Joint Explainability and Sensitivity-Aware Federated Deep Learning for Transparent 6G RAN Slicing

Swastika Roy[1,2], Farhad Rezazadeh[1,2], Hatim Chergui[3], and Christos Verikoukis[4,5]

[1] Telecommunications Technological Center of Catalonia (CTTC), Barcelona, Spain
[2] Technical University of Catalonia (UPC), Barcelona, Spain
[3] i2CAT Foundation, Barcelona, Spain
[4] University of Patras, Greece
[5] ISI/ATHENA, Greece
Contact Emails: {sroy, frezazadeh}@cttc.es, chergui@ieee.org, cveri@ceid.upatras.gr

Abstract—In recent years, wireless networks are evolving complex, which upsurges the use of zero-touch artificial intelligence (AI)-driven network automation within the telecommunication industry. In particular, network slicing, the most promising technology beyond 5G, would embrace AI models to manage the complex communication network. Besides, it is also essential to build the trustworthiness of the AI black boxes in actual deployment when AI makes complex resource management and anomaly detection. Inspired by closed-loop automation and Explainable Artificial intelligence (XAI), we design an Explainable Federated deep learning (FDL) model to predict per-slice RAN dropped traffic probability while jointly considering the sensitivity and explainability-aware metrics as constraints in such non-IID setup. In precise, we quantitatively validate the faithfulness of the explanations via the so-called attribution-based *log-odds metric* that is included as a constraint in the run-time FL optimization task. Simulation results confirm its superiority over an unconstrained integrated-gradient (IG) *post-hoc* FDL baseline.

Index Terms—6G, classification, FL , game theory, proxy-Lagrangian, SLA, stochastic policy, traffic drop, XAI, ZSM

I. INTRODUCTION

The most promising 6G network slicing technology insists on adopting autonomous management and orchestration of the end-to-end (E2E) network resources at the network domains because the isolation of slices may induce a high cost in terms of efficiency [1], [2]. So, ETSI standardized zero-touch network and service management (ZSM) framework has been considered [3].Here, zero-touch refers to the automation and management of resources without human interference. Besides, developing cognitive slice management solutions in 6G networks is essential to automatically orchestrate and manage network slices, particularly network resources across different technological domains (TDs), along with ensuring the end-user's QoE and QoS [4], [5]. Hence, the [6] has proposed an AI-native network slicing management solution of 6G networks to support emerging AI services. Also, AI algorithms should

Figure 1. RAN federated traffic drop classification in NS

be driven by the distributed nature of datasets to acquire the full potential of network slicing automation, which will solve the problematic behavior of the cloud-centric traditional ML schemes. Thus, a decentralized learning approach is required to handle distributed network slices efficiently. For this, we choose Federated learning (FL) [7], [8] to handle distributed network slices efficiently like our another research work [9]. Besides, even if DNN hold the state-of-the-art [10], [11], [12] in solving resource allocation and orchestration problems of network slicing, the black-box nature of such ML models impedes understanding of their decisions, any flaws in the datasets or the model's performance behavior. Moreover, the 6G network is going to be "machine-centric" technology which signifies that all the corresponding "smart things" in the 6G network will operate intelligently but as a smart black box [13]. Here, the smart black box is not transparent in its action or decision-making processes and could have adverse effects on the network's operations of the 6G technology. In this concern, XAI provides human interpretable methods to adequately explain the AI system and its decisions for gaining

the human's trust in the loop. Also, [14] indicates that it is a prerequisite of any ZSM-based AI models in 6G to enrich translucency of their models. Viewing this fact, zero-touch XAI-driven FL will be fetching a particular emphasis for its automation and unique advantages, which are essential for end-user trust and secured procedure. In contrast, the conventional XAI focuses only on the interpretability and transparency of any ML system. Some works of XAI [15], [16], [17] indicate the importance of explainability and present some research works on handover and resource allocation, etc., in the beyond 5G networks. In [18], XAI for physical/MAC layers in 6G networks are focused. In comparison, the authors of [19] present a trust-aware federated deep reinforcement learning-based device selection technique in an autonomous driving scenario. And, to evaluate the performance of XAI models, the paper [20] introduces some essential metrics. So, in this work, we will present a novel zero-touch Explainable Federated learning (FL) as the decentralized approach for traffic drop classification in 6G network slices [7].

