Explainable AI in 6G O-RAN: A Tutorial and Survey on Architecture, Use Cases, Challenges, and Future Research

Bouziane Brik, *Senior Member, IEEE*, Hatim Chergui, *Senior Member, IEEE*, Lanfranco Zanzi, *Member, IEEE*, Francesco Devoti, *Member, IEEE*, Adlen Ksentini, *Senior Member, IEEE*, Muhammad Shuaib Siddiqui, Xavier Costa-Pérez, *Senior Member, IEEE*, and Christos Verikoukis, *Senior Member, IEEE*

***Abstract*—The recent O-RAN specifications promote the evolution of RAN architecture by function disaggregation, adoption of open interfaces, and instantiation of a hierarchical closed-loop control architecture managed by RAN Intelligent Controllers (RICs) entities. This paves the road to novel data-driven network management approaches based on programmable logic. Aided by Artificial Intelligence (AI) and Machine Learning (ML), novel solutions targeting traditionally unsolved RAN management issues can be devised. Nevertheless, the adoption of such smart and autonomous systems is limited by the current inability of human operators to understand the decision process of such AI/ML solutions, affecting their trust in such novel tools. eXplainable AI (XAI) aims at solving this issue, enabling human users to better understand and effectively manage the emerging generation of artificially intelligent schemes, reducing the *human-to-machine* barrier. In this survey, we provide a summary of the XAI methods and metrics before studying their deployment over the O-RAN Alliance RAN architecture along with its main building blocks. We then present various use-cases and discuss the automation of XAI pipelines for O-RAN as well as the underlying security aspects. We also review some projects/standards that tackle this area. Finally, we identify different challenges and research directions that may arise from the heavy adoption of AI/ML decision entities in this context, focusing on how XAI can help to interpret, understand, and improve trust in O-RAN operational networks.**

***Index Terms*—6G, AI, ML, O-RAN, Survey, Trust, XAI**

I. Introduction

A. Context and Motivation

6G wireless networks are growing to revolutionize the way we connect, communicate, and share information, catalyzing smart services and innovative applications [1]–[4]. 6G is expected to transform mobile communication networks from the Internet of Things (IoT) to "connected intelligence", by leveraging Artificial Intelligence (AI) techniques and connecting billions of devices and people [5]–[8]. The promise of immense connected devices, ultra-low latency, low energy footprint, and extremely high data rates is expected to enhance the sustainability, connectivity, and trustworthiness of the next-generation mobile network, and support the development of innovative applications, such as truly immersive eXtended Reality (XR), smart grid 2.0, high-fidelity mobile hologram, and Industry 5.0 [9]–[12].

The co-existence of such a variety of applications, along with their specific requirements, demands a versatile mobile network capable of accommodating and guaranteeing the expected performances by means of accurate and smart management of network components and resources [13]–[15] across different technological domains, i.e., Radio Access Network (RAN), core network, cloud, and edge. To this end, both industry and academia are leveraging Network Slicing (NS), Software Defined Network (SDN), and Network Function Virtualization (NFV) paradigms to transform the mobile ecosystem into a more intelligent, energy-efficient, virtual, and software-focused ecosystem [16]–[19].

In this context, a global initiative was formed, consisting of over 200 companies from the telecommunication industry, who collaborated under the umbrella of the Open Radio Access Network (O-RAN) Alliance to introduce a novel RAN architectural design for the forthcoming generation of mobile networks (B5G and 6G) [20] [21]. The core concept of O-RAN revolves around the disaggregation of traditional RAN system functionalities and their conversion into software components, known as Virtual Network Functions Virtual Network Function (VNF), which are interconnected through standardized and open interfaces. Additionally, O-RAN introduces a novel hierarchical RAN Intelligent Controller (RIC) architecture [22], which includes two main building blocks namely Non Real-Time RAN Intelligent Controller (Non RT RIC) [23] and Near Real-Time RAN Intelligent Controller (Near RT RIC) [24], designed to enhance the capabilities and flexibility of the RAN

B. Brik is with the Computer Science Department, College of Computing and Informatics, University of Sharjah, UAE(e-mail: bbrik@sharjah.ac.ae).

H. Chergui and M.-S. Siddiqui are with i2CAT Foundation (e-mail: chergui@ieee.org, shuaib.siddiqui@i2cat.net).

L. Zanzi and F. Devoti are with NEC Laboratories Europe, 69115 Heidelberg, Germany (e-mails: lanfranco.zanzi@neclab.eu, francesco.devoti@neclab.eu).

A. Ksentini is with EURECOM, France (e-mail: adlen.ksentini@eurecom.fr).

X. Costa-Pérez is with NEC Laboratories Europe, i2CAT Foundation, and ICREA, 08034 Barcelona, Spain (e-mail: xavier.costa@neclab.eu).

C. Verikoukis is with the University of Patras, ATHENA/ISI, and IQUADRAT Informatica, Barcelona 08006, Spain (e-mail: cveri@ceid.upatras.gr).

ecosystem. The Non RT RIC is responsible for executing non-time-critical functions and tasks, such as policy management, network optimization, and long-term analytics, while the Near RT RIC focuses on time-critical operations and tasks that require low latency and quick decision-making.

It is easy to claim that AI will play a critical role in the development and implementation of future network management operations, pursuing better network performance, cost savings, and enhanced customer experience [25] [26] [27]. In this context, O-RAN envisions RIC entities to support programmable-based functions and logics, featured by the heavy usage of AI techniques, in particular, Machine Learning (ML) and Deep Learning (DL), to ease the development of intelligent and flexible RAN applications and reduce operational complexity [28]. Among others, the AI-based RICs aim to tackle traditionally hard-to-solve aspects of the RAN domain, such as spectrum management, mobility, radio resource assignment and scheduling, admission control, link management, and power allocation [29], [30]. This is particularly beneficial in the 6G landscape when considering various vertical industries and their corresponding networking requirements. Having said that, the recent European Union (EU)'s AI Act establishes XAI regulation that mandates transparency and human oversight for high-risk AI-driven systems, such as future 6G networks [31]. Moreover, the United States (US) focuses on maintaining global AI competitiveness while fostering trustworthy systems, with initiatives like the National AI Initiative Act [32]. The United Kingdom (UK)'s approach falls between the EU and US models, emphasizing responsible innovation and practical guidance, such as the ICO and Alan Turing Institute's AI decision explanation framework, alongside ambitions for global AI leadership [33], [34].

In this context, the widespread adoption of AI techniques in future 6G O-RAN should be accompanied by mechanisms that verify and explain the black-box models' decisions in a systematic and objective fashion [35], especially when they lead to Service Level Agreement (SLA) violations [36] or failures. This urges designers to clearly identify the operational boundaries of AI/ML models, characterize and understand their behaviour, and prioritize faithful and trustworthy decision-making processes to enable automated network service management while leaving the quality of service unaffected. On that account, new approaches are required to provide explainable and understandable decisions [37]. eXplainable AI (XAI) is an emerging paradigm that aims to shed light on the decision process that is performed by closed (black box) AI models. The main objective of XAI is to create a transparent and human-understandable model (white box) that clarifies the internal processes of AI models, e.g., by determining the contribution of each input feature to an AI decision or prediction [38]. XAI is crucial to demonstrate the accuracy, fairness, and transparency of AI models that drive decisions and operations in the network, thereby instilling trust and confidence in the deployment of AI-

List of Acronyms

3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
6G	Sixth Generation
A2C	Advantage Actor Critic
AI	Artificial Intelligence
ANN	Artificial Neutral Networks
AML	Adversarial Machine Learning
API	Application Programming Interface
B5G	Beyond Fifth-Generation
BBU	Baseband Unit
BS	Base Station
CAPEX	CAPital EXpenditures
CI/CD	Continuous Integration and Delivery
CNN	Convolutional Neural Network
CP	Control Plane
CT	Continuous Training
CU	Central Unit
DAG	Directed Acyclic Graph
dApp	Distributed Application
DARPA	Defense Advanced Research Projects Agency
DevOps	DEVelopment and IT Operations
DeepLIFT	Deep Learning Important FeaTures
DL	Deep Learning
DNN	Deep Neural Network
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
DU	Distributed Unit
eMBB	enhanced Mobile Broadband
eNB	eNodeB
ENI	Experiential Networked Intelligence
ETSI	European Telecommunications Standards Institute
FL	Federated Learning
GAN	Generative Adversarial Network
gNB	gNodeB
GNN	Graph Neural Network
HE	Horizon Europe
IEEE	Institute of Electrical and Electronics Engineers
IG	Integrated Gradients
IoT	Internet of Things
ISG	Industry Specification Group
KL	Kullback-Leibler
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Large Language Model
LO	Log-Odds
LSTM	Long Short Term Memory
MAC	Medium Access Control
MDP	Markov Decision Process
MEC	Multi-access Edge Computing
MLOps	ML system operations
ML	Machine Learning
mMTC	massive Machine Type Communications
MM	Mobility Management
MNO	Mobile Network Operator
MR	Machine Reasoning
MVNO	Mobile Virtual Network Operator
Near RT RIC	Near Real-Time RAN Intelligent Controller
NFV	Network Function Virtualization
NG RAN	New Generation RAN
Non RT RIC	Non Real-Time RAN Intelligent Controller
NR-MAC	New Radio Medium Access Control
NS	Network Slicing
O-Cloud	Open Cloud
O-CU-CP	Open RAN Central Unit Control Plane
O-CU-UP	Open RAN Central Unit User Plane
O-CU	Open RAN Central Unit
O-DU	Open RAN Distributed Unit

O-RAN	Open Radio Access Network
O-RU	Open RAN Radio Unit
OPEX	OPerational EXpenditures
OSC	Open RAN Software Community
PCA	Principal Component Analysis
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Control Protocol
PDSCH	Physical Downlink Shared Channel
PHY	Physical
PNF	Physical Network Function
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
QoE	Quality of Experience
QoS	Quality of Service
QoT	Quality of Transport
R2	R-squared
RAN	Radio Access Network
rApp	Non RT RIC Applicaiton
RAT	Radio Access Technologies
ReCo	Relative Consistency
RIC	RAN Intelligent Controller
RL	Reinforcement Learning
RM	Resource Management
RNN	Recurrent Neural Network
RSS	Received Signal Strenght
RT	Real Time
RU	Radio Unit
SCM	Structural Causal Models
SDN	Software Defined Network
SHAP	SHapley Additive exPlanations
SLA	Service Level Agreement
SMO	Service Management and Orchestration
SM	Spectrum Management
SVM	Support Vector Machines
TTI	Transmission Time Interval
UE	User Equipment
UP	User Plane
URLLC	Ultra Reliable Low Latency Communication
vBBU	Virtual BBU
VNF	Virtual Network Function
vO-CU	virtual Open RAN Central Unit
vO-DU	virtual Open RAN Distributed Unit
vRAN	Virtual RAN
WG	Working Group
XAI	eXplainable AI
xApp	Near RT RIC Application
XR	eXtended Reality

powered components in the O-RAN ecosystem by businesses and organizations [39], [40].

B. Review of Existing Related Surveys

Several studies already addressed the novel O-RAN architecture, highlighting its novel approach and investigating potential benefits and drawbacks. In [42], the authors provided a short review of both advantages and limitations of O-RAN, focusing on the O-RAN architecture and its main modules. The authors conducted a community survey on the benefits of O-RAN among 95 researchers from all around the world. Most of them agreed on the fact that O-RAN will be the foundation of next-generation networks. Finally, the authors discussed the benefits, current shortcomings, and future research directions of O-RAN. Similarly, [43] described the O-RAN architecture and its key concepts. In addition, the authors present a novel DL-based scheme for radio resource assignment, validating their performance using data collected from real mobile network deployments. The authors conclude their work by discussing open challenges and future research opportunities. Another review study is provided by [44]. The authors showcase how a DL-based scenario can be deployed on top of the O-RAN architecture, highlighting the main advantages and shortcomings of O-RAN.

The evolution of RAN architectures towards the O-RAN proposal both in terms of functionality and implementation is discussed in [45] [46]. In the same context, the support of B5G key concepts, such as network slicing and MEC, by the O-RAN architecture is elaborated by [47] [55] [56] [57].
In our previous work [48], we proposed a survey study on the O-RAN architecture, discussing the evolution of RAN architectures, and comparing different studies based on various perspectives. We focused our review on existing AI-based schemes dealing with the RAN challenges, and show how these schemes can be supported by O-RAN by considering the deployment of two realistic DL-based case studies. Similarly, in [41], the authors provided a tutorial on the O-RAN framework describing recent specifications in terms of architecture, design, and open interfaces. They also discuss the main open research challenges and the new innovation possibilities in the O-RAN architecture, focusing on AI and deep learning.

Besides, the XAI topic is attracting interest from research and industry domains. Currently, XAI is one of the main programs of the Defense Advanced Research Projects Agency (DARPA), expected to design efficiently the "third-wave AI systems" [58]. In [49] [50], the authors reviewed and analyzed several XAI approaches focusing on algorithmic aspects, classifications, and application domains, identifying several still open challenges and key future research directions. The main principles and practices of XAI are summarized in [40]. In particular, the authors target the specific pattern recognition models of machine learning in order to enhance the understanding of such models for industry practitioners (data scientists). In [51], the authors discussed a set of key measurement metrics that can help evaluate explainable AI systems. In 6G networks context, the authors of [38] discussed the use of XAI, targeting different 6G use cases (e.g., Industry 5.0). Similarly, in [53] the authors highlight existing tools in addition to their use to deal with 6G network challenges, discussing how to integrate XAI into 6G networks architecture through a real mobile traffic prediction use-case, and validating their findings on realistic traffic data. Conversely, the authors of [52] focused on XAI methods in low protocol layers of mobile networks, e.g., Physical (PHY) and Medium Access Control (MAC). In the same context, the authors of [54] describe the application of XAI related to security aspects, discussing how XAI can improve the interpretation of AI-based models for a wide range of security

TABLE I: Existing surveys on O-RAN, XAI, XAI for B5G. **H: High**, **M: Medium**, and **L: Low**.

Works	AI	XAI	B5G/6G support	O-RAN Architecture	O-RAN use cases	Future research directions (B5G, O-RAN, or XAI)	Contribution
[41]	L	L	H	H	L	H	A tutorial on O-RAN framework, by describing recent specifications in terms of architecture, design, and open interfaces.
[42]	L	L	H	H	L	H	A short survey on O-RAN's architecture, benefits, shortcomings, and future directions.
[43]	M	L	H	H	H	H	A concise paper on O-RAN architecture. It designed a DL-based resource allocation scheme and discussed future directions.
[44]	M	L	H	H	H	H	A short survey on O-RAN's architecture, benefits, and future directions. It showed the deployment of DL-based scenarios in O-RAN.
[45] [46]	L	L	H	H	L	L	Short review papers that discussed the evolution of RAN architectures towards O-RAN in terms of functionality and implementation.
[47]	L	L	H	M	L	L	A concise paper discussed the integration of emergent B5G concepts with O-RAN, such as network slicing and Multi-access Edge Computing (MEC).
[48]	H	L	H	H	H	H	A survey on DL/ML-based solutions for RAN/O-RAN. It includes O-RAN architecture description along with its use cases as well as future directions and open challenges.
[49] [50]	H	H	M	L	L	H	A review of XAI approaches, in terms of their algorithmic aspects, classifications, application domains, and future research directions.
[40]	H	H	M	L	L	H	A review of the main principles and practice of XAI. In particular, the specific pattern recognition models of machine learning are targeted.
[51]	H	H	M	L	L	H	A review on a set of key measurement metrics, which can help to measure and evaluate any explainable AI system.
[38]	H	H	H	L	L	M	A survey on the use of XAI for B5G/6G networks. It addresses how to design XAI systems for B5G use cases, as well as future research directions in such context.
[52]	H	H	H	L	L	L	A review of DL-based solutions in PHY and MAC layers and their performance vs XAI trade-off.
[53]	H	H	H	L	L	L	A review of existing XAI techniques and their applicability to deal with the 6G network challenges.
[54]	H	H	H	L	L	M	A survey on the application of XAI to the security aspects of B5G networks as well as future research directions.
This survey	**H**	**H**	**H**	**H**	**H**	**H**	**A comprehensive survey on the use of XAI to design transparent and trustworthy O-RAN architecture, covering architectural aspects, use cases, projects, standardization approaches, and future research directions.**

use-cases related to B5G/6G networks.

Table I summarizes the main topics discussed along the above works, and compares their contributions with respect to our work, in order to provide an easy understanding of the differentiation features with respect to the state-of-the-art. Despite the presence of several survey papers discussing XAI and O-RAN, there is a lack of comprehensive surveys jointly investigating XAI and O-RAN aspects able to effectively explore the potential of XAI for developing responsible, trustworthy, and transparent AI-powered O-RAN architecture. In addition, although the integration of XAI with B5G networks has been addressed e.g., in [38] [52], such studies do not focus either on the RAN part or consider the novel O-RAN architecture. Therefore, a comprehensive survey of XAI and its potential in designing the future O-RAN is greatly needed to guide the practitioners as well as researchers.

C. Main Contributions

The contributions of this paper can be summarized as follows:

- *Bridging the gap between O-RAN and XAI:* Existing surveys on O-RAN focused on its enabling technologies, such as hierarchical RAN Intelligent Controller, open interfaces, and programmable functions. To the best of our knowledge, there is no survey addressing the potential of human and O-RAN interactions, through XAI systems. Similarly, existing surveys on XAI targeted different XAI approaches and their taxonomies, and more recently their applications to B5G/6G networks. However, discussions on the potential of XAI for O-RAN are still missing. Therefore, this survey paper aims to bridge this gap by jointly exploring the key benefits of the introduction of XAI to O-RAN.

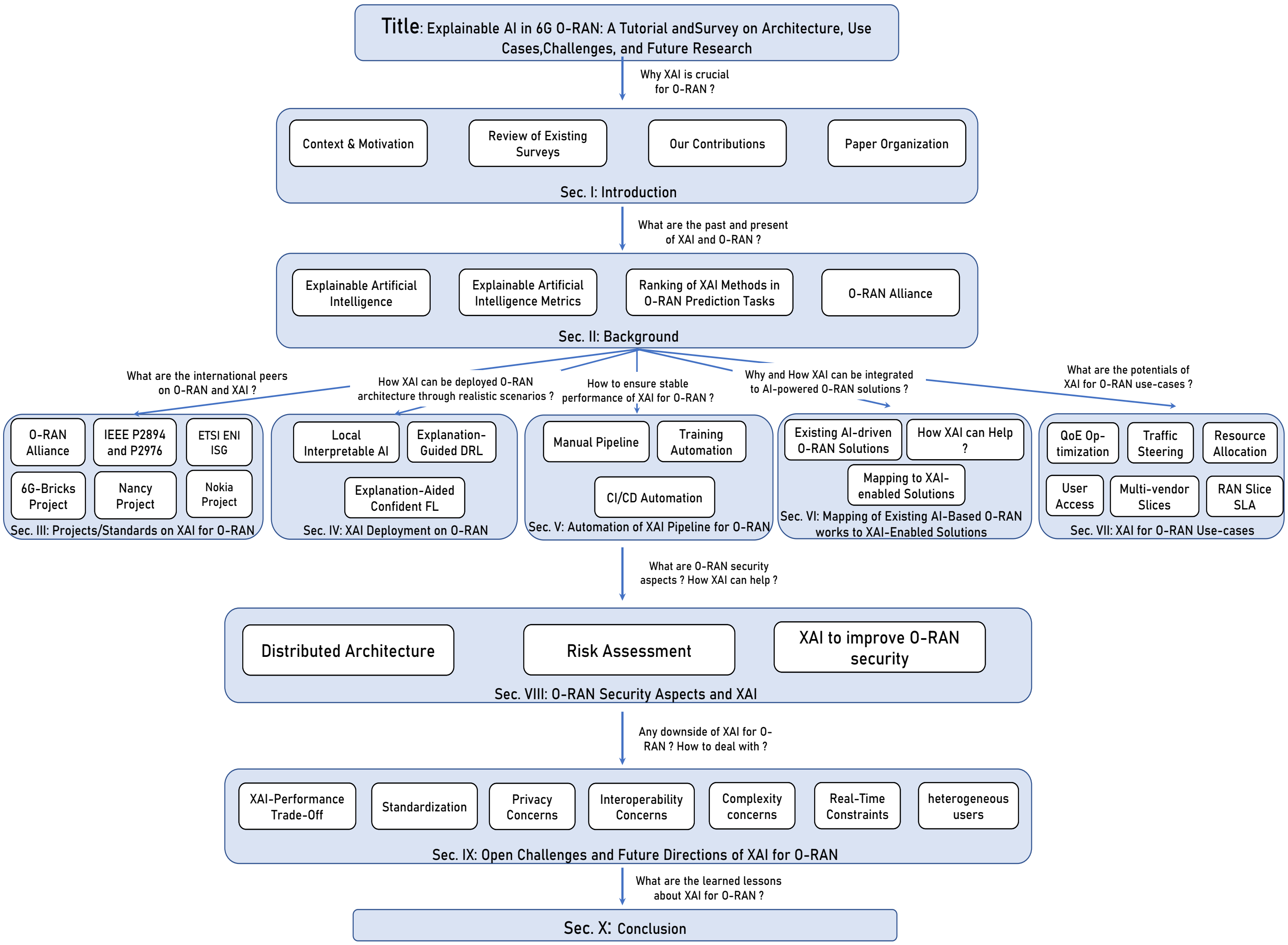


Fig. 1: The taxonomy of the article.

- *A comprehensive survey of XAI deployment on top of O-RAN:* Existing works studied both O-RAN and XAI separately, i.e. no work has combined both paradigms in its study. Hence, in this survey paper, we study the promising deployment of XAI on top of the AI-enabled O-RAN. This includes O-RAN architecture as well as O-RAN use cases. Furthermore, We study the mapping of existing O-RAN research solutions to XAI-supported solutions.
- *A depth analysis of XAI automation for O-RAN:* We provide an exhaustive analysis of how to automate the whole XAI pipeline on top of O-RAN, in order to ensure stable performance of deployed XAI techniques. To the best of our knowledge, no existing work has discussed the automation of XAI process for O-RAN. We also design new architectures showing the automated deployment of XAI, for different levels of automation.
- *O-RAN Security aspects and XAI:* We present the key findings of the official security risk assessments conducted on the O-RAN environment. We explore the potential of XAI to significantly improve the security layer of O-RAN, and how it could be used to build interpretable security threat detection mechanisms. Additionally, we discuss how XAI can help establish trust among stakeholders.
- *Identifying New XAI-related Issues and Promising Research Directions:* Integrating XAI with O-RAN will rise new issues, which should also be considered in future research studies. Thus, we exhaustively discuss new open challenges, along with future research directions.

