Co-funded by the European Union

SEamless integratioN of efficient 6G WirelesS tEchnologies for Communication and Sensing

D4.4 Design and implementation of the 6G-SENSES Intelligent Plane

July 2026

6G-SENSES project has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement 101139282

Project Start Date: 2024-01-01 **Duration**: 30 months

Call: HORIZON-JU-SNS-2023 **Date of delivery**: 2026-07-24

Topic: HORIZON-JU-SNS-2023-STREAM-B-01-02 **Version**: 1.0

Dissemination Level: Public (PU)

Grant Agreement Number:	101139282
Project Name:	SEamless integratioN of efficient 6G WirelesS tEchnologies for Communication and Sensing
Project Acronym:	6G-SENSES
Document Number:	**D4.4**
Document Title:	Design and implementation of the 6G-SENSES Intelligent Plane
Version:	1.0
Delivery Date:	2026-06-30 (**2026-07-24**)
Responsible:	*Accelleran* (**ACC**)
Editor(s):	German Castellanos (**ACC**)
Authors:	Pavlos Doanis, Markos Anastasopoulos, Anna Tzanakaki (**IASA**), Ramón Agüero, Luis Diez (**UC**), Khai Nguyen (**BR**), Jesús Gutiérrez (**IHP**), Muhammad Saadi (**NTU**), German Castellanos (**ACC**).
Keywords:	6G, Sensing, Intelligent Plane
Status:	Final
Dissemination Level	Public (PU)
Project URL:	https://www.6g-senses.eu

Revision History

Rev. N	Description	Author	Date
0.1	First draft layout with ToC	Simon Pryor (ACC)	2026-04-28
0.2	First contributions from partners in the definition of IP and SoTA	German Castellanos (ACC), ALL	2026-05-01
0.3	Complementation of the definition, implementation	German Castellanos (ACC), ALL Jesús Gutiérrez (IHP)	2026-06-01
0.4	Initial modification of contributions for the Enablers and implementation.	ALL	2026-06-12
0.5	Version ready for first full reviewers.	German Castellanos (ACC)	2026-06-17
0.6	Version including response from reviewers	Anna Tzanakaki (IASA), Navid Nikaein (BR)	2026-07-20
1.0	Last revision and submission of the document	Jesús Gutiérrez (IHP)	2026-07-24

Table of Contents

List of Figures

List of Tables

Executive Summary

This deliverable defines the architectural and implementation direction for introducing Artificial Intelligence (AI)/Machine Learning (ML)-driven intelligence into the 6G-SENSES system. The document positions the Intelligent Plane as the project framework for collecting network and sensing data, managing AI/ML models, executing inference, and connecting model outputs to network optimisation and control actions. Rather than defining the Intelligent Plane as an isolated component, the deliverable describes it as a distributed operational capability aligned with the 6G-SENSES 3GPP and Open Radio Access Network (O-RAN)-based architecture.

The starting point of the deliverable is the 6G-SENSES architectural vision, where communication, sensing, computing and orchestration functions must operate together across a heterogeneous infrastructure. The project combines 3GPP and non-3GPP (N3GPP) access technologies, with Integrated Sensing and Communication (ISAC) capabilities, Reconfigurable Intelligent Surface (RIS)-assisted operation and Cell-Free Multiple-Input Multiple-Output (CF-MIMO) concepts. These technologies generate heterogeneous telemetry and sensing information that can be exploited to improve network planning, energy efficiency, sensing-aware control and service optimisation.

The deliverable establishes the design principles of the Intelligent Plane around four main ideas. First, intelligence should be aligned with open and programmable network architectures. Second, the Intelligent Plane must provide multi-domain observability, collecting information from RAN, sensing, core, transport, edge/cloud and external data sources. Third, it must support the complete AI/ML lifecycle, from data preparation to inference and retraining. Fourth, intelligence must be distributed across different timescales and execution locations, from long-term analytics in the Service and Management Orchestrator (SMO) to near-real-time inference and control in the Near-Real-Time (RT) Radio Intelligent Controller (RIC) and, potentially, closer-to-RAN execution for future native AI/ML.

A significant part of the document reviews the relevant state of the art in 3GPP, O-RAN and related Smart Networks and Services Joint Undertaking (SNS JU) activities. In 3GPP, functions such as the Network Data Analytics Function (NWDAF) and the Analytics Data Repository Function (ADRF) provide analytics and data repository capabilities in the 5G Core Network (CN) and service domains. These functions complement the O-RAN RIC-based approach by enabling cross-domain visibility beyond the RAN. In O-RAN, the SMO, Non-RT RIC and Near-RT RIC provide the most direct framework for programmable network intelligence, with rApps and xApps supporting long-term analytics, policy generation, near-real-time inference and control. The deliverable also considers the evolution towards native AI, cross-domain AI, dApps, RT-RIC-like concepts and AI/ML lifecycle management. These developments are important because 6G-SENSES requires intelligence that can operate across communication, sensing, edge, transport and core domains, not only within isolated RAN optimisation loops.

A significant part of the document reviews the relevant state of the art in 3GPP, O-RAN and related SNS-JU activities. The deliverable also considers the evolution towards native AI, cross-domain AI, dApps, RT-RIC-like concepts and AI/ML lifecycle management. These developments are important because 6G-SENSES requires intelligence that can operate across communication, sensing, edge, transport and core domains, not only within isolated RAN optimisation loops. D4.4 also identifies the main technical enablers required to make the Intelligent Plane operational. These include the AI/ML techniques developed in WP4. The diversity of these techniques confirms that the Intelligent Plane must support AI/ML workflows.

Machine Learning Operations (MLOps) is identified as a central technical enabler. In the 6G-SENSES context, MLOps is not treated as a generic software engineering practice, but as the operational backbone that allows AI/ML models to be transformed into reusable network intelligence functions. D4.4 introduces MLRun as a relevant MLOps framework since it supports data preparation, feature management, model training, model

registry, model serving and monitoring. Together, the document maintains an implementation-flexible view: model lifecycle functions can either be implemented through a structured MLOps platform such as MLRun, or through lighter rApp-based workflows. The Intelligent Plane provides a path from telemetry collection to dataset generation, model training and deployment, inference and feedback. It also supports validation through the Network Digital Twin (NDT), reducing the risk of applying unsafe or sub-optimal configurations directly to the live system. In this way, **D4.4** acts as a bridge between the AI/ML methods developed in WP4, the infrastructure and sensing capabilities developed in WP3, and the demonstrator activities planned in WP5.

The deliverable concludes that the 6G-SENSES Intelligent Plane is both an architectural concept and an implementation framework. Architecturally, it aligns 6G-SENSES with the evolution of O-RAN, 3GPP analytics, native AI/ML and cross-domain intelligence. Technically, it identifies the data, MLOps, telemetry, model lifecycle and control mechanisms needed to make AI/ML operations. Implementation wise, it defines how dRAX, rApps, xApps, MLRun and NDT components can be combined to support sensing-assisted, AI/ML-driven and energy-aware network optimisation, providing a baseline for the next phase of the project, where the Intelligent Plane will be refined, validated and integrated into the 6G-SENSES demonstrators.

1 Introduction

This section introduces the context, scope and positioning of the deliverable within the 6G-SENSES project. It explains why an Intelligent Plane is needed to support the ambition of the project of combining communication, sensing, computing and Artificial Intelligence / Machine Learning (AI/ML)-driven optimisation within an O-RAN/3GPP-aligned architecture. The section also clarifies how **D4.4** relates to the WP4 objectives and project Key Performance Indicators (KPIs), and how the deliverable is organised to move from architectural positioning towards technical enablers and implementation aspects.

1.1 Scope and Objectives of the Deliverable

This deliverable defines the design and implementation direction of the 6G-SENSES Intelligent Plane. Its scope is to describe how AI/ML-based intelligence can be introduced into the 6G-SENSES architecture as an operational capability, supporting the joint optimisation of communication, sensing, computing and orchestration functions. The work is framed by the project objective of integrating 6G RAN technologies, including ISAC, Cell-Free Multiple-Input Multiple-Output (CF-MIMO), Reconfigurable Intelligent Surface (RIS)-assisted operation and multi-technology access, within an architecture aligned with 3GPP and Open Radio Access Network (O-RAN).

The deliverable does not aim to redefine the overall 6G-SENSES system architecture, which is addressed in WP2, nor to duplicate the detailed algorithmic work carried out in other WP4 tasks. Instead, it focuses on the architectural and implementation layer that allows those AI/ML techniques to be deployed, managed and connected to network control actions. In this sense, **D4.4** acts as the bridge between the AI/ML solutions developed in WP4, the infrastructure capabilities provided by WP3, and the integration and demonstration activities planned in WP5. A central objective is to clarify the role of the Intelligent Plane within the 6G-SENSES control framework. This includes identifying where intelligence functions are placed, how they interact with O-RAN entities such as the Service Management Orchestrator (SMO), Non-Real Time (RT) Radio Intelligent Controller (RIC) and Near-RT RIC, and how rApps, xApps and future closer-to-RAN execution mechanisms may support different control-loop timescales. Particular attention is given to the use of sensing information and network telemetry as inputs for AI/ML-driven optimisation. Another objective is to identify the technical enablers required to make the Intelligent Plane operational. These include data collection and monitoring mechanisms, AI/ML lifecycle management, MLOps support, 3GPP analytics functions such as Network Data Analytics Function (NWDAF) and Analytics Data Repository Function (ADRF), O-RAN evolution towards cross-domain intelligence, and the use of NDT capabilities for validation and optimisation. The deliverable also considers how these enablers can support energy-aware operation, sensing-assisted control and adaptive network management.

From an implementation perspective, **D4.4** provides the first consolidated view of how the Intelligent Plane can be realised in 6G-SENSES. The document describes the role of the dRAX RIC platform, the use of telemetry databases and live data streams, the interaction between rApps and xApps, and the possible integration of structured MLOps components such as MLRun. This implementation view is intended to support the project demonstrators, especially PoC#3, where real network data, sensing inputs and NDT models are combined to assess and optimise network behaviour before applying changes to the operational system.

Overall, the objective of this deliverable is to provide a coherent design and implementation baseline for the 6G-SENSES Intelligent Plane. This baseline will support the initial assessment required by the project milestones and will serve as the basis for further refinement towards the final integration and demonstration activities.

1.2 Position within the 6G-SENSES Architecture

Within the 6G-SENSES architecture, the Intelligent Plane is positioned as the functional layer that connects sensing, communication, computing and orchestration with AI/ML-driven decision-making. It is not introduced as a standalone architectural block, but as a set of capabilities distributed across the management, control and data-processing domains of the system. Its role is to make the information generated by the 6G-SENSES infrastructure usable for optimisation, automation and validation.

The architectural baseline of the project combines 3GPP and N3GPP access technologies, including 5G NR, Wi-Fi, Sub-6 GHz and mmWave systems, together with ISAC functions, RIS-assisted operation and CF-MIMO concepts. These technologies generate communication and sensing information that can be exploited to improve network behaviour. The Intelligent Plane provides the mechanisms to collect, expose, process and act on this information, while remaining aligned with the O-RAN-based control adopted by the project.

In this context, the O-RAN framework provides the main architectural anchoring point. The SMO and Non-RT RIC support longer-term functions such as analytics, policy generation, orchestration, model lifecycle management and interaction with NDT capabilities. The Near-RT RIC provides the execution environment for xApps that require faster access to telemetry and sensing information, enabling near-real-time inference and closed-loop optimisation. Where lower-latency operation is required, future extensions closer to the RAN, such as dApps or RT-RIC-like mechanisms, may be considered.

The Intelligent Plane also acts as an integration bridge between the different technical work streams of 6G-SENSES (Figure 1-1). Specifically, from **WP3**, it receives inputs from the communication and sensing technologies developed in the project. Subsequently **WP4**, provides the architectural context in which AI/ML methods, MLOps workflows and optimisation mechanisms can be deployed. Towards **WP5**, it offers the operational support needed for demonstrators, especially where real network telemetry, sensing information and NDT models are used to assess configurations before applying them to the live network.

Its position in the architecture is therefore both horizontal and hierarchical. Horizontally, it correlates information across RAN, sensing, transport, edge/cloud and core network (CN)-related domains. Hierarchically, it distributes intelligence across different control-loop timescales, from non-real-time model training and policy generation to near-real-time inference and control. This allows 6G-SENSES to move beyond isolated AI/ML algorithms and towards an operational framework where sensing-aware and energy-aware optimisation can be progressively introduced into the network.

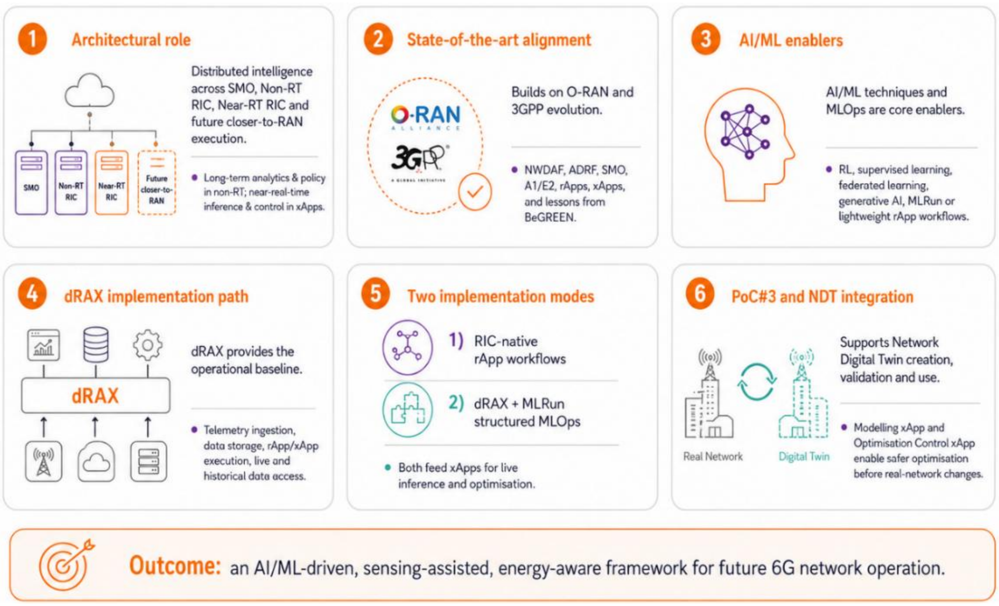


Figure 1-1 Summary of the Intelligent Plane contributions to the 6G-SENSES project

1.3 Relation to WP4 Objectives and Project KPIs

The 6G-SENSES proposal places the Intelligent Plane within the broader objective of bringing AI/ML capabilities into the operation of the RAN, user plane and data plane. This is not limited to the development of isolated algorithms. The project objective is to create the architectural and operational conditions under which AI/ML models can receive network information, interact with network functions, be deployed in the appropriate execution environment, and contribute to the optimisation of the 6G-SENSES infrastructure. Within this context, **WP4** provides the main technical work package for intelligence and energy-efficient AI/ML algorithms. The **WP4** objective most directly related to this deliverable is the provision of a Unified RAN Intelligence as a Service capability, deployable across the different 6G-SENSES building blocks. **D4.4** contributes to this objective by defining the Intelligent Plane as the framework that connects measurements, AI/ML model lifecycle management, O-RAN control functions and demonstrator integration.

Several **WP4** objectives are particularly relevant for this deliverable. **O4.1** requires the provision of fine-grain measurements that can be used as input features for AI/ML models supporting network management, configuration and optimisation. **D4.4** addresses this by identifying the telemetry and data-handling role of the Intelligent Plane, including the need to collect, expose and process network, sensing and infrastructure information before it can be used by AI/ML workflows. This is the starting point for any model-based optimisation, since the quality and availability of measurements directly condition the usefulness of the intelligence built on top of them. **O4.4** concerns the interfaces that allow ML models to interact with underlying network resources, with particular attention to O-RAN mechanisms and possible extensions that enable interaction with DU/CU elements. This deliverable supports that objective by positioning the Intelligent Plane around the SMO, Non-RT RIC, Near-RT RIC, rApps, xApps and related data/control interfaces. The intention is to clarify how model outputs can move from analytics or training environments towards operational network functions, instead of remaining detached from the control plane. **O4.7** introduces the need for co-sharing and co-development of ML models through suitable repositories. In D4.4, this is reflected in the discussion on MLOps and model lifecycle support. The Intelligent Plane must not only execute inference, but also support model storage, versioning, retraining and reuse across different applications and demonstrations. This is why the deliverable considers both lightweight rApp-based workflows and more structured MLOps approaches, such as MLRun, as possible implementation options. **O4.8** is the core objective for this deliverable, since it refers to the design and demonstration of the 6G-SENSES O-RAN-based Intelligent Plane supporting energy-efficient operation of the RAN infrastructure. **D4.4** directly addresses this by defining the architectural role of the Intelligent Plane and by describing how an implementation based on dRAX, rApps, xApps, telemetry databases and MLOps components can support AI/ML-driven optimisation. The focus is not only on energy efficiency as an isolated use case, but on establishing the reusable control-loop pattern required for sensing-aware and energy-aware network management. **O4.9** links the Intelligent Plane to demonstrators and testbed components for integration into WP5. This is an important boundary condition for **D4.4**: the deliverable must remain implementable, not only architectural. For this reason, the document connects the Intelligent Plane design to PoC#3, NDT support, live telemetry, historical datasets and the planned demonstrator activities. The purpose is to ensure that the Intelligent Plane can be exercised in realistic integration scenarios and can provide concrete support to WP5 validation.

These objectives are consolidated in Task T4.3, which is the **WP4** task directly associated with this deliverable. T4.3 provides the task-level bridge between the AI/ML objectives of **WP4** and the implementation-oriented content of **D4.4**. It frames the Intelligent Plane as the mechanism through which AI/ML methods, O-RAN control capabilities and energy-aware optimisation can be brought together in a deployable architecture.

The project KPIs most directly related to this deliverable are **KPI-O.3.3-1** and **KPI-O.3.3-2**. KPI-O.3.3-1 targets the integration of CU/DU with standardised interfaces for intelligent O-RAN support. **D4.4** contributes to this by identifying the role of O-RAN-aligned interfaces, RIC functions and dRAX-based components in exposing telemetry and enabling control interactions. KPI-O.3.3-2 targets the implementation of native AI/ML interfaces for the Intelligent user plane. This is addressed through the proposed model lifecycle, inference and control-loop mechanisms, where trained models can be consumed by operational applications and linked to network management decisions. Table 1-1, shows how the objectives and KPIs are addressed.