A. Contributions

In this paper, we present the following contributions

- We introduce a novel iterative explainable federated learning approach, where a constrained traffic drop detection classifier and an *explainer* exchange—in a closed loop way— attributions of the features as well as predictions to achieve a transparent zero-touch service management of 6G network slices at RAN in a non-IID setup.
- We adopt the integrated gradients XAI method to showcase features attributions.
- The generated attributions are then used to quantitatively validate the faithfulness of the explanations via the so-called *log-odds* metric which is included as a constraint in the FL optimization task.
- We formulate the corresponding joint recall and log-odds-constrained FL optimization problem under the *proxy-Lagrangian* framework and solve it via a non-zero sum two-player game strategy [21], while comparing with the unconstrained integrated-gradient post-hoc FL baseline.

II. RAN ARCHITECTURE AND DATASETS

A shown in Fig. 1, we consider a radio access network (RAN), which is composed of a set of K the base station (BSs), wherein a set of N parallel slices are deployed. Each BS runs a local control closed-loop (CL) which collects monitoring data and performs traffic drop prediction. Specifically, the collected data serves to build local datasets for 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 while $y_{k,n}^{(i)}$ represents the corresponding output. In this respect, Table I summarizes the features and the output of the local datasets. These accumulated datasets are non-IID due to the different traffic profiles induced by the heterogeneous

Feature	Description
Average PRB	Average Physical Resource Block
Latency	Average transmission latency
Channed Quality	SNR value expressing the wireless channel quality

Output	Description
Dropped Traffic	Probability of dropped traffic(%)

users' distribution and channel conditions. Moreover, since the collected datasets are generally non-exhaustive to train accurate anomaly detection classifiers, the local CLs take part in a federated learning task wherein an E2E slice-level federation layer plays the role of a model aggregator.

III. EXPLAINABLE FDL FOR TRANSPARENT TRAFFIC DROP CLASSIFICATION

Here, we describe the different stages of the joint explainability and sensitivity-aware FDL as summarized in Fig. 2.

A. Closed-Loop Description

We propose a federated deep learning architecture where the local learning is performed iteratively with run-time explanation in a closed loop way as shown in Fig. 2. We design a deep neural network FL model. 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 $\hat{y}_{k,n}^{(i)}$ to the Explainer. The latter first generates the features attributions using integrated gradients XAI method. The *Log-odds Mapper* then uses these attributions to select the top p features that are then masked. The corresponding soft probability outputs are afterward used to calculate the the log-odds (LO) metric that is fed back to the Learner to include it in the local constrained optimization in step 6. Similarly, the *Recall Mapper* calculate the recall score $\rho_{k,n}$ based on the predicated and true positive values at stage 3 and 4 to include it in the local constrained optimization in step 6. Indeed, for each local CL (k, n), the predicted traffic drop class $\hat{y}_{k,n}^{(i)}$, $(i = 1, \ldots, D_{k,n})$, should minimize the main loss function with respect to the ground truth $y_{k,n}^{(i)}$, while jointly respecting some long-term statistical constraints defined over its $D_{k,n}$ samples and jointly corresponding to recall and explainability log-odds.

As shown in steps 1 and 7 of Fig. 2, 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,

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

Figure 2. Explainable FDL building blocks

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 the K CLs that use it to start the next round of iterative 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.

B. Model Testing and Explanation

As 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 $\hat{\mathbf{y}}_{k,n}^{(i)}$. All the graph, test dataset and the predictions are fed to the Explainer at stage 3. After that, at stage 4, Explainer generates the attributions by leveraging the low-complexity Integrated Gradient (IG) scheme [22], which is based on the gradient variation when sampling the neighborhood of a feature. Attributions are 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.