D. Paper Organization

As shown in Fig. 1, the survey is organized as follows. Section. II provides background information related to the

topics considered in the survey.
Section. III presents the main ongoing projects and standards that are working to promote the adoption of AI/ML techniques in O-RAN, and show how they can be enhanced by XAI. Section. IV describes how XAI methods and models can be deployed on top of the O-RAN architecture, considering four realistic deployment scenarios, and communication interfaces. Section. V details an automation pipeline for XAI model training and deployment in the O-RAN context, involving multiple architectural components and communication interfaces.
Section VI gives a literature review of existing recent works, which leverage XAI techniques for the 6G O-RAN architecture, while Section. VII gives a literature review of existing related works in the field, focusing on AI techniques targeting RAN optimization and highlighting how these works can be mapped to XAI-enabled solutions to optimize multiple performances. Section. VIII provides an overview of O-RAN use cases taken from the literature and standard documentation, highlighting how XAI could bring benefits to the considered scenarios.
Section. IX provides an overview of security issues related to the O-RAN architecture, focusing on XAI-related aspects. Section. X highlights and discusses still open challenges along with their future research directions to deal with them. Finally, section. XI concludes this paper. Note that the used acronyms in this paper are described in the *List of Acronyms*, in alphabetical order, for ease of reference.

II. Background

This section provides background information on XAI and O-RAN topics which are required to fully understand the potential of XAI techniques in the O-RAN domain. Firstly, we describe the main concepts, techniques, and emergent applications of XAI. Secondly, we present the O-RAN architecture along with its main modules as designed by the O-RAN Alliance.

A. eXplainable AI (XAI)

In this subsection, we provide the background on XAI and its main concepts, applications, and ongoing studies.

1) Definitions and Key Concepts:

Definition II.1 (XAI)**.** XAI comprises methods and tools that help interpret, understand, and trust AI-based model results [40], [49], using objective metrics (see Subsection II-B). These tools assist in identifying and mitigating biases by revealing which features significantly impact predictions. This understanding promotes fairness and guides developers in refining algorithms, data, and features to improve model performance.

In other words, XAI aims to build a white-box model that provides insights into the inner workings of underlying ML/AI black-box models. This helps characterize model fairness, accuracy, and transparency in AI-enabled decisions, which is vital for businesses and organizations to have confidence and trust when deploying AI models [40]. More specifically, XAI leverages concepts of *explainability* and *interpretability* to expose information about the internal mechanisms of AI models.

Definition II.2 (Explainability and Interpretability)**.** *Explainability* refers to the extent to which the internal mechanisms of a machine learning model can be understood in human terms. It involves providing insights into how a model makes decisions, often through tools and methods that elucidate the model's logic. *Interpretability*, on the other hand, is the degree to which a human can consistently predict the model's output given a set of inputs. The key difference is that explainability focuses on the underlying workings and reasons behind a model's decisions, whereas interpretability emphasizes the ability to understand and predict a model's behavior based on its input-output relationship.

XAI enables to identify which features influence model predictions the most, shedding light on potential biases encoded in the data or model architecture.

Definition II.3 (Bias in telecom operations)**.** Bias in Deep Neural Network (DNN) predictions refers to systematic errors that lead to unfair outcomes for certain groups. This might be caused by i) *Unbalanced and Biased Data*, where the training dataset is not representative of the diverse network states, e.g., having more SLA violation samples leads to false alarms predictions; and ii) *Inappropriate Feature Selection*, where using features that encode sensitive attributes or their proxies can make a model inclined to allocate more resources to a service regardless of e.g., the traffic.

XAI models incorporate the so-called *explanation user interface* to generate a user-understandable explanation and/or interpretation of the rationale behind decisions taken by the model. Most AI models can be translated into an equivalent XAI counterpart, at the expense of integrating additional layers supporting the explanation user interface on top of the deployed model. Based on the design of the explanation user interface, the XAI model can provide both explainability and interpretability or only one, depending on the target human user [49].

2) Taxonomy of XAI Techniques, Applications, and Stakeholders: There are several existing taxonomies in the XAI realm, which can complement and/or overlap each other. Table II describes an XAI taxonomy that is mainly inspired by [40], [49], and is based on the following three main criteria:

- *Model Transparency:* XAI models can be classified based on the target ML models' transparency. In this regard, models are classified as interpretable or complex. Interpretable models are by themselves understandable for

TABLE II: XAI Taxonomy. A: Agnostic, S: Specific.

Transparency	Explain. Basis	Technique	Algorithm	Agno.	Pros and Cons	Reference
Black-Box Models	Attributions	Gradient	Saliency Maps	A	Pros: Simplicity, visual interpretability, widely applicable. Cons: Lack of context, sensitivity to input perturbations, limited to input gradients.	[59], [60]
			Gradient x Input	A	Pros: Simplicity, direct relevance, feature importance ranking. Cons: Input scaling sensitivity, limited to linear relationships, potential for misleading interpretations.	[61], [62]
			Integrated Gradients	A	Pros: Baseline comparison, path-based attribution, completeness, and sensitivity. Cons: Computationally intensive, baseline selection challenge, linearity assumption.	[63], [64]
			Smooth Gradient	A	Pros: Noise reduction, robustness to adversarial examples, gradient visualization. Cons: Interpretation challenges, hyperparameter sensitivity, computational overhead.	[65], [66]
			Epsilon-LRP	S	Pros: Deep model interpretability, conceptual clarity, attribution preservation. Cons: Complexity, parameter tuning, vulnerability to network architecture.	[67], [68]
		Perturbation	SHAP	A	Pros: Theoretical grounding based on game theory, global and local interpretability, and consistency. Cons: Computational complexity, high-dimensional data challenge, model approximation dependency.	[69], [70]
			DeepLIFT	A	Pros: Model-agnostic, captures interactions, relevance conservation. Cons: Computational overhead, baseline selection challenge, interpretation complexity.	[71]
			Occlusion	A	Pros: Intuitive visual interpretation, robustness to model architecture, spatial localization. Cons: Computational expense, coarseness of occlusion, interpretation subjectivity.	[72], [73]
		Importance Weights	GNNExplainer	A	Pros: Graph-specific interpretability, node and edge importance, feature relevance analysis. Cons: Complexity, model-specific, interpretation scalability.	[74]
	Surrogates	Local Techniques	LIME	A	Pros: Model-agnostic, local interpretability, simplicity. Cons: Interpretability limitation, instability, assumes linearity.	[75], [76]
		Global Techniques	TREPAN	S	Pros: Decision tree interpretability, human-readable explanations, transparent model behavior. Cons: Limited to decision tree models, model-specific, interpretation scalability.	[77]
		Rule-Based	RuleFit	S	Pros: combines decision trees and linear regression to provide interpretable insights into the model's decision-making process. Cons: may struggle to model highly intricate or complex nonlinear patterns in the data.	[78], [79]
	RL Reward	Rule-based	Reward Shaping	S	Pros: Provides explicit guidance to the RL/DRL agent, allowing it to focus on desired behaviors. Cons: Can introduce biases if the reward shaping is not carefully designed.	[80], [81]
	RL State	Model-based	Attention Mechanisms	S	Pros: Offers transparency by showing which parts of the input state the RL/DRL agent attends to. Cons: Attention mechanisms do not explicitly explain the agent's internal reasoning or decision-making process.	[82], [83]
	Symbolic	Machine Reasoning		S	Pros: Provides human-interpretable explanations for model decisions and shows explicit reasoning behind decisions, enhancing transparency. Cons: Requires expertise, computationally expensive, and may struggle with uncertain or probabilistic information.	[84], [85]
	Transformers' Attention Head	Attention Flow Analysis		S	Pros: Provides interpretability, enables fine-grained analysis, helps improve models, and offers domain-specific insights. Cons: Complexity, lack of unique interpretations, limited context, challenges in generalization.	[86], [87]
	Visual	SCM		S	Pros: Causal understanding, intuitive visualization, identifying confounding variables. Cons: Limited to causal modeling, simplified representation, expert knowledge required.	[88], [89]
	Text	Caption Generation		S	Pros: Contextual understanding, language comprehension, multimodal interpretation. Cons: Subjectivity and ambiguity, lack of fine-grained control, reliance on training data.	[90], [91]
	Graph	Knowledge Graphs		A	Pros: Structured representation, relationship understanding, integration, and interoperability. Cons: Knowledge acquisition and maintenance, incompleteness and accuracy, limited context and ambiguity.	[92], [93]
Transparent Models	Logistic / Linear Regression			S	Pros: Explainability, trust and accountability, debugging and error analysis. Cons: Performance limitations, vulnerability to adversarial attacks.	[94]
	Decision Trees					[95]
	K-Nearest Neighbors					[96]
	Rule-Based Learners					[97]
	Generative Additive Models					[98]
	Bayesian Models					[99]
	Self-Explainable Neural Networks					[100]

human users. In other words, such models are able to provide the rationale behind their decisions in an interpretable way to users [49]. Several proposed works succeeded in interpreting some relatively low-complex ML models, including logistic/linear regression, decision trees, K-Nearest neighbors, rule-based learners, etc. [49]. On the other hand, more complex models such as deep neural networks, in order to be interpretable, have to be approximated by generating simpler surrogate models that ease the explanation task by means of a technique known as *post-hoc explainability* [101]. The model complexity is a widely considered aspect in the literature related to XAI and is generally adopted to classify XAI approaches [49].

- *Model Agnosticity:* This criterion targets complex ML/DL models, where XAI models can be categorized based on the nature of their target explanations [49] [40]. In the paradigm of model-agnostic interpretability, the model is regarded as an opaque entity. This conceptualization dissociates interpretability from the specific characteristics and inner workings of the model, thereby liberating the model to exhibit maximum flexibility tailored to the requirements of the task at hand. This approach facilitates the utilization of diverse machine learning methodologies, encompassing even intricate deep neural networks. Furthermore, it affords the opportunity to manage the delicate balance between model complexity and interpretability, a crucial consideration delineated in the subsequent section. Importantly, this methodology allows for graceful handling of situations where achieving an interpretable explanation proves unattainable. Techniques such as SHAP and feature importance scores derived from permutation importance fall into this category. They work by analyzing the input-output relationship of the model without relying on its internal structure.
- *Explainability Methods:* When ML/DL models are considered complex models, some techniques should be devised and used to interpret such models. Thus, XAI models rely on several explanation types, to describe how these ML/DL models output their predictions for any input data.
  - Explanations by simplification refer to the techniques that simplify a complex model and approximate it to an interpretable model, which is easier to explain [75].
  - Feature relevance explanations study and quantify the impact of each input data, to explain a given ML model's prediction [102].
  - Local explanations focus on a single or particular prediction (output) of ML models to generate explanations [75].
  - Visual explanations aim to generate explanations in a visual way, describing the inner functioning of ML/DL models [88]. For instance, they could reveal which set of pixels is the most relevant to recognize content in image classification tasks. Visual explanations rely on several tools, e.g., graphs, heatmaps, scatter plots, etc.
  - Text explanations generate symbol interpretations of learning models using, for example, natural language text to explain their results [90]. For instance, they could be used to highlight which words (or forms) are leveraged in automatic email spam filtering.

Based on the above taxonomy criteria, several XAI approaches have been proposed in the literature. In what follows, we present the most popular ones, highlighting their main features:

- *SHapley Additive exPlanations (SHAP):* This approach relies on feature relevance explanation to interpret a particular prediction of supervised ML/DL models [111]. It computes an additive feature importance score with respect to a set of required properties (e.g., accuracy, consistency, and missingness). Hence, SHapley Additive exPlanations (SHAP) determines feature influence by applying the Shapley values method, which enables estimating the marginal contribution of one feature over the final reward function. In addition, combining several predictions can also be considered to build a global explanation. Several variants of SHAP have been proposed in the literature in order to optimize its computational complexity, such as DeepSHAP [111] and TreeSHAP [112].
- *Deep Learning Important FeaTures (DeepLIFT) [71]:* The purpose of Deep Learning Important FeaTures (DeepLIFT) is to clarify the output of a neural network by calculating the significance of each input feature to the output. This is accomplished by comparing the activation of each neuron in the network for a particular input to the activation that would have been obtained if a reference input had been used. The difference in the activations between the input and the reference is measured by DeepLIFT to compute the contribution of each input feature to the output. The contribution score obtained can be utilized to comprehend how the network reached its conclusion and to identify the most relevant input features. DeepLIFT has been effective in explaining the behavior of different neural network models, such as convolutional neural networks and recurrent neural networks, and has been applied to various fields, including drug discovery, image classification, and speech recognition.
- *Local Interpretable Model-Agnostic Explanations (LIME):* It is one of the most known solutions, that relies on local and simplification explanations, to explain supervised ML/DL models [75]. LIME is a model-agnostic approach targeting different types of data, e.g., tabular, text, graphs, and images. Local Interpretable Model-Agnostic Explanations (LIME) aims to approximate the learning models by developing locally linear models, which replace the black-box models to explain their individual predictions.

TABLE III: XAI Users

XAI Users	Needs	Key Application Areas	Reference
Data Scientists and Machine Learning Researchers	They require XAI techniques to understand and debug complex models, identify biases, and improve model performance	Model development, debugging, and optimization across various domains such as telecommunications, healthcare, finance, natural language processing, and computer vision	[75], [103]
End Users and Consumers	They need explanations to trust and understand AI systems in applications like recommender systems, personalized marketing, and decision support tools	E-commerce and personalized recommendation systems, healthcare decision support tools, financial advice platforms, and autonomous vehicles	[104], [105]
Managers and Decision Makers	They require transparent and interpretable AI models to make informed decisions, assess risks, and gain insights into the AI system's behavior	Business intelligence and analytics, risk assessment and management, fraud detection, and regulatory compliance across industries such as finance, healthcare, and manufacturing	[106]
Developers and Engineers	They need tools and methods to troubleshoot networks faults/SLA violations, build explainable AI systems, ensure reliability, and meet regulatory requirements	Building interpretable machine learning models, developing explainable AI frameworks and libraries, ensuring model reliability and security in domains like cybersecurity, telecommunications, and autonomous systems	[37], [107]
Auditors and Compliance Officers	They require XAI to assess the fairness, accountability, and compliance of AI systems and to identify potential biases or risks	Assessing the fairness and legality of AI systems in finance, hiring practices, loan approvals, credit scoring, and regulatory compliance in sectors such as finance and human resources	[108]
Legal Professionals and Judges	They need explanations to understand AI decisions, assess legal implications, and ensure transparency and fairness in legal proceedings	Interpreting AI-driven legal decisions, evaluating algorithmic fairness, ensuring transparency and accountability in legal proceedings, and addressing ethical concerns in areas like criminal justice and civil rights	[109]
Regulators and Policy Makers	They require XAI to establish guidelines, standards, and regulations around AI ethics, transparency, and accountability	Establishing guidelines, standards, and regulations for trustworthy AI in sectors including healthcare, finance, autonomous systems, and data privacy to protect public interests and ensure ethical AI deployment	[110]

- *Rulefit:* It integrates the benefits of decision trees and linear models. It first consists on the creation of a wide array of rules from an ensemble of decision trees, which capture intricate, non-linear patterns in the data. These rules are then utilized as features in a sparse linear model, combining high predictive performance with clear interpretability [113].
- *Integrated Gradients (IG):* also known as Path-Integrated Gradients or Axiomatic Attribution for Deep Networks. Integrated Gradients (IG) is an XAI technique that gives an importance value to each feature of the input using the gradients of the model output [114]. Specifically, it is a local method that consists of accumulating the gradients by sampling points at a uniform spacing along a straight line between the input and the baseline. This procedure avoids getting null gradients when, e.g., the deep learning model is flat in the proximity of the input feature. This method yields the specific positive or negative attributions of the input features.
- *Graph Neural Network (GNN) Explainer:* It is a technique that explains the predictions of Graph Neural Networks (GNNs) for graph-structured data. It identifies the most important nodes and edges contributing to the output by generating explanation vectors using an additional neural network. This generates an attention map that shows the relative importance of each node and edge. GNN Explainer can be applied to various GNN architectures and input graphs, without requiring changes to the model or training data. It is useful for understanding how GNNs perform predictions and identifying potential issues [74].
- *Reward Shaping:* It entails altering the reward function of the agent to offer supplementary feedback or incentives. This adjustment assists in steering the agent's learning process by molding the reward signal [80].
- *Attention Mechanism:* It enhances interpretability by identifying and highlighting the crucial elements in the input that significantly impact the decision-making process of the agent. They shed light on the specific features that capture the agent's attention and influence its decision [82].
- *Machine Reasoning (MR):* It utilizes logical reasoning and inference techniques to offer insights into the decision-making process of AI models, thereby improving transparency and trust. It generates explanations that are easily comprehensible to humans, fostering a deeper understanding and acceptance of AI systems. Nevertheless, applying machine reasoning in XAI necessitates expertise in logic and reasoning, and it may encounter difficulties when dealing with uncertain or probabilistic information. Nonetheless, the incorporation of machine reasoning in XAI contributes to the advancement of interpretable and accountable AI systems [84].
- *Attention Flow Analysis:* It assesses the individual contribution of attention heads in the encoder to the overall performance of the transformer's model. Specifically, it examines the roles played by these attention heads, with a

particular focus on the most important and confident ones. These heads often exhibit consistent and linguistically interpretable roles, providing valuable insights into the model's decision-making process [86].

- *Structural Causal Models (SCM):* It is another method that targets reinforcement learning models, aiming to show the causal link between the data variables. In [89], the authors leverage Structural Causal Models (SCM) method to explain the behavior of the reinforcement learning model. They are based on visual explanations through a Directed Acyclic Graph (DAG), where the nodes and edges reflect the model states and actions, respectively. By exploring the DAG, it can be extracted which actions take to move from one state to another. Once DAG is created, regression models are built to approximate the relationships using the minimum number of variables. Then, analyzing the DAG's variables will help in generating the explanations, in order to answer the question: "Why action X and not Y ?".
- *Caption generation:* It is a class of methods that aims to generate text interpretations to explain the outputs of DL models. In [91], the authors combined a Convolutional Neural Network (CNN) model and a bidirectional Long Short Term Memory (LSTM) encoder/decoder model. The LSTM encoder helps to extract video features, which are then used by the LSTM decoder to generate textual video captions.
- *Knowledge Graphs:* To produce human-understandable explanations, it is necessary to represent ideas in terms of concepts rather than numeric values. Concepts and the connection between them make what is called *knowledge graph*. It is a powerful way of representing data because Knowledge Graphs can be built automatically and can then be explored to reveal new insights about the domain, especially to find inferred concepts that were not asserted, along with being able to trace back all the steps, making it fully explainable [93].

As anticipated before, the selection of suitable explainability methods depends both on the complexity of the targeted model to be explained and on the target audience. Indeed, the type of explanation exposed and their level of detail depend mainly on the people who are getting such information. In this context, different user profiles may be targeted by XAI models, and XAI models' explanations should differ from one user to another [49]. Table. III illustrates the different objectives of XAI explainability, expected by different user profiles. For instance, users of the models look at trusting as well as understanding how the model works, while users affected by models' decisions aim to understand their decisions and the main reasons for conducting such decisions. Besides, developers and data scientists expect explanations related to the AI models' performance, in order to optimize them over time. However, both regulatory and manager users aim to get

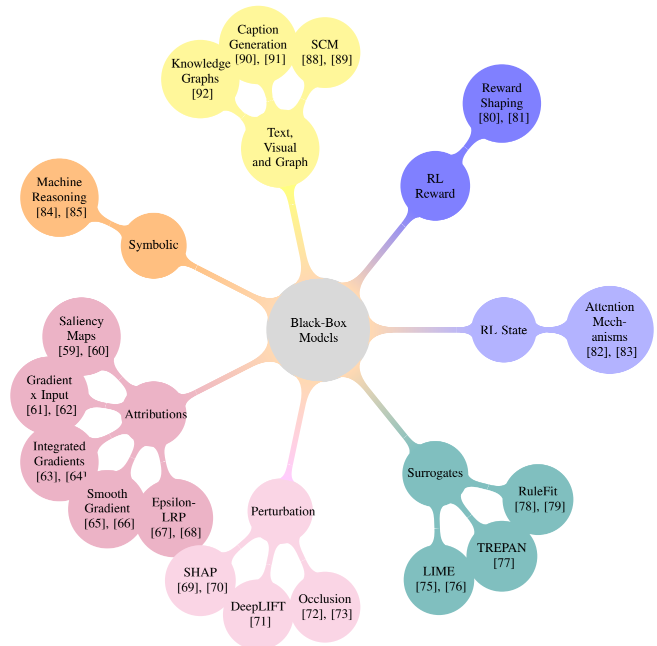


Fig. 2: Black-Box Models Mindmap.

more details related to the compliance of AI models with the legislation in force to check and assess them.

B. XAI Metrics

While human-in-the-loop (HITL) approaches can only yield subjective assessment of the trustworthiness of AI, the existence of objective metrics to characterize the transparency of AI models is a requirement to develop explanation-aware AI systems that exploit such XAI metrics through a feedback loop to assess the confidence of the models in run-time. We summarize relevant XAI metrics in Table IV, and compile a list of them in the following.

- *Confidence/Faithfulness:* A common approach to measuring the confidence of the explanation relies on the notion of feature relevance. Specifically, observing the effect of muting, i.e., replacing a feature with a baseline value—generally zero—helps to measure the effect on the prediction in both classification and regression tasks [115]. For instance, for a probabilistic classification model, we can obscure or remove features according to a policy defined as follows

$$\hat{x}_{i,k} = x_{i,k} \times (1 - p), \quad (1)$$

where p is a Bernoulli random variable, $p \sim \mathcal{B}(1, \pi_{i,k})$ and $\pi_{i,k}$ is a probability distribution of the features that can be computed as

$$\pi_{i,k} = \frac{\exp\left\{|a_{i,k}/x_{i,k}|\right\}}{\sum_{l=1}^{N} \exp\left\{|a_{l,k}/x_{l,k}|\right\}},\ i = 1, \ldots, N, \quad (2)$$

TABLE IV: Taxonomy of XAI metrics.

XAI Method	Basis	Metric	Type of Problem	Reference
Attributions-based	Features Mutation/Masking	Confidence/Faithfulness	Classification	[115]
		Log-odds		[116]
		Comprehensiveness		[117]
		Sufficiency		[117]
	Raw features	Interpretability	Regression	[118]
		Ambiguity	Prediction/Decision	[119], [120]
Surrogates-based	Perturbation	Robustness/Sensitivity	Regression/Decision	[100]
		(in)fidelity		[121], [122]
	LIME Explainer R2 Score		Classification and Regression	[75]
	Relative Consistency			[123]
White-box baseline	Explainer Recall		Classification	[124]
	Explainer Precision			[125]

where N is the number of features and $a_{i,k}$ is the attribution of feature i in a sample of class k. It is obtained using any attribution-based XAI method, such as IG or SHAP. The confidence score in this case is

$$c_k = \frac{\Delta_k^{(c)}}{\Delta_k}, \tag{3}$$

where $\Delta_k^{(c)}$ is the number of samples that conserve their class k after the mutation of the dataset and Δ_k stands for the original count of samples with class label k. For regression tasks, however, the classes are replaced with the notion of groups, which are defined by comparing the continuous prediction output with one or several thresholds.

