidence score has remained almost the same even if the attribution scores have been generated by various XAI methods, which makes our proposed algorithm more reliable.

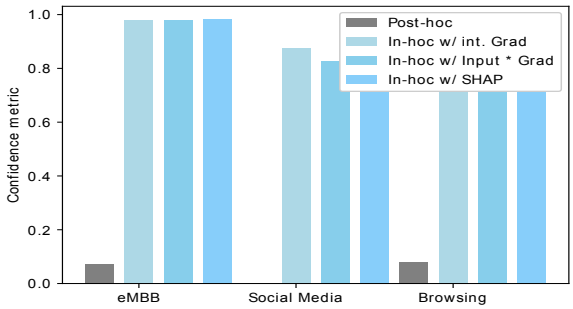
Overall, based on the presented results in Fig. 4 and Fig. 5, the constrained *in-hoc* FL ensures a trade-off between convergence and confidence, i.e., while the model loss decreases within an allowable confidence threshold, the model's confidence increases and significantly outperforms the post-hoc baseline. This indicates that the model becomes more confident in its predictions as it converges. This is achieved thanks to the explanation-guided *in-hoc* constrained FL optimization. On the contrary, in the state-of-the-art post-hoc scenario, the model confidence degrades as it starts to converge.

C. Explaining the Impact of Network Parameters on Slice Resource Allocation

In this subsection, we first examine the post-hoc feature attributions plots of the eMBB slice, which are generated via SHAP as illustrated in Fig 6. The analysis reveals that CQI has the highest impact on the CPU allocation, as it is widely concentrated towards positive values. This concentration indicates that better channel quality is leading to lower CPU load, which can be interpreted by the reduction of retransmissions and queuing time. Consequently, the model can anticipate efficient data transmission, increased network capacity, and potential resource optimization. Additionally, the concentration of OTT traffic per TRP around the lowest positive values highlights varying traffic levels that impact CPU load. Lower levels of OTT traffic typically require less processing resources, leading to decreased CPU load. The FL model can also leverage this information to optimize resource allocation and guide network planning and scaling efforts. Furthermore, the negative concentration of MIMO full rank suggests signal degradation and interference, which could potentially lead to



(a) Model confidence metric vs. FL rounds with slices' confidence bounds $\nu = [0.82, 0.83, 0.85]$.



(b) Confidence metric with the proposed in-hoc scheme on a test dataset.

Figure 5. Confidence metric using different XAI methods.

higher CPU load predictions. However, in our specific case, where the value is very small, the impact of MIMO full rank on CPU load can be considered negligible. Nonetheless, this analysis still allows the FL model to identify and address potential issues related to signal degradation and interference, providing valuable insights for network optimization in other scenarios where MIMO full rank plays a more significant role. By considering these features, the FL model can make accurate predictions and enable resource optimization, leading to improved network performance and enhanced management of CPU load in network slices in a transparent and explainable way.

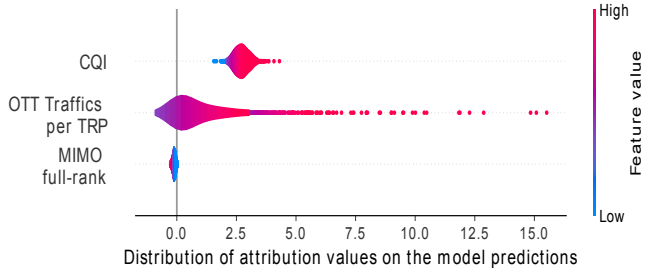


Figure 6. Distribution plot of attribution for eMBB slice

D. Time complexity & overhead

Attention has been also paid to the training time having in mind future actual deployment of our solution. The efficiency of the solution's training process directly impacts its real-world usability and effectiveness. Faster training times enable the system to respond promptly to dynamic changes, ensuring real-world efficiency and cost-effectiveness. Additionally, efficient training times facilitate scalability, allowing the system to handle increasing data volumes and user demands without compromising performance. For evaluating the time complexity of the proposed solution, the *in-hoc* FL explainer can be implemented with various feature attribution generators, we use IG as the primary method. Fig. 7 shows the convergence time of the browsing slice under our *in-hoc* FL strategy with SHAP, IG and Input $\times$ Grad as attribution scores generators as well as the baseline unconstrained post-hoc method. Since Fig. 4 indicates that the browsing slice with *in-hoc* FL starts to converge at around round 13, the computation time has been calculated until that convergence round. It can observe from the Fig. 7 that *in-hoc* FL with IG takes lower time than others but nearly close to *post-hoc* one while presenting the highest confidence. This finding suggests that utilizing IG as the primary feature attribution method in the *in-hoc* FL explainer enables faster convergence without compromising the confidence of the predictions. This is a significant advantage for real-world deployment, as it indicates efficient allocation of computational resources while maintaining the trustworthiness of the model.