C. Log-odds Mapping

To characterize the trustworthiness of the local model, we calculate the log-odds metric, $\theta_{k,n}$ [23]. It measures the influence of the top-attributed features on the model's prediction. Specifically, the log-odds score is defined as the average difference of the negative logarithmic probabilities on the predicted class before and after masking the top $p\%$ features with zero padding [23]. In this respect, the *log-odds Mapper* at stage 5 of Fig. 2 starts by selecting top $p\%$ features based on their attributions which is collected from stage 4 and replace them with zero padding. That is,

$$\theta_{k,n} = -\frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \log \frac{\Pr\left(\hat{y}_{k,n}^{(i)}|\hat{\mathbf{x}}_{k,n}^{(i)}\right)}{\Pr\left(\hat{y}_{k,n}^{(i)}|\mathbf{x}_{k,n}^{(i)}\right)}, \quad (2)$$

where, $\hat{y}_{k,n}^{(i)}$ is the predicted class, $\mathbf{x}_{k,n}^{(i)}$ are the features in the original dataset and $\hat{\mathbf{x}}_{k,n}^{(i)}$ denotes the features in the modified dataset with top $p\%$ features zero-padded. Finally, the log-odds Mapper reports the log-odds score, which is used as one of the constraints for the constrained FL optimization task.

D. Joint Recall and Explainability-Aware Traffic Drop Classification

Besides the log-odds score used for explainability, as shown in steps 3 and 4, we invoke the *recall* as a measure of the sensitivity of the FL local classifier, which we denote $\rho_{k,n}$, i.e.,

$$\rho_{k,n} = \pi^+\left(\mathcal{D}_{k,n}\left[\hat{y}_{k,n}^{(i)} = 1\right]\right) \quad (3)$$

Where, $\pi^+(\mathcal{D}_{k,n})$ defines the proportion of $\mathcal{D}_{k,n}$ classified positive, and $\mathcal{D}_{k,n}[*]$ is the subset of $\mathcal{D}_{k,n}$ satisfying expression $*$.

In order to trust the traffic drop anomaly detection/classification, a set of AI SLA is established between the slice tenant and the infrastructure provider, where a lower bound α_n is imposed to the recall score, while an upper bound β_n is set for the log-odds score. This translates into solving a constrained local classification problem in iterations specified by the epochs as well as in FL rounds t ($t = 0, \ldots, T-1$) i.e.,

$$\min_{\mathbf{W}_{k,n}^{(t)}} \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \ell\left(y_{k,n}^{(i)}, \hat{y}_{k,n}^{(i)}\left(\mathbf{W}_{k,n}^{(t)}, \mathbf{x}_{k,n}\right)\right), \quad (4a)$$

$$\text{s.t.} \quad \rho_{k,n} \geq \alpha_n, \quad (4b)$$

$$\theta_{k,n} \leq \beta_n, \quad (4c)$$

which is solved by invoking the so-called *proxy Lagrangian* framework [24], since the recall is not a smooth constraint. This consists first on constructing two Lagrangians as follows:

$$\mathcal{L}_{\mathbf{W}_{k,n}^{(t)}} = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \ell\left(y_{k,n}^{(i)}, \hat{y}_{k,n}^{(i)}\left(\mathbf{W}_{k,n}^{(t)}, \mathbf{x}_{k,n}\right)\right)$$
$$+ \lambda_1 \Psi_1\left(\mathbf{W}_{k,n}^{(t)}\right) + \lambda_2 \Psi_2\left(\mathbf{W}_{k,n}^{(t)}\right), \quad (5a)$$

$$\mathcal{L}_\lambda = \lambda_1 \Phi_1\left(\mathbf{W}_{k,n}^{(t)}\right) + \lambda_2 \Phi_2\left(\mathbf{W}_{k,n}^{(t)}\right) \quad (5b)$$

where $\Phi_{1,2}$ and $\Psi_{1,2}$ represent the original constraints and their smooth surrogates, respectively. In this respect, the recall surrogate is given by,