- *Log-Odds (LO):* Similarly to the confidence, this 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 [116]. Given the attribution scores generated by an explanation algorithm, we select the top $p\%$ features based on their attributions and replace them with zero padding. More concretely, for a dataset with L samples, it is defined as:

$$\text{log-odds}(p) = -\frac{1}{L}\sum_{i=1}^{L} \log \frac{\Pr\left(\hat{y}|\mathbf{x}_i^{(p)}\right)}{\Pr\left(\hat{y}|\mathbf{x}_i\right)} \tag{4}$$

where $\hat{y}$ is the predicted class, $\mathbf{x}_i$ is the ith sample, and $\mathbf{x}_i^{(p)}$ is the modified samples with top $p\%$ features replaced with zero padding. Lower scores are better.

- *Comprehensiveness:* is the average difference of the change in predicted class probability before and after removing the top $p\%$ features. Similar to Log-odds, this measures the influence of the top-attributed words on the model's prediction. It is defined as [117]:

$$\text{Comp}(p) = \frac{1}{L}\sum_{i=1}^{L}\left[\Pr\left(\hat{y}|\mathbf{x}_i^{(p)}\right) - \Pr\left(\hat{y}|\mathbf{x}_i\right)\right] \tag{5}$$

Here $\mathbf{x}_i^{(p)}$ denotes the modified dataset with top $p\%$ samples deleted. Higher scores are better.

- *Sufficiency:* is defined as the average difference of the change in predicted class probability before and after keeping only the top $p\%$ features. This measures the adequacy of the top $p\%$ attributions for the model's prediction. Its definition follows the one of comprehensiveness, except for the fact that the $x_i^{(p)}$ is defined as the samples containing only the top $p\%$ features. Lower scores are better [117].
- *Robustness/Sensitivity:* A crucial property that interpretability methods should satisfy to generate meaningful explanations is that of robustness with respect to local perturbations of the input. This is not the case for popular interpretability methods; even adding minimal white noise to the input introduces visible changes in the explanations [100]. To formally quantify the stability of an explanation generation model, one can estimate the Lipschitz constant λ for a given input x_i and a neighborhood B_ϵ of size ϵ as,

$$\lambda(x_i) = \underset{x_j \in B_\epsilon(x_i)}{\arg\max} \frac{\|\Phi(x_i) - \Phi(x_j)\|_2}{\|x_i - x_j\|_2}, \tag{6}$$

where the evaluation of the explaining function Φ for methods like LIME and SHAP is expensive as it involves model estimation for each query. In contrast, gradient-based attribution methods present a lower complexity. On the other hand, computing (6) for post-hoc explanation frameworks is much more challenging, since they are not end-to-end differentiable. Thus, one needs to rely on black-box optimization instead of gradient ascent. This continuous notion of local stability in (6) might be inadequate for discrete inputs or settings where adversarial perturbations are overly restrictive. In such cases, one can instead define a (weaker) sample-based notion of stability. For any x in a finite set $X = \{x_i\}_{i=1}^n$ one replace $B_\epsilon(x_i)$ with an ϵ-neighborhood within X, i.e.,

$$\mathcal{N}_\epsilon(x) = \{x' \in X|\ \|x - x'\| \leq \epsilon\}. \tag{7}$$

- *Ambiguity:* It indicates how concise is the explanation, i.e., characterized by few prominent features, facilitating interpretation and potentially including higher informational value with reduced noise [118], compared to an ambiguous uniform importance distribution. Indeed, let N denote the number of features. If we map the attributions to a probability space (using e.g., Eq. (2)), the resulting entropy,

$$\mathcal{H}_k = -\sum_{i=1}^{N} \pi_{i,k} \log(\pi_{i,k}), \tag{8}$$

measures the uncertainty of the output (prediction or decision) with respect to the input (features or states) [119]. On the other hand, when the number of features is very high, one can characterize the uncertainty by comparing the distributions of both the attributions and a reference uniform probability density function. This can be done by invoking the discrete Kullback-Leibler (KL) divergence. The larger the KL divergence, the higher the certainty yield by the XAI method.

- *Infidelity:* In XAI surrogate methods, i.e., the schemes that consist of approximating the original model with a low-complexity more interpretable surrogate such as LIME, the fidelity of the surrogate to the original model can be quantified. Indeed, given a black-box function f, explanation functional Φ, a random variable $\mathbf{I} \in \mathbb{R}^n$ with probability measure $\mu_{\mathbf{I}}$, which represents meaningful perturbations of interest, the explanation infidelity can be defined as [120]

$$\mathcal{I}(\Phi, f, \mathbf{x}) = \mathbf{E}_{\mathbf{I}\sim\mu_{\mathbf{I}}} \left[\mathbf{I}^T \Phi(f, \mathbf{x}) - (f(\mathbf{x}) - f(\mathbf{x} - \mathbf{I}))^2\right] \tag{9}$$

  where $\mathbf{I}$ represents significant perturbations around $\mathbf{x}$ and can be specified in various ways, such as the difference to a baseline $\mathbf{I} = \mathbf{x} - \mathbf{x}_0$.
- *Fidelity and Soundness:* Two metrics can be applied to evaluate fidelity. Firstly, [75] used recall ($\mathcal{R}$) as a measure of fidelity for this method, which is defined as follows,

$$\mathcal{R} = \frac{|\mathcal{T} \cap \mathcal{E}|}{|\mathcal{T}|} \tag{10}$$

  where the term True Features $\mathcal{T}$ represents the relevant features as extracted directly from the white box model and Explanation Features $\mathcal{E}$ represents the features characterized as most relevant by the explanation [121]. This measure indicates how well the explanation captures the most relevant features from the predictive model, i.e., as a measure of the completeness of the explanation. Additionally, to understand how well the explanation excludes irrelevant features (soundness of the explanation), precision ($\mathcal{P}$) can be measured,

$$\mathcal{P} = \frac{|\mathcal{T} \cap \mathcal{E}|}{|\mathcal{E}|} \tag{11}$$

- *R-squared (R2) Score:* Behind the workings of LIME lies the assumption that every complex model is linear on a local scale. LIME tries to fit a simple model around a single observation that will mimic how the global model behaves at that locality. The simple model can then be used to explain the predictions of the more complex model locally. In this respect, R-squared (R2) score is used to measure the performance of the surrogate local model.
- *Relative Consistency:* Let f_i denote a predictor trained over dataset $\mathcal{D}_i$. Explanations arising from different predictors are said to be consistent if they are close when the predictions agree with one another, i.e., given the sets

$$\begin{aligned} \mathcal{S}' &= \{\delta_{i,j}(x) | f_i(x) = y \cup f_j(x) = y\} \\ \mathcal{S}'' &= \{\delta_{i,j}(x) | f_i(x) = y \oplus f_j(x) = y\}, \end{aligned} \tag{12}$$

  where $\delta_{i,j}(x)$ is a similarity measure of the explanations $f_i(x)$ and $f_j(x)$, and γ is a fixed threshold. We aim at making the gap between the set of consistent explanations $\mathcal{S}'$ and inconsistent ones $\mathcal{S}''$ visible. In this respect, we invoke the true positive rate,

$$\mathrm{TPR}(\gamma) = \frac{|\{\delta \in \mathcal{S}' : \delta \leq \gamma\}|}{|\{\delta \in \mathcal{S} : \delta \leq \gamma\}|}, \tag{13}$$

  where $\mathcal{S} = \mathcal{S}' \cup \mathcal{S}''$. In addition, we also consider the true negative rate,

$$\mathrm{TNR}(\gamma) = \frac{|\{\delta \in \mathcal{S}'' : \delta > \gamma\}|}{|\{\delta \in \mathcal{S} : \delta > \gamma\}|}. \tag{14}$$

  The quality of these explanations can be assessed independently of the accuracy of the predictor via the *Relative Consistency (ReCo)* metric [122]:

$$\mathrm{ReCo} = \max_{\gamma} \mathrm{TPR}(\gamma) + \mathrm{TNR}(\gamma) - 1, \tag{15}$$

  with a score of 1 indicating perfect consistency of the predictors' explanations, and a score of 0 indicating complete inconsistency.
- *BLEU Score:* Bilingual Evaluation Understudy Score [123] is used to evaluate the quality of text generated by a language model by comparing it to one or more reference texts. It measures the overlap of n-grams between the generated text and the reference texts.
- *ROUGE Score:* Recall-Oriented Understudy for Gisting Evaluation [124] is a set of metrics for evaluating automatic summarization and machine translation by comparing the overlap of n-grams, word sequences, and word pairs between the generated summary/translation and reference texts.
- *Perplexity:* It is a measurement of how well a language model predicts a sample [125]. It is defined as the exponentiated average negative log-likelihood of a sequence. Lower perplexity indicates better predictive performance.

TABLE V: Benchmarking of XAI methods for a NeSy-based O-RAN CPU usage prediction task

Epoch	Metric	SHAP	Saliency	Grad × Input	Int. Grad	LRP
100	Confidence	0.8981	0.8681	0.8912	0.8588	0.8681
	Ambiguity	1.6092	1.4833	1.6086	1.6076	1.6076
	Time complexity (s)	2.1999	0.0768	0.0815	0.1207	0.1067
500	Confidence	0.8900	0.8923	0.8684	0.8828	0.8732
	Ambiguity	1.6092	1.4166	1.6085	1.6084	1.6084
	Time complexity (s)	1.8113	0.0540	0.0530	0.0935	0.0780

C. Ranking of XAI Methods in O-RAN Prediction Tasks

In this subsection, we present a comparative study of the common classes of XAI methods in the specific task of resource prediction in O-RAN, where Neuro-Symbolic (NeSy) models are an efficient solution [126]. Specifically, we make use of *Logic Tensor Network* to predict CPU usage in a virtual BS (vBS) by leveraging well-established O-RAN experimental datasets [127]. We then assess the explanation *ambiguity* (i.e., lack of evidence) and *confidence* metrics; already described in the previous subsection, as well as the processing time for 1 epoch. Table V summarizes the benchmarking results, which reveal that SHAP presents the higher processing time due to its foundation in game theory, which involves calculating the contribution of each feature to the prediction by considering all possible combinations of features. Gradient methods, including saliency maps, Gradient × Input, and Integrated Gradients, are less time-consuming compared to SHAP due to their computational efficiency. These methods compute feature attributions using straightforward gradient calculations, involving only a single backward pass or a few integration steps, without the need to evaluate the model on all possible feature subsets. This avoids the combinatorial explosion and significantly reduces computation time. As a result, gradient methods leverage efficient backpropagation algorithms, making them much faster while still providing valuable insights into model predictions.

D. O-RAN Alliance Specifications

The O-RAN Alliance aims to lead the telecom Industry toward designing an intelligent and open RAN [20] [21] leveraging and extending 3rd Generation Partnership Project (3GPP) reference RAN architecture towards greater flexibility in network deployment and scalability for new services. The O-RAN Alliance aims to foster a more modular and flexible RAN ecosystem by disaggregating software from hardware and establishing open and interoperable interfaces. This approach allows for greater compatibility and interchangeability among different vendors' equipment, enabling network operators to avoid vendor lock-in and embrace a wider range of technology solutions.

The new O-RAN architecture leverages NFV and SDN technologies to define new open interfaces and disaggregate the RAN functional blocks, to allow the deployment of new services and applications. O-RAN divides the Baseband Unit (BBU) of RAN into three functional blocks, Central Unit (CU), Distributed Unit (DU), and Radio Unit (RU). To support control user plane separation, the CU block is also divided into control plane CU-Control Plane (CP) and user plane CU-User Plane (UP) sub-blocks. The radio frequency signals are received, transmitted, amplified, and digitized at RU, which is located near the antenna, while CU and DU represent the base station's computation parts and are in charge of transmitting the digitalized radio signal to the network.
We note that in Release 15 [128], 3GPP introduced a flexible architecture for the 5G RAN. This architecture splits the base station (gNodeB or gNB) into three logical nodes: CU, Responsible for higher-layer functions, coordination, and management. DU, Handles mid-layer functions and connects to the Radio Unit (RU). RU, Deals with lower-layer RF functions. The functional split allows network engineers to optimize performance based on factors like latency, cost, and specific use cases. On the other hand, O-RAN defines the Open RAN concept, which aims for horizontal openness through open interfaces connecting various RAN functions (from RU to DU-CU, controller, and orchestrator). Specifically, O-RAN has standardized the Lower Layer Split (LLS) by defining split option 7-2x [129]. This split results in the Open RU (O-RU) and Open DU (O-DU). Additionally, O-RAN integrates 3GPP-defined interfaces (such as F1, W1, E1, and Xn) within its architecture.

The DU block may be deployed near or at the RU block, while the CU block may be deployed near the core network part. It is also worth noting that 3GPP has defined different RAN deployment scenarios and functional split options, which are described in [48] [130]. The two main components introduced by the O-RAN architecture are summarized below:

- *Non Real-Time RAN Intelligent Controller (Non RT RIC):* it supports non-RT functions (i.e., with a time granularity greater than 1s) such as policy-based guidance. The Non RT RIC is located at the Service Management and Orchestration (SMO) and comprises two sub-functions: Non RT RIC Applicaitons (rApps) and Non RT RIC framework. The latter is in charge of providing all required services to rApps via the R1 interface, whether from Non RT RIC framework or SMO, while rApps leverage the functionality provided by the Non RT RIC framework, such as data

monitoring via O1 interface (stored in a database), to perform intelligent RAN optimization functions at non-RT scale. Such functionality enables rApps to get information and trigger actions, e.g., re-configuration and policies. Hence, Non RT RIC enables exposing an intelligent RAN policy to Near RT RIC, through A1 interface, based mainly on data analytics and ML/DL inference.
We note that SMO plays a crucial role as an intelligent automation platform that simplifies network complexity, enhances performance, and minimizes operational costs for the RAN domain. Specifically, SMO manages RAN as a service by applying automation at scale, SMO abstracts RAN functions and applications, making them easier to handle. Additionally, SMO interfaces with O1, A1, and O2, overseeing orchestration, management, and automation of RAN elements.
- *Near Real-Time RAN Intelligent Controller (Near RT RIC):* it is in charge of controlling and optimizing the O-RAN nodes (CU and DU) and their resources through fine-grained data monitoring and actions over E2 interface, at a near RT scale (i.e., from 10ms to 100ms). It hosts several Near RT RIC Applications (xApps), which may collect near RT information (e.g., at a User Equipment (UE) or Cell basis) through E2 interface, and provide value-added services, with respect to the Non RT RIC's policies received via the A1 interface. xApps include Mobility Management (MM), Resource Management (RM), Spectrum Management (SM), etc.

III. Projects/Standards on XAI for O-RAN

XAI is increasingly becoming critical for the adoption of ML/DL in O-RAN. To achieve trustworthiness and transparency in ML/DL models in O-RAN, there are some ongoing standardization activities and research projects targeting XAI and O-RAN aspects. Some of them include:

- *O-RAN Alliance:* As we describe in Subsection. II-D, the O-RAN Alliance is a global organization that is working to promote an intelligent and open RAN for mobile cellular networks. The O-RAN Alliance comprises 11 Working Groups (WGs) and three focus groups dedicated to RAN cloudification, automation, and disaggregation. In particular, WG2 in [131] describes lifecycle management of AI/ML models on O-RAN including learning model design, composition, training, runtime, and deployment solutions. It also highlights the main criteria for determining multiple ML training and inference host deployment options. In this context, the focus of WG2 can be extended to implementing XAI in O-RAN. To promote XAI adoption in O-RAN, the WG2 can work on various initiatives, ranging from XAI models' specifications and requirements to implementation and deployment. This may also include the creation of XAI platforms and tools, the development of interfaces and standards for XAI, and the promotion of XAI best practices.
- *IEEE P2894 and P2976:* these standards aim to deliver specifications on XAI in order to facilitate its adoption in real-world scenarios. The IEEE P2894 standard aims to design an architectural framework and define application guidelines for XAI, including the definition and description of XAI, the main classes of XAI techniques, the main application scenarios of XAI techniques, and performance evaluations of XAI in real systems such as telecommunication networks [132]. Besides, the IEEE P2976 standard is working to achieve interoperability and clarity of AI systems design through leveraging XAI techniques [133]. Specifically, IEEE P2976 defines optional and mandatory constraints and requirements that should be satisfied for an AI algorithm, method, or system to be considered explainable. In this context, these specifications can be leveraged by O-RAN standards such as O-RAN Alliance in order to develop and advance the adoption of XAI in the O-RAN ecosystem.
- *ETSI Experiential Networked Intelligence (ENI):* The ETSI Industry Specification Group (ISG) is working on defining a cognitive network management architecture based on context-aware policies and leveraging AI techniques. This effort aspires to adjust provided services in Fifth Generation (5G) networks and beyond based on changes in business goals, environmental conditions, and user requirements. Thus, it aims to provide automated service operation, provisioning, and assurance, along with efficient resource management and orchestration. Besides, recently, ETSI has released its first specifications on O-RAN called "O-RAN Fronthaul Control, User and Synchronization Plane Specification v7.02" [134]. This specification focuses on Open Fronthaul as one of the interfaces in the O-RAN Architecture. It specifies the synchronization plane protocols, user plane, and control plane used over the fronthaul interface to link the O-RU and O-RU components. This specification has been submitted to ETSI as a publicly available specification (PAS) produced by the O-RAN WG4 and approved by the ETSI Technical Committee. Therefore, considering this first specification of ETSI about O-RAN, the ETSI Experiential Networked Intelligence (ENI) Industry Specification Group (ISG) can also focus on adopting XAI on top of the designed cognitive network architecture in order to create an AI framework that is explainable, transparent, and thus can be used to ensure the accountability of AI-enabled systems in O-RAN.
- *6G-Bricks:* is a Horizon Europe project that explores novel unified control paradigms based on Explainable AI and Machine Reasoning, which will be delivered in the form of a reusable component with open Application Programming

Interfaces (APIs), termed "bricks" [135]. Initial integration with O-RAN will be performed, aiming for the future-proofing and interoperability of 6G-BRICKS outcomes.
- *NANCY:* it is the acronym of *An Artificial Intelligent Aided Unified Network for Secure Beyond 5G Long Term Evolution*; a Horizon Europe project which partly investigates the design of an XAI engine, to provide transparency and trustworthiness [136]. It also aims to identify the key factors that affect the system's local and overall performance.
- *Hexa-X:* developed a user-friendly support to Federated Learning (FL) of explainable-by-design models termed OpenFL-XAI which extends the open-source framework OpenFL [137]. Specifically, Hexa-X showed the benefits of building XAI models in a federated manner, with a specific focus on an automotive use case, namely Tele-operated Driving (ToD), which is one of the innovative services envisioned in 6G.

Overall, these standards and projects are working to promote the adoption of AI techniques, particularly machine learning and deep learning in O-RAN, while ensuring that these technologies are interpretable, accountable, and transparent. By doing so, they can help build trust in AI systems deployed in O-RAN. Thus, they encourage competition and innovation in the telecommunication industry.

IV. XAI Deployment on O-RAN

In this section, we describe how XAI methods can be deployed in the O-RAN framework and architecture by means of three realistic reference scenarios that are derived from XAI literature.

A. Introduction and Motivation

As described in Section. II-D, the basic idea of O-RAN is not only to disaggregate RAN functions exploiting the flexibility brought by virtualization techniques, but also to design RICs that locally host specific RAN applications (e.g., rApps and xApps), addressing several and heterogeneous control tasks such as handover management, energy management, fault detection, and radio resource allocation. The O-RAN framework has been devised to natively support a heavy usage of machine/deep learning (ML/DL) techniques to enhance the development and operations of intelligent RAN applications to pave the road for future B5G network services. For instance, as shown in [29], enabling cooperation among several xApps can help to optimize network performance, both in terms of data throughput and packet delivery ratio. However, one of the main challenges of AI-based O-RAN management is the lack of transparency on the decision-making processes that govern AI algorithms, which makes it difficult for network operators and engineers to diagnose problems, and further optimize the network behavior.

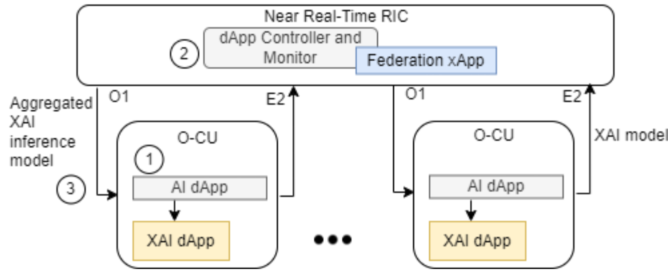


Fig. 3: Deployment of Federated AI and XAI in O-RAN. Adapted from [131], [138]–[140].

Therefore, there is a pressing need to integrate XAI into the O-RAN management operations, as to gain more detailed information about the decision-making processes of ML and DL algorithms. Specifically, XAI techniques should be incorporated into the running AI-based rApps/xApps to provide transparent explanations of their outputs. This would not only improve the accuracy and transparency of the decisions made by these systems but also increase the trust of network operators and engineers in the performance of the network.

B. Local Interpretable AI Deployment

The availability of open interfaces and the distributed nature of RAN deployments allows for the design and implementation of advanced federated and distributed schemes that aim to can overcome traditional RAN management scalability issues.

Indeed, to reduce monitoring overhead, reaction time and single point of failure risk, it is always beneficial to process data locally, where they are made available from dedicated monitoring functions. Therefore, raw control plane information generated by end-users at a given cell (or multiple cells) can be processed locally, in the Open RAN Central Unit (O-CU), and used to train AI/ML-based *dApps* [138] and their corresponding local XAI dApps (**Step 1**). As depicted in Fig. 3, to leverage the distributed nature of RAN deployments, such local information can be transferred to the Near RT RIC, exploiting the E2 interface, (**Step 2**). By combining multiple local models trained over a particular portion of the input space, the Near RT RIC aims to derive more generalized and advanced models to the O-CU and the corresponding Distributed Application (dApp). This information can be provided as feedback (**Step 3**) via the O1 interface. Hence, leveraging collected data from distributed nodes via the O1 interfaces, predictions along with their corresponding explanations can be performed in real-time. Favoured by a continuous learning process, both AI and XAI's outputs should be considered to perform management decisions and improve network performance. For instance, such outputs can help to update users' scheduling policies or radio resource assignments. In this context, different XAI techniques can be leveraged. For instance, RuleFit is one of the most used

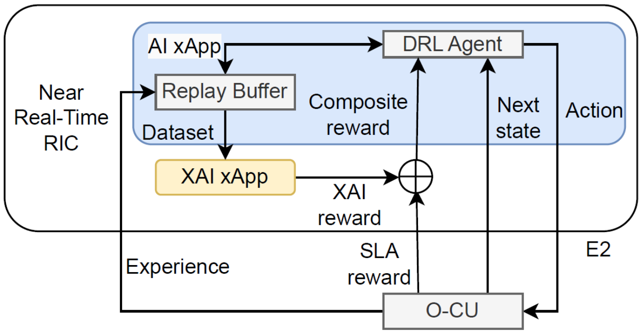


Fig. 4: Deployment of explanation-guided deep reinforcement learning in O-RAN. Adapted from [119].