Table IV
OVERHEAD COMPARISON (KB) FOR BROWSING SLICE

Scheme	Convergence round	Overhead(KB)
In-hoc	13	16510
Post-hoc	20	77714

Moreover, the results of Table IV reveal that the complexity introduced by XAI has a notable impact on the overall system performance. A significant overhead disparity can be observed in comparing the proposed *in-hoc* FL and the baseline *Post-hoc* FL. The *in-hoc* FL method exhibited substantially lower overhead (16,778 KB) at the outset of convergence in round 12, in contrast to the *Post-hoc* FL (77,776 KB), which commenced convergence at round 17. This substantial difference suggests that the *in-hoc* FL approach offers notable advantages regarding communication efficiency. Notably, overhead is directly proportional to energy consumption. Hence, it is reasonable to infer that the *in-hoc* FL method likely entails a lower computational load and potentially higher energy efficiency due to minimized data exchange. Specifically, the reduction in overhead for the *in-hoc* FL method suggests that the complexity introduced by XAI in this context leads to improved communication efficiency, which translates into potential benefits in computational load and energy efficiency. Such an approach would provide a more holistic understanding of the performance trade-offs associated with XAI complexity. In summary, our analysis provides critical insights into our solution's complexity and overhead. The in-hoc FL approach with IG achieves a 16.67% reduction in computation time compared to the Post-hoc FL, ensuring faster convergence without compromising prediction confidence, which is vital for real-world adaptability. Additionally, we observed a significant reduction in overhead with our in-hoc FL approach, signifying improved communication efficiency and potential gains in computational load and energy efficiency. This balanced approach positions our solution as a strong prospect for real-world deployment, offering efficiency and trustworthiness.

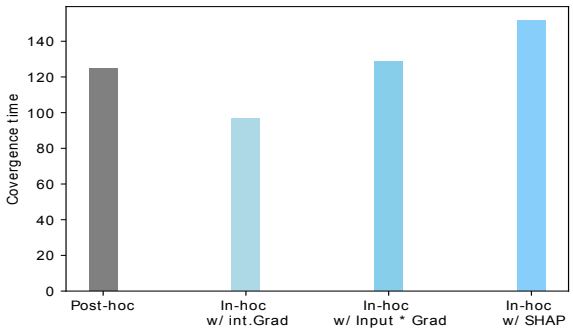


Figure 7. Convergence time of the proposed *in-hoc* FL under different explainers vs. the post-hoc baseline.

E. Scalability analysis

Scalability is paramount in FL, ensuring consistent system performance regardless of the number of participating clients. Fig. 8 shows that our *in-hoc* FL approach exhibits robust scalability. Firstly, it consistently exhibits stable convergence behavior, irrespective of changes in the number of participating clients. Secondly, the confidence metric, essential for assessing the prediction reliability, consistently remains above the 90% threshold within our in-hoc federated learning approach. This adaptability ensures reliable performance in dynamic real-world situations, making our FL solution well-suited for diverse deployment scenarios.

V. CONCLUSION

In this paper, we have presented a novel explanation-guided *in-hoc* federated learning approach to achieve a transparent zero-touch service management of 6G network slices while bridging the trade-off between performance and explainability. We have considered the XAI-based confidence metric in a closed loop way to solve the underlying joint optimization task using a proxy-Lagrangian two-player game strategy. In particular, we have used both integrated-gradient and perturbation-