$$\Psi_1 = \frac{\sum_{i=1}^{D_{k,n}} y_{k,n}^{(i)} \times \min\left\{\hat{y}_{k,n}^{(i)}, 1\right\}}{\sum_{i=1}^{D_{k,n}} y_{k,n}^{(i)}} - \alpha_n \quad (6)$$

while $\Psi_2 = \Phi_2 = \beta_n - \theta_{k,n}$ since the negative logarithm is already a convex function. It also confirms that the solutions of the optimization problem are equivalent to those obtained if only the original constraints were used.

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 $\mathcal{L}_\lambda$ [21, Lemma

Algorithm 1: Explainable Federated Deep Learning

Input: K, m, η_λ, T, L. # See Table II
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 $\left(\mathcal{M}_{k,n}\left(\mathbf{W}_{k,n,l}, \mathbf{x}_{k,n}\right)\right)$
        # Mask the top p% dataset based on the attributions with zero padding
        # Calculate the log-odds metric
        $\theta_{k,n} = \frac{1}{D_{k,n}} \sum_{i=1}^{D_{k,n}} \log \frac{\Pr\left(\hat{y}_{k,n}^{(i)} | \check{\mathbf{x}}_{k,n}^{(i)}\right)}{\Pr\left(\hat{y}_{k,n}^{(i)} | \mathbf{x}_{k,n}^{(i)}\right)}$
        # Calulate the recall metric
        $\rho_{k,n} = \pi^+ \left(\mathcal{D}_{k,n}\left[\hat{y}_{k,n}^{(i)} = 1\right]\right)$
        Let $\lambda^{(l)}$ be the top eigenvector of $\mathbf{A}^{(l)}$
        # Solve problem (4) 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

8]. While optimizing the first Lagrangian w.r.t. $\mathbf{W}_{k,n}$ requires differentiating the constraint functions $\Psi_1(\mathbf{W}_{k,n}^{(t)})$ and $\Psi_2(\mathbf{W}_{k,n}^{(t)})$, to differentiate the second Lagrangian w.r.t. λ we only need to evaluate $\Phi_1\left(\mathbf{W}_{k,n}^{(t)}\right)$ and $\Phi_2\left(\mathbf{W}_{k,n}^{(t)}\right)$. 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,

$$\min_{\mathbf{W}_{k,n}\in\Delta} \max_{\lambda, \|\lambda\|\leq R_\lambda} \mathcal{L}_{\mathbf{W}_{k,n}^{(t)}} \tag{7a}$$

$$\max_{\lambda, \|\lambda\|\leq R_\lambda} \min_{\mathbf{W}_{k,n}\in\Delta} \mathcal{L}_\lambda, \tag{7b}$$

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 [25]. These steps are all summarized in Algorithm 1.

IV. RESULTS

This section analyzes the proposed Closed loop EFL framework in detail. To build the explainability-aware constrained traffic drop classification model, we use feature attributions which is the pillar of this approach. After that, we present the impact of considering jointly the recall and log-odds metrics as constraints for optimizing the FL classification problem by showing results of FL convergence and log-odds score. Finally, we study the correlation between features attributions, observed predictions, and *true* predictions and draw some important conclusions. Specifically, to implement the model Tester and Explainer, we invoke DeepExplain framework, which includes state-of-the-art gradient and perturbation-based attribution methods [26]. It provides an attribution score based on the feature's contribution to the model's output, which we integrate with our proposed constrained traffic drop classification FL framework in a closed-loop iterative way.

A. Parameter Settings and Baseline

Three primary slices eMBB, uRLLC and mMTC are considered to analyze the proposed Explainable FL policy. Here, the datasets are collected from the BSs and the overall summary of those datasets are presented in Table II. We use vector β for the explainability lower bound threshold and α for the upper bound of recall score corresponding to the different slices. As a baseline, we adopt a vanilla FL [27] with post-hoc integrated gradient explanation, that is, a posterior explanation performed upon the end of the FL training.