Table 1-1 How D4.4 addresses the 6G-SENSES WP4objectives

Obj.	Objective / KPI focus	Relevance for D4.4	How it is addressed in the deliverable
O4.1	Provision of fine-grain measurements as input features for AI/ML models supporting network management, configuration and optimisation.	Establishes the data foundation of the Intelligent Plane. Without reliable telemetry and sensing data, AI/ML workflows cannot support operational decisions.	D4.4 identifies the role of telemetry collection, data exposure, historical storage and live data streams. It links radio KPIs, infrastructure metrics, sensing information, events and logs to dataset creation, model training and inference.
O4.4	Definition of interfaces allowing ML models to interact with underlying network resources, including O-RAN mechanisms and possible DU/CU interaction.	Connects AI/ML outputs with network control, avoiding isolated model execution outside the operational architecture.	D4.4 positions the Intelligent Plane around SMO, Non-RT RIC, Near-RT RIC, rApps, xApps and O-RAN-aligned interfaces. It explains how model outputs can be consumed by control applications and translated into optimisation actions.
O4.7	Co-sharing and co-development of ML models through suitable repositories.	Supports reuse, governance and lifecycle management of models across different applications and demonstrators.	D4.4 introduces MLOps as a key Intelligent Plane capability, covering model storage, versioning, retraining and reuse. It considers both rApp-based workflows and structured MLOps support through MLRun.
O4.8	Design and demonstration of the 6G-SENSES O-RAN-based Intelligent Plane for energy-efficient RAN operation.	Core objective of D4.4. It defines the architectural and implementation basis of the Intelligent Plane.	D4.4 provides the Intelligent Plane design, its O-RAN alignment, the dRAX-based implementation direction, and the split between non-real-time learning and near-real-time inference/control.
O4.9	Development of demonstrators and testbed components related to the Intelligent Plane for WP5 integration.	Ensures that the Intelligent Plane is implementable and connected to validation activities, not only described architecturally.	D4.4 links the Intelligent Plane to PoC#3, WP5 demonstrators, NDT support, real telemetry, sensing inputs and validation-before-deployment workflows.
T4.3	Intelligent Plane for energy optimisation.	Provides the task-level home of D4.4 within WP4 and frames its implementation focus.	D4.4 translates T4.3 into an implementation path using dRAX, rApps, xApps, telemetry databases and optional MLRun-based MLOps. It also describes how the same control-loop pattern supports adaptive energy optimisation.
KPI-O.3.3-1	Integration of CU/DU with standardised interfaces for intelligent O-RAN support.	Measures whether the Intelligent Plane can interact with the RAN through standardised or O-RAN-aligned mechanisms.	D4.4 addresses this through the analysis of O-RAN control functions, dRAX RIC capabilities, CU/DU integration points, telemetry exposure and the role of rApps/xApps in intelligent control.
KPI-O.3.3-2	Implementation of native AI/ML interfaces for the Intelligent user plane.	Measures the ability to expose AI/ML capabilities as operational interfaces rather than standalone algorithms.	D4.4 supports this by defining model lifecycle, inference and control-loop mechanisms where trained models can be stored, exposed, consumed by xApps/rApps and used to support network management decisions.

Overall, this deliverable acts as the connective deliverable between the project-level ambition for AI-native 6G operation, the **WP4** objectives on measurements, interfaces, repositories and O-RAN intelligence, and the KPIs that require concrete integration with CU/DU and native AI/ML interfaces. Its main contribution is to translate those objectives into an Intelligent Plane design that can be refined, implemented and validated through the remaining 6G-SENSES integration activities.

1.4 Organization of the Document

The document is organised to move from architectural context to implementation detail. After the introductory material, Chapter 2 establishes the basis for the 6G-SENSES Intelligent Plane by recalling the elements of the project architecture that are relevant for AI/ML-driven operation. Particular attention is given to the distribution of intelligence across the SMO, Non-RT RIC, Near-RT RIC and related control functions, as well as to the design principles that guide this distribution.

The technical background is then addressed in Chapter 3, which reviews the main developments in 3GPP, O-RAN and related SNS-JU activities. This includes the role of network analytics, AI/ML lifecycle management, rApps, xApps, cross-domain intelligence and previous Intelligent Plane work, such as BeGREEN. The purpose of this chapter is not to repeat the full standardisation landscape, but to identify the references that are most relevant for the 6G-SENSES implementation.

Building on this background, Chapter 4 identifies the enablers required to make the Intelligent Plane operational. The discussion covers the AI/ML techniques investigated in WP4, the need for MLOps support, the relevance of 3GPP NWDAF/ADRF, the evolution of O-RAN intelligence, telemetry collection, NDT support and future AI-RAN considerations.

The implementation view is presented in Chapter 5. This chapter describes how the Intelligent Plane can be realised in 6G-SENSES using dRAX, rApps, xApps, telemetry databases and, where appropriate, MLRun-based MLOps support. It also connects the implementation to PoC#3 and to the preparation of WP5 demonstrator activities, where real network data, sensing inputs and digital-twin mechanisms are expected to support validation and optimisation.

Finally, Chapter 6 summarises the main achievements, open challenges and contribution of the Intelligent Plane to the overall 6G-SENSES vision. The document concludes with references and acronyms.

2 6G-SENSES Intelligent Plane Architecture

This chapter establishes the architectural basis for the 6G-SENSES Intelligent Plane. It starts from the overall project architecture and focuses on the elements that are relevant for AI/ML-driven operation, sensing-aware optimisation and O-RAN-based control. The aim is not to repeat the full system architecture, but to clarify where the Intelligent Plane fits, which functions it is expected to support, and how intelligence can be distributed across management, orchestration, RIC-based control and, where needed, closer-to-RAN execution environments.

2.1 Review of 6G-SENSES WP2 Architecture & Context

The 6G-SENSES architecture, initially defined in **D2.2** [1] and further refined in **D2.3** [2], provides a unified framework compliant with 3GPP and O-RAN standards that integrates multiple Radio Access Technologies (RATs), computing resources, high capacity transport networks, and AI/ML-driven automation to jointly support communication and sensing capabilities. In what follows, we provide a brief recap of the architecture to highlight its key components and provide the context for the remainder of this deliverable.

Figure 2-1 illustrates the overall 6G-SENSES architecture, which integrates 3GPP (5G NR) and N3GPP (Sub-6 GHz, Wi-Fi, mmWave) technologies within an ISAC framework, following a multi-layer, disaggregated design. Consistent with the O-RAN architecture, traditional RAN functions are disaggregated into the Radio Unit (O-RU), Distributed Unit (O-DU), and Centralized Unit (O-CU), interconnected through standardised fronthaul and F1 interfaces. Building upon this architecture, 6G-SENSES extends the O-RAN framework to support shared O-RUs capable of jointly serving communication and sensing functions. Beyond the RAN, the architecture comprises a cloud–edge infrastructure, a CN, and a Software Defined Networking (SDN)-controlled multi-technology transport network supporting fronthaul, midhaul, and backhaul connectivity.

Intelligence and control across the architecture are enabled through a hierarchical O-RAN framework comprising the Near-RT RIC, the Non-RT RIC, and the Service Management and Orchestration (SMO). The Near-RT RIC hosts xApps that operate at 10 ms to 1 s timescales, enabling closed-loop control of communication and sensing functions based on real-time measurements. The Non-RT RIC and SMO operate at longer timescales (above 1 s), hosting rApps and AI/ML functions for analytics, policy generation, network management, and e2e slice optimisation.

The architecture supports multiple mechanisms for exposing N3GPP sensing data to the O-RAN control framework. Non-3GPP sensing systems may deliver sensing information to the RIC through an extension of the 5G CN Non-3GPP InterWorking Function (N3IWF) with an E2 interface. Alternatively, both 3GPP and N3GPP sensing nodes can be abstracted as sensing Radio Units (sRUs) and expose sensing information through a sensing-oriented E2 service model, denoted in Figure 2-1 by E2S, supporting different data representations, such as I/Q samples, Channel State Information (CSI), and heatmaps. Furthermore, external sensing data from third-party platforms or domain-specific systems can be ingested at the SMO level through a gateway, processed by the Non-RT RIC to derive higher-level context, and forwarded to the Near-RT RIC for use by xApps. Additional integration options are also supported, including connectivity through UEs acting as aggregators or relays and through Wi-Fi access points belonging to a N3GPP access network.

3GPP-based sensing leverages cellular communication signals to generate sensing measurements that are exposed to the Near-RT RIC through the E2 interface. Depending on the deployment, sensing data may be forwarded as raw I/Q samples, with sensing processing performed by dedicated xApps, or as locally processed sensing information through the E2S described previously. At the Near-RT RIC, sensing-aware xApps may combine 3GPP sensing measurements with other data sources, such as N3GPP sensing data, to support sensing-aware RAN optimisation. Sensing outputs can also be forwarded to the SMO for longer timescale analytics and optimisation, ensuring that both communication and sensing requirements are fulfilled.

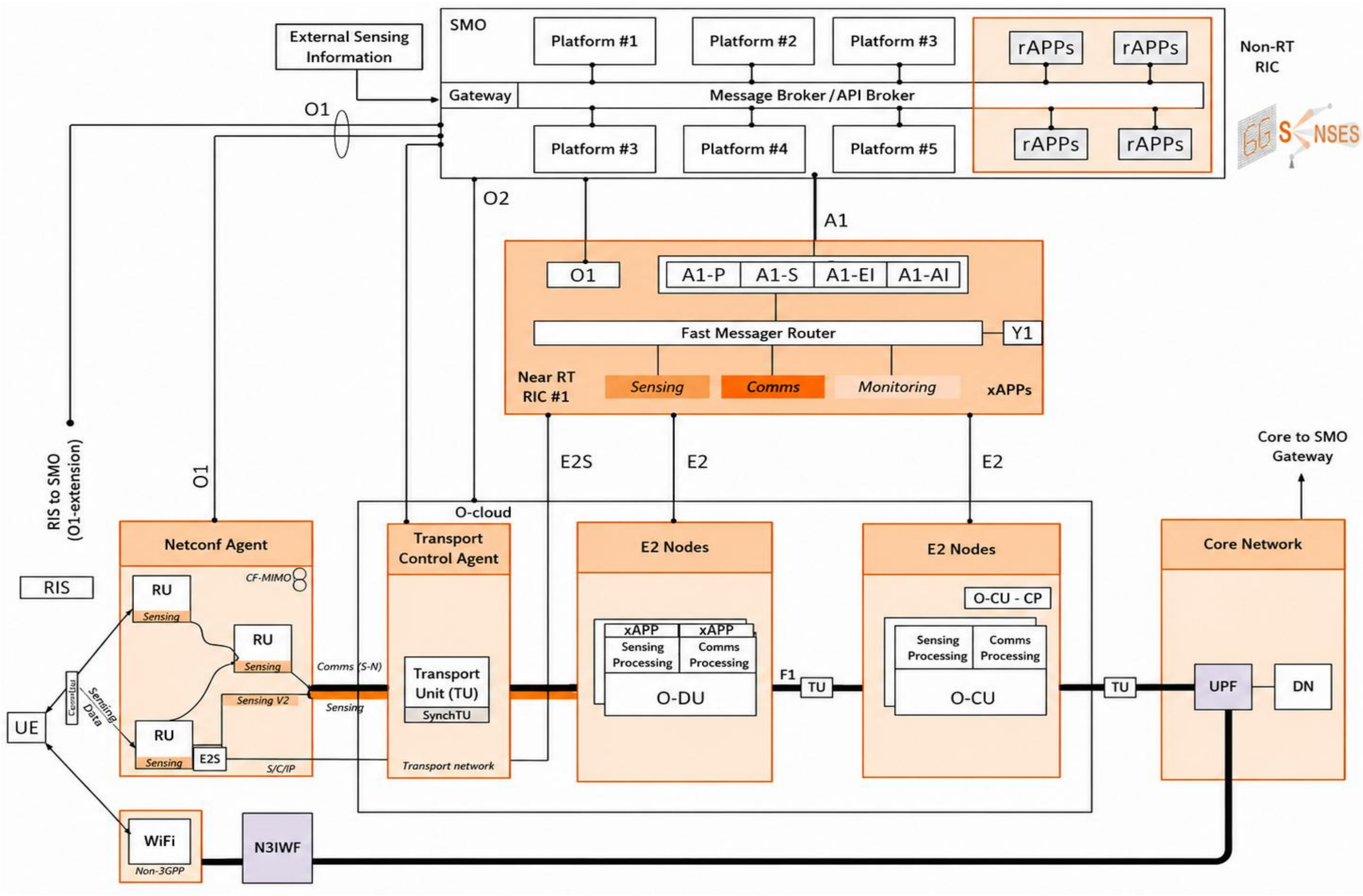


Figure 2-1 6G-SENSES Architecture Overview

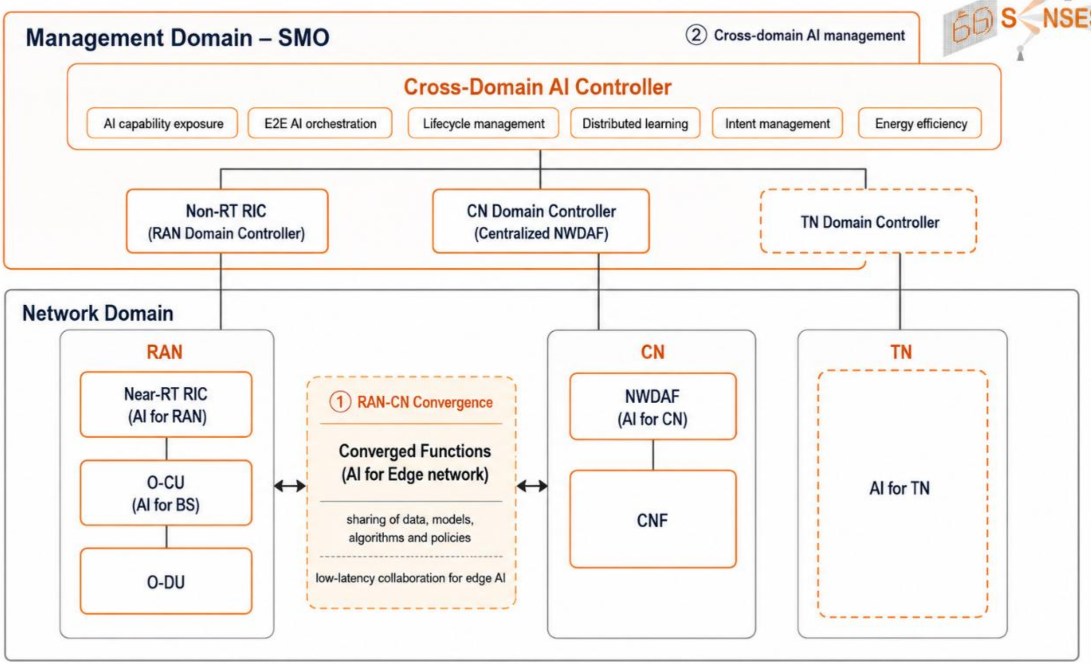


Figure 2-2 Cross domain AI/ML structure

The 6G-SENSES architecture supports flexible deployment of AI/ML functions across the SMO, Non-RT RIC, and Near-RT RIC, depending on latency, computing, and data availability requirements. Training and inference can be performed centrally in a non-real-time manner, split between centralised training and edge inference, executed entirely at the edge for real-time adaptation, or distributed across multiple entities through FL. This flexibility enables the architecture to support a wide range of AI-native use cases, balancing responsiveness, scalability, computational efficiency, and data privacy requirements.

The cross-domain AI architecture proposed in the O-RAN next Generation Research Group (nGRG) Research Report RR-2024-02 identifies two complementary directions to extend O-RAN intelligence beyond isolated RAN optimisation. The first direction is a converged RAN-CN architecture, where RAN and CN AI functions collaborate more directly, enabling shared use of data, models, algorithms and policies for use cases such as mobility optimisation, radio resource management, slice resource allocation and e2e Quality of Service (QoS) assurance. This convergence can reduce the latency and overhead caused by indirect information exchange through existing interfaces and can support more efficient AI execution at the edge. The second direction is an enhanced SMO-based architecture, where a Cross-Domain AI Controller is introduced in the management domain to coordinate AI capabilities across RAN, CN, Transport Network and other domains. This controller would support AI capability exposure, cross-domain data mapping, AI task decomposition, model lifecycle management, distributed learning, intent management, conflict resolution, closed-loop control and energy-efficiency optimisation. Together, these two solutions show how O-RAN can evolve from domain-specific AI functions towards a coordinated cross-domain Intelligent Plane a presented in Figure 2-2.

2.2 Design Principles and Functional Requirements

The design of the 6G-SENSES Intelligent Plane is driven by the need to introduce AI/ML capabilities as operational functions of the network rather than as isolated optimisation tools. In the context of 6G-SENSES, intelligence is required to support the joint operation of communication, sensing and computing resources across a heterogeneous infrastructure composed of 3GPP and N3GPP access technologies, O-RAN components, transport resources, edge/cloud resources, ISAC functions, RIS elements and orchestration entities. The Intelligent Plane therefore needs to provide a coherent framework where telemetry, sensing information, AI/ML models, inference outputs and control decisions can be exchanged across different domains and time scales.

A first design principle is alignment with open and programmable network architectures, in particular O-RAN and 3GPP-based systems. The Intelligent Plane shall not be defined as a monolithic control function, but as a distributed framework that exploits existing control and management entities, such as the SMO, Non-RT RIC, Near-RT RIC, xApps, rApps and the relevant O-RAN interfaces. This principle is important to ensure that the proposed architecture can evolve from current 5G/O-RAN deployments towards 6G-native operation without requiring a complete redesign of the network stack. The Intelligent Plane should therefore support the progressive introduction of AI/ML-based control loops, from long-term policy and model management in the SMO/Non-RT RIC domain to near-real-time inference and control actions in the Near-RT RIC domain.

A second design principle is multi-domain observability. The 6G-SENSES use cases require decisions to be taken based not only on classical RAN performance indicators, but also on sensing data, mobility information, traffic demand, transport conditions, edge resource availability and energy-related metrics. The Intelligent Plane must therefore provide the means to collect, harmonise and expose heterogeneous telemetry streams originating from 3GPP RAN nodes, N3GPP sensing sources such as Wi-Fi, transport controllers, Multi-access Edge Computing (MEC) platforms, CN analytics functions and external sensing systems. This requires a common approach to data ingestion, metadata handling and feature preparation, so that AI/ML models can consume consistent inputs independently of the underlying technology domain.

A third design principle is AI/ML lifecycle management. For the Intelligent Plane to be operationally useful, it must support the full lifecycle of AI/ML models, including data collection, feature engineering, training, validation, deployment, inference, monitoring and retraining. From this perspective, MLOps frameworks such as MLRun provide a useful reference model for the 6G-SENSES requirements. The project requires capabilities similar to those offered by modern MLOps platforms, namely the ability to define reproducible training pipelines, manage feature sets, register and version models, deploy models as online inference services, monitor model behaviour during execution and trigger updates when data drift or performance degradation is detected. In the 6G-SENSES context, these capabilities need to be adapted to telecom-specific constraints, including latency sensitivity, distributed deployment, strict resource limitations, and the need to coordinate AI/ML decisions with network control interfaces.