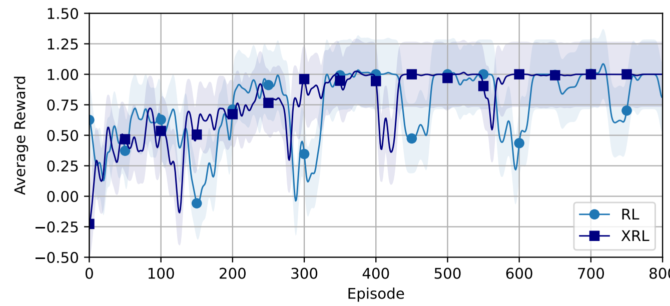


Fig. 5: Explanation guided DRL maximize the decision confidence compared to DRL. Adapted from [119].

XAI techniques [78] [79]. Its fundamental idea is to capture interactions between the original dataset features in order to create new features in the form of decision rules. Then, RuleFit learns a new transparent learning model using the original features, and also a number of new features that are decision rules.

Furthermore, the XAI explainability (outputs) may target different user profiles (**Step 5**). For instance, users of the models may want to trust and understand how the model works, while explanations related to the AI models' performance are sent to developers and data scientists to optimize their behavior over time. In addition, more details about AI models' compliance with the legislation in force should be communicated to both regulatory and manager users to check and assess them.

C. Explanation-Guided Deep Reinforcement Learning Deployment

Undoubtedly, Reinforcement Learning (RL) will play a significant role in enabling smart and automated O-RAN operations [141]. In the context of explanation-guided learning [142]–[144], explanation-guided deep reinforcement learning is a branch of artificial intelligence that combines deep learning and reinforcement learning with human-interpretable explanations. The goal of this approach is to enable humans to better understand the decision-making process of RL agents. In this method, the RL agent learns from its environment while considering human knowledgeable inputs providing explanations for the agent's behavior. The explanations can be in the form of natural language, visualizations, or other means. By providing contextual and external information, a human expert in the field of application can guide the agent toward better decision-making and improve its overall performance. As shown in Fig. 4, a Deep Reinforcement Learning (DRL) agent at the Near RT RIC performs resource allocation under latency constraints and interacts with the O-CU environment through the E2 interface [145]. The agent temporarily stores its experiences and observations in a replay buffer, which is continually updated.

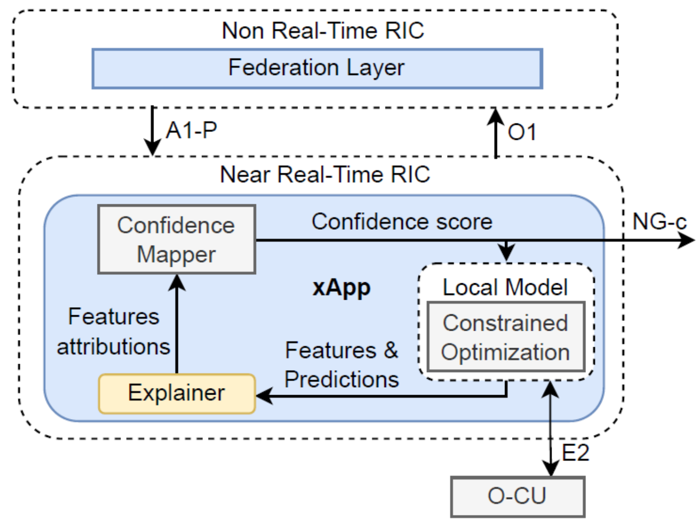


Fig. 6: Deployment of explanation-aided confident federated learning in O-RAN. Adapted from [146].

Then, the XAI xApp co-located with the Near RT RIC derives the SHAP importance values from a batch state-action dataset extracted from the buffer. To quantify the uncertainty of an action given a specific input state, the obtained SHAP values are afterwards converted to a probability distribution via *softmax* and used to calculate the entropy that measures the uncertainty, as formulated in Eq. (8). The multiplicative inverse of the maximum entropy value is used as an XAI reward (to minimize uncertainty and therefore maximize confidence). Combining the SLA reward (e.g., the multiplicative inverse of the latency) with the XAI reward results in a composite reward that reduces the uncertainty of state-action pairs and guides the agent to select the best and most explainable actions for specific network state values as illustrated in Fig. 5.

D. Explanation-Aided Confident Federated Learning Deployment

Explanation-aided confident Federated Learning (FL) is a type of machine learning that combines federated learning with human-interpretable explanations. In FL, data is collected and processed locally on individual devices, and only the necessary information is shared with a central server for model training [147] [148]. The goal of explanation-aided confident FL is to enable individuals and organizations to collaborate on training models while maintaining privacy and security.

To achieve a confident FL-based resource allocation/prediction, the local learning is performed iteratively with a run-time explanation as detailed in [146]. The overall working principle of the scheme is manifested in Fig. 6. For each local epoch, the dataset collected through the E2 interface is used to train a local resource allocation model via constrained optimization, which yields the features and the corresponding predictions to the XAI xApp where an *explainer* generates the features attributions using one of the feature attribution XAI methods (e.g., SHAP, Integrated Gradient, etc.). The *confidence mapper* then converts these attributions to a soft probability distribution and translates it afterwards into a confidence metric according to Eq. (3), and feeds it back to the optimizer to include it as an additional constraint in the local optimization. Moreover, the confidence metric is sent via the NG-c interface to the peer O-CUs. In this respect, each O-CU uses the gathered set of confidence scores to assess its priority, where only the K O-CUs with the largest confidence scores out of the available N O-CUs take part in the FL training to guarantee better confidence. Upon the termination of the local optimization, the model weights are reported to the federation layer—located at the Non RT RIC—to perform model aggregation and broadcast it via the A1-P interface. This iterative procedure results in highly confident FL in O-RAN compared to vanilla post-hoc FL as depicted in Fig. 7.

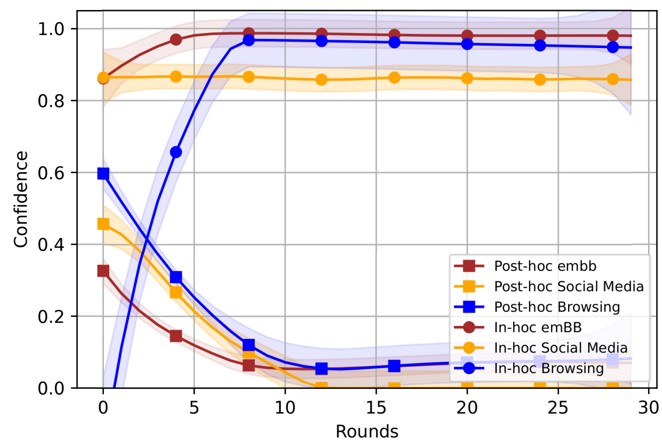


Fig. 7: FL confidence vs. rounds. Adapted from [146].

V. Automation of AI/XAI Pipeline for O-RAN

As discussed in Subsection. IV, XAI tools can be leveraged to assess the trustworthiness of the ML/DL models on top of the O-RAN architecture. In such context, the MLOps pipeline [149] will be augmented by a model transparency check block in the form of a closed loop that leverages the XAI objective metrics to evaluate the confidence of O-RAN AI xApps on the fly, as shown in Fig. 8.

DevOps paradigm includes a set of practices that combines software development (Dev) and IT operations (Ops). DevOps aims not only to reduce the systems' development life cycle but also to provide continuous software delivery with high quality, by leveraging paradigms and concepts like Continuous Integration and Delivery (CI/CD). When dealing with machine learning operations, and automation of the learning process, the paradigm can also be called ML system operations (MLOps) [150].

It is worth noting that O-RAN specification [131] introduces three control loops that facilitate the deployment of AI/ML (Artificial Intelligence/Machine Learning) functionalities within the O-RAN framework. These control loops are designed to operate at different time scales, enabling efficient integration and utilization of AI/ML capabilities in the network. Loop 1 is deployed at Open RAN Distributed Unit (O-DU) level to deal with per Transmission Time Interval (TTI) scheduling and operates at a timescale of the TTI or above. Loop 2 deployed at the Near RT RIC to operate within the range of $10ms - 1s$ and above. Loop 3 at the Non RT RIC at greater than 1 sec (ML/DL training, orchestration, etc.). In what follows, we focus more on both loops 3 and 2 for XAI models training, inference, and performance monitoring. Indeed, three main levels of automation have been categorized [150]: Manual (no MLOps), training pipeline automation, and CI/CD pipeline automation. A typical architecture integrating XAI with the MLOps pipeline is introduced in [119].

A. Manual Pipeline

It corresponds to the basic level of maturity, where all the ML steps, including data collection and preparation, model training, and validation, are performed manually (cf. Fig. 8). Hence, it is called no MLOps. At this level, data scientists usually use a rapid application development tool to build learning models, such as Jupyter Notebooks. In this case, the different steps of ML are released at the Non RT RIC module (ML), while the trained models are deployed at the Near RT RIC through the A1 interface, to provide prediction services (Ops). Note that the transitions from one step to another are also performed manually, and driven by a source code, developed interactively, till an executable model is created.

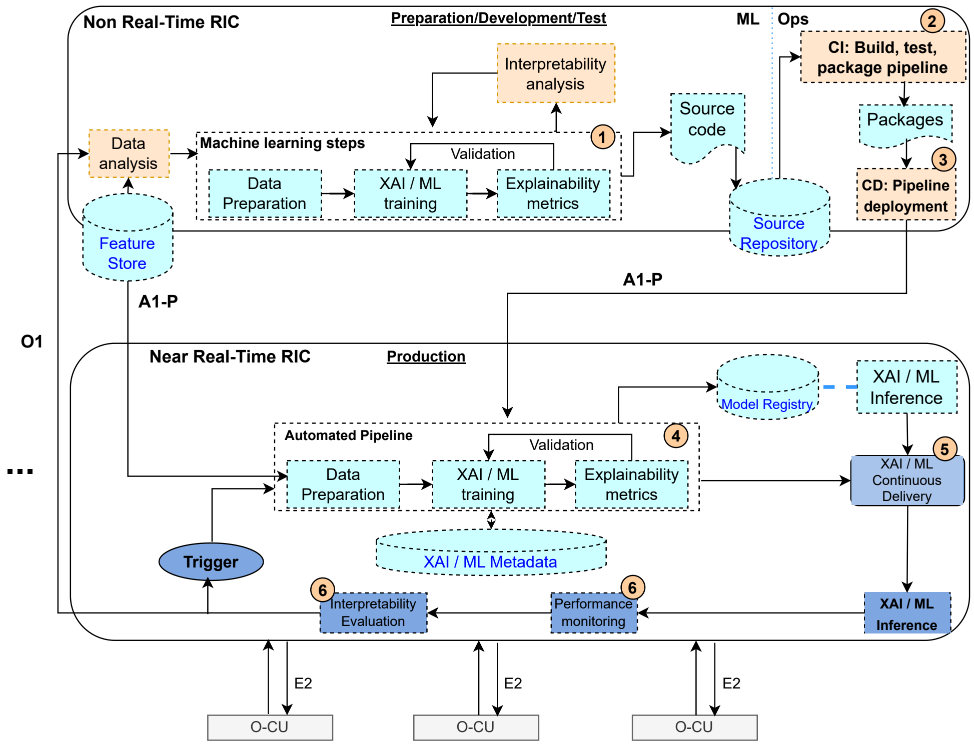


Fig. 8: XAI-driven Automated Continuous Integration and Delivery Pipeline.

In practice, this pipeline corresponds to the learning models, which are rarely updated and often break when they are deployed (models) in the real world. In addition, the performance of learning models at the RAN environment may degrade, due mainly either to the dynamic evolving of data profiles describing the environment or to the very dynamic changes that may occur in the radio access environment. Hence, automating the whole learning process becomes primordial.

B. Training Pipeline Automation

This level introduces a continuous training of the models and thus consists of performing the model training steps automatically. In particular, when new data profiles are monitored, the model retraining process is triggered. This process also includes data and model validation phases to achieve continuous delivery of the learning models. This level introduces two new components, named *feature store* as a centralized repository to store features and enable access to new features for training serving, and *machine learning metadata* to store information about the execution of ML pipeline (cf. Fig. 8).
We note that the interface A1-P is used to deploy the trained learning model at the near Real-Time RIC. In addition, when new data profiles appeared and thus the learning model should be updated through the automated pipeline, the A1-P is also used to transfer the new data, stored in the Feature Store database, from the Non Real-Time RIC to the Near Real-Time RIC to updated the learning model with its corresponding XAI model in the Near Real-Time RIC entity.

C. Continuous Integration and Delivery Pipeline Automation

At this level, a complete CI/CD system is introduced to enable reliable and fast learning model deployments in production. Thus, this level achieves the highest degree of automation in ML Ops, by enabling data scientists and developers to efficiently explore new ideas about feature engineering, model hyperparameters, and architecture. The main difference with the previous level is that CI/CD enables building, validating, and deploying the data, learning models, and model training pipeline components automatically.
Fig. 8 shows the automation of the ML pipeline using CI/CD in O-RAN context, which mainly features automated both ML pipelines and CI/CD routines.
In this context, in [151], the authors introduce principles for applying reinforcement learning (RL) in the O-RAN stack, emphasizing its integration into wireless network research. It reviews current research in this area and applies it to the RAN framework within the O-RAN architecture. The paper proposes a taxonomy to address challenges faced by

ML/RL models across their lifecycle–from system specification to production deployment–including data acquisition, model design, testing, and management. To tackle these challenges, the paper integrates existing MLOps principles tailored for RL agents, introducing a systematic model development and validation lifecycle termed RLOps. Key components of RLOps discussed include model specification, development, deployment in production environments, operational monitoring, and ensuring safety/security. The paper concludes by proposing best practices for RLOps to achieve automated and reproducible model development, all within a holistic data analytics platform embedded in O-RAN deployments.

VI. Taxonomy of XAI for 6G O-RAN

In this section, we give a literature review of existing recent works, which leverage XAI techniques for the 6G O-RAN architecture.

A recent work is leveraging XAI for DRL on top of the O-RAN architecture [155]. It addresses resource allocation problems at the O-RAN level and leverages XAI to provide network-oriented explanations based on an attributed graph, which forms a link between different DRL agents (graph nodes) and the state space input (the attributes of each graph node). This new scheme, termed EXPLORA, explains the wireless context in which the reinforcement learning agents operate. It shows XAI can be leveraged in DRL to perform optimal actions leading to median transmission bitrate and tail improvements of 4% and 10%, respectively.
In [152], the authors discuss XAI-based security architecture for the Open RAN in 6G, named XcARet. This architecture aims to provide transparent and cognitive security solutions for O-RAN while ensuring energy efficiency. They first describe the new security issues of O-RAN due mainly to its open interface and data flow features. Then, they provide recommendations for a dynamic policy of security adjustments, while considering the energy efficiency of the O-RAN architecture. Additionally, they also discussed about how to ensure the transparency of their dynamic security policy by explaining the adjustment decisions. In this context, another work discussed about the security challenges of the O-RAN architecture was proposed in [153]. The authors discussed about reliable AI and how to design and train rApps and xApps which are robust and secure against attacks. They also discussed on how to prevent, detect, and react to attacks that may target different components of O-RAN. Once an attack is performed and detected, the authors recommend to leverage XAI in order to understand what caused the attack, how it was performed, and eventually learn to recover from it. In this context, the XAI techniques can be applied to provide the non-RT RIC with information on which rApps and xApps were impacted, what type of input caused the attack, and why some applications gave unexpected outputs. This information can then be exploited by the non-RT RIC to re-train the AI models and thus deal with the observed vulnerabilities.
In [154], the authors address the misconfiguration issues in O-RAN. They present an depth analysis of the potential misconfiguration issues in O-RAN with respect to the use of NFV and SDN, specifically, the use of AI/ML. They investigated how AI/ML can be used to identify the O-RAN misconfigurations. A case study is proposed to show the impact on the UEs of conflicting policies amongst xApps, along with a potential AI-based solution. As AI finds use at different levels of O-RAN, the authors stress the need for XAI for O-RAN, especially for safety-critical use cases such as transportation automation, vital infrastructure operation (e.g., nuclear energy and water), human-machine brain interfaces, and healthcare.
In [144], [159], the authors got inspired by XAI and closed-loop automation to design an Explainable Federated deep learning (FDL) model to predict per-slice RAN dropped traffic probability in a non-IID setup, while jointly considering the explainability and sensitivity-aware metrics as constraints. Specifically, they quantitatively validate the explanations of faithfulness through the so-called attribution-based log-odds metric that is included as a constraint in the run-time FL optimization problem.
A novel multi-agent deep reinforcement learning (MADRL) framework, named standalone explainable protocol (STEP), for 6G O-RAN slicing was proposed in [156]. STEP enables Slices' orchestration agents to learn and adapt resource allocation while ensuring their post-hoc explainability, thanks to XAI. It is based on an information bottleneck framework to extract the most relevant information from running network slices at the O-RAN level, thus ensuring efficient decision-making and communication.

Moreover, in [157], the paper demonstrates how XAI can enhance the design of xApps by presenting a case study on an ML model trained for traffic classification using O-RAN KPIs. Utilizing SHAP, the study identifies the most influential KPIs for the model's predictions. By training the model with these selected KPIs, the research aims to reduce the overhead of transmitting all KPIs while observing the impact on model accuracy. Unlike existing works focusing solely on feature contribution, this paper uniquely leverages XAI to refine the training features based on their contribution. The contributions include: a SHAP-based XAI framework to identify key KPIs for traffic classification; two methods to reduce the number of KPIs-top K overall and top K per class-resulting in only a 7% accuracy drop with fewer KPIs; and an analysis showing a 33% reduction in control traffic data rate while maintaining high accuracy.

Additionally, [158] advances mobile traffic forecasting by introducing AIChronoLens, which links XAI explanations with temporal input properties. This approach addresses shortcom-

TABLE VI: Taxonomy of XAI works for the 6G O-RAN Architecture.

Ref.	XAI Class	Target AI Techniques	O-RAN's challenges	Impacted O-RAN Component
[152]	Post-Hoc Explanation	Deep Neural Networks	Security	rApp, Non Real-Time RIC
[153]			Energy-efficiency and security	Near Real-Time RIC
[154]			Misconfiguration and xApps conflict	Near Real-Time RIC
[155]		Deep Reinforcement Learning	Resource Allocation	Near Real-Time RIC
[156]		Multi-Agent Deep Reinforcement Learning	Resource allocation in O-RAN Slicing	Non Real-Time RIC
[157]		Convolutional Neural Networks	O-RAN monitoring overhead reduction	E2, vBS, Non Real-Time and Near Real-Time RICs
[158]		DLinear, PatchTST and LSTM	O-RAN traffic forecasting	Non Real-Time RIC
[144], [159]	Explanation-Guided Learning	Federated Deep Learning	Per-slice RAN dropped traffic detection	Near Real-Time and Non Real-Time RICs
[126]	Neuro-Symbolic Reasoning		O-RAN CPU resource provisioning	vBS, O-CU, O-DU

ings in legacy XAI techniques and enables a direct comparison of different AI models on the same dataset, enhancing their integration into the O-RAN architecture. Specifically, the AIChronoLens is used to explain DLinear, PatchTST, and LSTM in a prediction task of vBS mobile traffic and RRC connected users.

Finally, [126] introduces the Federated Machine Reasoning (FLMR) framework, a neuro-symbolic approach tailored for federated reasoning. FLMR enhances CPU demand prediction by leveraging contextual data and vBS configuration specifics from local monitoring within a shared O-Cloud platform of O-RAN. The framework ensures transparency in AI/ML decisions, addressing while comparative analysis against the DeepCog baseline demonstrates superior performance, achieving a six-fold reduction in resource under- and over-provisioning.

As we can observe from Table VI, there are multiple works leveraging XAI for the O-RAN architecture to provide more trust, transparency, and robustness to the AI-empowered solutions. The proposed works addressed the resource allocation, security, and misconfiguration of O-RAN xApps at the non Real-Time and Near Real-Time RICs. They leverage mainly post-hoc explanation and explanation-guided learning to explain different AI algorithms such as Deep Reinforcement Learning and supervised Deep Federated Learning.
The above works motivate our study and show the need to provide a comprehensive survey of XAI and its potential in designing the future O-RAN to guide the practitioners as well as researchers.

VII. Mapping of Existing AI-based O-RAN works to XAI-enabled solution

In this section, we first give a literature review of existing works, which leverage AI (ML/DL) techniques on top of the O-RAN architecture, in order to optimize RAN functions. We then discuss how these works can be mapped to XAI methods.

A. Existing AI-driven O-RAN Works

- *User Access Control:* The user access control or user association challenge is addressed in [160] [161], in order to ensure load balancing among Base Stations (BSs) and avoid frequent handovers. The authors designed a federated deep reinforcement learning. The UEs collaboratively trained their local models and then aggregated them at the RIC level. The designed model succeeded in maximizing the overall UEs' throughput and reducing frequent handovers.
- *Attack Detection:* In [162], the authors tackle security vulnerabilities in RAN cellular networks, focusing on the lack of integrity protection in the Radio Resource Control (RRC) layer. They propose a real-time anomaly detection framework using distributed applications in 5G Open RAN networks. By leveraging AI, they identify legitimate message sources and detect suspicious activities through Physical Layer features, which generate reliable fingerprints, infer the time of arrival of unprotected uplink packets, and handle cross-layer features. Their approach, validated in emulation environments with over 85% accuracy in attack prediction, is integrated into a real-world prototype with a large channel emulator. It meets the 2 ms low-latency real-time constraint, making it suitable for real-world deployments.
- *Energy-Aware RAN scalability:* In [163], the authors introduce ScalO-RAN, an optimization-based control framework designed as an O-RAN rApp to allocate and scale AI-based O-RAN applications (xApps, rApps, dApps). This framework ensures application-specific latency requirements are met, monetizes shared infrastructure, and reduces energy consumption. ScalO-RAN is prototyped on an OpenShift cluster with base stations, RIC, and AI-based xApps deployed as micro-services. Numerical and experimental evaluations show that ScalO-RAN optimally allocates and distributes O-RAN applications within com-

puting nodes to meet stringent latency requirements. The study highlights that scaling O-RAN applications is primarily a time-constrained issue, necessitating policies that prioritize AI applications' inference time over resource consumption.