TABLE II
SETTINGS

Parameter	Description	Value
DNN	Deep neural network size	2-hidden layers with 10 nodes
N	# Slices	3
K	# BSs	50
$D_{k,n}$	Local dataset size	800 samples
T	# Max FL rounds	50
U	# Total users (All BSs)	15000
L	# Local epochs	100
R_λ	Lagrange multiplier radius	Constrained: 10^{-5}
η_λ	Learning rate	0.02

B. Result Analysis

In this scenario, resources allocated to slices according to their traffic patterns and radio conditions while ensuring a long term isolation via the constraints.

- **Convergence:** As depicted in Fig. 3, we can conclude that the proposed constrained EFL resource allocation models of the different slices have converged faster than the baseline unconstrained IG post-hoc case. Here, the optimizer of EFL considers the relationships between the objectives and constraints of the two-player optimization problem, leading to improved performance compared to the uncon. IG post-hoc one, which accounts for only the objective function during optimization.

- **Sensitivity analysis:** To analyze our proposed model's sensitivity, we choose the recall metric, which is the rate of actual positive values for measuring the performance of our binary classification model. From Fig. 4, we can

Figure 3. Analysis of FL training loss vs FL rounds of Proposed EFL with Lower bound of Recall score, $\alpha = [0.9, 0.95, 0.95]$ and Upper bound of log-odds score, $\beta = [-0.01, -0.01, -0.01]$

Figure 4. Analysis of Recall score with Lower bound of Recall score, $\alpha = [0.9, 0.95, 0.95]$ and Upper bound of log-odds score, $\beta = [-0.01, -0.01, -0.01]$

(a) log-odds Score vs. top p %

(b) log-odds Score

Figure 5. Analysis of log-odds score with Lower bound of Recall score, $\alpha = [0.9, 0.95, 0.95]$ and Upper bound of log-odds score, $\beta = [-0.01, -0.01, -0.01]$

observe that the recall score of the proposed one for all slices is in close proximity to the target threshold γ (i.e., around 0.88%), which is an acceptable value for operators and slices' tenants.

- **Trustfulness:** In Fig. 5-(a), we observe the effect of changing the value top $p\%$ on the log-odds, considering proposed model for all slices. Also we present a comparative analysis of log-odds score in Fig.5-(b) for both cases which proof the superiority of the proposed constrained EFL model. So, the statistics of the log-odds score give us an approximate idea of our model's reliability and trustworthiness. It shows that the log-odds score is decreasing with respect to the top $p\%$ value, which conveys that our model is explainable and trustworthy in the training phase.

Furthermore, in Fig. 6, the correlation heatmaps of the proposed XAI method of the eMBB slice has presented for further analysis. It helps us visualize the strength of relationships between different variables and, in our case, identify which feature variation impacts the most for SLA variation. To plot

correlation matrix heatmap, we consider one matrix, $\mathbf{R}_{k,n} = [\mathbf{a}_{k,n}, \hat{\mathbf{y}}_{k,n}, \mathbf{y}_{k,n}]$, where, $\mathbf{a}_{k,n}$ is the attribution score of features variable with dimensions $D_{k,n} \times Q$ and $\hat{\mathbf{z}}_{k,n}$ is the predicted output variable with dimensions $D_{k,n} \times 1$ and $\mathbf{y}_{k,n}$ is the true predicted value with dimensions $D_{k,n} \times 1$. From the heatmap we see that the third feature, which is the channel quality, has the most impact on the recall value. If the third feature increases, the recall value will increase and vice versa.

V. CONCLUSION

This paper has presented a novel closed-loop explainable federated learning (EFL) approach to achieve transparent zero-touch service management of 6G network slices at RAN in a non-IID setup. We have jointly considered explainability and sensitivity metrics as constraints in the traffic drop prediction task, which we have solved using a proxy-Lagrangian two-player game strategy. From the results, we conclude that the proposed EFL scheme is reliable and trustful compared to state-of-the-art unconstrained post-hoc FL. Finally, the heatmaps of the attributions correlation matrix are presented to showcase the features whose variation influence more the traffic drop.