A fourth design principle is functional distribution across time scales and locations. Not all intelligence functions have the same latency, data availability or compute requirements. Some AI/ML processes, such as model training, policy optimisation, traffic prediction and long-term energy optimisation, can operate at non-real-time time scales and are naturally located in the SMO or Non-RT RIC domain. Other functions, such as dynamic scheduling support, sensing-aware control, mobility-driven adaptation or fast RAN optimisation, require near-real-time operation and are better suited to xApps hosted by the Near-RT RIC. In specific cases, inference or preprocessing may also be pushed closer to the DU, RU, MEC node or specialised accelerator when latency, bandwidth or privacy constraints require it. The Intelligent Plane shall therefore support flexible placement of AI/ML functions across cloud, edge and RAN locations, including split deployments where model training and inference are executed in different parts of the system.

From these design principles, a set of functional requirements can be derived. The Intelligent Plane shall provide mechanisms for collecting telemetry and sensing information from multiple network domains, including RAN, transport, edge, core and external sensing sources. It shall support the preparation and storage of data and features for AI/ML workflows, including both offline datasets for training and online features for inference. It shall enable the deployment of AI/ML models as operational functions, either as rApps, xApps, dApps, service-based components or edge inference services, depending on the required control loop and deployment constraints. It shall support the exchange of policies, model outputs and control recommendations between the SMO/Non-RT RIC and Near-RT RIC domains, enabling coordinated optimisation across long-term and near-real-time control loops. It shall also support monitoring of deployed models and services, including the detection of performance degradation, abnormal behaviour or changes in input data distributions that may require retraining or redeployment. A comprehensive resume of these principles is showcased in Figure 2-3.

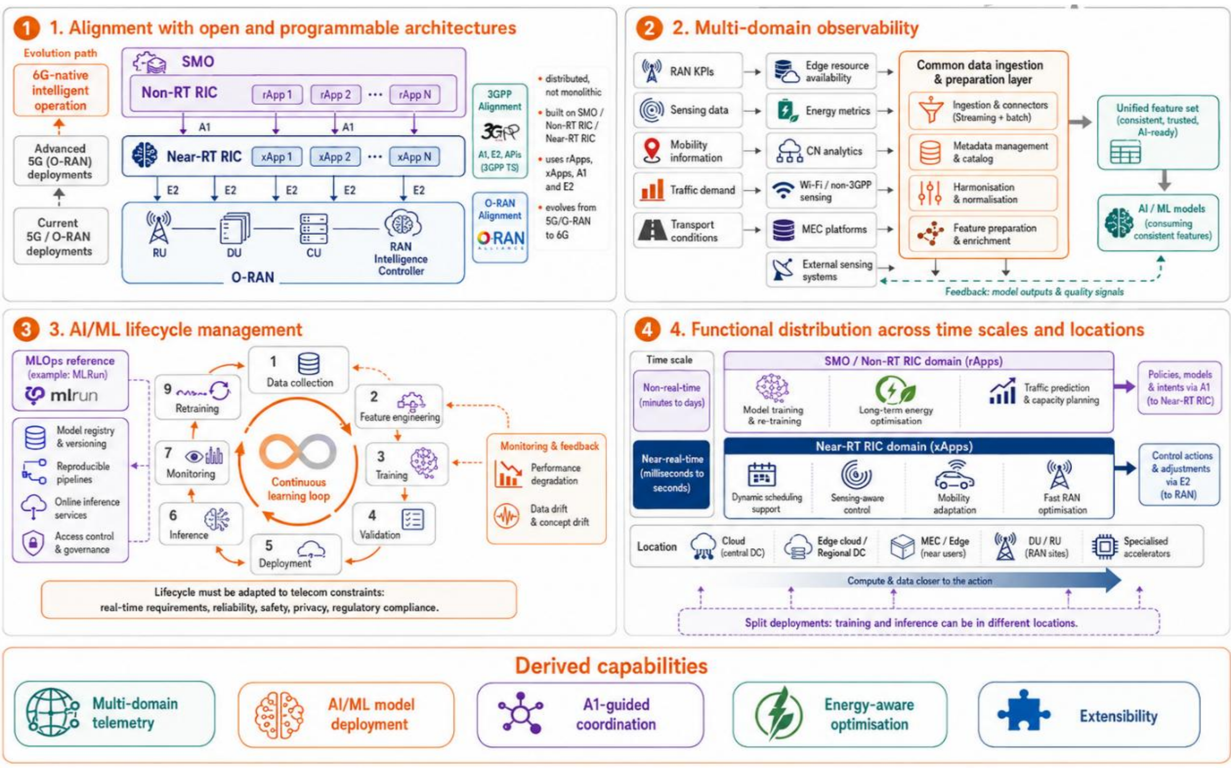


Figure 2-3 Design principles for the Intelligent Plane

In addition, the Intelligent Plane shall be capable of supporting energy-aware and multi-objective optimisation. In 6G-SENSES, energy efficiency is not treated as an isolated KPI, but as a network-wide optimisation target that must be balanced against coverage, latency, sensing availability, quality of service and resource utilisation. The Intelligent Plane must therefore provide the data, model execution environment and control hooks required to evaluate trade-offs between performance and energy consumption. This is particularly relevant for the evolution of the SMO intelligence plane, where higher-level policies can coordinate RAN, transport and edge decisions while near-real-time functions enforce local adaptations through the appropriate O-RAN control mechanisms.

Finally, the Intelligent Plane shall be designed for extensibility. As 6G-SENSES evolves towards more advanced ISAC, AI-RAN and native AI/ML concepts, the architecture must allow new telemetry sources, service models, AI agents, repositories, orchestration functions and control interfaces to be integrated without disrupting the overall framework. This extensibility is essential for supporting future RT-RIC concepts, advanced dApps, hardware acceleration, NDTs and cross-domain automation. Therefore, the Intelligent Plane is not only a set of AI/ML algorithms, but the operational framework that allows those algorithms to be trained, deployed, monitored and connected to real network control actions within the 6G-SENSES architecture.

2.3 Integration with O-RAN Architecture

The 6G-SENSES Intelligent Plane is integrated with O-RAN by mapping intelligence functions to the existing SMO, Non-RT RIC and Near-RT RIC domains instead of introducing a separate management layer. This preserves alignment with the O-RAN control model while allowing 6G-SENSES-specific extensions for sensing, energy optimisation and NDT support. The SMO and Non-RT RIC host long-term intelligence functions such as analytics, model lifecycle management, policy generation, energy-aware optimisation and cross-domain orchestration. These functions rely on historical telemetry, topology information, configuration data, sensing context and, where available, DT outputs. The Near-RT RIC hosts xApps for near-real-time inference and control. In 6G-SENSES, xApps may consume both classical RAN telemetry and sensing-derived information, enabling sensing-aware optimisation actions through the E2 interface under policy or enrichment guidance received from the non-real-time domain.

The main O-RAN interfaces define the operational boundaries of this integration: A1 supports policy and enrichment exchange between Non-RT and Near-RT domains; E2 exposes measurements and control towards O-RAN nodes; O1 supports Fault, Configuration, Accounting, Performance, and Security (FCAPS) and configuration management; and O2 supports interaction with the O-Cloud infrastructure.

A key contribution of 6G-SENSES is the extension of this O-RAN-based intelligence model towards ISAC and multi-technology sensing. Traditional O-RAN deployments mainly focus on RAN telemetry generated by O-CU, O-DU and O-RU functions. However, 6G-SENSES requires the Intelligent Plane to incorporate sensing information from several sources, including 5G NR-based sensing, Wi-Fi sensing, RIS-assisted sensing and potentially external sensing systems. This requires the O-RAN integration to be expanded with sensing-aware service models and data exposure mechanisms. In this approach, sensing-capable entities can be abstracted as sources of measurements for the RIC, while xApps and rApps consume this information according to their respective time scale and function. The E2 interface and the related service models are therefore essential enablers, because they provide the mechanism through which sensing information can become actionable within O-RAN closed-loop control.

The integration with O-RAN also defines how the Intelligent Plane interacts with the lower-layer RAN functions. Through the E2 interface, the Near-RT RIC can receive measurements from O-DU and O-CU functions and send control actions or policy-driven commands back to the RAN. Through O1, the SMO can support configuration, fault, performance and lifecycle management of O-RAN network functions. Through O2, the SMO can interact with the O-Cloud infrastructure where virtualised and containerised RAN functions are deployed. Through A1, the Non-RT RIC can guide the Near-RT RIC by providing policies and enrichment information. These interfaces define the practical boundaries of the Intelligent Plane: long-term intelligence is anchored in SMO/Non-RT RIC functions, near-real-time intelligence is implemented through xApps in the Near-RT RIC, and infrastructure-level deployment is managed through O-Cloud and orchestration mechanisms.

This architectural view is also consistent with lessons learned in related O-RAN-based research activities such as BeGREEN, 6G-TWIN and 6G-SENSES. In BeGREEN, the Intelligent Plane concept was strongly linked to energy efficiency, where AI/ML mechanisms and O-RAN control loops are used to optimise network operation and reduce energy consumption. In 6G-TWIN, the emphasis is closer to the use of digital representations of the network to support prediction, planning, testing and optimisation before applying actions to the live infrastructure. In 6G-SENSES, these two directions converge: the Intelligent Plane must support energy-aware optimisation, while also being able to consume sensing information and potentially interact with Network Digital Twin capabilities for safer and more informed decision-making. The common architectural pattern across these projects is that intelligence is not placed in a single function, but distributed across the SMO, Non-RT RIC, Near-RT RIC and, where needed, closer-to-RAN execution environments.

The evolution of O-RAN towards 6G also opens the discussion on real-time intelligence and dApps. While the current O-RAN architecture defines the Near-RT RIC for control loops typically operating from tens of milliseconds to seconds, some 6G use cases require even faster interaction with lower-layer RAN functions. Examples include beam management, ultra-fast scheduling, sensing signal adaptation, RIS control, and tight PHY/MAC optimisation. For these cases, ongoing O-RAN nGRG discussions introduce the concept of distributed applications, or dApps, which can be deployed closer to O-DU or O-CU functions and complement the existing xApp/rApp model. In the 6G-SENSES architecture, this direction is relevant because several ISAC and native AI/ML functions may require access to lower-layer radio and sensing information before it is abstracted or delayed by higher-level control loops. For example, sensing-aware beam adaptation, RIS reconfiguration, positioning-assisted radio control, fast interference mitigation or PHY/MAC parameter adjustment may need to process I/Q samples, channel estimates, scheduling information, sensing measurements or user-plane context close to the O-DU/O-CU execution environment. In these cases, routing all inference and control decisions through the conventional Near-RT RIC or Non-RT RIC path may introduce excessive latency, bandwidth overhead or loss of data granularity. DApps are therefore considered as a possible future extension of the Intelligent Plane, especially for functions that require direct access to user-plane or lower-layer data and very short reaction times.

From the 6G-SENSES point of view, the Intelligent Plane should therefore be understood as an O-RAN-aligned intelligence fabric. At the management and non-real-time level, the SMO and Non-RT RIC provide the environment for functions that depend on historical data, wider network context and longer decision windows. These include telemetry aggregation, service and slice analytics, energy-efficiency assessment, model training and retraining, model catalogue management, policy generation, intent translation and coordination with orchestration functions. In this layer, rApps can analyse long-term RAN, sensing, infrastructure and service information and produce policies, enrichment information or model updates for downstream control functions. At the near-real-time level, the Near-RT RIC hosts xApps that consume live RAN and sensing telemetry through RIC data streams and E2-based measurements, execute inference or optimisation logic, and translate the resulting decisions into near-real-time control actions towards O-CU/O-DU functions. It may use dApps or similar near-user-plane mechanisms for future real-time AI/ML functions that cannot be efficiently implemented in the current RIC domains. It relies on O-RAN interfaces such as A1, E2, O1 and O2 as the primary integration mechanisms, while extending them where needed to support sensing, ISAC and multi-domain telemetry.

This integration approach allows the 6G-SENSES Intelligent Plane to remain compatible with the O-RAN architectural direction while addressing requirements that go beyond current 5G/O-RAN deployments. The objective is not to redefine O-RAN, but to use its programmable and open control framework as the foundation for 6G-native intelligence. In this sense, the 6G-SENSES Intelligent Plane becomes the project-specific realisation of an AI-enabled O-RAN system, where sensing, communication, energy optimisation, orchestration and model lifecycle management are jointly coordinated across SMO, RIC and edge/cloud execution environments.

2.4 AI/ML Framework and Functional Distribution

MX-AI is BubbleRAN's automation and intelligence platform and development kit designed to assist in mobile-network operations and optimisation. MX-AI provides a set of built-in, reusable and extendable multi-agent blueprints in support of network observability, intent-driven automation, and optimisation by packaging agents, models, tools, and data services into deployable resources, defined as AIFabrics (Figure 2-4). It provides a comprehensive AI-RAN Platform and DevKit designed to enable autonomous operations across the entire RAN lifecycle. Serving as a powerful architectural add-on, it provides an e2e substrate for MLOps that significantly extends the traditional SMO execution environment. To ensure continuous innovation and modularity, MX-AI integrates with the TelcoFabric portal, establishing a shared ecosystem where developers and operators can seamlessly catalogue, reuse, and deploy AI agents, ML models, datasets, rApps, and xApps. Within the 6G-SENSES architecture, this platform provides the foundational intelligence layer required to orchestrate complex, multi-domain network tasks [3], [4], [5].

Functionally, MX-AI maps directly onto the core design principles of multi-domain observability, functional distribution, and continuous lifecycle management. To achieve multi-domain observability, the platform ingests and pairs historical infrastructure KPIs with streaming radio telemetry, transforming complex, high-fidelity signals—such as raw I/Q data from uplink Sounding Reference Signals (SRS)—into compact spatial features [4]. Continuous lifecycle management is maintained natively through structured MX-AI pipelines that continuously retrain models to mitigate data drift, ensuring the Intelligent Plane dynamically adapts to changing environmental and traffic behaviours [6], [7].

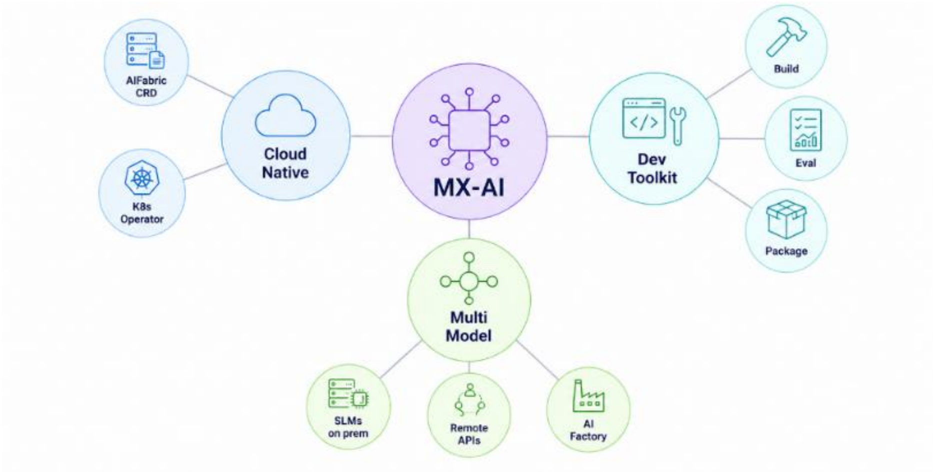


Figure 2-4 MX-AI Components

3 Intelligent Plane in 3GPP/O-RAN Networks SotA

With the evolution from 5G IMT-2020 usage scenarios (eMBB, URLLC & mMTC), to IMT-2030 (6G), mobile networks are "shifted" from high-speed connectivity to ubiquitous intelligence, integrated sensing, and extreme performance

6G-SENSES targets two of the additional usage scenarios that were devised at the IMT-2020, namely "ISAC" and "AI and Communication". ISAC is one of the main cornerstones of 6G-SENSES, while the latter having strong correlation to **WP4** work, defining the 6G-SENSES Intelligent Plane, as presented in Figure 3-1.

3.1 AI/ML SotA in 3GPP

The rapid integration of AI and ML techniques within communication systems led the 3GPP to introduce AI/ML technology for RAN in Release 18 [8] as part of the first 5G-Advanced release. Nowadays, nearly all 3GPP WGs, including both SA and RAN WGs, have include, in one way or another, AI/ML capabilities within their corresponding domains. In this sense, the SA5 WG [9] plays a key role in the harmonisation of the AI/ML capabilities regarding management and deployment. The 3GPP WG SA5 provided initial specifications related to AI/ML in Release 17, including Management Data Analytics (MDA) [10] and general AI/ML management in different segments of the 5GS, such as NWDAF. Furthermore, some recent works, such as [11], have address the embedding of AI/ML into 5G from a systemic perspective, including security aspects at the architectural level. For the exploitation of the AI/ML architectural concepts, existing works focus on mobility and positioning, either analysing the possibilities brought by the standard [12] or by proposing extensions to it [13].

Regarding learning methods, in [14] they are divided within supervised, semi-supervised, unsupervised and reinforcement learning (RL), and for each type and training data and category of inference is identified. Regardless of the adopted learning technique, the 3GPP defines a domain-agnostic AI/ML Life Cycle Management (LCM) that covers all main steps of the AI/ML operations in the 5GS, including the phases shown in Figure 3-2.

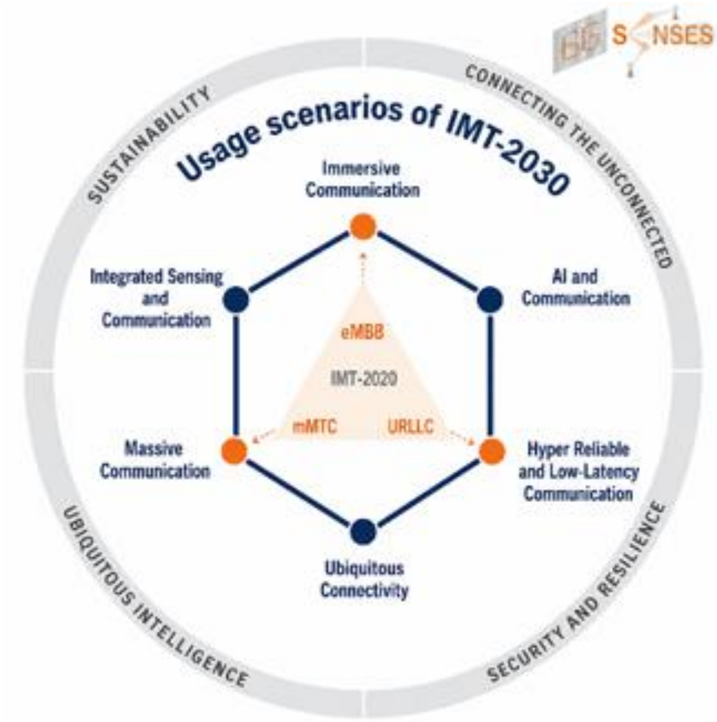


Figure 3-1 IMT-2023 Usage Scenario relevant for 6G-SENSES

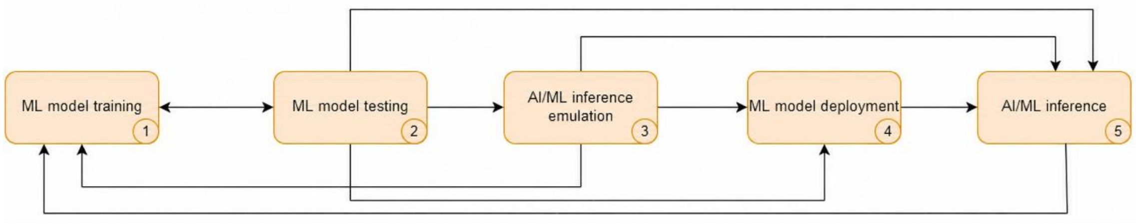


Figure 3-2 3GPP ML model lifecycle

The management capabilities defined by the 3GPP can be divided according to the location of the training and inference functions. In addition, it covers the different placement options according to the system segment (i.e. RAN and 5G CN) and data analytics:

- AI/ML capabilities for RAN: it covers scenarios with both the ML model training and inference are located in the RAN node, as well as
- AI/ML capabilities the 5G CN: it covers scenarios where both the ML model training and inference are located within the 5G CN, which allows leveraging the NWDAF [14].
- AI/ML capabilities for the MDA: it requires that the inference if located in the MDA function, while the training can be either inside or outside.