- *Channel State Information (CSI):* a novel research platform for real-time inference using AI-enabled CSI feedback, closely simulating real-world scenarios, is designed in [164]. The framework is validated by integrating a CSI auto-encoder into the OpenAir-Interface (OAI) 5G protocol stack. The authors demonstrate real-time functionality with the encoder at the User Equipment (UE) and the decoder at the Next Generation Node Base (gNB). The experiments are conducted on both an Over-the-Air (OTA) indoor testbed platform, ARENA, and on the Colosseum wireless network emulator.
- *Total Cell Throughput:* An online training environment of a reinforcement learning model is deployed at the RIC level in [165]. The developed model controlled function parameters in DU, to maximize total cell throughput. Thanks to the deployed learning model, the total cell throughput increased by 19.4%.
- *SLA-Aware Network Slicing:* The authors in [166] propose a Deep Reinforcement Learning (DRL) agent for O-RAN applications, specifically for RAN slicing with Service Level Agreements (SLAs) focused on end-to-end latency. Using the OpenRAN Gym environment, the DRL agent adapts to varying SLAs and outperforms state-of-the-art methods, achieving significantly lower SLA violation rates and resource consumption without the need for re-training.
- *Function Placement:* The O-RAN architecture leverages virtualization and disaggregation of RAN functionalities among three key units (RU, DU, and CU). The authors of [167] studied the placement of resource allocation function based on service requirements, by dynamically selecting CU-DU units. Thus, they generated two reinforcement learning models based on Actor-Critic. The first one is used to assign resource blocks to UEs according to traffic types, delay budget, and UEs priorities, while the second one is leveraged to optimize function placement and hence the decisions of resource allocation. The authors showed that through this dynamic placement, both latency and throughput are highly improved.
- *RAN Orchestration:* In [168], the authors present OrchestRAN, a network intelligence orchestration framework for next-generation systems based on the Open Radio Access Network (RAN) paradigm. Designed to function in the non-Real-time (RT) RAN Intelligent Controller (RIC) as an rApp, OrchestRAN allows Network Operators (NOs) to specify high-level control and inference objectives, such as scheduling adjustments and near-RT capacity forecasting for specific base stations. OrchestRAN automatically selects the optimal set of data-driven algorithms and their execution locations (cloud or edge) to fulfill the NOs' objectives, ensuring timing requirements are met and preventing conflicts between algorithms managing the same parameters.
- *Resource allocation:* In [169] [170] [171], the authors studied the multi-agent team learning deployment on top of the O-RAN architecture by deciding about each agent placement and the required AI feedback. As a case study, the authors addressed the challenge of how to coordinate several running and independent xApps in O-RAN. They designed two xApps, called resource allocation xApp and power control xApp, and then used federated deep reinforcement learning to enhance learning efficiency as well as network performance in terms of throughput and latency. Similarly, in [172], the authors aimed to deal with the conflicts that may occur among running xApps when deployed by different vendors. Leveraging Q-learning, they proposed a team learning algorithm for resource allocation, to increase cooperation between xApps and hence optimize the performance of the network.
  Another distributed RL model was generated in [140], to manage RAN slice resource orchestration on top of the O-RAN architecture. The distributed RL architecture is composed of multiple intelligent agents, one for each network slice, performing local radio allocation decisions. Similarly, in [173], the authors leveraged federated distributed RL to manage the radio resource allocation among multiple Mobile Virtual Network Operators (MVNOs) for two different network slices (Ultra Reliable Low Latency Communication (URLLC) and enhanced Mobile Broadband (eMBB)). In [174], the challenge of how to optimally assign DU resources for various RUs is studied. A deep reinforcement learning model is built to achieve efficient management of RUs-DU resources. Experimental results showed that the proposed scheme improves highly resource usage efficiency. In the same context of resource allocation, in [175], the authors present PandORA, a framework for automatically designing and training DRL agents for Open RAN applications, packaging them as xApps, and evaluating them in the Colosseum wireless network emulator. They benchmark 23 xApps embedding DRL agents trained with various architectures, reward designs, action spaces, and decision-making timescales, enabling hierarchical control of different network parameters. These agents are tested on the Colosseum testbed under diverse traffic and channel conditions, both static and mobile. The experimental results show that fine-tuning RAN control timers and selecting appropriate reward designs and DRL architectures can significantly enhance network performance based on conditions and demand.

TABLE VII: Mapping of Existing AI-based O-RAN works to XAI-enabled Solutions.

Works	Addressed RAN Function	AI Technique	XAI Deployment at RIC as xApps				
			XAI Technique	Metrics	O-RAN Module	Functional Blocks	Interfaces
Yang et al. [160] [161]	User access control	Federated deep reinforcement learning	Reactive/Proactive Explanations	Confidence	O-CU-CP	UE and gNB procedure management	O1 (Monitoring), A1 (Analytics and Policies), E2 (Realization).
Hoejoo et al. [165]	Total cell throughput	Deep reinforcement learning		State-action certainty	O-DU	Resource assignment (NR-MAC)	
Shahram et al. [167]	Resource allocation and function placement	Actor-Critic learning		State-action certainty	O-DU	Resource assignment (NR-MAC)	
Rivera et al. [169] [170] Han et al. [171]	Resource allocation and power control	Federated deep reinforcement learning		Confidence, Log-odds	O-DU	Resource assignment (NR-MAC) and PDSCH (High-PHY)	
Han et al. [172]	Resource allocation	Q-learning		State-action certainty	O-DU	Resource assignment (NR-MAC)	
Farhad et al. [140]	Resource allocation	Distributed deep reinforcement learning		Robustness	O-DU	Resource assignment (NR-MAC)	
Wang et al. [174]	Resource allocation	Deep reinforcement learning		State-action certainty	O-DU O-RU	Resource assignment (NR-MAC)	
Abouaomar et al. [173]	Resource allocation	Federated Deep reinforcement learning		Confidence, Log-odds	O-DU	Resource assignment (NR-MAC)	
Scalingi et al. [162]	Attack Detection	CNN, LSTM DRL		Confidence, Log-odds	O-RU	Low Physical	
Maxenti et al. [163]	Energy-Aware scalability	CNN, LSTM DRL		Confidence, Log-odds	O-CU and O-DU	UE and gNB procedure and NR-MAC	
Cheng et al. [164]	Channel State Information	Encoder Decoder		Confidence, Log-odds	O-DU	PUSCH and PDSCH	
Raftopoulos et al. [166]	SLA-Aware Network Slicing	DRL		Confidence, Log-odds	O-CU, O-DU	Resource assignment (NR-MAC)	
D'Oro et al. [168]	RAN Orchestration	General		Confidence, Log-odds	O-CU, O-DU, O-RU	All Blocks	
Tsampazi et al. [175]	Resource Allocation	DRL		Confidence, Log-odds	O-DU	Resource assignment (NR-MAC)	

B. How XAI can Help

Integrating ML/DL-based algorithms with RAN functionalities has been found to address many challenging tasks: power control, user access control, handover, resources management, etc., which accordingly helps to optimize the performance of the RAN part. This was highly motivated in O-RAN, especially with the introduction of RIC modules. Indeed, the RAN functions are usually formulated as Markov Decision Process (MDP) [176], which explains the wide application of reinforcement learning, e.g., Q-learning, deep Deep Q-Network (DQN), and Actor-Critic, either in a centralized or a federated way, in order to derive the optimal policy about the corresponding RAN function. In addition, team learning is also an emerging paradigm to optimize the coordination and control of the running xApps at the O-RAN's RICs.

It is worth noting that resource management is the most studied RAN function using feature engineering approaches, such as feature extraction and feature selection, in addition to reinforcement learning algorithms (DQN and Q-learning). With this strategy of RAN functions analysis, it is possible to determine the contribution of every feature, e.g., higher-order cumulants, related to the RAN performances. This helps to adjust the features' values to optimize the ML/DL predictions. In fact, humans/users' understanding of Q-learning models is limited to small scenarios involving a few states and actions. However, these models may become complex, especially with a high number of features, states, and actions, making them less interpretable by humans. The challenge here is the accuracy-interpretability trade-off, which means that the greater the accuracy, the less likely the model is interpretable and vice versa.

For example, the Q-learning model can improve the overall performance of radio resource management by exploiting more descriptive frequency and time features, but its complexity increases when considering more features including network density and network/user power, service requirements, and the trust and security of the wireless communications, and thus this will introduce more states and actions in the system. Besides, despite being a more advanced algorithm as compared to Q-learning, DQN gives black-box models that output a lack of explainability. For instance, radio resource allocation using a DQN can introduce many ambiguous points, which should be explained, such as which layers/neurons in the DQN architecture can help to improve the accuracy, and why some UEs get the same number of radio block than others, even with different service requirements (URLLC, eMBB, massive Machine Type Communications (mMTC)). In this context, XAI is highly recommended since it provides profound insights into automated decisions and predictions. These details can help different users, as well as the network operators, to deal with unexpected problems, either related to the ML/DL models or to the corresponding xApps of the O-RAN's RICs. Therefore, the performance of the different RAN functions can be highly enhanced.

Within this context, a recent work is leveraging XAI for DRL on top of the O-RAN architecture [155]. This work addresses resource allocation and control challenges on top of the O-RAN architecture. It leverages XAI to provide network-oriented explanations based on an attributed graph which forms a link between different DRL agents (graph nodes) and the state space input (the attributes of each graph node). This new scheme explains the wireless context in which the reinforcement learning agents operate. It shows XAI can be leveraged in DRL to perform optimal actions leading to median transmission bitrate and tail improvements of 4% and 10%, respectively.

C. Mapping to XAI-enabled Works

In Table VII, we compare the existing AI-driven O-RAN works according to several criteria, including the addressed RAN function and the leveraged AI techniques. In addition, we illustrate how XAI can be deployed, as xApps, on top of these works, to explain their AI-based decisions. We observe that most of the existing works are based on reinforcement learning to manage RAN functions, especially power and resource allocation, user access control, function placement, and total cell throughput. According to [177], two main groups of XAI techniques can be applied to reinforcement learning strategies, to give both local and global explanations.

- *Reactive XAI techniques* imply that explanations are given in an immediate time horizon. It includes three main approaches: i) *Policy simplification*, which aims to simplify the policy of transition function into a form that is interpretable/understandable by itself, by using for instance decision trees and fuzzy rule-sets [178]. ii) *Reward decomposition* into understandable components, which aims to better understand the reasons for certain reward values [179]. iii) *Feature contribution and visual methods*, to determine features' contribution to the decision-making process, and then generate explanations based on that contribution. Examples of such techniques are both LIME [75] and SHAP [111].

- *Proactive XAI techniques* focus on a longer time horizon to provide the required explanations. These techniques can be classified into four main classes: i) *Structural causal model*, that aims to learn the relationships between variables/features. This technique generates human explanations since it considers the world as a causal lens [89] [180]. ii) *Explanation in terms of consequences*, which enables agents in reinforcement learning to answer a question about their policy in terms of consequences [181]. In other words, it enables to determine what each agent can get from a state, and which outputs it expects from a visited state or corresponding action. iii) *Hierarchical policy* by decomposing a given task into multiple sub-tasks, located at different abstraction levels [182]. Then, a prerequisite for executing a subsequent task is given as an interpretation of a particular action. iv) *Relational reinforcement learning*, which is based on a set of rules to provide background knowledge to the decision agent [183]. In this way, actions, states, and policies are represented by relation language, which helps to understand the reinforcement learning model's outputs. Moreover, the deployed XAI techniques can be evaluated according to several metrics, as shown in Table VII. Specifically, in DRL-based resource management use-cases, the state-action mapping *certainty* can be measured via the entropy, which is computed from the attributions of input state features. Moreover, the *confidence and Log-odds* metrics serve to quantify the trust in AI predictions/decisions by using the input XAI attributions as a basis to mask the impactful features in either offline or DRL-based on-the-fly datasets, and measure the corresponding deviation of the output, which is fed back to the optimizer/agent for accountability.

It is noteworthy that different data types can be monitored at the Non RT RIC from Open RAN Radio Unit (O-RU) modules, via the **O1 interfaces**, in order to feed AI and XAI xApps that are deployed at the Near RT RIC and which collaborate by exchanging model and predictions/decisions XAI metrics, respectively as exemplified in IV to achieve transparent AI operation. The resulting trustworthy output AI-based action is enforced at different levels (O-DU and O-CU), via the E2 interface.

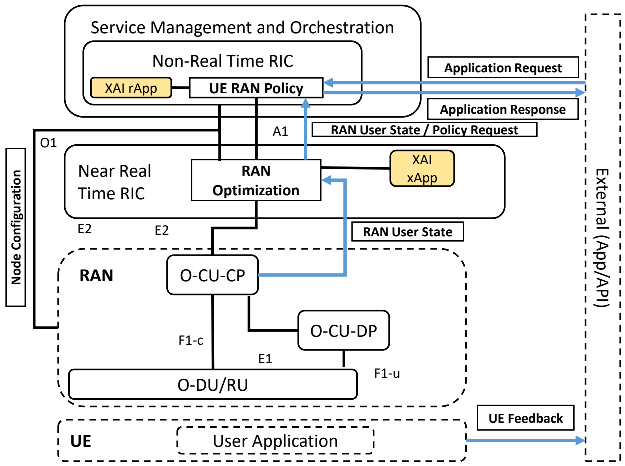


Fig. 9: Use case: User-Centric QoE Optimization. Adapted from [184]

VIII. XAI for O-RAN Use-cases

In the following, we collect a list of use-cases in the context of O-RAN and network slicing, highlighting how they would benefit from the introduction of XAI methods.

A. Quality of Experience (QoE) Optimization

Modern applications in the 5G ecosystem demand large bandwidth and low-latency communication to provide an adequate level of QoE, which can hardly be achieved by current semi-static Quality of Service (QoS) frameworks devoted to concurrently supporting heterogeneous broadband applications like in the Fourth Generation (4G) era. Radio fluctuations impair radio transmission capabilities, especially when adopting higher carrier frequencies like mm waves, leading to variable application requirements even within the same communication session. In order to improve QoE, estimation and prediction tasks performed at the application level can help in dealing with such a dynamic environment, favouring both user experience and efficient use of RAN resources [184].

Several works have addressed QoE modeling with traditional ML methods [185]. However, the prevalent black-box nature of ML models limits insights into QoE influence factors. Differently, XAI tools can provide contextual information for QoE assurance xApp [155] by e.g., identifying the relevant network environment factors that lead to under-provisioning decisions to underweight them, reducing thereby SLA violation. Moreover, XAI models such as Fuzzy decision trees have shown their suitability in identifying stall events in data transmissions impairing the resulting QoE, while providing interpretability of such events that can be leveraged to identify their cause [186], while SHAP have been used successfully to interpret Deep Neural Network (DNN) and random forest black-box models specifically trained to the QoE modelling task. In this regard, the interpretable output of XAI approaches becomes a valuable source of information to define strategies to improve QoE delivered by the network.

The open interfaces introduced by the O-RAN architecture significantly ease the per-user flow modification and configuration utilizing proactive closed-loop network optimization. Fig. 9 depicts a possible deployment addressing this use case. It involves Non RT RIC, Near RT RIC, E2 Nodes, and external applications running on the UE. The open interface allows external applications to interface with the O-RAN domain, which, empowered by ad-hoc optimization logic, would be capable of dynamically re-configuring the networking settings in response to real-time triggering events. By integrating XAI tools in the form of a standalone xApp into O-RAN, objective metrics measuring the trustworthiness of AI functions, including features attributions, can be computed on the fly (see Section II-B), providing context to the O-RAN reconfiguration decisions. This not only enables the transparency of the reconfiguration process but also facilitates the identification of patterns, root causes, and potential improvements in the overall performance of the network.

B. Traffic Steering

Imbalances in the traffic load across cells of different access technologies may lead to performance degradation. [184]. In this context, O-RAN A1 interface would allow enforcing desired optimization policies and utilizing the appropriate performance criteria to manage user traffic across different radio access technologies proactively. The Non RT RIC monitors the user experience by UE level performance measurements on a cell level and may decide to relocate one or more users to other cells based on global optimization objectives, e.g., fairness in bandwidth sharing, QoE maximization, load-balancing. In all these scenarios, attribution-based XAI methods such as SHAP and Integrated-Gradient can provide context to a traffic steering xApp in the form of attributions pointing out the impactful factors to consider in e.g., an offloading decision, which contributes to its transparency [155]. Besides, in [52], LIME is applied to DRL-driven traffic offloading in wireless networks, aiming at helping radio engineers better understand the consequences of the model choices and better monitor and configure the network. Whereas in [187], authors leverage XAI methods to improve the Quality of Transport (QoT) estimation process in optical transport networks, which is a key component in driving traffic steering decisions.

C. RAN Slice Service Level Agreement (SLA) Assurance

The 5G infrastructure has been designed to cope with highly diverse performance requirements coming from heterogeneous services and vertical applications. In this context, network slicing arises as a key technology to efficiently support tailored

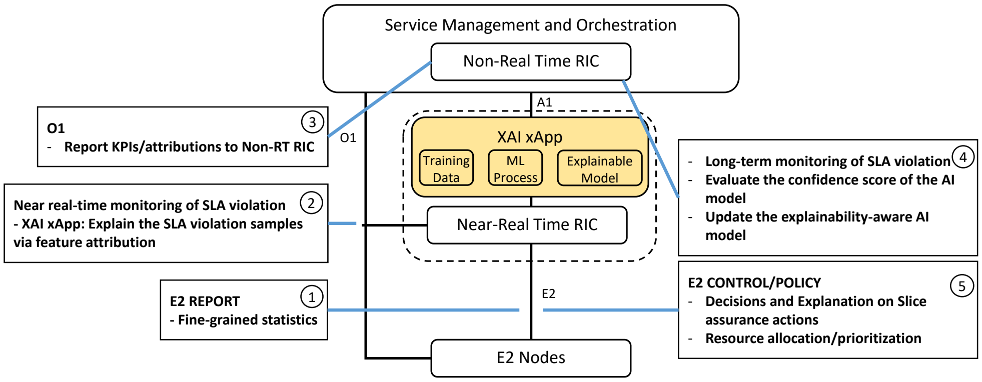


Fig. 10: Use case: Explainable slice SLA assurance. Adapted from [188]

end-to-end connectivity satisfying specific business requirements. In general, the business parties and the infrastructure provider define the set of networking capabilities required to successfully run the service in a SLA, e.g., in terms of data rate, latency, and resource availability [189]. Perhaps not surprisingly, this introduced the need for ad-hoc mechanisms able to efficiently measure and expose such information to 3^{rd} party entities traditionally alien to the telecommunication market. In this context, O-RAN's open interfaces and AI/ML-based architecture will enable such mechanisms, enabling the operators to take full advantage of the business opportunities brought by the network slicing concept.

More in detail, specific slice configuration settings derived from the SLA requirements can be easily enforced by initiating the procedure from the SMO layer, and finely adjusted over time thanks to measurements feedback and zero-touch XAI-based mechanisms applied at the different layers of the architecture, especially at the Non RT RIC and Near RT RIC. XAI xApp would build context for traditional SLA assurance xApps [36], [140], by providing feature attributions derived from e.g., SHAP, to point out the network factors that lead to SLA violation in the event of a resource under-provisioning decision [190]. XAI would also assess the trustworthiness of AI functions via various metrics (see II-B), which is part of their SLA [37], [191]. Fig. 10 summarizes the main workflow to achieve the solution.

D. Multi-vendor Slices

The coexistence of different network functions provided by different vendors to instantiate operators' services is one of the key enablers for flexible and efficient use of radio resources and CAPital EXpenditures (CAPEX)/OPerational EXpenditures (OPEX) optimization. To this extent, the O-RAN architecture enables the deployment of multiple slices

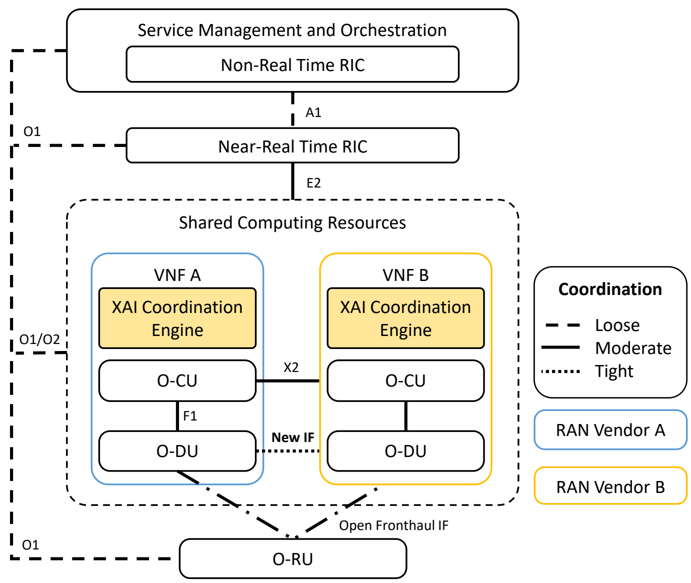
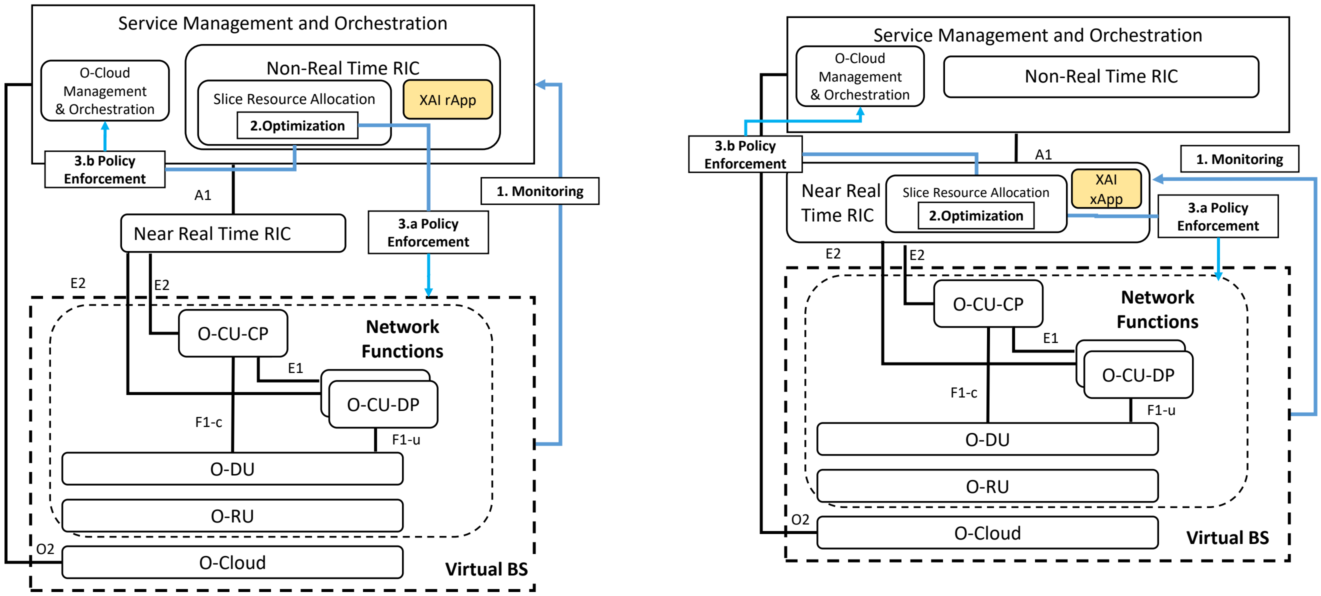


Fig. 11: Use case: Multi-vendor deployment with explainable coordination. Adapted from [184].

comprising functions provided by different vendors offering a variety of virtual Open RAN Distributed Unit (vO-DU) and virtual Open RAN Central Unit (vO-CU) options, specifically optimized to meet the requirements of a certain service. This brings several advantages such as a more flexible and time-to-market slice deployment, where operators can select from the available options the most suitable vO-DU and vO-CU to deploy their services and huge business opportunities for the vendors. To deploy multi-vendor slices, vO-DUs and vO-CUs must coordinate to coexist and share the radio environment efficiently and avoid conflicts among the deployed services [188]. Fig. 11 depicts three possible ways of coordination: *i) loose*



(a) Resource allocation over Non RT RIC.