Figure 6. Correlation heatmap of eMBB slices based on attribution scores of features generated by XAI.

VI. ACKNOWLEDGMENT

REFERENCES

[1] C. Marquez, M. Gramaglia, Fiore *et al.*, "How should i slice my network? a multi-service empirical evaluation of resource sharing efficiency," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 191–206. [Online]. Available: https://doi.org/10.1145/3241539.3241567

[2] X. Foukas, G. Patounas *et al.*, "Network slicing in 5g: Survey and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.

[3] (2019) Zero-touch network and service management (zsm); reference architecture. [Online]. Available: https://www.etsi.org/technologies/zero-touch-network-service-management

[4] V. Sciancalepore, F. Z. Yousaf, and X. Costa-Perez, "z-torch: An automated nfv orchestration and monitoring solution," *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1292–1306, 2018.

[5] R. Wen, G. Feng, J. Tang, T. Q. S. Quek, G. Wang, W. Tan, and S. Qin, "On robustness of network slicing for next-generation mobile networks," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 430–444, 2019.

[6] W. Wu, C. Zhou, M. Li, H. Wu *et al.*, "Ai-native network slicing for 6g networks," *IEEE Wireless Communications*, vol. 29, pp. 96–103, 2022.

[7] B. Brik and A. Ksentini, "On predicting service-oriented network slices performances in 5g: A federated learning approach," 11 2020.

[8] F. Rezazadeh, L. Zanzi, F. Devoti, H. Chergui, X. Costa-Pérez, and C. Verikoukis, "On the specialization of fdrl agents for scalable and distributed 6g ran slicing orchestration," *IEEE Transactions on Vehicular Technology*, pp. 1–15, 2022.

[9] S. Roy, H. Chergui, L. Sanabria-Russo, and C. Verikoukis, "A cloud native sla-driven stochastic federated learning policy for 6g zero-touch network slicing," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 4269–4274.

[10] F. Fossati, S. Moretti, and S. Secci, "Multi-resource allocation for network slicing under service level agreements," in *2019 10th International Conference on Networks of the Future (NoF)*, 2019, pp. 48–53.

[11] 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.

[12] 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.

[13] 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.

[14] S. Wang, M. Qureshi, L. Miralles-Pechua'an, T. Huynh-The, T. R. Gadekallu, and M. Liyanage, "Explainable ai for b5g/6g: Technical aspects, use cases, and research challenges," *ArXiv*, vol. abs/2112.04698, 2021.

[15] C. Li, W. Guo, S. C. Sun *et al.*, "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.

[16] A. Terra, R. Inam, Baskaran *et al.*, "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.

[17] Y. Wu, G. Lin, and J. Ge, "Knowledge-powered explainable artificial intelligence for network automation toward 6g," *IEEE Network*, vol. 36, no. 3, pp. 16–23, 2022.

[18] W. Guo, "Explainable artificial intelligence for 6g: Improving trust between human and machine," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 39–45, 2020.

[19] 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.

[20] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *ArXiv*, vol. abs/1812.04608, 2018.

[21] A. Cotter, H. Jiang, and K. Sridharan, "Two-player games for efficient non-convex constrained optimization," *CoRR*, vol. abs/1804.06500, 2018. [Online]. Available: http://arxiv.org/abs/1804.06500

[22] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 3319–3328.

[23] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 3145–3153.

[24] A. Cotter, M. R. Gupta, H. Jiang, N. Srebro *et al.*, "Training well-generalizing classifiers for fairness metrics and other data-dependent constraints," in *ICML*, 2019.

[25] 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. [Online]. Available: https://doi.org/10.1145/1390156.1390202

[26] [Online]. Available: https://github.com/marcoancona/DeepExplain

[27] H. B. McMahan, E. Moore *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *International Conference on Artificial Intelligence and Statistics*, 2016.