For each phase of the lifecycle, which is depicted in Figure 3-2, its capability is defined according to its interaction with system Management Services (MnS) as follows:

- ML model training: it permits the MnS to request, control and monitor the ML model training. The training can be of a single model or of a set of models. It may include the validation of the model performance, based on validation data, and re-training if validation fails.
- ML model testing: it allows the MnS to request the testing of trained models, using defined testing data. It also includes the selection of performance metrics, and potentially the triggering of re-training if deemed necessary.
- AI/ML inference emulation: before the deployment of a trained model, this phase allows the validation of the model in an emulation environment, which aims to mimic the real environment (i.e. similar to a DT). If the emulation result does not achieve the expected performance, the model needs to the retrained.
- ML model deployment: the phase enables the MnS to trigger, control and monitor the loading the ML model, embracing all atomic functions composing the model. In some cases, for instance when training and inference are co-located, this step would not be necessary.
- AI/ML inference: this final stage permits the MnS to control the execution of inference, configure output parameters and monitor its performance.

In the particular case of model training the 3GPP identifies a set of relevant training schemes. The first one is transferring learning, which allows determining at which extend parts of a deployed ML model are shared with another domain. In this sense, it is particularly well suited for multi-vendor environments. In particular the specifications highlight that the ML model itself cannot be transferred, since models are not expected to be multi-vendor objects. Instead, the ML knowledge, embracing information such as statistics or summaries, will be shared, leading to the defined ML-knowledge-based transfer learning (MLKTL) [10].

The 3GPP also considers the training of ML models for multiples contexts that share similarities with respect to the type of learning, performance, involved tasks, etc. In particular, it is considered the creation of clusters of ML models with common contexts. In this regard, models of a cluster can be all trained from a baseline model or from a model of similar context, allowing a substantial reduction of the training time.

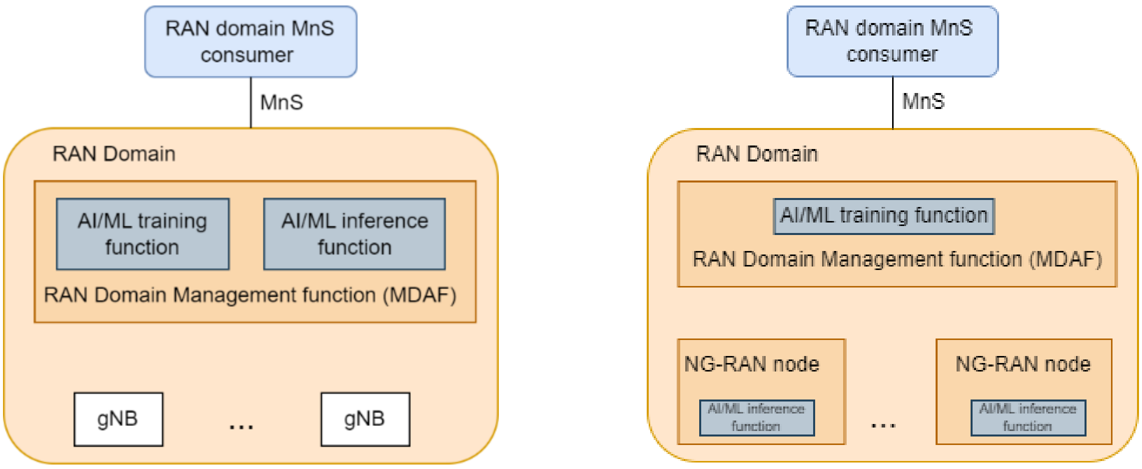
Similar to the previous case, but from a different angle, it is also envisaged the support for pre-specialised training. It refers to the partial training of models with some specific dataset, but for not specific inference type. Opposed to that, it supposes a partial training to support commonalities among multiple use cases. Following the specialised, the training functions of the 3GPP are also to support fine-tuning of pre-specialised models to support a specific inference type.

Within the same domain, it is also foreseen the management of distributed training, where workloads of model training if split among multiples training functions. In the 5GS training functions can be located at different nodes with heterogeneous capacities (processing, memory, storage, etc.). To optimise the training process and resources management, efficient training splitting decisions will be necessary. It is worth noting that this type of training split involves the exchanging/transport of data samples.

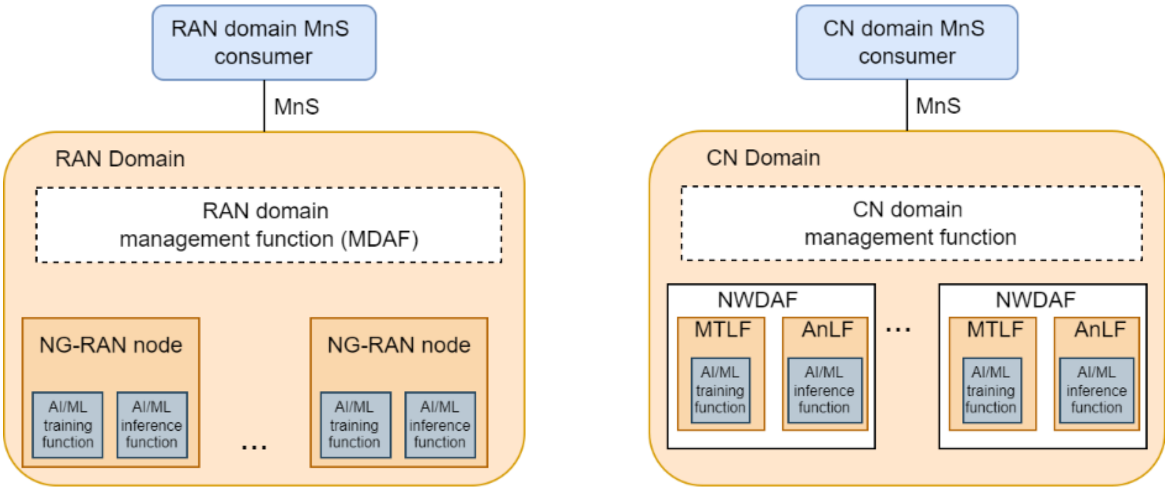
Related to the previous case, but without exchanging data, the 3GPP also considers FL. In this case, FL clients collaboratively train a model using local datasets, and afterwards interim ML models are exchanged. FL can be categorised into horizontal (HFL) and vertical (VFL), according to the data information. In the HFL case, the FL clients handle different datasets but with the same features. On the other hand, in the vertical case the FL clients handle the same datasets but exploit different features the samples. In both cases, the 3GPP assumes an architecture embracing an FL server to orchestrate the ML model aggregation and distribution among the FL clients.

The final training scenario considered by the 3GPP is RL, where RL agents develop a policy over time based in a trial-and-error approach. Considering the particular training paradigm of RL, two different schemes are foreseen: online and offline. In the former, the RL agent interacts directly with the RL environment (i.e. an emulation framework), which provides the reward and environment transition. In the latter case, the RL agent is trained using a pre-collected dataset, which must contain rewards and transitions collected from the environments during a period of time.

Regardless of the adopted learning scheme, the 3GPP [10] has identified 4 AI/ML management scenarios, involving both RAN and CN domains, which are depicted in Figure 3-3, where the first 3 scenarios correspond to the RAN domain and the last one to the CN domain.



a. RAN domain both at MDAF **b. RAN MDAF/NR-RAN node split**



c. **RAN domain both at NG-RAN node** d. **CN at NWDAF**

Figure 3-3 Location of AI/ML training and inference function at RAN in CN domains in 3GPP [2]

In the first one, both AI/ML functions (training and inference) are located at a RAN management function, such as the MDA function. The second scenario considers the case where training and inference are separated. In particular, in this case, shown in Figure 3-3 (b), the training is performed in a RAN management function while the trained model is deployed for inference in a next-generation RAN (NG-RAN) node. Furthermore, in the third scenario both training and inference functions are deployed in the NG-RAN. As for the CN domain, the 3GPP considers the case of deploying the AI/ML functions within the NWDAF, particularly the training and inference functions would be deployed at the Model Training Logical Function (MTLF) and the Analytics Logical Function (AnLF), correspondingly.

3.2 AI/ML SotA in O-RAN

O-RAN provides the most direct architectural reference for the 6G-SENSES Intelligent Plane because it already separates programmable intelligence across SMO, Non-RT RIC and Near-RT RIC domains[15], [16]. The Non-RT RIC supports long-term analytics, policy generation, enrichment information and AI/ML lifecycle functions through rApps, while the Near-RT RIC supports near-real-time inference and control through xApps interacting with E2 nodes. The O-RAN entities are depicted in Figure 3-4.

The A1, E2, O1, O2 and R1 interfaces are therefore central to AI/ML operation in O-RAN. A1 enables policy and enrichment exchange between Non-RT and Near-RT domains [17]; E2 exposes measurements and control procedures [18]; O1 and O2 support management and infrastructure interaction; and R1 supports rApp interaction with the Non-RT RIC framework [15].

The state of the art has moved beyond simple model inference. Current O-RAN research and specifications consider model lifecycle management, data preparation, model deployment, inference services, monitoring, conflict mitigation, explainability, security, verification and xApp/rApp coexistence [19], [20], [21]. These topics are especially important because open programmability increases both innovation and the risk of unsafe or conflicting control decisions. O-RAN nGRG work on native AI, cross-domain AI and dApps further extends this direction. Native AI treats AI/ML as an inherent network capability; cross-domain AI coordinates data, models and policies across RAN, CN, transport, cloud and application domains; and dApps explore closer-to-RAN execution for functions that require tighter latency than the Near-RT RIC can provide.

The gap addressed by 6G-SENSES is the extension of this AI-ready O-RAN framework towards ISAC, multi-technology sensing, sensing-aware closed loops, edge/transport-aware optimisation, NDT support and energy-efficient operation.

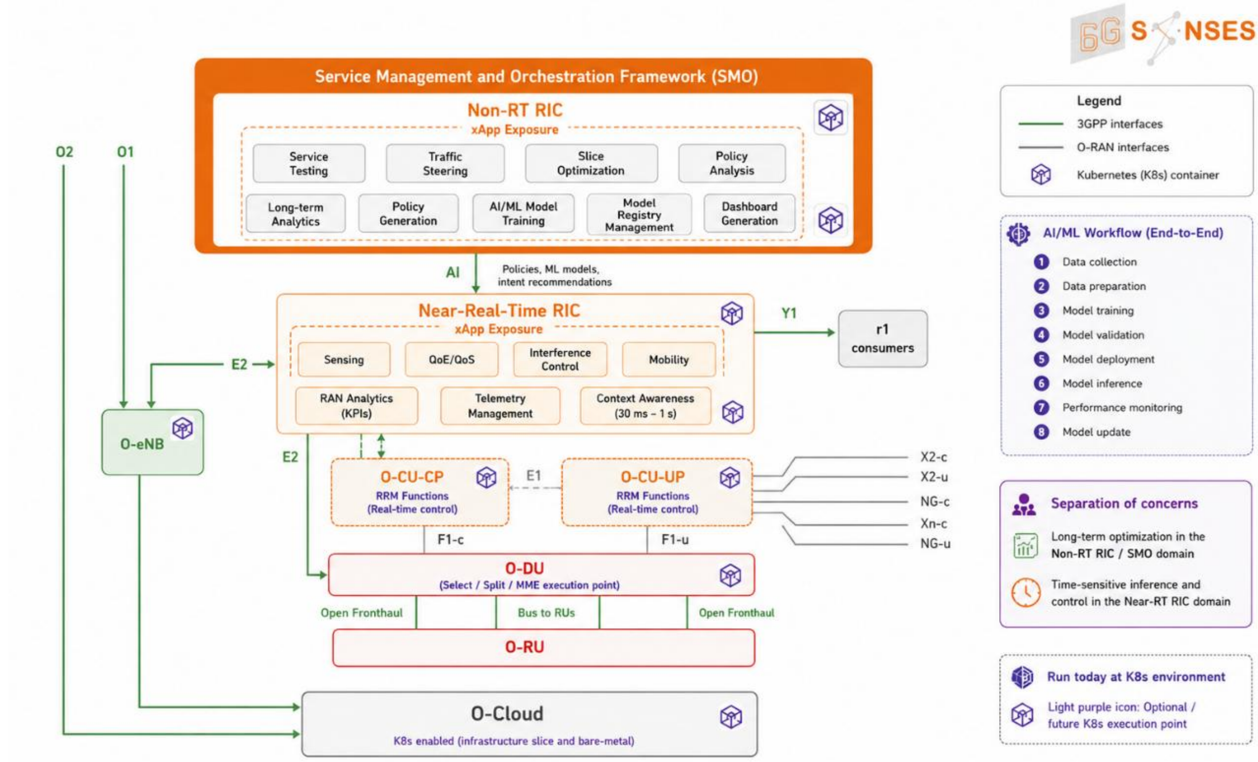


Figure 3-4 O-RAN entities identified to support AI/ML services for the Intelligent Plane

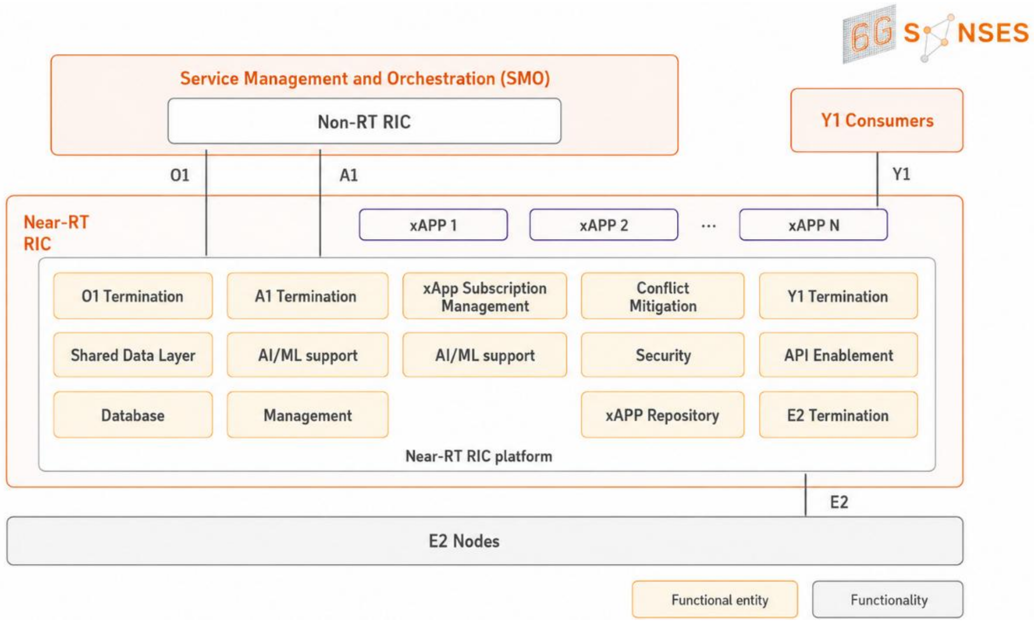


Figure 3-5 Function entities and functionalities that support AI/ML services

Figure 3-5 summarises the internal functional structure of the Near-RT RIC and its interaction with the SMO/Non-RT RIC, Y1 consumers and E2 nodes. In this architecture, the Near-RT RIC hosts xApps and provides the supporting platform capabilities required for near-real-time control, including O1, A1, Y1 and E2 termination, shared data handling, subscription management, conflict mitigation, security, API enablement, xApp repository functions and AI/ML support. This view is aligned with the direction described by the nGRG Contributed Research Report on Cross-domain AI, RR-2024-02, which identifies cross-domain AI as a key research direction for next-generation networks and highlights the need for collaboration across data, computing resources and AI models. In this context, the Near-RT RIC becomes the execution point for time-sensitive AI/ML inference and control through xApps, while the SMO/Non-RT RIC provides longer-term analytics, policy guidance and model lifecycle support via A1 and O1 [22]. This separation allows AI/ML functions to be distributed across domains while maintaining a coherent control framework between management-level intelligence, near-real-time optimisation and the E2-controlled RAN nodes.

A further development in O-RAN research is the discussion around native AI and cross-domain AI in nGRG [20]. This work recognises that 6G systems will require intelligence that is not limited to the RAN domain alone, but able to interact with transport, edge, cloud, CN and application domains. Native AI refers to the integration of AI/ML as an inherent capability of the network architecture, rather than as an external management tool. Cross-domain AI extends this view by considering how data, models and control decisions can be coordinated across several technological and administrative domains. For 6G-SENSES, this direction is highly relevant because ISAC, RIS, multi-WAT sensing, edge processing and transport-aware service delivery require AI/ML functions that can reason across heterogeneous sources of telemetry and different control time scales.

The nGRG activity has also introduced the need for AI/ML lifecycle management and testing methodologies suitable for next-generation networks [23]. Unlike conventional software components, AI/ML systems may change their behaviour depending on training data, retraining procedures, runtime drift and feedback from the controlled environment [24]. This introduces challenges for validation, reproducibility, fairness, robustness and operational safety. Recent O-RAN-oriented research has therefore started to address AI verification, explainable AI, adversarial robustness and misconfiguration detection. These topics are particularly important in Open RAN because openness and programmability increase innovation but also expand the potential attack and failure surface. A malicious, poorly trained or conflicting xApp/rApp may degrade network performance or violate operational constraints, especially when multiple applications share control over the same RAN resources.

Explainable AI (XAI) is another major research direction for AI/ML in O-RAN [25]. Since xApps and rApps can influence RAN behaviour, operators require mechanisms to understand why a model recommends or applies a given action. This is essential for trust, auditability and operational acceptance. XAI techniques can help expose feature importance, decision boundaries, confidence levels and causal relationships between network states and AI decisions. In future O-RAN deployments, explainability is expected to become relevant not only for human interpretation, but also for automated conflict resolution, policy validation and safe closed-loop control.