(b) Resource allocation over Near RT RIC.

Fig. 12: Use case: Explainable RAN slice resource allocation decisions taken at Non-Real Time and Near-Real Time RICs.

coordination where there is no direct coordination between deployed services, and the radio resource is fully controlled by the RICs through the O1, A1, and E1 interfaces. *ii*) *moderate coordination* where different network functions are allowed to communicate with each other through the X2 and the F1 interfaces to negotiate radio resources without directly involving the RICs. In this case, the negotiation must cope with the time frame allowed by the X2 interface communication exchange, which is in the order of seconds; *iii*) the WG1 and WG4 of O-RAN Alliance envision also a so-called *tight coordination* allowing faster radio resource negotiation among slices, which would require a new interface, dubbed as *New IF* in Fig. 11, for direct communication between vO-DUs. In this context, distributed AI/ML models are particularly suitable to smartly perform the negotiation task [192], [193]. In this regard, an XAI-enabled component can be deployed to take control of the coordination and negotiation of resources between different vendors, while the heightened transparency and interpretability offered by XAI enhance the efficacy of resource management and coordination among vendors in this complex scenario. We report in the figure an example of a deployment of this component, suitable for both the moderate and the tight coordination case.

E. Resource Allocation Optimization

The need to concurrently support multiple heterogeneous slices characterized by service-tailored networking requirements exacerbates the setup of efficient and dynamic resource allocation mechanisms able to cope with highly different spatiotemporal traffic distributions.

For example, end-user mobility towards public events causes spatially localized peaks of traffic in eMBB kind of slices, or IoT smart sensors sporadically generating data volumes from highly distributed deployments in mMTC settings. Compared to traditional RAN deployments characterized by monolithic architecture and private management interfaces, the O-RAN's paradigm would allow for easier and more flexible control of the radio resources. In addition, the possibility to devise a data-driven ML-based optimization algorithm would help to automatize the process, exploiting the closed-loop management framework and previous historical information to perform the best allocation decisions. Additionally, AI/ML model can be used to perform proactive management of the radio resources, predicting recurrent traffic demand patterns of 5G networks in different time epochs and spatial locations, and for each network slice, therefore anticipating the slice's networking needs favoring a better end-users QoE, and limiting the overall energy consumption. All these methods are traditionally based on RL algorithms and agents interacting with the environment and learning by trial and error. More advanced solutions adopt federated learning techniques to improve the performances of the agents, gaining global knowledge of the system from the collection of multiple locally-trained models [140]. Such enriched information is then sent back to the single agents, improving their training, and speed, and allowing more general management policy definitions.

In both these scenarios, XAI methods can further extend the potential of the RL management solutions. On the one side, they will allow for better control of the learning procedure, and guide the agent towards the definition of a safe decision process by adding confidence and trust in the devised management policies, as demonstrated in [194]. On the other side, they may help in limiting the information exchange required by the federated learning approach [195] [196]. Only being able to map the context-decision space uniquely, would allow sharing to the federation layer only local models carrying insightful information while filtering out erroneous or redundant items. Fig. 12 depicts two possible deployment options, one assuming the main optimization and computing effort running within the Non RT RIC entity, and the other envisioning such task running within the Near RT RIC. The final deployment choice would depend on multiple factors, including the type of use-case and machine-learning model to be run and its timescale and complexity, also considering the different computing capabilities of the RICs. An interesting example of integration of XAI engine running as an xApp on an O-RAN compliant Near RT RIC is provided in [155].

F. User Access Control

The O-RAN vision aims at evolving the current RAN architecture providing open interfaces and ML-based optimization to attract new business entities and ease overall management and reduce operational costs. Current RAN deployments are composed of thousands of nodes [197]. In such a complicated deployment, it is expected that the network assigns each UE to a serving BS, maximizing the overall throughput and the single end-user QoE. This problem is also known as *user access control*. Traditional user access control schemes imply that user associations are based on networking metrics such as the Received Signal Strenght (RSS), which guides the UE towards the base station providing the best channel. Handover ping-pong effect and load balancing have been identified as two main issues brought by RSS-based schemes [198].

IX. O-RAN Security Aspects and XAI

Due to the central role of the 5G network in providing communication to backbone society infrastructures, security, and security risk awareness play a key role in network deployment. The O-RAN Alliance has recently created a Security Working Group (WG11) which identified a list of stakeholders responsible for ensuring the security of the RAN. This goes beyond the parties involved in traditional 4G and 5G networks, such as vendors, operators, and system integrators. In fact, operators will play a central role in securing the infrastructure, given the platform's openness and the use of multi-vendor components, which allows them to customize and secure the infrastructure. This also enables them to evaluate and verify the security of the open components that are introduced in the network, which may not be possible in fully vendor-driven closed architectures. In addition, according to [199], network functions and virtualization platform vendors, as well as third-party xApp and rApp developers, Open Cloud (O-Cloud) providers, and administrator profiles, that manage virtualized and disaggregated components, are all new stakeholders. However, due to the plethora of heterogeneous components forming the O-RAN ecosystem, and the high exploitation of AI-driven network management and third parties components running services, securing the O-RAN infrastructure is still a challenge. In this regard, XAI can strongly enhance security in O-RAN deployments by providing insights, explanations, and transparency into the decision-making process of AI models. Moreover, XAI helps with threat detection, model transparency, accountability, analysis of training data, and human-in-the-loop security, leading to improved threat detection, increased trust, and compliance with security regulations. However, it could also be the target of cyberattacks that could vanish its benefits.[1]

A. Distributed Architecture

The open architecture defined in the O-RAN specifications has been identified as a possible security issue due to its distributed nature, which expands the attack surface to malicious entities. The WG11 has recently identified the possible vulnerabilities coming from the openness of the platform and classified them into different threat categories [201]. Such categories include *i)* threats against the O-RAN system including architectural elements and interfaces that can compromise the availability, data, infrastructure integrity, and data confidentiality of the infrastructure *ii)* Threats against the O-Cloud, which could compromise virtual network functions, misuse containers or virtual machines, or spoof underlying networking or auxiliary services [202] *iii)* Threats in open-source code, which could potentially contain backdoors [200], [203], *iv)* Physical threats against the hardware infrastructure, *v)* Threats against the protocol stack and threats against AI/ML, including poisoning attacks that exploit unregulated access to the data stored in the O-RAN system to inject altered and misleading data. To counteract such threats, different security principles have been defined [201] to provide requirements, recommendations, and potential countermeasures, including mutual authentication (embracing the *zero-trust* paradigm), access control, robust cryptography, trusted communication, secure storage, secure boot and self-configuration, secure update processes, recoverability and backup mechanisms, and effective management of security risks posed by open-source components. As of the latest update, there is no explicit reference to leveraging eXplainable Artificial Intelligence (XAI) to bolster security within the context of Open Radio Access Network (O-RAN).

[1] Section IX provides a high-level overview of security in O-RAN, tailored to enhance understanding of related XAI approaches. For a more detailed survey on generic O-RAN security aspects, we refer readers to [200].

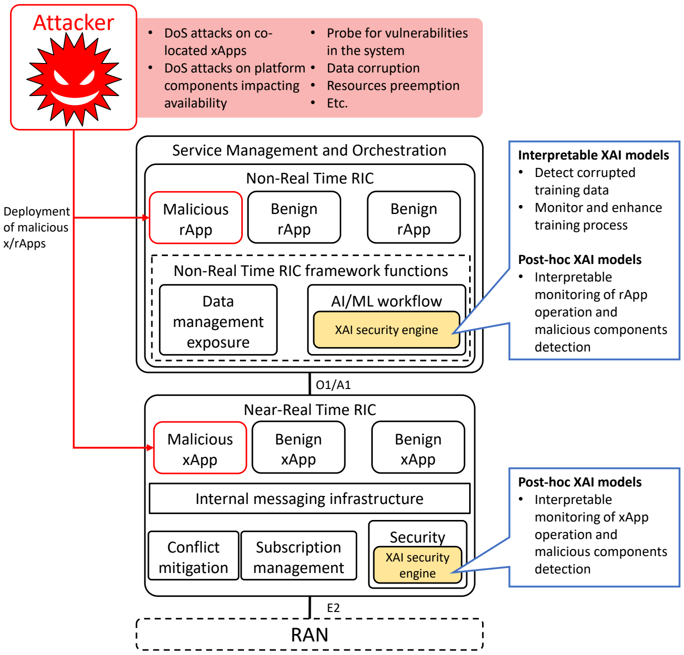


Fig. 13: O-RAN architecture with additional XAI-based security components. Adapted from [203].

However, XAI holds the potential to offer transparency and insights into the decision-making processes of Artificial Intelligence (AI) models, empowering stakeholders with a deeper understanding of how these security principles manifest in practice. This heightened clarity can facilitate enhanced validation, compliance, and trust in the security protocols deployed within O-RAN setups, thereby bolstering the overall security posture of the network.

B. Risk Assessments

The Security Working Group (WG11) has conducted comprehensive analyses on the Non RT RIC (Non RT RIC), the O-Cloud, and the Near Real-Time RIC (Near RT RIC) frameworks [204]–[206]. These assessments have evaluated the likelihood of attacks and their potential impact on protection goals, which can be categorized as follows: i) *Confidentiality*: Ensuring that sensitive information remains inaccessible to unauthorized entities. ii) *Integrity*: Safeguarding data from unauthorized manipulation, ensuring its integrity and preventing corruption or outdatedness. iii) *Availability*: Guaranteeing the availability of data, information, and services to authorized entities. Specifically, the Security Technical Report identifies 26 distinct threats to the Non RT RIC Framework, rApps, R1 interface, and A1 interface, along with corresponding recommended security controls [204].

The security analysis of the O-Cloud encompasses critical services, cloud service and deployment models, stakeholder roles and responsibilities, threat models, and best practices for mitigating threats [205]. Similarly, the Security Technical Report for the Near RT RIC and xApps addresses 11 key security issues and proposes 13 solutions, including modifications to existing documents and specifications maintained by WG3, with a mapping table illustrating how each solution corresponds to the identified issues [206]. With the increasing interest in deploying O-RAN, third-party entities and government bodies have conducted parallel and independent security risk assessments. [207] evaluates the integrity, availability, and confidentiality aspects alongside two additional protection goals: Accountability, concerning the traceability of actions to specific entities, and privacy, safeguarding sensitive data through anonymity, unlinkability, and unobservability. The analysis reveals a deficiency in adopting a "security/privacy by design/default" approach within O-RAN specifications, leading to multiple security vulnerabilities. This underscores the urgent need for a thorough revision of O-RAN specifications with a stronger security emphasis before productive applications are deployed. Another study [208] highlights risks associated with multiple suppliers, new network functions, and expanded interfaces, thereby increasing the attack surface. Furthermore, it identifies potential risks arising from the integration of AI and ML) in network functions, which could compromise network integrity. Additionally, reliance on cloud platforms for hosting base station software in O-RAN deployments could heighten dependency on cloud service providers, potentially exposing vulnerabilities, especially if multiple Mobile Network Operators (MNOs) utilize the same cloud provider.

At the time of writing, and to the best of our knowledge, there are no risk assessments specifically targeting security threats introduced by the use of XAI, nor suggesting XAI-based solutions/recommendations to enhance security in open networks. Nonetheless, we will discuss them shortly in the upcoming subsections.

C. XAI to Improve O-RAN Security

The utilization of XAI in the security domain of O-RAN could help enhance the transparency and comprehensibility of the operations and decision-making processes of third-party deployed components as to enable stakeholders to fully understand the decision process of such elements, finally helping to catch malicious behaviors and ensuring accountability, therefore reducing the risk of errors or malicious actions. This is particularly relevant when considering the high number of AI- and ML-driven components that will be deployed in the network, which due to their black-box nature pose a significant challenge to reveal malicious behavior and security threats [203]. By implementing XAI techniques, the complex algorithms used in ML-/AI-based systems can be made more interpretable. This enhances transparency and enables stakeholders to identify potential biases or shortcomings in the system, allowing for continuous improvement and optimization.

The O-RAN architecture foresees embedding the so-called *security subsystem* in the Near RT RIC. Such a component is in charge of detecting malicious xApps [24], e.g., preventing them from data leakage or overall network performance reduction. Differently, at the time of writing, the Non RT RIC architecture does not include a functional block solely dedicated to monitoring and detecting malicious rApps for security purposes. However, although not yet detailed, security is mentioned among the non-functional requirements of the AI/ML workflow module, which supports the development and implementation of AI/ML models for tasks such as self-optimization, automation, and data-driven decision-making within the Non RT RIC [131]. In this context, the surface of attacks extends to the AI/ML models running in RICs, i.e., the models that are used for intelligent operations for inference and control in the deployed rApps and xApp [41].

In the O-RAN architecture, AI/ML models are mainly deployed in the RIC as xApps/rApps [131], as depicted in Fig. 13. Such elements bring in the autonomous operation of several vital network functions including mobility management, resource allocation, etc. Hence, the deployment of malicious AI/ML models or the manipulation of benign ones by attackers could disrupt RAN node functionalities, resulting in severe network failures [41].

To overcome this issue, the use of XAI-enabled security engines has been recently proposed in [203]. Regarding the AI/ML workflow module, the work suggests embedding transparent XAI models such as Principal Component Analysis (PCA) and clustering to characterize input data and filter out corrupted or misleading samples. Other self-explanatory transparent models such as decision trees and random forests, are proposed to monitor and enhance the training process. Finally, post-hoc models are employed to further refine threat detection during validation. Similarly, the use of post-hoc XAI models is proposed to facilitate interpretable monitoring of xApp/rApp operation and detection of malicious components deployed in the RICs.

Nonetheless, employing XAI techniques O-RAN will require additional effort to build and define operational pipelines to generate feedback resulting in explanations. The integration of both ML and XAI models will most likely impact computational power requirements. Nevertheless, these additional expenses are warranted by the bolstering of O-RAN security and management functionalities [209].

D. Security threats related to XAI

As O-RAN continues to gain traction in the telecommunications industry, it becomes imperative to address the various security challenges inherent in its open and disaggregated framework. From concerns surrounding data confidentiality and integrity, to the need for robust authentication mechanisms, the security landscape of O-RAN is both complex and dynamic. Several attacks targeting ML/AI-enabled functions can be found in the literature. For example, in [210] authors realize and demonstrate an Adversarial Machine Learning (AML) attack on the traffic steering function, exploiting the query-based evasion attack technique proposed in [211]. In particular, the AML provides corrupted received signal strength samples to hinder the QoE classification and in turn, perform wrong traffic steering decisions. Whereas in [212], authors develop a malicious xApp designed to execute AML attacks on the data stored in the O-RAN RIC database, threatening the operation of ML-based interference classification xApp. In this regard, the authors design also a data distillation technique to mitigate cyber threats effectively.

In this context, XAI methods can be employed to improve security in O-RAN deployments. However, as the adoption of XAI gains momentum, so does the prevalence of cyberattacks targeting these models, as noted in recent research [213]. [214] evaluates two different attacks to the XAI layer: the first wherein the underlying ML model and XAI interpreted are simultaneously corrupted, while the second aims to construct adversarial samples that are sparse, interpretable, and close to the model's training data distribution. This is achieved by using a manifold approximation algorithm on a small set of test data to find data and explanation distributions and inducing minimal distortions on the input to steer the output toward the target (distorted) explanation.

Similar attacks can vanish the security advantage of an overlaying XAI layer [215]. For example, Adversarial and Data poisoning attacks involve intentional tampering of input data to mislead an AI system's predictions and explanation, aiming to bias the outcome of the model and leading to misinterpretation of the model's behavior [214]. Evasion attacks, involve revealing sensitive information through the explanations or interpretations provided by an AI system. This can be achieved by using the explanations to infer sensitive information about individuals, such as their health status, financial situation, or other private information, even if the AI system was not explicitly trained on such data [215]. Finally, social engineering attacks, involve manipulating users or human interpreters of the AI system's explanations to make incorrect or biased interpretations. This can be achieved by providing misleading or persuasive explanations that influence the human interpreter's decision-making or perception of the AI system's behavior [216]. However, the research in the area of security attacks specifically targeting XAI models is still in its infancy.

X. Open Challenges and Future Research Directions

In this section, we analyze different open questions and research topics for efficient deployment of XAI in future-proof

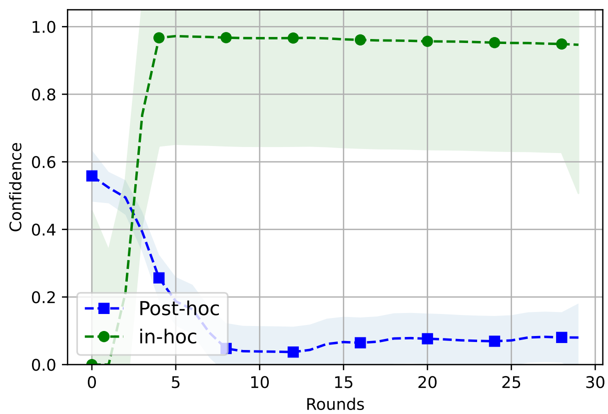


Fig. 14: Confidence vs. Federated Learning rounds [146].

6G O-RAN architecture, while pointing out various challenges in this area.

A. Explainability-Performance Trade-Off

The flip side of the highly performing yet complex AI models, such as DNNs and Transformers, is being ill-disposed to direct interpretation. They consist of numerous layers and billions of parameters, making it difficult to explain how specific inputs are transformed into outputs. In contrast, simpler models like decision trees or linear regressors are more interpretable, as their decision-making processes are more transparent. Therefore, it turns out that there is a trade-off between performance and explainability, especially when the type of training data justifies the use of complex models [49], [217]. Such observation is also valid when it comes to model optimization. To exemplify this, we plot the confidence metric described in Subsection. II-B in two scenarios, *i*) In each round, an FL model is optimized via a vanilla loss function such as mean square error, and then a post-hoc explanation in the form of attributions is generated and used to calculate the confidence metric and *ii*) In each round, an FL model is trained through a constrained optimization approach, where the confidence metric is evaluated in run time and jointly enforced as a constraint along with the original loss during training. According to Fig. 14, the model confidence degrades as it gradually converges in the post-hoc scenario, which highlights the abovementioned trade-off. Interestingly, the *in-hoc* strategy can maintain the model confidence across the training rounds, as thoroughly studied in [146]. Striking a balance between performance and interpretability is therefore an open research direction, especially to guarantee a successful deployment of AI in critical 6G use cases under O-RAN architectures.

B. LLMs in O-RAN: An Explainability Perspective

As anticipated in [218], the potential of Large Language Model (LLM)s to transform the Telecom domain lies in their ability to harness generative capabilities and leverage the multimodal nature of wireless network data, thereby enhancing contextual, situational, and temporal awareness. This advancement holds the promise of significantly improving wireless network operations, including localization, beamforming, power allocation, handover, and spectrum management, while also eliminating the requirement for task-specific AI models. Although this is still a far cry from achievement, future-proof 6G O-RAN might incorporate LLMs into the design of the different radio functions. In such a scenario, the complex architecture of LLMs that is mainly based on transformers with billions of parameters raises new challenges concerning their explainability and opens a new research line to tackle it.

C. Lack of Standardization

One of the still open challenges related to the adoption of XAI to O-RAN is the lack of standardization efforts across different components and interfaces, except the O-RAN Alliance and some ongoing projects initiated by network operators [20] [21] [219] that are more focusing on the main O-RAN's components. This makes it difficult to design XAI tools that can be deployed across different O-RAN components. Indeed, standardization is critical for enabling interoperability among various components of O-RAN and facilitating the development and deployment of XAI tools, on top of O-RAN. In this context, the research and industry communities should work towards developing common standards for open interfaces, data formats, APIs, etc.

D. Privacy Concerns of Distributed XAI Models for the Complex Multi-Vendor O-RAN

As mentioned before, one of the main features of O-RAN is to disaggregate the RAN functions and manage them through the running xApps at the Near RT RIC, to fulfil the QoS requirements of the envisioned B5G network services. In addition, the different O-RAN components are supplied and supported by various isolated vendors/operators. However, XAI tools typically need large amounts of data to train and test their models for O-RAN systems, and the availability of data may be limited or difficult to access, due to security and privacy concerns in a multi-vendor context. Therefore, these vendors/operators should collaborate to ensure stable RAN performance/cost and deal with the limited available data.
In this context, distributed/collaborative deep learning is expected to be widely leveraged. For instance, FL is one of the promising collaborative learning techniques that consists of generating learning models collaboratively while preserving the privacy of involved learners such as vendors/operators [220].

However, generating learning models (XAI or AI) in a federated way is still challenging since it still presents privacy concerns. Indeed, even FL avoids sharing learners' private data, however, it was demonstrated that FL is still vulnerable to some attacks, such as poisoning and Byzantine attacks, and sharing model updates during the training process can reveal private information [221]. Privacy-preserving XAI techniques can be used to provide explanations without revealing sensitive information. These techniques include methods such as differential privacy, blockchain, and homomorphic encryption [222].

E. Interoperable XAI Models for the Complex Multi-Vendor O-RAN

Introducing open interfaces and RAN functions disaggregation have led to split gNB into multiple CUs and DUs, which may belong to different vendors-operators and are connected through the F1 interface. However, designing XAI models while considering the interoperability among multiple vendors can be extremely challenging as they may have different implementations, capabilities, and requirements. In this context, a potential solution is leveraging collaborative and distributed learning techniques in building XAI models, as described in Subsection. IV-D. This will enable the generation of local XAI models specific to each vendor, aggregating them at the central Non RT RIC level to compose a global and interoperable XAI model. This can be achieved by leveraging and extending the WG5 activities through the definition of open interfaces between base stations (such as E1 and F1), and between gNBss (i.e., Xn interface), and between gNBs and eNodeB (eNB)s (i.e, X2 interface).