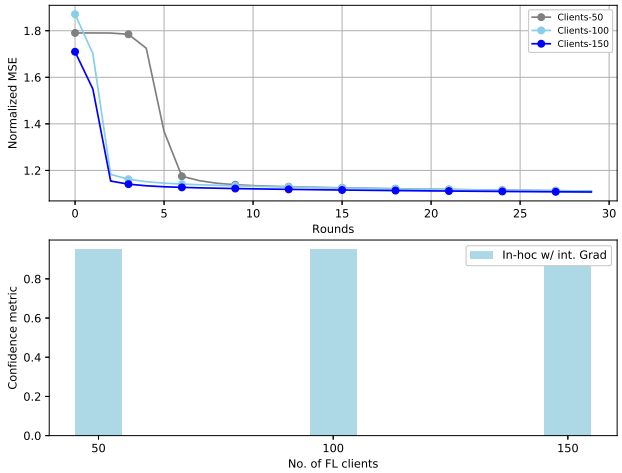


Figure 8. Scalability of the proposed In-hoc FL approach

based XAI schemes to generate the attributions consumed by our *in-hoc* FL, which present almost the same superior performance compared to an unconstrained *post-hoc* FL baseline. We have also provided a post-hoc analysis of the impact of networks parameters on the slice-level resource allocation, which points out that CQI is the key influencing feature. Finally, computational complexity (up to the convergence round) has been assessed which demonstrates that our *in-hoc* FL has less complexity than the *post-hoc* while presenting a clearly superior confidence. In the future, we aim to enhance model explainability through symbolic reasoning, which employs explicit rules and logic to make the decision-making process more understandable and trustworthy. Besides, we plan to contribute to benchmarking and standards, conduct real-world testing in 6G networks, explore user-centric and energy-efficient solutions, and ensure interoperability with existing technologies. These initiatives will improve the effectiveness and practicality of our approach in the dynamic 6G network landscape.

VI. Acknowledgment

This work has been supported in part by the projects 6G-BRICKS (101096954) HORIZON-JU-SNS-2022 and ADROIT6G (101095363) HORIZON-JU-SNS-2022. Additionaly, this work was partially funded by MINECO (Spain) and the EU NextGenerationEU/PRTR (Call UNICO I+D 5G 2021, ref. number TSI-063000-2021-10)

References

[1] (2019) Zero-touch network and service management (zsm); reference architecture. [Online]. Available: https://www.etsi.org/technologies/zero-touch-network-service-management

[2] H. Chergui, L. Blanco, and C. Verikoukis, "Statistical federated learning for beyond 5G SLA-constrained RAN slicing," *IEEE Transactions on Wireless Communications*, vol. 21, no. 3, pp. 2066–2076, 2022.

[3] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Frontiers in Big Data*, vol. 4, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2021.688969

[4] S. Wang, M. A. Qureshi, L. Miralles-Pechuán, T. Huynh-The, T. R. Gadekallu, and M. Liyanage, "Applications of explainable ai for 6g: Technical aspects, use cases, and research challenges," *arXiv preprint arXiv:2112.04698*, 2021.

[5] C. Benzaid and T. Taleb, "AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions," *IEEE Network*, vol. 34, no. 2, pp. 186–194, 2020.

[6] Y. Xiao, G. Shi, and M. Krunz, "Towards ubiquitous ai in 6g with federated learning," *arXiv preprint arXiv:2004.13563*, 2020.

[7] P. Muñoz, I. de la Bandera, E. J. Khatib, A. Gómez-Andrades, I. Serrano, and R. Barco, "Root cause analysis based on temporal analysis of metrics toward self-organizing 5G networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2811–2824, 2017.

[8] C. Li, W. Guo, S. C. Sun, S. Al-Rubaye, and A. Tsourdos, "Trustworthy deep learning in 6G-enabled mass autonomy: From concept to quality-of-trust key performance indicators," *IEEE Vehicular Technology Magazine*, vol. 15, no. 4, pp. 112–121, 2020.

[9] A. Terra, R. Inam, S. Baskaran, P. Batista, I. Burdick, and E. Fersman, "Explainability methods for identifying Root-cause of SLA violation prediction in 5G network," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–7.

[10] P. Barnard, I. Macaluso, N. Marchetti, and L. A. DaSilva, "Resource reservation in sliced networks: An explainable artificial intelligence (XAI) approach," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 1530–1535.

[11] M. S. Munir, K. T. Kim, A. Adhikary, W. Saad, S. Shetty, S.-B. Park, and C. S. Hong, "Neuro-symbolic explainable artificial intelligence twin for zero-touch ioe in wireless network," *IEEE Internet of Things Journal*, 2023.

[12] A. Renda, P. Ducange, F. Marcelloni, D. Sabella, M. C. Filippou, G. Nardini, G. Stea, A. Virdis, D. Micheli, D. Rapone, and L. G. Baltar, "Federated learning of explainable AI models in 6G systems: Towards secure and automated vehicle networking," *Inf.*, vol. 13, p. 395, 2022.