A more recent direction is the extension of AI/ML control closer to the real-time domain. The current O-RAN architecture provides the Near-RT RIC for control loops down to approximately 10 ms, but some RAN functions, such as PHY/MAC adaptation, beam management, positioning, sensing signal configuration and ultra-fast spectrum sharing, may require sub-10 ms or even sub-millisecond reactions. To address this gap, O-RAN nGRG and the research community have discussed concepts such as distributed applications (dApps) [21], [26], real-time RIC, µApps, zApps and tApps. Among these, dApps have gained attention as a possible framework for real-time AI-based observability and programmability closer to O-DU/O-CU functions. This direction is particularly relevant for future 6G ISAC and AI-RAN use cases, where inference may need direct access to lower-layer information, user-plane data or sensing measurements that cannot be efficiently transported to a central RIC before action is required.

From a 6G-SENSES perspective, the state of the art in AI/ML for O-RAN can be summarised through four complementary layers. First, O-RAN provides an AI-ready control architecture through SMO, Non-RT RIC, Near-RT RIC, A1, E2, O1, O2 and R1. Second, it provides an application ecosystem through rApps and xApps, enabling third-party and operator-specific AI/ML functions. Third, it is evolving towards richer AI/ML lifecycle management, including model management, inference services, monitoring and retraining. Fourth, research beyond current specifications is moving towards native AI, cross-domain AI, explainability, verification, security and real-time dApps. These directions collectively define the technical context in which the 6G-SENSES Intelligent Plane is positioned.

The main gap that remains is the integration of AI/ML across communication, sensing and orchestration domains in a unified operational framework. Existing O-RAN work has made strong progress in RAN optimisation, xApp/rApp programmability and AI/ML workflow definition, but 6G-SENSES requires these concepts to be extended towards ISAC, multi-technology sensing, sensing-aware control loops, edge/transport-aware optimisation and energy-efficient operation. Therefore, the project builds on the O-RAN AI/ML state of the art while extending it towards a broader Intelligent Plane where sensing data, communication telemetry, AI/ML models and orchestration actions are jointly managed across the 6G-SENSES architecture.

3.3 AI/ML SotA within EU SNS-JU

The SNS JU framework provides an attractive ground on how AI and ML are becoming foundational enablers for future 6G systems. Across the SNS JU programme, AI/ML is no longer considered only as a set of isolated optimisation techniques, but as a cross-cutting capability supporting autonomous network operation, intelligent resource management, data-driven service orchestration, security, sustainability, and trustworthy decision-making. The landscape also highlights the importance of aligning AI/ML developments with emerging standardisation frameworks, particularly those from 3GPP, O-RAN, ETSI, and ITU-T, while emphasising the need for lifecycle management, model monitoring, explainability, and MLOps practices to move AI-enabled solutions from research prototypes towards operational 6G networks.

Within this context, 6G-SENSES contributes directly to the SNS JU AI/ML vision by applying intelligence to sensing-enabled, multi-technology 6G radio environments. The project's work on an evolved O-RAN-based Intelligent Plane, cross-domain sensing and telemetry integration, and AI-assisted optimisation of heterogeneous access technologies is closely aligned with the AI-native networking direction identified in the SNS JU landscape. In particular, 6G-SENSES explores how AI/ML can support ISAC, CF-MIMO, RIS-assisted operation, edge intelligence, and proactive RAN control, while also addressing the practical requirements of interoperability, automation, scalability, and validation in realistic experimental environments. As such, the project provides a concrete implementation pathway for several of the AI/ML priorities identified across the SNS JU ecosystem.

The following sections position the 6G-SENSES work within the broader AI/ML state of the art emerging from the SNS JU ecosystem, with particular attention to how intelligence is being embedded into future 6G network architectures.

3.3.1 SNS-JU TB White Paper: AI/ML as a Key Enabler of 6G Networks

This white paper is produced by the SNS JU Technology Board (TB) [27], which is the main collaborative body of SNS JU technical experts, and aims to introduce the work taking place within the SNS JU projects, relevant to AI/ML. An overview of the projects is provided while details and statistics about the AI/ML based solutions developed within the SNS JU are also presented including the goal of the AI mechanisms, the learning type & method, the network segment they are implemented on, characteristics of the AI models used and information about the training data sets and the output of the mechanisms. It is the hope of the SNS JU TB that this paper will assist in the understanding of the work performed within the SNS JU on the very important topic of AI, promote the AI solutions developed in Europe and foster synergies with other global AI and 6G experts.

3.3.2 The AI/ML landscape for Smart Networks and Services

The SNS JU Reliable Software Networks Working Group has created a new white paper [28] exploring the role of AI and ML in advancing smart networks and services, particularly in the context of 6G. This paper serves as a primer that clarifies key AI/ML terminology, summarises major current standardisation efforts, and explores open implementations and innovation pathways. It provides a foundational resource for researchers, industry stakeholders, and policymakers, with the primary goal of aligning efforts and accelerating the development of AI-native networks and systems, particularly within the framework of SNS-JU research projects.

The White Paper underscores the importance of AI/ML in driving innovation, enhancing network performance, and addressing challenges such as security, trust, and sustainability in next-generation networks. It defines AI-native systems as systems where AI is embedded into design, deployment, operation, and maintenance, rather than added afterward. It links this to self-management, self-optimisation, and self-repair capabilities.

For 6G-SENSES, this supports positioning the Intelligent Plane as more than an orchestration add-on. It can be described as the project's AI-native control and optimisation layer across heterogeneous RAN domains: Sub-6, mmWave, Wi-Fi, 5G NR, RIS-assisted links, and CF-MIMO/ISAC components. This is particularly aligned with the project objective to ingest cross-technology sensing, telemetry, and control into an evolved 6G RIC. The White Paper describes the O-RAN AI/ML framework with training management, data management and exposure, AI Training Host Platform, AI Inference Host Platform, model onboarding, model serving, and monitoring in Near-RT and Non-RT RIC contexts. This aligns strongly with 6G-SENSES' plan to extend O-RAN RIC capabilities for heterogeneous wireless access technologies and sensing support, including xApp/rApp extensibility and "6G native AI/ML automation and zero-touch optimisation."

The SNS JU AI/ML landscape White Paper reinforces the relevance of the 6G-SENSES approach by identifying AI-native, interoperable, and governable AI/ML operations as key enablers for 6G. 6G-SENSES contributes to this direction through an evolved O-RAN-based Intelligent Plane that ingests cross-technology sensing, telemetry, and control from Multi-WAT ISAC environments, enabling AI-assisted optimisation of CF-MIMO, RIS-assisted links, edge caching, and proactive RAN management. The white paper further highlights the need for MLOps, monitoring, explainability, and sustainability-aware lifecycle management, which can guide the validation and operationalisation of 6G-SENSES PoCs.

3.3.3 SNS BeGREEN Intelligence Plane

The BeGREEN project [29] is a relevant SNS JU reference because it implemented an O-RAN-based Intelligence Plane focused on energy efficiency. Its main value for 6G-SENSES is not that the architecture can be copied directly, but that it validates a practical design pattern: combine SMO/Non-RT RIC functions, Near-RT RIC xApps, an AI Engine, telemetry, model lifecycle management and measurable energy-efficiency indicators within a single operational framework [30].

BeGREEN showed that AI/ML lifecycle management should be decoupled from RAN control logic. Instead of embedding models directly inside each xApp or rApp, the AI Engine provides model training, serving, monitoring and reusable inference outputs, while RIC applications remain focused on control and optimisation logic. This separation improves reusability, maintainability and independent evolution of models and applications [31].

The project also demonstrated that energy optimisation requires coordination across multiple domains: RAN, edge, O-Cloud, relays, RIS and infrastructure resources. It introduced explicit energy-efficiency metrics, such as Energy Score and Energy Rating, and validated workflows involving AI Engine integration, rApps, xApps, A1-based coordination and conflict mitigation [32].

For 6G-SENSES, the transferable lessons are clear: decouple model lifecycle management from control applications; expose model outputs through reusable interfaces; coordinate long-term and near-real-time decisions through SMO/RIC functions; include conflict-aware control; and evaluate the net operational benefit of AI/ML, including its own computational and energy cost. The main difference is that 6G-SENSES extends this pattern from energy-centric optimisation towards ISAC, multi-RAT sensing, Network Digital Twin support and sensing-aware network control.

The Figure 3-6 presents the BeGREEN Intelligence Plane [30] as an AI/ML-enhanced control architecture spanning the SMO/Non-RT RIC, Near-RT RIC, 5G CN, Edge Server, and O-Cloud. Its main objective is to introduce energy-aware intelligence across the network by combining long-term AI model management, near-real-time inference, and distributed execution of AI functions. The BeGREEN AI Engine acts as the central AI/ML environment. It includes datalakes, AI/ML services and pipelines, and a model catalogue and selection function, supporting data collection, model development, training, selection, and lifecycle management. It also includes model inference and energy score/rating functions, which can be offloaded in a serverless manner to support flexible deployment of AI workloads.

Within the SMO Framework, the Non-RT RIC hosts BeGREEN-specific rApps for AI-assisted automation and energy efficiency. These rApps interact with A1 functions and SMO exposure mechanisms to generate policies, recommendations, and control inputs for lower network layers. The SMO also coordinates AI-driven control for edge, relay, RIS, O1, and O2 functions. The Near-RT RIC provides the near-real-time AI execution layer. It hosts BeGREEN xApps, generic xApps, a Telemetry Gateway, and AI-related functions such as A1 termination, E2 functions, conflict management, ISAC functions, and RIS functions. This layer enables time-sensitive inference and control decisions based on telemetry and policies received from the SMO / Non-RT RIC. The lower part of the Figure 3-6 shows how AI/ML functions interact with the execution infrastructure, including the Edge Server, 5G CN, and O-Cloud. AI applications may run at the edge, while O-CU, O-DU, O-RU, RIS, relay, and ISAC components provide the operational environment where AI-driven decisions are applied. Overall, the Figure 3-6 highlights a distributed AI/ML architecture where the AI Engine manages models and intelligence resources, the Non-RT RIC performs long-term optimisation and policy generation, and the Near-RT RIC executes near-real-time inference and control, enabling energy-aware optimisation across RAN, edge, RIS, relay, and ISAC domains.

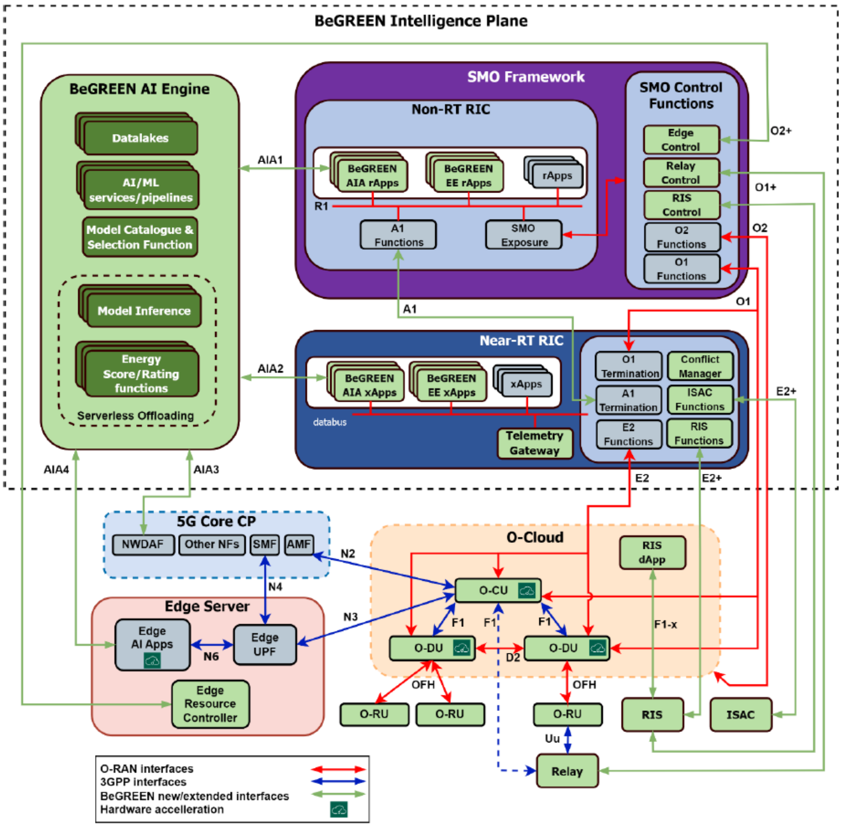


Figure 3-6 Final BeGREEN Intelligence Plane [30]

The final BeGREEN implementation validated several key workflows and technologies. The AI Engine was benchmarked under different training and model-serving workloads, including energy measurement. The integration between the AI Engine and the Non-RT RIC was validated through AI Engine Assist (AIA) rApps and the Information Coordination Service, supporting the exposure of inference outputs as information types. Dataset creation and model training workflows were automated, and the interaction between AI Engine, rApps and RIC components was demonstrated through a representative cell on/off control loop. The Near-RT RIC integration also included conflict mitigation, which is a critical requirement when several xApps or A1 policies attempt to control overlapping resources. BeGREEN validated a conflict management approach where conflicting energy-saving and QoS policies can be detected and managed, showing that conflict-aware control is necessary for safe AI-driven operation.

Final results in BeGREEN show that the Intelligence Plane can support practical energy-efficiency optimisation across multiple network domains. In the vRAN domain, BeGREEN investigated cache isolation, CPU/cache allocation and lightweight control strategies to reduce platform power without degrading throughput or latency. For carrier on/off switching, compact classifiers were used to identify energy-saving opportunities while preserving SLA/QoS guardrails. For relay-enhanced RAN control, BeGREEN evaluated fixed relay and Relay UE mechanisms to improve coverage and reduce power consumption in challenging propagation scenarios. In the Edge domain, traffic-aware compute management for UPF functions was validated using realistic traffic traces and ML-based forecasting, while joint orchestration of vRAN and Edge AI services demonstrated how multi-objective algorithms can reduce total energy while maintaining service performance. These results confirm that the Intelligent Plane is not only a conceptual management framework, but a practical architecture capable of hosting and coordinating several AI/ML-assisted optimisation loops [33].

For 6G-SENSES, the impact of BeGREEN is twofold. First, it provides a concrete architectural precedent for implementing an Intelligent Plane on top of O-RAN, using the SMO, Non-RT RIC, Near-RT RIC, rApps, xApps, MLOps services and serverless inference as the main integration points. This directly supports the 6G-SENSES need to define where AI/ML models are trained, deployed, exposed and monitored across the architecture. Second, it provides lessons for extending the Intelligent Plane beyond classical RAN optimisation. BeGREEN already considered RIS, ISAC, relays, edge resources and CF concepts, which are also central to 6G-SENSES. However, while BeGREEN focused primarily on energy efficiency, 6G-SENSES must extend this intelligence framework towards integrated communication and sensing, multi-RAT sensing data exposure, sensing-aware control, cross-domain orchestration, and future native AI/ML in the RAN. Therefore, BeGREEN can be seen as a strong baseline for the 6G-SENSES Intelligent Plane, but not as a final template. The main contribution for 6G-SENSES is the proven design pattern: decouple AI/ML lifecycle management from control applications, expose model outputs through reusable interfaces, coordinate long-term and near-real-time decisions through SMO/RIC functions, and always evaluate the net operational benefit of AI/ML, including its own computational and energy cost.

3.4 Additional Relevant SotA

BubbleRAN MX-AI platform presented in Figure 3-7, is an AI-RAN Platform and DevKit that significantly extends the SMO execution environment and could complements the Cross-Domain AI Controller in section 2 to enable autonomous operations across the network lifecycle (Day 0/1/2+).

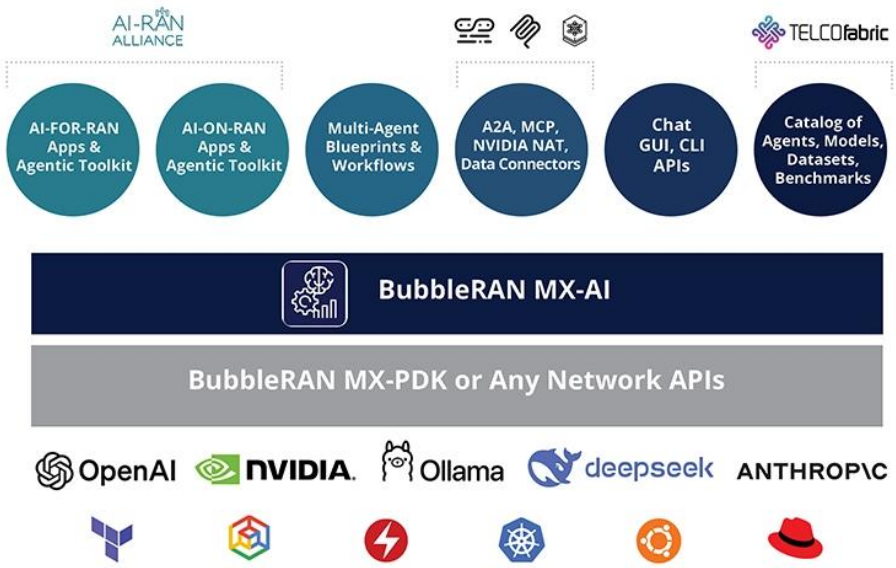


Figure 3-7 BubbleRAN MX-AI platform

The framework provides an e2e toolkit for MLOps anchored in the Non-RT RIC with effective intelligence requires three foundational capabilities: continuous access to heterogeneous data, lifecycle management, and actionable inference. The platform handles dataset preparation by consolidating both historical data and streaming telemetry—such as radio performance metrics: Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), throughput; and infrastructure KPIs such as Central Processing Unit (CPU), Random Access Memory (RAM). It supports continuous model retraining to mitigate data drift, ensuring that AI/ML models dynamically adapt to evolving network behaviours and traffic patterns. This lifecycle is managed either natively via structured MLOps pipelines (e.g., integrating MLRun) or directly through customised rApps.

A defining feature of the MX-AI framework is its use of Multi-Agent Workflows and Blueprints to implement AI-for-RAN and AI-on-RAN applications. Powered by the BubbleRAN Agentic Toolkit (BAT), the platform orchestrates collaborative AI agents—such as Supervisor, SMO, SLA, and API Agents. These agents communicate via Agent-to-Agent (A2A) and Model Context Protocol (MCP) frameworks to break down complex network intents, negotiate resources, and execute closed-loop decision-making. This collaborative intelligence transforms raw network observability into coordinated, system-wide optimisations.

While the Non-RT RIC and SMO manage long-term analytics and model training, time-sensitive execution is distributed to the Near-RT RIC. Data is collected in a near-real-time execution environment through the deployment of specialised xApps. These xApps tap into live telemetry streams to perform immediate inference. Utilising the models trained and exposed by the MLOps layer, these xApps enforce dynamic optimisation and control policies directly on the underlying network nodes.