F. Complexity of the O-RAN Systems

O-RAN systems can be highly complex, including several layers of software and hardware components. This complexity can complicate the development of efficient XAI tools, that can provide a concise interpretation and explanation of the decisions made by the AI-based O-RAN systems. One way to deal with the complexity of AI models is to simplify them by applying techniques such as rule-based systems, decision trees, or linear models. These models are easier to explain and more transparent than complex deep neural networks. However, this may come at the cost of lower performance and accuracy. Furthermore, model-agnostic XAI approaches can also be leveraged to provide explanations for any AI model, regardless of its structure and complexity. These approaches include local surrogate models, partial dependence plots, and feature importance. The development and deployment of such approaches will enable XAI for a wider range of AI-based models deployed on top of the O-RAN system.

G. Real-Time Constraints

Some components of the O-RAN system are designed to operate in real-time, emphasizing the need of tailored XAI tools for this time-sensitive context. For instance, perturbation methods like SHAP are often computationally intensive, involving iterative evaluations of feature perturbations, which can be impractical for real-time applications with strict latency requirements. In contrast, gradient-based methods offer more efficient computation by leveraging model gradients with respect to input features. This efficiency makes gradient methods better suited for real-time deployment, providing interpretable explanations with reduced computational overhead. While perturbation methods excel in comprehensive feature importance analysis, their practicality in real-time settings is limited compared to gradient-based approaches, which prioritize responsiveness and low latency. These approaches include model compression, feature selection, and online learning. The in-depth development of such approaches will enable the deployment of real-time XAI models on top of the O-RAN systems without compromising their performance.

H. Heterogeneity of target audiences in XAI

One of the main challenges related to XAI models is to provide understandable explanations and interpretations to different user profiles (data scientists, developers, managers, etc.). One way to deal with this issue is to design human-centered XAI models, which are easy to understand and use for different end-users. This can be achieved by developing interactive explanations, visualization tools, and user-friendly interfaces. Moreover, human-centred XAI can also involve co-design and user studies with end-users to ensure that XAI models' outputs meet their expectations and needs.

Overall, the future of XAI for O-RAN will require interdisciplinary collaboration and research between network engineers, human-computer, AI researchers, and data scientists interaction experts. Potential solutions will consist of developing further data formats and standardized interfaces, of the use of simpler and more interpretable/explainable AI models, and the development of real-time XAI models that can operate within the time constraints of O-RAN systems. By addressing these open challenges of XAI for O-RAN, we can ensure that these systems are accountable, trustworthy, and transparent.

I. Security

As discussed in Section IX, in O-RAN security enforcement mechanisms are crucial for strengthening the resilience of network deployments to heterogeneous types of security threads. In this regard, the usage of XAI can be explored to develop and deploy intelligent and interpretable security monitoring systems, which enable more effective detection and mitigation of malicious components. In O-RAN, XAI-aided security components can be deployed in different O-RAN architectural layers. Research efforts in this direction are needed to optimize the design of such components enabling them to provide transparent insights into third-party decision-making, and effectively facilitating the identification of malicious behaviors. At the same time, efforts should be made in the definition of Human-in-the-Loop Security enforcement protocols, that thanks to XAI will empower human operators to interpret automated decisions, ensuring robust security oversight.

XI. Conclusion

By providing insights into how these systems work and making their decisions more transparent, XAI can help to improve the reliability, performance, and security of future O-RAN networks, playing a crucial role in their development and helping mobile network operators build and manage more effective and efficient networks. In this survey paper, we presented a comprehensive overview of XAI techniques in the context of O-RAN networks. First, we describe how XAI methods can be deployed in the O-RAN framework and architecture by means of three realistic reference scenarios. We then give a literature review of existing works, which leverage AI (ML/DL) techniques on top of the O-RAN architecture, in order to optimize RAN functions. We also discuss how these works can be mapped to XAI-enabled solutions. In addition, we collect a list of use-cases in the context of O-RAN and network slicing, highlighting how they would benefit from the introduction of XAI methods. Besides, to ensure good performance of the intelligent RAN functions over time, we show how to perform continuous monitoring of both model and data profiles, and how to automate the whole AI/XAI learning models development, including data collection/extraction, model training, validation, and deployment. Moreover, we explore the potential of XAI to significantly improve the security layer of O-RAN and envision how it could be used to build interpretable security threat detection mechanisms. Furthermore, we describe ongoing standardization activities and research projects targeting XAI and O-RAN aspects. Finally, we discuss the main open challenges related to XAI for O-RAN in addition to suggesting potential solutions to such challenges. With this work, we aim to foster and inspire the research on XAI, aiming at a new level of interpretable and human-understandable O-RAN network management operations.

Acknowledgment

This work has been partially supported by the Competitive Research Project at Sharjah university in UAE under the DeepORAN project (Grant No. 2402150256), the European Union's Horizon Program under the 6G-Bricks project (Grant No. 101096954), Sunrise-6G project (Grant No. 101139257), NANCY (Grant No. 101096456) and 6G-GOALS (Grant No. 101139232). This paper is also part of the Grant TSI-063000-2021-10 funded by the Ministry for Digital Transformation and of Civil Service and by the "European Union NextGenerationEU/PRTR".

References

[1] M. Z. Chowdhury, M. Shahjalal, S. Ahmed, and Y. M. Jang, "6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 957–975, 2020.

[2] X. Xie, B. Rong, and M. Kadoch, "6G Wireless Communications and Mobile Networking," *Bentham Books: Sharjah, United Arab Emirates*, 2021.

[3] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE network*, vol. 34, no. 3, pp. 134–142, 2019.

[4] F. Fang, Y. Xu, Q.-V. Pham, and Z. Ding, "Energy-Efficient Design of IRS-NOMA Networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 14 088–14 092, 2020.

[5] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, J. S. Thompson, E. G. Larsson, M. D. Renzo, W. Tong, P. Zhu, X. Shen, H. V. Poor, and L. Hanzo, "On the Road to 6G: Visions, Requirements, Key Technologies and Testbeds," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2023.

[6] D. Zhou, M. Sheng, J. Li, and Z. Han, "Aerospace Integrated Networks Innovation for Empowering 6G: A Survey and Future Challenges," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2023.

[7] B. Mao, J. Liu, Y. Wu, and N. Kato, "Security and Privacy on 6G Network Edge: A Survey," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2023.

[8] E. Coronado, R. Behravesh, T. Subramanya, A. Fernàndez-Fernàndez, M. S. Siddiqui, X. Costa-Pérez, and R. Riggio, "Zero touch management: A survey of network automation solutions for 5G and 6G networks," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2535–2578, 2022.

[9] C. D. Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, and M. Liyanage, "Survey on 6G Frontiers: Trends, Applications, Requirements, Technologies and Future Research," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 836–886, 2021.

[10] L. U. Khan, I. Yaqoob, M. Imran, Z. Han, and C. S. Hong, "6G Wireless Systems: A Vision, Architectural Elements, and Future Directions," *IEEE Access*, vol. 8, pp. 147 029–147 044, 2020.

[11] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A Speculative Study on 6G," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 118–125, 2020.

[12] B. Brik, K. Dev, Y. Xiao, G. Han, and A. Ksentini, "Guest Editorial Introduction to the Special Section on AI-Powered Internet of Everything (IoE) Services in Next-Generation Wireless Networks," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 5, pp. 2952–2954, 2022.

[13] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System," *IEEE Access*, vol. 7, pp. 70 371–70 421, 2019.

[14] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[15] Y. Wu, H.-N. Dai, H. Wang, Z. Xiong, and S. Guo, "A Survey of Intelligent Network Slicing Management for Industrial IoT: Integrated Approaches for Smart Transportation, Smart Energy, and Smart Factory," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1175–1211, 2022.

[16] Z. A. El Houda, B. Brik, and L. Khoukhi, "Ensemble Learning for Intrusion Detection in SDN-Based Zero Touch Smart Grid Systems," in *2022 IEEE 47th Conference on Local Computer Networks (LCN)*, 2022, pp. 149–156.

[17] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, "A Survey on Software-Defined Networking," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 27–51, 2015.

[18] S. Wijethilaka and M. Liyanage, "Survey on Network Slicing for Internet of Things Realization in 5G Networks," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 957–994, 2021.

[19] S. Ben Saad, A. Ksentini, and B. Brik, "An end-to-end trusted architecture for network slicing in 5G and beyond networks," *SECURITY AND PRIVACY*, vol. 5, no. 1, p. e186, 2022.

[20] O. R. Alliance, "O-RAN: Towards an open and smart RAN," *White paper*, vol. 19, 2018.

[21] S. Abeta, T. Kawahara, A. Umesh, and R. Matsukawa, "O-RAN Alliance Standardization Trends." *NTT DOCOMO Technical Journal*, vol. 21, no. 1, 2019.

[22] O-RAN Alliance, "O-RAN-WG1-O-RAN Architecture Description (ORAN.WG1.O-RAN-Architecture-Description-v04.00)," Technical Specification, 2020.

[23] O. Alliance, "O-RAN Non-RT RIC: Functional Architecture 1.01-March 2021 (O-RAN. WG2. Non-RT-RIC-ARCH-TR-v01. 01)," *Technical Specification*, 2021.

[24] O.-R. W. G. 3, "O-RAN near-RT RAN intelligent controller near-RT RIC architecture 2.00," *O-RAN. WG3. RICARCH-v02. 00*, 2021.

[25] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[26] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, "Federated Learning for Internet of Things: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[27] M. Shen, K. Ye, X. Liu, L. Zhu, J. Kang, S. Yu, Q. Li, and K. Xu, "Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 791–824, 2023.

[28] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Intelligent Closed-loop RAN Control with xApps in OpenRAN Gym," in *European Wireless 2022; 27th European Wireless Conference*, 2022, pp. 1–6.

[29] H. Zhang, H. Zhou, and M. Erol-Kantarci, "Team Learning-Based Resource Allocation for Open Radio Access Network (O-RAN)," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 4938–4943.

[30] M. Sulaiman, A. Moayyedi, M. Ahmadi, M. A. Salahuddin, R. Boutaba, and A. Saleh, "Coordinated slicing and admission control using multi-agent deep reinforcement learning," *IEEE Transactions on Network and Service Management*, 2022.

[31] E. Commission, "Communication on Fostering a European approach to Artificial Intelligence," 2021. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence

[32] E. O. of the President of United States of America, "Maintaining American Leadership in Artificial Intelligence - Executive Order 13859 of February 11, 2019," 2019. [Online]. Available: https://www.federalregister.gov/d/2019-02544

[33] D. for Digital, Culture, M. . Sport, D. for Business, E. . I. Strategy, and O. for Artificial Intelligence of the United Kingdom, "National AI Strategy - AI Action Plan," 2022. [Online]. Available: https://www.gov.uk/government/publications/national-ai-strategy-ai-action-plan/national-ai-strategy-ai-action-plan

[34] M. Aitken, D. Leslie, F. Ostmann, J. Pratt, H. Margetts, and C. Dorobantu, "Common Regulatory Capacity for AI," 2022.

[35] H. Elayan, M. Aloqaily, F. Karray, and M. Guizani, "Internet of Behavior (IoB) and Explainable AI Systems for Influencing IoT Behavior," *IEEE Network*, pp. 1–8, 2022.

[36] M. A. Habibi, B. Han, M. Nasimi, and H. D. Schotten, "The structure of service level agreement of slice-based 5G network," *arXiv preprint arXiv:1806.10426*, 2018.

[37] 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 *IEEE Global Communications Conference (GLOBECOM)*, 2020, pp. 1–7.

[38] S. Wang, M. A. Qureshi, L. Miralles-Pechuán, T. Huynh-The, T. R. Gadekallu, and M. Liyanage, "Explainable AI for B5G/6G: Technical Aspects, Use Cases, and Research Challenges," 2021.

[39] S. Ben Saad, B. Brik, and A. Ksentini, "A Trust and Explainable Federated Deep Learning Framework in Zero Touch B5G Networks," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 1037–1042.

[40] V. Belle and I. Papantonis, "Principles and Practice of Explainable Machine Learning," 2020.

[41] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2023.

[42] A. S. Abdalla, P. S. Upadhyaya, V. K. Shah, and V. Marojevic, "Toward Next Generation Open Radio Access Networks–What O-RAN Can and Cannot Do!" *IEEE Network*, pp. 1–8, 2022.

[43] S. Niknam, A. Roy, H. S. Dhillon, S. Singh, R. Banerji, J. H. Reed, N. Saxena, and S. Yoon, "Intelligent O-RAN for Beyond 5G and 6G Wireless Networks," 2020.

[44] A. Garcia-Saavedra and X. Costa-Perez, "O-RAN: Disrupting the Virtualized RAN Ecosystem," *IEEE Communications Standards Magazine*, pp. 1–8, 2021.

[45] S. K. Singh, R. Singh, and B. Kumbhani, "The Evolution of Radio Access Network Towards Open-RAN: Challenges and Opportunities," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2020, pp. 1–6.

[46] W. Diego, "Evolution Toward the Next Generation Radio Access Network," in *IFIP Networking Conference (Networking)*, 2020, pp. 685–685.

[47] C. L. I, S. Kuklinski, T. Chen, and L. L. Ladid, "A Perspective of O-RAN Integration with MEC, SON, and Network Slicing in the 5G Era," *IEEE Network*, vol. 34, no. 6, pp. 3–5, 2020.

[48] B. Brik, K. Boutiba, and A. Ksentini, "Deep Learning for B5G Open Radio Access Network: Evolution, Survey, Case Studies, and Challenges," *IEEE Open Journal of the Communications Society*, pp. 1–1, 2022.

[49] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information fusion*, vol. 58, pp. 82–115, 2020.

[50] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," 2020.

[51] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for Explainable AI: Challenges and Prospects," 2019.

[52] W. Guo, "Explainable Artificial Intelligence for 6G: Improving Trust between Human and Machine," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 39–45, 2020.

[53] C. Fiandrino, G. Attanasio, M. Fiore, and J. Widmer, "Toward native explainable and robust AI in 6G networks: Current state, challenges and road ahead," *Computer Communications*, vol. 193, pp. 47–52, 2022.

[54] T. Senevirathna, Z. Salazar, V. H. La, S. Marchal, B. Siniarski, M. Liyanage, and S. Wang, "A Survey on XAI for Beyond 5G Security: Technical Aspects, Use Cases, Challenges and Research Directions," 2022.

[55] B. Brik and A. Ksentini, "Toward Optimal MEC Resource Dimensioning for a Vehicle Collision Avoidance System: A Deep Learning Approach," *IEEE Network*, vol. 35, no. 3, pp. 74–80, 2021.

[56] B. Brik, P. A. Frangoudis, and A. Ksentini, "Service-Oriented MEC Applications Placement in a Federated Edge Cloud Architecture," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[57] U. Fattore, M. Liebsch, B. Brik, and A. Ksentini, "AutoMEC: LSTM-based user mobility prediction for service management in distributed MEC resources," in *Proceedings of the 23rd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2020, pp. 155–159.

[58] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, Jun. 2019.

[59] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," 2013. [Online]. Available: https://arxiv.org/abs/1312.6034

[60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[61] 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, pp. 3145–3153.

[62] Luisa M Zintgraf and Taco S Cohen and Tameem Adel and Max Welling, “Visualizing Deep Neural Network Decisions: Prediction Difference Analysis,” in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=BJ5UeU9xx

[63] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[64] Yi Huang and Adams Wai-Kin Kong, “Transferable Adversarial Attack based on Integrated Gradients,” in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=DesNW4-5ai9

[65] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “Smoothgrad: removing noise by adding noise,” *arXiv preprint arXiv:1706.03825*, 2017.

[66] Alexander Levine and Sahil Singla and Soheil Feizi, “Certifiably Robust Interpretation in Deep Learning,” 2020. [Online]. Available: https://openreview.net/forum?id=rkxVz1HKwB

[67] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[68] M. P. Ayyar, J. Benois-Pineau, and A. Zemmari, “Review of white box methods for explanations of convolutional neural networks in image classification tasks,” *Journal of Electronic Imaging*, vol. 30, pp. 050 901 – 050 901, 2021.

[69] J. Castro, D. Gómez, and J. Tejada, “Polynomial Calculation of the Shapley Value Based on Sampling,” *Computers & Operations Research*, vol. 36, no. 5, pp. 1726–1730, 2009, selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

[70] N. Jethani, M. Sudarshan, I. C. Covert, S.-I. Lee, and R. Ranganath, “FastSHAP: Real-Time Shapley Value Estimation,” in *International Conference on Learning Representations*, 2022.

[71] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[72] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[73] X. Kong and X. Zhang, “Understanding Masked Image Modeling via Learning Occlusion Invariant Feature,” 2023.

[74] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, *GNNExplainer: Generating Explanations for Graph Neural Networks*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[75] M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why should I trust you?" Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[76] H. Jia, H. Chen, J. Guan, A. S. Shamsabadi, and N. Papernot, “A Zest of LIME: Towards Architecture-Independent Model Distances,” in *International Conference on Learning Representations*, 2022.

[77] M. W. Craven and J. W. Shavlik, “Extracting Tree-Structured Representations of Trained Networks,” in *NIPS*, 1995.

[78] Z. A. El Houda, B. Brik, and S.-M. Senouci, “A Novel IoT-Based Explainable Deep Learning Framework for Intrusion Detection Systems,” *IEEE Internet of Things Magazine*, vol. 5, no. 2, pp. 20–23, 2022.

[79] Z. A. E. Houda, B. Brik, and L. Khoukhi, “"Why Should I Trust Your IDS?": An Explainable Deep Learning Framework for Intrusion Detection Systems in Internet of Things Networks,” *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1164–1176, 2022.

[80] A. Gupta, A. Pacchiano, Y. Zhai, S. Kakade, and S. Levine, “Unpacking Reward Shaping: Understanding the Benefits of Reward Engineering on Sample Complexity,” in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 15 281–15 295.

[81] Y. Hu, W. Wang, H. Jia, Y. Wang, Y. Chen, J. Hao, F. Wu, and C. Fan, “Learning to Utilize Shaping Rewards: A New Approach of Reward Shaping,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 15 931–15 941.

[82] A. Mott, D. Zoran, M. Chrzanowski, D. Wierstra, and D. Jimenez Rezende, “Towards Interpretable Reinforcement Learning Using Attention Augmented Agents,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[83] A. Manchin, E. Abbasnejad, and A. van den Hengel, “Reinforcement Learning with Attention that Works: A Self-Supervised Approach,” in *Neural Information Processing*, T. Gedeon, K. W. Wong, and M. Lee, Eds. Cham: Springer International Publishing, 2019, pp. 223–230.

[84] K. Cyras, R. Badrinath, S. K. Mohalik, A. Mujumdar, A. Nikou, A. Previti, V. Sundararajan, and A. V. Feljan, “Machine Reasoning Explainability,” in *AAMAS 2021*. IFAAMAS, 2021.

[85] D. A. Hudson and C. D. Manning, “Compositional Attention Networks for Machine Reasoning,” in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=S1Euwz-Rb

[86] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, “Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned,” in *Annual Meeting of the Association for Computational Linguistics*, Jul. 2019, pp. 5797–5808.

[87] R. Achtibat, S. M. V. Hatefi, M. Dreyer, A. Jain, T. Wiegand, S. Lapuschkin, and W. Samek, “AttnLRP: Attention-Aware Layer-wise Relevance Propagation for Transformers,” *ArXiv*, vol. abs/2402.05602, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267547813

[88] J. Krause, A. Perer, and K. Ng, “Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models,” in *ACM Conference on Human Factors in Computing Systems*, 2016, pp. 5686–5697.

[89] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, “Explainable Reinforcement Learning Through a Causal Lens,” 2019.

[90] A. Bennetot, J.-L. Laurent, R. Chatila, and N. Diaz-Rodriguez, “Towards Explainable Neural-Symbolic Visual Reasoning,” 2019.

[91] Y. Dong, H. Su, J. Zhu, and B. Zhang, “Improving Interpretability of Deep Neural Networks with Semantic Information,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 975–983.

[92] I. Tiddi, F. Lécué, and P. Hitzler, Eds., *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, ser. Studies on the Semantic Web. IOS Press, 2020, vol. 47.

[93] F. Bianchi, G. Rossiello, L. Costabello, M. Palmonari, and P. Minervini, “Knowledge graph embeddings and explainable AI,” *CoRR*, vol. abs/2004.14843, 2020.

[94] A. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O’Reilly Media, Incorporated, 2016. [Online]. Available: https://books.google.es/books?id=q5pnAQAACAAJ

[95] T. Chen and C. Guestrin, “XGBoost,” in *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, aug 2016. [Online]. Available: https://doi.org/10.11452F2939672.2939785

[96] P. Cunningham and S. J. Delany, “k-Nearest Neighbour Classifiers - A Tutorial,” *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–25, Jul. 2021. [Online]. Available: https://doi.org/10.11452F3459665

[97] Z. Wang, W. Zhang, N. Liu, and J. Wang, “Scalable rule-based representation learning for interpretable classification,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 30 479–30 491, 2021.

[98] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton, “Neural additive models: Interpretable machine learning with neural nets,” *Advances in neural information processing systems*, vol. 34, pp. 4699–4711, 2021.

[99] M. C. A. Clare, M. Sonnewald, R. Lguensat, J. Deshayes, and V. Balaji, “Explainable Artificial Intelligence for Bayesian Neural Networks: Toward Trustworthy Predictions of Ocean Dynamics,” *Journal of Advances in Modeling Earth Systems*, vol. 14, no. 11, oct 2022. [Online]. Available: https://doi.org/10.10292F2022ms003162

[100] D. Alvarez-Melis and T. Jaakkola, “Towards Robust Interpretability with Self-Explaining Neural Networks,” *ArXiv*, vol. abs/1806.07538, 2018.

[101] D. Vale, A. El-Sharif, and M. Ali, “Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law,” *AI and Ethics*, pp. 1–12, 2022.

[102] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” 2015.

[103] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Interpretable machine learning: definitions, methods, and applications,” *ArXiv*, vol. abs/1901.04592, 2019.

[104] W. Jin, J. Fan, D. Gromala, P. Pasquier, and G. Hamarneh, “EUCA: A Practical Prototyping Framework towards End-User-Centered Explainable Artificial Intelligence,” *ArXiv*, vol. abs/2102.02437, 2021.

[105] N. Tamani, B. Brik, N. Lagraa, and Y. Ghamri-Doudane, “Vehicular Cloud Service Provider Selection: A Flexible Approach,” in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[106] J. Gerlach, P. Hoppe, S. Jagels, L. Licker, and M. H. Breitner, “Decision support for efficient XAI services - A morphological analysis, business model archetypes, and a decision tree,” *Electron. Mark.*, vol. 32, no. 4, pp. 2139–2158, 2022. [Online]. Available: https://doi.org/10.1007/s12525-022-00603-6

[107] A. H. Mohammadkhani, N. S. Bommi, M. Daboussi, O. Sabnis, C. Tantithamthavorn, and H. Hemmati, “A Systematic Literature Review of Explainable AI for Software Engineering,” 2023.