[13] G. Rjoub, J. Bentahar, and O. A. Wahab, "Explainable AI-based federated deep reinforcement learning for trusted autonomous driving," in *2022 International Wireless Communications and Mobile Computing (IWCMC)*, 2022, pp. 318–323.

[14] Y. Li, A. Huang, Y. Xiao, X. Ge, S. Sun, and H.-C. Chao, "Federated orchestration for network slicing of bandwidth and computational resource," *arXiv preprint arXiv:2002.02451*, 2020.

[15] M. Naylor, C. French, S. Terker, and U. Kamath, "Quantifying explainability in nlp and analyzing algorithms for performance-explainability tradeoff," *arXiv preprint arXiv:2107.05693*, 2021.

[16] Y. Gao, S. Gu, J. Jiang, S. R. Hong, D. Yu, and L. Zhao, "Going beyond xai: A systematic survey for explanation-guided learning," *arXiv preprint arXiv:2212.03954*, 2022.

[17] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 3319–3328.

[18] S. Jha, S. Raj, S. Fernandes, S. K. Jha, S. Jha, B. Jalaian, G. Verma, and A. Swami, "Attribution-based confidence metric for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[19] A. Cotter, M. R. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. L. Wang, B. E. Woodworth, and S. You, "Training well-generalizing classifiers for fairness metrics and other data-dependent constraints," in *International Conference on Machine Learning*, 2018.

[20] A. Cotter, H. Jiang, and K. Sridharan, "Two-player games for efficient non-convex constrained optimization," in *Algorithmic Learning Theory*. PMLR, 2019, pp. 300–332.

[21] G. J. Gordon, A. Greenwald, and C. Marks, "No-regret learning in convex games," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 360–367.

[22] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (cccp)," in *Advances in Neural Information Processing Systems*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14. MIT Press, 2001.

[23] [Online]. Available: https://github.com/marcoancona/DeepExplain
[24] G. Wang, "Interpret federated learning with shapley values," *arXiv preprint arXiv:1905.04519*, 2019.
[25] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *Int. J. Appl. Earth Obs. Geoinformation*, vol. 103, p. 102520, 2021.

Swastika Roy received her Bachelor's degree in Electronics and Telecommunication Engineering from Rajshahi University of Engineering and Technology (RUET), Bangladesh, in 2015. Later, she completed her M.Tech degree in Electrical Engineering with a specialization in Communication Engineering from the Indian Institute of Technology Bombay (IITB), India, in 2019. Currently, she is a Researcher at Iquadrat Informatica, Barcelona, Spain and concurrently pursuing her Ph.D. degree at the Signal Theory and Communication Engineering Department of Universitat Politècnica de Catalunya (UPC), Barcelona, Sapin. Previously, she worked as an Early Stage Researcher (ESR) at CTTC, Barcelona, Spain, as part of the H-2020 Marie Sklodowska Curie ITN SEMANTIC project.

Hatim Chergui (IEEE SM'22) received the Ph.D. degree (summa cum laude) in electrical engineering and telecommunications from IMT-Atlantique (Télécom-Bretagne), Brest, France. He is a Senior Researcher at i2CAT Foundation, Barcelona, Spain. He was the project manager of the H2020 MonB5G European project and a researcher at CTTC, Spain. He served as a RAN expert at both INWI and Huawei Technologies, Morocco. He was the recipient of the IEEE ComSoc CSIM 2021 Best Journal Paper Award and the IEEE ICC 2020 Best Paper Award. He chaired several workshops at IEEE ICC and GLOBECOM. He is an Associate Editor of IEEE Networking Letters.

Christos Verikoukis Christos Verikoukis [SM'07] (chverik@gmail.com) received the Ph.D. degree from Technical University of Catalonia (UPC), Barcelona, Spain, in 2000. He is currently an Associate Proferssor with the Univerisity of Patras. He has authored 158 journal papers and over 200 conference papers. He is also a co-author of three books, 14 chapters in other books, and two patents. He has participated in more than 40 competitive projects, and has served as a project coordinator of several funded projects from the European Comission and of national projects in Greece and Spain. He is currently the IEEE ComSoc GITC member and the editor-in-chief of the IEEE Networking Letters.