The AI/ML framework leverages a split-workflow distribution where functions are placed according to their specific latency, compute, and data requirements. Long-term policy generation and training reside centrally, while near-real-time inference is pushed to the edge. Furthermore, the platform orchestrates functions across diverse hardware and services. A prime example is the distribution of ISAC monitoring tasks, leveraging O-RAN 7.2 split architectures on specialised radio hardware, including machines equipped with Liteon or Benetel devices.

4 6G-SENSES Intelligent Plane: Technical Enablers & Drivers

This chapter identifies the main technical enablers that were described in deliverable D4.3 [34] required to move the 6G-SENSES Intelligent Plane from architectural concept to operational capability. It brings together the AI/ML techniques, data and telemetry mechanisms, MLOps support, 3GPP analytics functions, O-RAN evolution and NDT aspects that are needed to support intelligent network operation. The chapter therefore acts as the bridge between the architectural positioning described earlier and the implementation choices presented in Chapter 5.

4.1 Supported AI/ML Techniques

Different AI/ML solutions have been investigated in T4.2 for the optimisation of communication and sensing services, as well as network operation, within the 6G-SENSES architecture. These solutions employ different AI/ML techniques and span various network domains and operational timescales, ranging from near-real-time resource optimisation in the RAN and Edge domains to non-real time e2e slice orchestration, intent-driven slicing, and predictive analytics. Consequently, the Intelligent Plane must provide the capabilities required to manage the lifecycle of diverse AI/ML models, including data collection, training, deployment, inference, monitoring, and continuous adaptation. In what follows, we summarize the AI/ML techniques proposed and evaluated in D4.3, highlighting some of the key requirements they impose on the Intelligent Plane.

4.1.1 DRL-based resource allocation and orchestration.

Various RL and Deep Reinforcement Learning (DRL) techniques were proposed for both RAN optimisation and e2e orchestration. These techniques support closed-loop network optimisation by continuously interacting with the network environment, observing its state through telemetry information, taking exploration and exploitation actions, and using feedback from the environment (reward) to improve their policy towards maximising long-term performance objectives. Therefore, they are well suited to dynamic and uncertain network environments.

Within the RAN domain, a Deep Q-Network (DQN) agent was employed for sensing-aware MAC scheduling to dynamically adjust scheduler parameters based on predictive environmental information enabled by ISAC. In addition, a Meta-RL agent was employed to optimise RIS parameters with the objective of maximising the communication sum rate. Beyond single-agent methods, a Multi-Agent Deep Deterministic Policy Gradient (MADDPG) framework was proposed for RIS optimisation, where individual RIS elements act as agents that collaboratively optimize communication and sensing parameters, including beamforming and radar processing variables.

At a higher level, involving e2e cross-domain operation, a hierarchical DQN framework was proposed at the SMO layer for the joint orchestration of communication and sensing slices. The framework jointly optimizes VNF placement and sensing parameter configuration to minimize energy consumption while satisfying Service Level Agreements (SLAs). It employs a hierarchical decision-making structure comprising a high-level agent responsible for capacity provisioning decisions and a low-level agent responsible for fine-grained orchestration actions.

Supporting these solutions requires the Intelligent Plane to provide access to network telemetry, AI model training and inference capabilities, and interfaces enabling interaction with network control and orchestration functions.

4.1.2 Supervised Learning for predictive analytics and sensing intelligence.

Another category of AI/ML techniques investigated is Supervised Learning. These techniques learn patterns and relationships from labelled datasets and use the resulting models to infer information from previously unseen data. In 6G-SENSES, they are employed to predict network behaviour, estimate service performance, and extract information from sensing data.

A Graph Neural Network (GNN)-based framework was proposed for e2e performance prediction in dynamic RAN environments. The model captures topological relationships among network elements to predict key performance indicators such as delay, jitter, and packet loss under varying and previously unseen traffic conditions. Such predictive capabilities enable proactive decision-making without relying on computationally expensive simulations.

Beyond network performance prediction, supervised learning techniques were also employed for sensing-related tasks that extract high-level contextual information from wireless signals. One example is the WiRD-Gest framework, which utilizes a monostatic full-duplex sensing pipeline to transform raw Wi-Fi signals into high-quality Range-Doppler representations and leverages Deep Neural Networks (DNNs) for accurate gesture classification. Different neural network architectures can be employed, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid CNN-RNN models, with the latter achieving the highest classification accuracy during evaluation. Another example is the application of supervised learning to the trajectory prediction of moving targets, where Range-Doppler maps collected from multiple sensing nodes are fused at a centralised sensing aggregation function to exploit multi-static sensing gains. A hybrid CNN-LSTM architecture processes multi-view sensing data, combining spatial feature extraction and temporal learning to estimate the future trajectory of the target.

4.1.3 Federated Learning for distributed trajectory prediction.

Another AI/ML technique investigated in 6G-SENSES is FL, which enables the collaborative training of AI models across multiple distributed sensing nodes without requiring the exchange of raw sensing data. Instead, each node trains a local model using its own sensing observations and periodically exchanges only model updates, such as weights or gradients, with a federated aggregation function. The aggregation function combines the received updates to construct a global model, which is subsequently redistributed to the participating nodes and used as the basis for further local training.

In 6G-SENSES, FL was applied to the trajectory prediction of moving targets using Range-Doppler radar data generated by distributed sensing nodes an reported in detail in **D4.3** [34]. By avoiding centralised data aggregation, the FL-based approach significantly reduces fronthaul and backhaul bandwidth requirements while improving scalability, robustness, and data privacy. The resulting global model provides trajectory predictions that can be exploited by higher-layer control and orchestration functions to proactively allocate communication, sensing, transport, and compute resources. Furthermore, the availability of both centralised and FL-based sensing modes enables intelligent agents to dynamically balance sensing accuracy, network traffic load, and resource utilisation according to the current operational conditions.

4.1.4 Generative AI for intent-driven network automation.

Another AI/ML technique investigated in 6G-SENSES is Generative AI (GenAI), which leverages Large Language Models (LLMs) to translate high-level operator intents into actionable network policies. In particular, GenAI was applied to automated RAN slicing through an LLM-powered rApp that enables intent-based network control. The proposed framework allows operators to express service requirements in a high-level manner, which are automatically translated into slicing policies and deployed through the O-RAN control framework. These policies are enforced through the SMO, Non-RT RIC, and Near-RT RIC, enabling e2e slice configuration and closed-loop network operation. Such solutions significantly reduce operational complexity and manual intervention.

The diversity of AI/ML techniques considered within 6G-SENSES demonstrates that future 6G systems should support heterogeneous AI workflows. Rather than targeting a specific learning paradigm, the Intelligent Plane should be flexible enough to support centralised and distributed learning, online and offline training and inference, efficient data collection and management, as well as future AI-driven network automation functionalities.

4.2 MLOps

MLOps has emerged as a key technical enabler for moving AI/ML models from experimental development to reliable operational deployment [35]. In the context of communication networks, and specifically in O-RAN-based architectures, MLOps are particularly relevant because AI/ML models are expected to support operational decisions that may affect radio resource management, energy optimisation, mobility management, sensing-aware control, traffic prediction, slicing and orchestration. These models cannot be treated as static software components. Their performance depends on input data distributions, training conditions, feature quality, deployment location, runtime constraints and feedback from the controlled system. Therefore, the Intelligent Plane requires not only AI/ML algorithms, but also an operational framework to manage their complete lifecycle. At a conceptual level, MLOps extends DevOps principles to ML systems. While DevOps focuses on the automation, integration, testing and deployment of software, MLOps adds specific mechanisms for dataset management, feature engineering, experiment tracking, model training, model validation, model registry, model serving, drift detection, monitoring and retraining. This distinction is important because an ML model deployed in a live network may degrade over time even if the surrounding software remains unchanged. Traffic patterns, user mobility, radio propagation, sensing conditions and service demands may evolve, causing data drift or concept drift. Without monitoring and lifecycle management, such changes may lead to incorrect predictions, unstable policies or sub-optimal control actions [36].

For the 6G-SENSES Intelligent Plane, MLOps should be understood as the operational backbone that supports AI/ML-based control loops across the SMO, Non-RT RIC, Near-RT RIC, edge/cloud infrastructure and potentially lower-layer RAN execution points. The AI/ML lifecycle in this context begins with the collection of telemetry and sensing data from RAN, transport, core, edge and external sources. This data must then be cleaned, transformed, aligned with metadata and converted into features that can be reused by different models. After model training and validation, the selected model must be registered, versioned and deployed to an appropriate execution environment. Depending on the control-loop requirements, inference may be performed in the SMO/Non-RT RIC domain, in the Near-RT RIC through xApps, in an external AI Engine, at the edge, or in future AI-native RAN functions. Once deployed, the model must be monitored to detect performance degradation, excessive latency, drift, abnormal behaviour or excessive resource consumption. These observations may trigger retraining, model replacement, policy adaptation or rollback.

The O-RAN architecture provides a natural framework for this lifecycle [16]. The Non-RT RIC and SMO domain is suitable for long-term analytics, model training, model governance, policy generation and model lifecycle control. The Near-RT RIC is suitable for near-real-time inference and control, where xApps consume telemetry and apply actions through the E2 interface under A1 policy guidance. This separation aligns well with MLOps principles, since training and model governance can be centralised or semi-centralised, while inference can be distributed closer to the control point. However, O-RAN alone does not define a complete implementation of the MLOps toolchain. Practical deployment requires additional components such as feature stores, object storage, model registries, workflow engines, serving runtimes, monitoring services and interfaces to expose model outputs to rApps, xApps or other consumers.

A useful reference implementation for this type of architecture is provided by **MLRun** (see Figure 4-1) [37]. MLRun is an open-source AI/MLOps orchestration framework designed to manage machine learning and generative AI applications across their lifecycle. It supports the automation of data preparation, training, validation, deployment, serving and monitoring workflows over elastic infrastructure, including Kubernetes-based environments. From the perspective of the Intelligent Plane, MLRun is attractive because it combines several capabilities that are otherwise often distributed across separate tools: project and pipeline management, function execution, model tracking, feature store integration, serverless serving, monitoring and integration with runtime engines such as Nuclio [38].

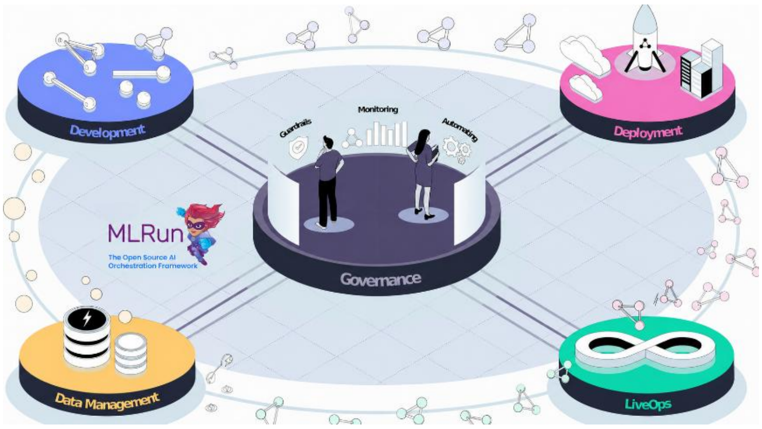


Figure 4-1 MLRun architecture [38]

One of the most relevant MLRun capabilities is its feature store. In network intelligence scenarios, models often require consistent access to historical and real-time features, such as cell load, PRB utilisation, SINR, RSRP/RSRQ, throughput, latency, energy consumption, user mobility, slice demand or sensing outputs. A feature store allows these features to be defined once and reused across training and serving, reducing the risk of training-serving skew. In MLRun, feature sets can ingest data from offline or online sources, apply transformations, store the resulting features and expose them for batch or real-time access. This is particularly relevant for 6G-SENSES, where the same telemetry or sensing-derived features may be consumed by several AI/ML models supporting different use cases.

Another relevant capability is model serving. MLRun supports real-time serving through Nuclio, a high-performance serverless runtime designed for data, I/O and compute-intensive workloads. This allows ML models to be exposed as scalable inference services, rather than being embedded directly into the application logic of an rApp or xApp. In an Intelligent Plane architecture, this decoupling is valuable: AI/ML developers can manage the model pipeline, while rApp/xApp developers can focus on control logic and consume model outputs through well-defined interfaces. This principle was already demonstrated in the BeGREEN AI Engine, where MLRun and Nuclio were used to host AI/ML services and expose inference outputs to RIC applications. For 6G-SENSES, a similar pattern can be used to expose traffic predictions, sensing predictions, energy-efficiency estimates, trajectory predictions or model recommendations to the Intelligent Plane.

MLRun also supports model monitoring, which is critical for safe operation in dynamic telecom environments. Monitoring should cover not only software-level metrics such as request rate, error rate, serving latency and resource utilisation, but also ML-specific indicators such as input feature drift, prediction drift, model confidence, accuracy degradation and retraining triggers. In a 6G-SENSES deployment, monitoring must also account for operational constraints: an AI/ML model that is accurate but too slow, too energy-hungry, or unstable under traffic bursts may not be suitable for near-real-time control. Therefore, model monitoring should be connected to both the MLOps layer and the network observability layer, so that model performance can be assessed together with network KPIs, energy consumption and service-level objectives. An integration with the MLOps model for the 6G-SENSES is described in Figure 4-2, showcasing the interactions among the lifecycle and the O-RAN architecture [39], [40].

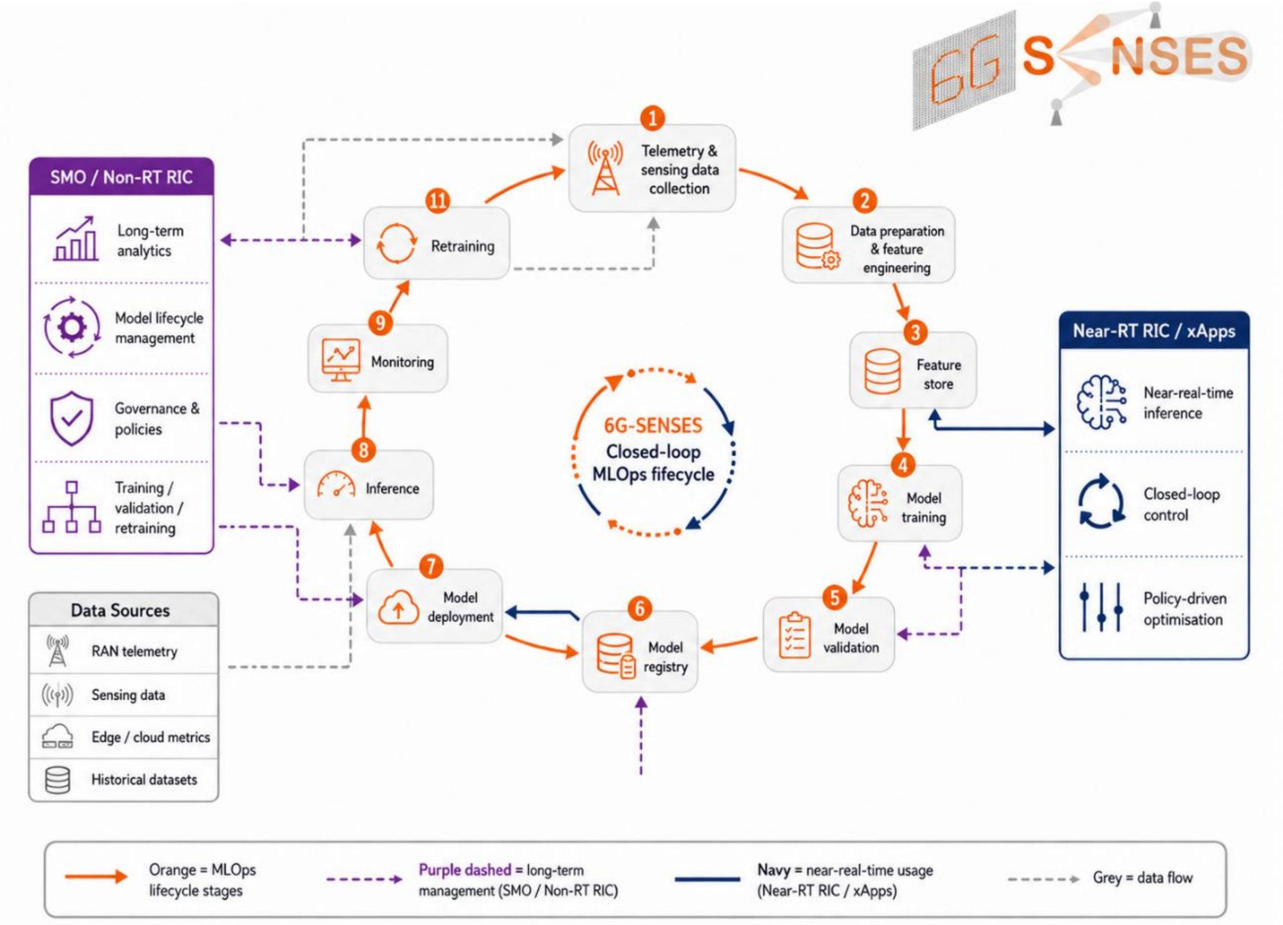


Figure 4-2 Generic MLOps lifecycle for the 6G-SENSES Intelligent Plane

From a SotA perspective, MLOps for telecom and O-RAN is still evolving. The most mature commercial and open-source MLOps platforms were originally designed for enterprise AI workloads, not for highly distributed, latency-sensitive RAN environments. In telecom networks, additional requirements arise models may need to be deployed across multiple edge locations, consume streaming telemetry, react to strict control-loop time scales, operate under limited compute resources, and satisfy safety constraints before influencing network behaviour. Furthermore, O-RAN introduces specific integration points through A1, R1, E2, O1 and O2, meaning that model outputs must be mapped into policies, enrichment information, RIC indications, control actions or orchestration decisions. This makes the combination of O-RAN and MLOps a relevant research and engineering challenge.

For 6G-SENSES, MLOps should therefore not be presented only as a generic software practice. It should be positioned as a technical enabler of the Intelligent Plane as depicted in Figure 4-3. Its role is to ensure that AI/ML models are reproducible, reusable, deployable, observable and governable. This is especially important because 6G-SENSES targets advanced scenarios involving ISAC, RIS, multi-RAT sensing, AI/ML-assisted optimisation, DTs and cross-domain orchestration. These scenarios require models to be trained and validated on heterogeneous data, deployed at different control-loop time scales, monitored continuously, and updated when the environment changes. A lightweight MLOps framework based on concepts demonstrated in MLRun can provide the necessary bridge between research algorithms and operational AI/ML services.