[108] C. A. Zhang, S. Cho, and M. Vasarhelyi, “Explainable Artificial Intelligence (XAI) in auditing,” *International Journal of Accounting Information Systems*, vol. 46, p. 100572, 2022.

[109] J. Collenette, K. Atkinson, and T. Bench-Capon, “Explainable AI tools for legal reasoning about cases: A study on the European Court of Human Rights,” *Artificial Intelligence*, vol. 317, p. 103861, 2023.

[110] N. A. Smuha, “The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence,” *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, 2019. [Online]. Available: https://ssrn.com/abstract=3443537

[111] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *International Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 4768–4777.

[112] S. M. Lundberg, G. G. Erion, and S.-I. Lee, “Consistent Individualized Feature Attribution for Tree Ensembles,” 2019.

[113] J. H. Friedman and B. E. Popescu, “Predictive Learning via Rule Ensembles,” *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 916–954, 2008.

[114] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” *CoRR*, vol. abs/1703.01365, 2017. [Online]. Available: http://arxiv.org/abs/1703.01365

[115] S. Jha, S. Raj, S. L. Fernandes, S. K. Jha, S. Jha, B. Jalaian, G. Verma, and A. Swami, *Attribution-Based Confidence Metric for Deep Neural Networks*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[116] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning Important Features Through Propagating Activation Differences,” *CoRR*, vol. abs/1704.02685, 2017.

[117] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, “ERASER: A Benchmark to Evaluate Rationalized NLP Models,” *CoRR*, vol. abs/1911.03429, 2019.

[118] U. Bhatt, A. Weller, and J. M. F. Moura, “Evaluating and Aggregating Feature-based Model Explanations,” in *International Joint Conference on Artificial Intelligence*, 2020 . [Online]. Available: https://api.semanticscholar.org/CorpusID:218486810

[119] F. Rezazadeh, H. Chergui, L. Alonso, and C. Verikoukis, “SliceOps: Explainable MLOps for Streamlined Automation-Native 6G Networks,” 2023. [Online]. Available: https://arxiv.org/abs/2307.01658

[120] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, “On the (in)Fidelity and Sensitivity of Explanations,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[121] M. Velmurugan, C. Ouyang, C. Moreira, and R. Sindhgatta, “Developing a Fidelity Evaluation Approach for Interpretable Machine Learning,” *CoRR*, vol. abs/2106.08492, 2021.

[122] T. Fel and D. Vigouroux, “Representativity and Consistency Measures for Deep Neural Network Explanations,” *CoRR*, vol. abs/2009.04521, 2020.

[123] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002 , pp. 311–318.

[124] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[125] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, R. L. Mercer, and P. Roossin, “A statistical approach to language translation,” in *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*, 1988 .

[126] S. Roy, H. Chergui, A. Ksentini, and C. Verikoukis, “Federated Machine Reasoning for Resource Provisioning in 6G O-RAN,” 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270371415

[127] J. X. Salvat Lozano, J. A. Ayala-Romero, L. Zanzi, A. Garcia-Saavedra, and X. Costa-Perez, “O-RAN experimental evaluation datasets,” 2022. [Online]. Available: https://dx.doi.org/10.21227/64s5-q431

[128] 3rd Generation Partnership Project (3GPP), “Technical Specification Group Radio Access Network; Study on New Radio Access Technology: Radio Access Architecture and Interfaces,” 3GPP, Technical Specification (TS) 38.300, 2018. [Online]. Available: https://www.3gpp.org/ftp/Specs/html-info/38300.htm

[129] O.-R. Alliance, “O-RAN Alliance Specification: Lower Layer Split (LLS) Option 7-2x,” Specification, 2023. [Online]. Available: https://www.o-ran.org/specifications

[130] K. Boutiba, A. Ksentini, B. Brik, Y. Challal, and A. Balla, “NRflex: Enforcing network slicing in 5G New Radio,” *Computer Communications*, vol. 181, pp. 284–292, 2022.

[131] O.-R. W. G. 2, “O-RAN AI/ML workflow description and requirements 1.03,” *O-RAN. WG2. AIML-v01. 03 Technical Specification*, 2021.

[132] IEEE P2894, “Guide for an Architectural Framework for Explainable Artificial Intelligence,” https://standards.ieee.org/ieee/2894/10284/.

[133] IEEE P2976, “Standard for XAI, eXplainable Artificial Intelligence, for Achieving Clarity and Interoperability of AI Systems Design,” https://standards.ieee.org/ieee/2976/10522/.

[134] O. Alliance, “O-RAN fronthaul control, user and synchronization plane specification 8.0,” *Specification WG4: Open Fronthaul Interfaces Workgroup*, 2022.

[135] 6G-BRICKS, 2023. [Online]. Available: https://6g-bricks.eu/

[136] NANCY, 2023. [Online]. Available: https://cordis.europa.eu/project/id/101096456

[137] H.-X. project, “OpenFL-XAI,” 2023 . [Online]. Available: https://github.com/Unipisa/OpenFL-XAI

[138] S. D’Oro, M. Polese, L. Bonati, H. Cheng, and T. Melodia, “dApps: Distributed Applications for Real-Time Inference and Control in O-RAN,” *IEEE Communications Magazine*, vol. 60, no. 11, pp. 52–58, 2022.

[139] J. L. C. 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)*. IEEE, 2022, pp. 1–8.

[140] 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.

[141] Z. A. E. Houda, H. Moudoud, and B. Brik, “Federated deep reinforcement learning for efficient jamming attack mitigation in o-ran,” *IEEE Transactions on Vehicular Technology*, vol. 73, no. 7, pp. 9334–9343, 2024.

[142] A. A. Ismail, H. Corrada Bravo, and S. Feizi, "Improving deep learning interpretability by saliency guided training," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 726–26 739, 2021.

[143] J. Sun, S. Lapuschkin, W. Samek, Y. Zhao, N.-M. Cheung, and A. Binder, "Explanation-Guided Training for Cross-Domain Few-Shot Classification," *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 7609–7616, 2020.

[144] S. Roy, H. Chergui, and C. Verikoukis, "Explanation-guided fair federated learning for transparent 6g ran slicing," *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, 2024.

[145] O. Adamuz-Hinojosa, L. Zanzi, V. Sciancalepore, A. Garcia-Saavedra, and X. Costa-Pérez, "ORANUS: Latency-tailored Orchestration via Stochastic Network Calculus in 6G O-RAN," *IEEE International Conference on Computer Communications (INFOCOM)*, 2024.

[146] S. Roy, H. Chergui, and C. Verikoukis, "Towards Bridging the FL Performance-Explainability Trade-Off: A Trustworthy 6G RAN Slicing Use-Case," *IEEE Transactions on Vehicular Technology*, pp. 1–11, 2024.

[147] Z. A. E. Houda, B. Brik, A. Ksentini, L. Khoukhi, and M. Guizani, "When Federated Learning Meets Game Theory: A Cooperative Framework to Secure IIoT Applications on Edge Computing," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 7988–7997, 2022.

[148] B. Brik, M. Messaadia, M. Sahnoun, B. Bettayeb, and M. A. Benatia, "Fog-Supported Low-Latency Monitoring of System Disruptions in Industry 4.0: A Federated Learning Approach," *ACM Trans. Cyber-Phys. Syst.*, vol. 6, no. 2, may 2022. [Online]. Available: https://doi.org/10.1145/3477272

[149] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young, "Machine Learning: The High Interest Credit Card of Technical Debt," in *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, 2014.

[150] "MLOps: Continuous delivery and automation pipelines in machine learning," Dec. 2022. [Online]. Available: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

[151] P. Li, J. Thomas, X. Wang, A. Khalil, A. Ahmad, R. Inacio, S. Kapoor, A. Parekh, A. Doufexi, A. Shojaeifard, and R. J. Piechocki, "RLOps: Development Life-Cycle of Reinforcement Learning Aided Open RAN," *IEEE Access*, vol. 10, pp. 113 808–113 826, 2022.

[152] P. Porambage, J. Pinola, Y. Rumesh, C. Tao, and J. Huusko, "XcARet: XAI based Green Security Architecture for Resilient Open Radio Access Networks in 6G," in *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2023, pp. 699–704.

[153] J. Groen, S. DOro, U. Demir, L. Bonati, M. Polese, T. Melodia, and K. Chowdhury, "Implementing and Evaluating Security in O-RAN: Interfaces, Intelligence, and Platforms," 2023.

[154] N. M. Yungaicela-Naula, V. Sharma, and S. Scott-Hayward, "Misconfiguration in O-RAN: Analysis of the impact of AI/ML," *Computer Networks*, vol. 247, p. 110455, 2024.

[155] C. Fiandrino, L. Bonati, S. D'Oro, M. Polese, T. Melodia, and J. Widmer, "EXPLORA: AI/ML EXPLainability for the Open RAN," *Proceedings of the ACM on Networking*, vol. 1, no. CoNEXT3, pp. 1–26, 2023.

[156] F. Rezazadeh, H. Chergui, S. Siddiqui, J. Mangues, H. Song, W. Saad, and M. Bennis, "Intelligible Protocol Learning for Resource Allocation in 6G O-RAN Slicing," 2023.

[157] C. Tassie, B. Kim, J. Groen, M. Belgiovine, and K. R. Chowdhury, "Leveraging Explainable AI for Reducing Queries of Performance Indicators in Open RAN," in *IEEE International Conference on Communications (ICC)*, June 2024.

[158] P. F. Pérez, C. Fiandrino, M. Fiore, and J. Widmer, "An In-Depth Analysis of Advanced Time Series Forecasting Models for the Open RAN," in *IEEE INFOCOM 2024*, 2024 .

[159] S. Roy, F. Rezazadeh, H. Chergui, and C. Verikoukis, "Joint Explainability and Sensitivity-Aware Federated Deep Learning for Transparent 6G RAN Slicing," in *ICC 2023 - IEEE International Conference on Communications*, 2023, pp. 1238–1243.

[160] Y. Cao, S.-Y. Lien, Y.-C. Liang, K.-C. Chen, and X. Shen, "User Access Control in Open Radio Access Networks: A Federated Deep Reinforcement Learning Approach," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 3721–3736, 2022.

[161] Y. Cao, S.-Y. Lien, Y.-C. Liang, and K.-C. Chen, "Federated Deep Reinforcement Learning for User Access Control in Open Radio Access Networks," in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–6.

[162] A. Scalingi, S. D'Oro, F. Restuccia, T. Melodia, and D. Giustiniano, "Det-RAN: Data-Driven Cross-Layer Real-Time Attack Detection in 5G Open RANs," in *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications*, May 2024.

[163] S. Maxenti, S. D'Oro, L. Bonati, M. Polese, A. Capone, and T. Melodia, "ScalO-RAN: Energy-aware Network Intelligence Scaling in Open RAN," 2024. [Online]. Available: https://arxiv.org/abs/2312.05096

[164] H. Cheng, P. Johari, M. A. Arfaoui, F. Periard, P. Pietraski, G. Zhang, and T. Melodia, "Real-Time AI-Enabled CSI Feedback Experimentation with Open RAN," in *2024 19th Wireless On-Demand Network Systems and Services Conference (WONS)*, 2024, pp. 121–124.

[165] H. Lee, Y. Jang, J. Song, and H. Yeon, "O-RAN AI/ML Workflow Implementation of Personalized Network Optimization via Reinforcement Learning," in *2021 IEEE Globecom Workshops (GC Wkshps)*, 2021, pp. 1–6.

[166] R. Raftopoulos, S. D'Oro, T. Melodia, and G. Schembra, "DRL-based Latency-Aware Network Slicing in O-RAN with Time-Varying SLAs," 2024. [Online]. Available: https://arxiv.org/abs/2401.05042

[167] S. Mollahasani, M. Erol-Kantarci, and R. Wilson, "Dynamic CU-DU Selection for Resource Allocation in O-RAN Using Actor-Critic Learning," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[168] S. D'Oro, L. Bonati, M. Polese, and T. Melodia, "OrchestRAN: Network automation through orchestrated intelligence in the open RAN," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 270–279.

[169] P. E. Iturria-Rivera, H. Zhang, H. Zhou, S. Mollahasani, and M. Erol-Kantarci, "Multi-Agent Team Learning in Virtualized Open Radio Access Networks (O-RAN)," *Sensors*, vol. 22, no. 14, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/14/5375

[170] P. E. I. Rivera, S. Mollahasani, and M. Erol-Kantarci, "Multi Agent Team Learning in Disaggregated Virtualized Open Radio Access Networks (O-RAN)," *CoRR*, vol. abs/2012.04861, 2020. [Online]. Available: https://arxiv.org/abs/2012.04861

[171] H. Zhang, H. Zhou, and M. Erol-Kantarci, "Federated Deep Reinforcement Learning for Resource Allocation in O-RAN Slicing," 2022. [Online]. Available: https://arxiv.org/abs/2208.01736

[172] ——, "Team Learning-Based Resource Allocation for Open Radio Access Network (O-RAN)," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 4938–4943.

[173] A. Abouaomar, A. Taik, A. Filali, and S. Cherkaoui, "Federated Deep Reinforcement Learning for Open RAN Slicing in 6G Networks," 2022. [Online]. Available: https://arxiv.org/abs/2206.11328

[174] X. Wang, J. D. Thomas, R. J. Piechocki, S. Kapoor, R. Santos-Rodríguez, and A. Parekh, "Self-play learning strategies for resource assignment in Open-RAN networks," *Computer Networks*, vol. 206, p. 108682, 2022.

[175] M. Tsampazi, S. D'Oro, M. Polese, L. Bonati, G. Poitau, M. Healy, and T. Melodia, "PandORA: Automated Design and Comprehensive Evaluation of Deep Reinforcement Learning Agents for Open RAN," *ResearchGate Preprint*, pp. 1–18, May 2024.

[176] Y. Chen, Y. Gao, C. Jiang, and K. J. R. Liu, "Game theoretic Markov decision processes for optimal decision making in social systems," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 268–272.

[177] A. Krajna, M. Brcic, T. Lipic, and J. Doncevic, "Explainability in reinforcement learning: perspective and position," 2022.

[178] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy Gradient Methods for Reinforcement Learning with Function Approximation," in *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller, Eds., vol. 12. MIT Press, 1999.

[179] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. Doshi-Velez, "Explainable Reinforcement Learning via Reward Decomposition," in

in proceedings at the International Joint Conference on Artificial Intelligence. A Workshop on Explainable Artificial Intelligence., 2019.

[180] S. Sloman, *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press, 08 2005.

[181] J. van der Waa, J. van Diggelen, K. v. d. Bosch, and M. Neerincx, "Contrastive explanations for reinforcement learning in terms of expected consequences," *arXiv preprint arXiv:1807.08706*, 2018.

[182] T. Shu, C. Xiong, and R. Socher, "Hierarchical and Interpretable Skill Acquisition in Multi-task Reinforcement Learning," 2017.

[183] V. Zambaldi, D. Raposo, A. Santoro, V. Bapst, Y. Li, I. Babuschkin, K. Tuyls, D. Reichert, T. Lillicrap, E. Lockhart, M. Shanahan, V. Langston, R. Pascanu, M. Botvinick, O. Vinyals, and P. Battaglia, "Relational Deep Reinforcement Learning," 2018.

[184] O-RAN Alliance, "O-RAN Working Group (O-RAN.WG1.Use-Cases-Analysis-Report-v09.00)," Technical Specification, Oct. 2022.

[185] O. Izima, R. de Fréin, and A. Malik, "A Survey of Machine Learning Techniques for Video Quality Prediction from Quality of Delivery Metrics," *Electronics*, vol. 10, no. 22, p. 2851, 2021.

[186] A. Renda, P. Ducange, G. Gallo, and F. Marcelloni, "XAI models for quality of experience prediction in wireless networks," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2021, pp. 1–6.

[187] O. Ayoub, S. Troia, D. Andreoletti, A. Bianco, M. Tornatore, S. Giordano, and C. Rottondi, "Towards explainable artificial intelligence in optical networks: the use case of lightpath QoT estimation," *Journal of Optical Communications and Networking*, vol. 15, no. 1, pp. A26–A38, 2022.

[188] O-RAN Alliance, "O-RAN-WG1-O-RAN Slicing Architecture v03.00," Technical Specification, 2021.

[189] S. B. Saad, A. Ksentini, and B. Brik, "A Trust architecture for the SLA management in 5G networks," in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–6.

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

[191] A. Terra, R. Inam, P. Batista, and E. Fersman, "Using Counterfactuals to Proactively Solve Service Level Agreement Violations in 5G Networks," in *IEEE International Conference on Industrial Informatics (INDIN)*. IEEE, 2022, pp. 552–559.

[192] S.-p. Yeh, S. Bhattacharya, R. Sharma, and H. Moustafa, "Deep Learning for Intelligent and Automated Network Slicing in 5G Open RAN (ORAN) Deployment," *IEEE Open Journal of the Communications Society*, 2023.

[193] M. Q. Hamdan, H. Lee, D. Triantafyllopoulou, R. Borralho, A. Kose, E. Amiri, D. Mulvey, W. Yu, R. Zitouni, R. Pozza *et al.*, "Recent advances in machine learning for network automation in the o-ran," *Sensors*, vol. 23, no. 21, p. 8792, 2023.

[194] N. Khan, S. Coleri, A. Abdallah, A. Celik, and A. M. Eltawil, "Explainable and Robust Artificial Intelligence for Trustworthy Resource Management in 6G Networks," *IEEE Communications Magazine*, 2023.

[195] Z. A. El Houda, L. Khoukhi, and B. Brik, "A Low-Latency Fog-based Framework to secure IoT Applications using Collaborative Federated Learning," in *2022 IEEE 47th Conference on Local Computer Networks (LCN)*, 2022, pp. 343–346.

[196] B. Brik and A. Ksentini, "On Predicting Service-oriented Network Slices Performances in 5G: A Federated Learning Approach," in *2020 IEEE 45th Conference on Local Computer Networks (LCN)*, 2020, pp. 164–171.

[197] A. Okic, L. Zanzi, V. Sciancalepore, A. Redondi, and X. Costa-Pérez, "π-ROAD: a Learn-as-You-Go Framework for On-Demand Emergency Slices in V2X Scenarios," in *IEEE International Conference on Computer Communications*, Jul. 2021.

[198] Y. Cao, S.-Y. Lien, Y.-C. Liang, K.-C. Chen, and X. Shen, "User Access Control in Open Radio Access Networks: A Federated Deep Reinforcement Learning Approach," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 3721–3736, 2022.

[199] O-RAN Alliance, "O-RAN-WG11-O-RAN Security Requirements Specifications v04.00)," Technical Specification, 2022.

[200] S. B. Saad, B. Brik, and A. Ksentini, "Toward Securing Federated Learning against Poisoning Attacks in Zero Touch B5G networks," *IEEE Transactions on Network and Service Management*, pp. 1–1, 2023.

[201] O-RAN Alliance, "O-RAN-WG11-O-RAN Security Threat Modeling and Remediation Analysis (O-RAN.WG11.Threat-Model-v04.00)," Technical Specification, 2022.

[202] J. Pérez-Valero, A. Virdis, A. G. Sanchez, C. Ntogkas, P. Serrano, G. Landi, S. Kuklinski, C. Morin, I. L. Pavon, and B. Sayadi, "AI-driven Orchestration for 6G Networking: the Hexa-X vision," in *IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 1335–1340.

[203] T. Senevirathna, Z. Salazar, V. H. La, S. Marchal, B. Siniarski, M. Liyanage, and S. Wang, "A survey on XAI for beyond 5G security: technical aspects, use cases, challenges and research directions," *arXiv preprint arXiv:2204.12822*, 2022.

[204] O-RAN Security Focus Group, "Study on Security for Non-RT-RIC (O-RAN.WG11.Security-Requirements-Specifications-v04.00)," Technical Specification, 2022.

[205] ——, "Study on Security for O-Cloud (O-RAN.WG11.Security-Requirements-Specifications-v04.00)," Technical Specification, 2022.

[206] ——, "Study on Security for Near Real-Time RIC and xApps (O-RAN.SFG.Security-for-Near-RT-RIC-xApps-v01.00)," Technical Specification, 2022.

[207] Köpsell, Stefan and Ruzhanskiy, Andrey and Hecker, Andreas and Stachorra, Dirk and Franchi, Norman, "Open RAN Risk Analysis," Report, Feb. 2022.

[208] NIS Cooperation Group, "Report on the cybersecurity of Open RAN," Report, May 2022.

[209] 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.

[210] E. Habler, R. Bitton, D. Avraham, D. Mimran, E. Klevansky, O. Brodt, H. Lehmann, Y. Elovici, and A. Shabtai, "Adversarial Machine Learning Threat Analysis and Remediation in Open Radio Access Network (O-RAN)," *arXiv preprint arXiv:2201.06093*, 2022.

[211] J. Chen, M. I. Jordan, and M. J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," in *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 1277–1294.

[212] A. Chiejina, B. Kim, K. Chowhdury, and V. K. Shah, "System-level Analysis of Adversarial Attacks and Defenses on Intelligence in O-RAN based Cellular Networks," *arXiv preprint arXiv:2402.06846*, 2024.

[213] Y. Mirsky, A. Demontis, J. Kotak, R. Shankar, D. Gelei, L. Yang, X. Zhang, M. Pintor, W. Lee, Y. Elovici *et al.*, "The threat of offensive AI to organizations," *Computers & Security*, p. 103006, 2022.

[214] A. Kuppa and N.-A. Le-Khac, "Black box attacks on explainable artificial intelligence (XAI) methods in cyber security," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[215] X. Zhao, W. Zhang, X. Xiao, and B. Lim, "Exploiting explanations for model inversion attacks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 682–692.

[216] P. Giudici and E. Raffinetti, "Explainable AI methods in cyber risk management," *Quality and reliability engineering international*, vol. 38, no. 3, pp. 1318–1326, 2022.

[217] M. Naylor, C. French, S. Terker, and U. Kamath, "Quantifying Explainability in NLP and Analyzing Algorithms for Performance-Explainability Tradeoff," in *International Conference on Machine Learning*, 2021.

[218] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large Language Models for Telecom: The Next Big Thing?" 2023. [Online]. Available: https://arxiv.org/abs/2306.10249

[219] Nokia, "Nokia opens new O-RAN Collaboration and Testing Center in the U.S." https://www.nokia.com/about-us/news/releases/2021 /06/16/nokia-opens-new-o-ran-collaboration-and-testing-center-in-the-us/.

[220] W. Hammedi, B. Brik, and S. M. Senouci, "Federated Deep Learning-Based Framework to Avoid Collisions Between Inland Ships," in *2022 International Wireless Communications and Mobile Computing (IWCMC)*, 2022, pp. 967–972.

[221] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, “Learning Differentially Private Recurrent Language Models,” in *International Conference on Learning Representations*, 2017.
[222] K. Zhang, X. Song, and C. e. a. Zhang, “Challenges and future directions of secure federated learning: a survey.” *Front. Comput. Sci.*, vol. 16, no. 5, Dec 2022.