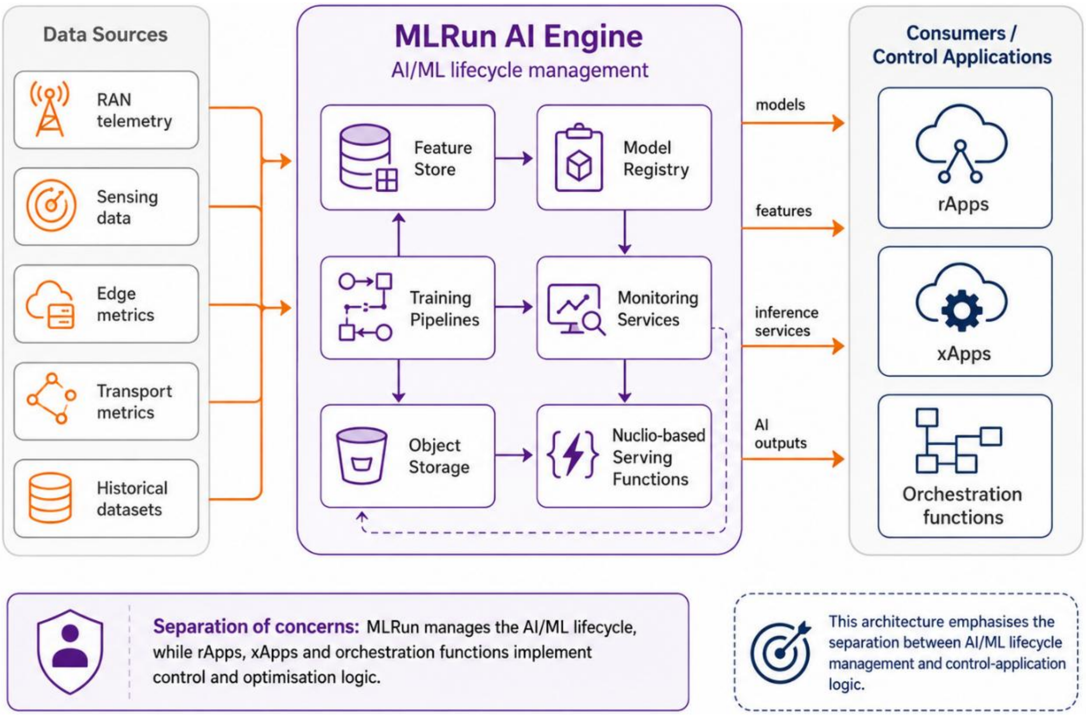


Figure 4-3 MLRun-based AI Engine concept for rApp/xApp model consumption

Finally, it is useful to clarify the relationship between MLOps and adjacent “Ops” concepts that are increasingly used in the AI and networking domains. AIOps, or Artificial Intelligence for IT Operations, refers to the use of AI/ML to automate operational tasks such as anomaly detection, event correlation, root-cause analysis, fault prediction and remediation [41]. In contrast, MLOps refers to the engineering discipline required to build, deploy, monitor and maintain the ML models themselves. In other words, MLOps operates the models; AIOps uses models to operate the system[42]. ModelOps is often used as a broader governance-oriented term covering model lifecycle management across different types of models and risk contexts. DataOps focuses on the reliable and automated delivery of data pipelines, while LLMOps focuses on the specific lifecycle requirements of large language models, including prompt management, retrieval augmentation, evaluation, guardrails and cost control [41].

For the 6G-SENSES deliverable, the recommended terminology is to use MLOps as the primary concept for AI/ML lifecycle management in the Intelligent Plane. AIOps can be introduced as a complementary future direction for autonomous operation of the 6G-SENSES infrastructure, especially for anomaly detection, self-healing, root-cause analysis and intent-driven operations. LLMOps should only be mentioned if the project explicitly incorporates large language models or generative AI agents for network management. DataOps can be referenced as a supporting discipline for telemetry and sensing data pipelines, while ModelOps can be used when discussing governance, auditability, model catalogues and policy-based lifecycle control. This avoids unnecessary buzzword accumulation while still positioning 6G-SENSES consistently with current AI-native network management trends.

4.3 3GPP NWDAF/ADRF

3GPP has progressively introduced analytics and AI/ML capabilities into the 5G System, creating a complementary intelligence framework to the O-RAN RIC-based approach. The main component is the NWDAF, which provides analytics services to 5G CN functions and other authorised consumers. Alongside NWDAF, the ADRF enables the storage and retrieval of analytics-related data, supporting the reuse of historical information and analytics outputs [43].

NWDAF operates within the 3GPP Service-Based Architecture (SBA) and exposes analytics to functions such as the Access and Mobility Management Function (AMF), Session Management Function (SMF), Policy Control Function (PCF), Network Slice Selection Function (NSSF), Network Exposure Function (NEF) and Application Function (AF). Typical analytics include network load, slice load, service experience, QoS sustainability, UE mobility, congestion and abnormal behaviour. These analytics support functions such as policy management, mobility optimisation, traffic steering and service assurance. ADRF complements NWDAF by providing persistent storage for analytics data. This allows analytics information to be shared across multiple consumers and reused over time, improving continuity and consistency in distributed deployments. From an Intelligent Plane perspective, ADRF acts as a repository of historical analytics that can support data-driven optimisation and AI/ML workflows.

For 6G-SENSES, NWDAF and ADRF represent the 3GPP counterpart of the O-RAN intelligence framework. While the SMO, Non-RT RIC and Near-RT RIC focus on RAN-domain intelligence, NWDAF and ADRF provide visibility into the CN and service domains. Combining these sources enables a broader view of network behaviour and supports cross-domain optimisation involving radio, core, edge and sensing information.

Beyond analytics, 3GPP has expanded its AI/ML activities through Release 18 and Release 19. These releases introduce AI/ML lifecycle management capabilities covering model training, testing, deployment and inference, as well as support for AI-assisted network optimisation use cases. This evolution aligns with the O-RAN vision, where programmable control through rApps and xApps is complemented by analytics, data exposure and AI/ML management functions across the wider 5G ecosystem [44].

In the context of 6G-SENSES, NWDAF/ADRF should be viewed as enablers of cross-domain intelligence rather than replacements for the RIC. The Intelligent Plane can combine RAN telemetry and sensing information from the O-RAN ecosystem with core-network and service-level analytics provided by NWDAF and ADRF. This integration is particularly valuable for Network Digital Twin and energy optimisation use cases, where accurate decisions require both radio-domain measurements and e2e network context.3GPP AI/ML evolutions, as depicted in Figure 4-4.

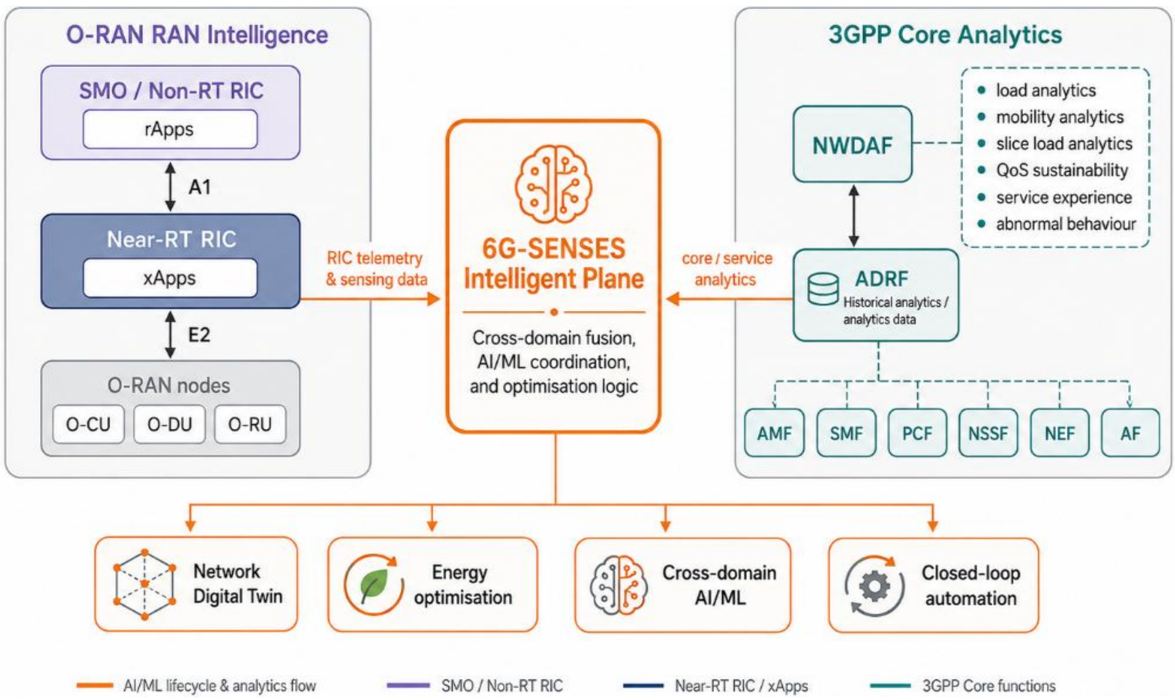


Figure 4-4 3GPP NWDAF/ADRF and O-RAN Intelligent Plane alignment

4.4 Evolution of the Intelligent Plane under O-RAN

The evolution of the 6G-SENSES Intelligent Plane follows the broader O-RAN direction towards open, programmable and AI-assisted network operation. In this context, the Intelligent Plane is not introduced as a separate monolithic entity, but as an evolution of existing O-RAN control and management functions, complemented with additional data, AI/ML and cross-domain coordination capabilities. The main architectural anchors are the SMO, the Non-RT RIC and the Near-RT RIC, which together provide the framework for long-term analytics, model lifecycle management, policy generation, near-real-time inference and control. In the 6G-SENSES implementation, this evolution is reflected through the use of the dRAX RIC platform as the operational basis for telemetry collection, rApp/xApp execution and AI/ML-assisted control loops [45].

A key direction in this evolution is the move towards a common cross-domain SMO. Future intelligent network operation requires visibility beyond the RAN alone. The SMO is expected to coordinate information coming from O-RAN telemetry, 3GPP analytics functions such as NWDAF/ADRF, N3GPP integrated access such as Wi-Fi, and non-integrated sensing sources such as radar, lidar or other external sensing systems. For 6G-SENSES, this is particularly relevant because the project combines communication and sensing information to support NDT creation, energy optimisation and sensing-aware control. The Intelligent Plane must therefore be able to ingest and correlate heterogeneous information sources, while exposing them in a usable form to AI/ML workflows and control applications [19], [46].

The RIC functions also evolve accordingly. The Non-RT RIC becomes the natural location for long-term intelligence, including data analysis, model training, policy generation, model selection and retraining. The Near-RT RIC remains the main execution environment for near-real-time inference and control, where xApps consume live telemetry and apply decisions through O-RAN interfaces [23]. This separation is important because not all AI/ML functions have the same latency, compute or data requirements. Some models require historical data and can be trained or updated over longer time windows, while others need to operate close to the live network state and produce timely outputs for control decisions.

The xApp/rApp framework provides the extensibility mechanism required by this evolution. rApps can be used to implement non-real-time analytics, model lifecycle functions and policy-oriented optimisation, while xApps can implement near-real-time inference and control logic. In a dRAX-based implementation, these applications can consume telemetry from the RIC data bus and historical databases, interact with MLOps capabilities such as MLRun, and exchange AI-related information across the non-real-time and near-real-time domains. This creates a flexible implementation path where lightweight AI/ML workflows can be implemented directly through rApps and xApps, while more structured model lifecycle management can be supported by an external MLOps component.

Beyond the current RIC split, O-RAN is also evolving towards more native AI/ML concepts in the RAN. This includes the discussion of real-time or closer-to-RAN intelligence, where AI/ML functions may need to operate with tighter latency constraints and closer interaction with DU/CU or lower-layer functions. Concepts such as RT-RIC and dApps point in this direction, enabling more distributed AI/ML execution where the Near-RT RIC may not be sufficiently close to the control target. For 6G-SENSES, this evolution is relevant for future sensing-aware control, RIS coordination, beam management, advanced radio optimisation and other functions that may require fast processing of radio and sensing information [47].

Telemetry, monitoring and data collection mechanisms are the foundation of this evolution. The Intelligent Plane depends on the ability to collect live and historical data, organise it into datasets, monitor model behaviour and feedback performance information for retraining or policy adaptation. In practice, this requires the combination of streaming telemetry, historical databases, logs, event information, sensing data and model monitoring outputs. These mechanisms allow the Intelligent Plane to move from static model deployment towards continuous model operation and actualisation.

Overall, the evolution of the Intelligent Plane under O-RAN can be understood as the gradual integration of data collection, AI/ML lifecycle management, rApp/xApp programmability, cross-domain analytics and near-real-time control into a coherent operational framework. The technical enablers described in this chapter define the architectural direction. Chapter 5 builds on this foundation and describes how these concepts are implemented in 6G-SENSES through the dRAX platform, MLRun-based or rApp-based MLOps workflows, RIC telemetry mechanisms, and their integration with the PoC#3 NDT scenario.

4.5 Network Automation and NDT, NBI's

The concept of NDT and the interfaces to interact with it are under consideration by different standardisation bodies. Within the IETF [48] a 3-tier architecture is being considered which embraces the (i) physical network, (2) NDT and (3) application layer. In this architecture the north-bound interface (NBI) allows the communication between the application and the NDT, and lightweight extensible Restful interface is considered as a candidate. Furthermore in [49] the properties of the NBI are further analyzed, highlighting the following features:

- Openness: to meet business requirements of different network applications.
- Scalable: to support a growing number of network applications.
- Portable: network applications can be deployed on top of the NDT of different sizes and with different properties (i.e. representation scale, precision, etc.). Thus, the NBI should be easily portable and deployable in different NDT.
- Flexible: to reduce deployment time and cost.

Within the 3GPP context, the NDT concept is first introduced in [50] as a key element to enhance network automation functions. The network automation comprises both analytics services and decisions functions, which assume the availability of the network behavior information. In this sense, the NDT is responsible on providing such behavioral model. In [3] it is considered that the NDT can be deployed inside or outside the network function that consumes the model, so that NDT NBI interface would be only present in the second case. Regardless of the deployment options, three use cases are envisaged for NDT supporting network automation within 3GPP. In the first use case the NDT would help with the evaluation of high-risk network operations, such as network congestion or network breakdown. In this regard the NDT could be used to evaluate known risky network configurations and to verify optimisation policies that may hinder the network operation. In the second use case, the NDT would be adopted to evaluate responses to failure events, where signaling storm is especially relevant. In this sense, the automation function would develop a response to the failure event, which would be first evaluated within the NDT to anticipate its impact. The last use case takes a different approach, where failures are induced in the NDT to evaluate the response of the automation functions. This use case considers three scenarios, embracing service degradation, coverage issue and fault injection. In [3] initial data structure of the interface for NDT Life Cycle Management (LCM) and queries are provided. Furthermore, the NDT interface operates as a Management Service (MnS), so it is assumed to be a service-based interface, similar to the approach proposed by the IETF.

4.6 AI-RAN Considerations

The integration of AI RAN into the 6G SENSES framework marks a fundamental shift from conventional RAN architectures towards a converged compute communication platform that directly advances the project objectives in sensing, communication, and energy optimisation [51]. This transformation brings significant challenges related to spectrum allocation, network architecture, and resource management, which must be overcome through the adoption of advanced machine learning and edge computing technologies. The convergence of communication and computing in AI RAN manifests in three complementary forms. AI for RAN improves network spectral and operational efficiencies. AI on RAN enables new AI applications hosted on the RAN infrastructure. AI and RAN aids higher asset utilisation by sharing compute resources between AI and RAN functions. For 6G SENSES, this multi-faceted integration is particularly relevant as the project combines communication, sensing, computing, and orchestration across a heterogeneous infrastructure that includes 3GPP and non 3GPP access technologies, ISAC capabilities, RIS assisted operation, and Cell Free massive MIMO concepts [52]. AI for RAN directly influences the physical and medium access control layers by enabling adaptive beamforming, dynamic interference mitigation, and intelligent scheduling, all of which are essential for managing the dense, multi-RAT sensing environment targeted by the project. Concurrently, AI on RAN unlocks the potential to host third party or vertical specific AI applications directly at the edge, transforming the RAN from a pure connectivity pipe into a service delivery platform that can provide low latency, context aware services while generating new operational value. Finally, AI and RAN emphasises the dynamic pooling of heterogeneous compute resources such as GPUs and DPUs so that unused RAN processing capacity can be repurposed for AI inference or training during off peak hours, a principle that directly supports the 6G SENSES objective of improving asset utilisation without compromising carrier grade performance [53].

The impact of AI RAN on the 6G SENSES project is twofold, addressing both the operational intelligence layer and the underlying infrastructure economics. First, AI RAN enables the Intelligent Plane to move beyond isolated AI and ML algorithms toward a unified operational framework where sensing data, communication telemetry, and AI models are jointly managed across the SMO, Non-RT RIC, and Near RT RIC domains. This aligns directly with the 6G SENSES objective of establishing a cross domain AI architecture that coordinates data, models, and policies across RAN, core, transport, cloud, and application domains. The emergence of AI RAN also introduces advanced orchestration frameworks that can seamlessly manage compute and communication resources while accounting for latency, reliability, and energy constraints. For 6G SENSES, this means the Intelligent Plane can leverage AI RAN principles to support distributed AI and ML deployment strategies from long term analytics in the SMO and Non-RT RIC to near real time inference in the Near RT RIC, and potentially closer to RAN execution for future native AI and ML functions. The joint orchestration mandated by AI RAN aligns perfectly with the MLOps lifecycle described earlier in this deliverable, as it requires continuous model training, validation, deployment, and monitoring across distributed edge nodes. This synergy allows the Intelligent Plane to automate the entire feedback loop from collecting raw telemetry and sensing data, through updating models via structured MLOps frameworks, to enforcing the resulting policies through the Near-RT RIC, thus reducing the mean time to resolution for network anomalies and adapting swiftly to evolving radio conditions. Second, AI RAN transforms the economic and operational viability of the architecture by enabling concurrent processing of RAN workloads and AI inference tasks on accelerated hardware, thereby improving asset utilisation and reducing the overall energy footprint of network operations [54]. For PoC#3 and the NDT scenarios, AI RAN provides the computational headroom required to run high fidelity DT simulations concurrently with live network operations, allowing the Intelligent Plane to safely stress test candidate optimisation strategies such as RIS configuration adjustments or energy saving cell switch offs against a virtual replica of the physical environment before committing them to the live infrastructure. Additionally, the native AI support within AI RAN facilitates the use of explainability and drift detection mechanisms, ensuring that models deployed on the Intelligent Plane remain robust against data distribution shifts caused by changing user mobility, interference patterns, or environmental conditions, thereby enhancing the overall trustworthiness and resilience of the 6G SENSES control loops.

5 6G-SENSES Intelligent Plane implementation

This chapter moves from the architectural and technical enabler discussion towards the concrete implementation of the 6G-SENSES Intelligent Plane. It describes how the Intelligent Plane can be realised through distributed AI/ML deployment strategies, sensing-aware and native AI/ML functions, dRAX-based RIC capabilities, rApps, xApps, telemetry databases, MLOps workflows and Network Digital Twin integration.

The chapter focuses on the operational path required to transform raw network and sensing data into reusable datasets, trained models, live inference and optimisation actions. It also clarifies how different implementation options can coexist, from lightweight rApp/xApp-based workflows to more structured MLOps-supported deployments, and how these capabilities contribute to PoC#3 and the wider WP5 demonstrator integration.

5.1 Distributed AI/ML Deployment Strategies

Distributed AI/ML deployment in the 6G-SENSES Intelligent Plane is driven by the fact that not all intelligence functions have the same latency, data-volume, privacy, compute or control requirements. Model training, dataset preparation, model validation, policy generation and retraining are naturally placed in non-real-time or management-domain environments, such as the SMO, Non-RT RIC or an external MLOps platform. These functions require access to historical datasets, larger compute resources and model lifecycle services, but they do not usually require immediate interaction with the live radio interface. By contrast, inference functions that support near-real-time optimisation need to be placed closer to the operational control point, for instance in xApps running in the Near-RT RIC, where live telemetry can be consumed and translated into predictions, recommendations or control actions. This separation between training and inference is aligned with both the O-RAN AI/ML workflow and the 3GPP AI/ML management view, where different deployment options are considered for model training, model deployment and inference execution [16], [55], [56].

For 6G-SENSES, this leads to a multi-tier deployment strategy. Long-term intelligence, including traffic prediction, energy modelling, model selection, policy generation and NDT model actualisation, can be executed in the SMO/Non-RT RIC or in an associated MLOps environment such as MLRun. Near-real-time intelligence, such as radio-condition estimation, sensing-aware optimisation, mobility-driven adaptation or energy-aware control, can be executed through xApps consuming live telemetry from the RIC data bus. In future extensions, some functions may need to move even closer to the RAN, edge or DU/CU domain when strict latency or bandwidth constraints make centralised inference impractical. This cloud-edge continuum is particularly relevant for AI/ML-assisted RIC deployments, where performance-critical components and selected xApps may need to run at the edge while less time-sensitive functions remain in cloud or management-domain locations [57].

5.2 Support for ISAC and Native AI/ML Evolution

By leveraging the BubbleRAN MX-PDK, the architecture implements a complete e2e pipeline for streaming, processing, and executing inference on sensing data extracted directly from the radio access network infrastructure (Figure 5-1) [4].

The physical validation framework is established on a fully disaggregated, standards-compliant Open RAN (O-RAN) stack utilising a Lite-On 7.2-split Radio Unit (O-RU). Rather than relying on secondary, specialised sensor overlays, the network natively multi-purposes existing cellular communication signals for passive environmental sensing. Within the Near-Real-Time RAN Intelligent Controller (Near-RT RIC), specialised xApps are deployed to manage the collection and lifecycle of sensing telemetry. These xApps ingest high-fidelity signal metrics via an extension of the O-RAN service model paradigm, specifically utilising a low-latency service model to extract upstream channel measurements.

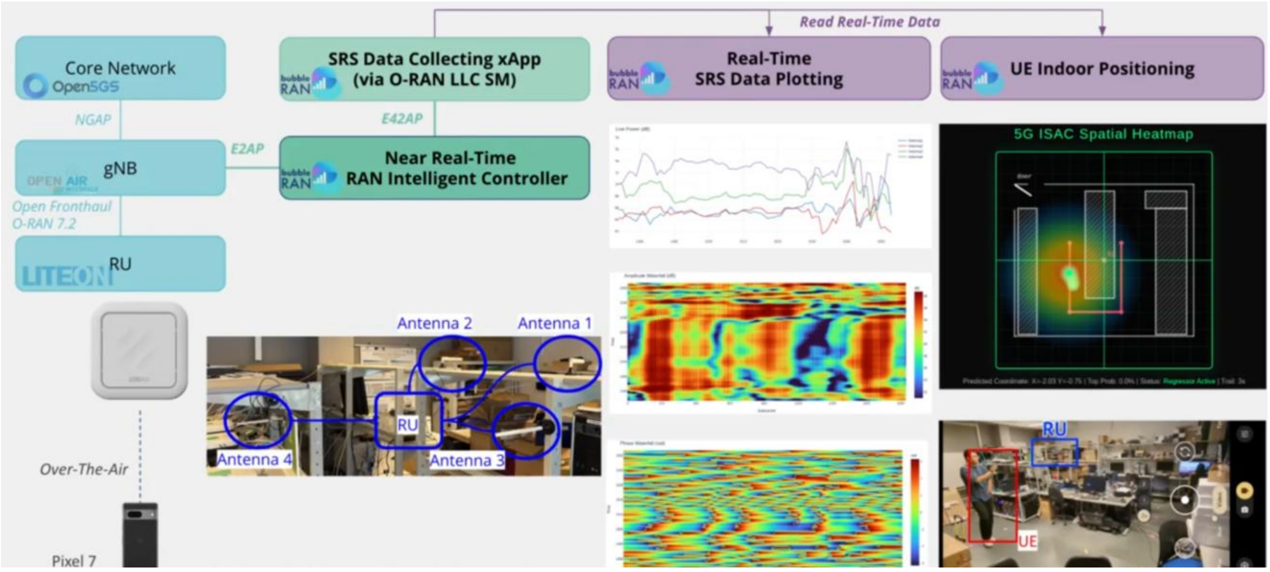


Figure 5-1 E2e physical validation framework

The raw telemetry collected by the ingestion fabric consists of complex-valued raw In-Phase and Quadrature (I/Q) data samples captured from uplink SRSs. To render this high-volume data ingestible for edge-bound ML models, the xApp executes localised physics-aware feature engineering, transforming the raw channel matrices into a compact, 32-dimensional (32-D) spatial feature vector

The resulting 32-D spatial feature vectors are fed into an edge-ready ML model (such as a Random Forest regressor or a localised deep neural network) executing near-real-time inference on standard host CPUs to support three primary spatial intelligence applications:

- **Real-Time UE Positioning and Tracking:** Maps the 32-D channel fingerprints to continuous 2D spatial coordinates (x,y), providing deterministic user location estimation alongside a dynamic spatial probability distribution map.
- **Passive Object and Human Detection:** Monitors localised anomalies, shadowing effects, and sharp power drops within the amplitude/phase waterfall representations to detect, classify, and alert on physical line-of-sight obstructions without active device participation.
- **Indoor Multi-UE Collision Risk and Density Estimation:** Aggregates location probabilities across multiple concurrent device identifiers (IDs) to map spatial density thresholds. By tracking overlapping probability clouds, the Intelligent Plane proactively flags localised activity hotspots and outputs preventative collision alerts for autonomous or industrial indoor environments.

5.3 AI/ML-Driven Energy Optimisation and Implementation

The implementation of the 6G-SENSES Intelligent Plane is based on the integration of AI/ML capabilities into the operational network framework, rather than deploying them as external tools. From an implementation perspective, three capabilities are required: access to heterogeneous network data, support for MLOps functions such as data preparation, model training, storage and retraining, and the use of trained models for live inference and intelligent network management. Together, these capabilities form the practical implementation pillars of the Intelligent Plane [16], [35].

The first pillar is data availability. The dRAX RIC environment provides access to radio performance metrics such as RSRP, RSRQ, throughput and PRB usage; infrastructure metrics such as CPU, RAM and network usage; network events such as UE connection and disconnection; and logs from CU, xApps and platform components. This data is available both as streaming telemetry for near-real-time consumption and as historical data for offline analysis, dataset construction and model training.

The second pillar is MLOps. Raw telemetry collected by the RIC must be selected, cleaned, aligned, transformed and converted into training-ready datasets before it can be used by AI/ML algorithms. For example, a model estimating UE location from RSRP and RSRQ measurements requires radio measurements to be extracted, correlated with location information and prepared in a consistent format. MLOps therefore provides the workflow needed to move from raw network data to trained and reusable models[35].

The third pillar is inference and control. Once trained and stored, models must be used by the network to support operational decisions. In the dRAX implementation, this role is mainly associated with xApps in the Near-RT RIC, which consume live data from the dRAX data bus, execute inference and use the model output to support prediction, classification, optimisation or control actions. For energy optimisation, this pattern can be used to predict traffic, estimate radio conditions, forecast energy demand or support energy-aware configuration decisions [16], [58].

Model retraining is also essential, since traffic patterns, user mobility, radio propagation and environmental conditions evolve over time. A model trained under one scenario may become less accurate as network conditions change. The Intelligent Plane must therefore support retraining or fine-tuning based on updated historical data, ensuring that deployed models remain aligned with current network behaviour.

In the dRAX-based implementation, historical data is stored in dedicated databases. Time-series performance metrics are stored in InfluxDB, counter-based KPIs and infrastructure metrics are stored in Prometheus, and events and logs are stored in Loki. These repositories provide the data foundation required for dataset generation, training and retraining, while the retention period can be configured according to the needs of the ML workflow.

Two implementation approaches are considered. The first is an rApp-based workflow, where a custom rApp queries the dRAX databases, prepares datasets, trains models using Python libraries, stores the results and exposes them through an API or internal mechanism. This approach is lightweight and suitable for prototyping or narrowly scoped use cases. The second approach integrates MLRun as a dedicated MLOps platform deployed alongside dRAX. In this case, dRAX collects and exposes the data, while MLRun manages dataset preparation, model training, storage and lifecycle management. This approach is more appropriate when several models, datasets and retraining cycles must be maintained [58], [59].

MLRun should therefore be understood as the MLOps component of the Intelligent Plane, not as the Intelligent Plane itself. The Intelligent Plane is the broader framework composed of the RIC, telemetry sources, databases, rApps, xApps, A1-based workflows, MLOps capabilities and inference/control loops. In this architecture, dRAX provides the operational network intelligence framework, while MLRun contributes structured AI/ML lifecycle management.

For energy optimisation, the preferred implementation pattern is a split workflow. Historical data is used in the non-real-time domain to create or update models through MLRun or custom rApps. In the near-real-time domain, xApps use trained models and streaming telemetry to execute inference and support optimisation decisions. Model performance can then be fed back to the non-real-time domain to trigger retraining or updates, creating a closed AI/ML operational loop for adaptive energy optimisation.

5.4 Intelligent Plane Implementation and Integration in 6G-SENSES

The implementation of the 6G-SENSES Intelligent Plane is defined around the RIC as the central operational framework. In this view, the RIC is not only an entity connected to the Intelligent Plane, but the practical foundation through which the Intelligent Plane is realised. The RIC provides the essential execution environment for network intelligence: it collects and stores telemetry, exposes live data streams, hosts non-real-time and near-real-time applications, and enables the interaction between analytics, AI/ML workflows and control actions. Therefore, the 6G-SENSES Intelligent Plane is implemented as a RIC-based system extended with MLOps capabilities, AI/ML-driven applications and NDT functions.

This implementation approach is particularly relevant for PoC#3, which focuses on the creation and use of a NDT via RAN sensing. PoC#3 requires the integration of real network telemetry, sensing information, model generation, simulation environments and optimisation actions. As a result, it provides the most complete context for exercising the Intelligent Plane within 6G-SENSES, and relies on the ability of the RAN and associated sensing sources to generate data that can be used to model the radio environment, refine NDT models and evaluate candidate network configurations before applying optimisation decisions to the real system.

5.4.1 RIC-based implementation baseline

The dRAX platform provides the baseline implementation environment for the Intelligent Plane described in Figure 5-2. Its role is to expose the network data, host the applications that consume this data, and support the execution of AI/ML-assisted control loops. The data layer is one of the most important parts of this implementation. dRAX provides live data through its near-real-time data bus, which can be consumed by xApps for inference and control. At the same time, it stores historical information in dedicated databases that can be used for offline analysis, dataset generation, model training and retraining [58].

Different categories of data are managed through different storage components. Time-series performance metrics, such as RSRP, RSRQ, throughput and PRB usage, are stored in InfluxDB. Counter-based KPIs and infrastructure metrics, such as CPU usage, RAM usage, disk usage, handover success rate, handover failure rate and Radio Resource Control (RRC) connection statistics, are stored in Prometheus. Event data and logs, including UE connection events, User Equipment (UE) disconnection events, handover events and logs from system components such as the CU and xApps, are stored in Loki. Together, these components provide the historical and operational data foundation required by the Intelligent Plane.

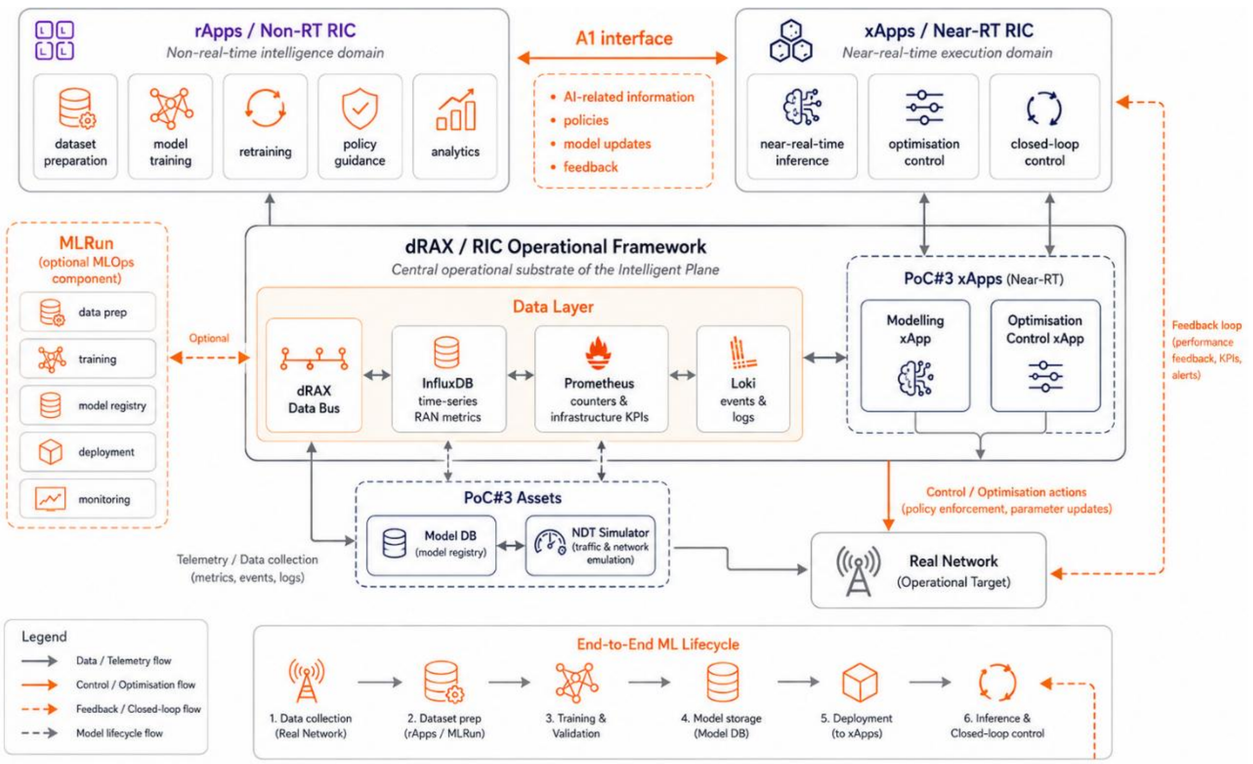


Figure 5-2 Intelligent Plane implementation architecture

This data foundation is important not only for classical AI/ML training, but also for NDT generation. An NDT requires both historical and live network information to represent, simulate and update the behaviour of the real environment. In PoC#3, this means that the RIC data layer becomes the source from which the radio environment can be characterised and translated into NDT-compatible models.

5.4.2 MLOps integration and model lifecycle support

The availability of raw telemetry is only the first step. To become useful for AI/ML, this data must be selected, cleaned, aligned, transformed and converted into datasets suitable for training and validation. In the current implementation strategy, this MLOps capability can be realised in two complementary ways.

The first option is to implement the required AI/ML workflow directly through rApps [58], [60]. In this mode, a custom rApp retrieves historical data from the dRAX databases, prepares the dataset, performs model training using standard Python libraries, stores the trained model and exposes the result to other applications. This approach is lightweight and suitable for fast prototyping, because it does not require the deployment of an additional MLOps platform. It also provides flexibility, as the rApp developer can use the most appropriate AI/ML framework depending on the specific use case.

The second option is to integrate MLRun as a dedicated MLOps component [59]. In this case, MLRun is deployed alongside dRAX and connects directly to the RIC databases, without requiring an rApp to act as an intermediate bridge. The RIC remains responsible for collecting and exposing the network data, while MLRun provides a more structured framework for dataset preparation, model training, model storage, model lifecycle management and potentially model monitoring. This option is more suitable when the Intelligent Plane must manage several models, datasets, retraining cycles and application consumers in a reusable and maintainable way.

It is important to clarify the architectural role of MLRun in this implementation. MLRun is not the Intelligent Plane by itself. Instead, MLRun provides the MLOps part of the Intelligent Plane. The Intelligent Plane is the broader operational framework composed of the RIC, telemetry sources, historical databases, rApps, xApps, control interfaces, AI/ML workflows and model-driven control loops. This distinction allows 6G-SENSES to maintain a flexible implementation path: simple workflows can be implemented directly through rApps, while more mature or reusable workflows can be supported through MLRun.

5.4.3 Non-real-time learning and near-real-time inference

The Intelligent Plane implementation follows a functional split between non-real-time learning and near-real-time inference. In the non-real-time domain, historical data is used to create, train, validate and update models. This process can be carried out by an rApp or by MLRun, depending on the chosen deployment mode. The non-real-time domain is therefore responsible for model lifecycle functions such as dataset preparation, training, model storage, model selection and retraining.

In the near-real-time domain, xApps consume live data from the dRAX data bus and use trained models to perform inference. This is where the trained AI/ML model becomes operational. The xApp retrieves or receives the relevant model, processes live measurements and produces an output that can support intelligent network management. Depending on the use case, this output may correspond to a prediction, classification, recommendation, optimisation decision or control action. For example, an xApp could use live RSRP and RSRQ measurements to support UE localisation, forecast future radio conditions, or trigger optimisation logic based on the current network state.

The interaction between non-real-time and near-real-time functions is enabled through the A1 interface. In the 6G-SENSES implementation, A1 is considered a key mechanism for AI workflows because it supports the exchange of information between rApps and xApps. This information may include updated ML models, enrichment information, AI-related metadata or feedback on model performance. As a result, the Intelligent Plane can establish a loop where models are trained or updated in the non-real-time domain, consumed by xApps in the near-real-time domain, monitored during operation, and retrained when the observed performance indicates that the model is no longer representative of the current network conditions.

This model is directly relevant for 6G-SENSES because the project targets dynamic radio and sensing environments. Network conditions may change due to traffic evolution, UE mobility, propagation changes, weather, interference, environmental modifications or sensing-driven context updates. Therefore, the Intelligent Plane must support not only model creation, but also model actualisation. Retraining is a core capability, ensuring that deployed models can remain aligned with the evolving network.

5.4.4 Integration with PoC#3 Network Digital Twin

PoC#3 applies this Intelligent Plane implementation to the creation and use of NDT models. The PoC#3 architecture is organised around two main application components: the Modelling xApp and the Optimisation Control xApp.

The Modelling xApp is responsible for the lifecycle of the NDT models. It includes model management, model serving and implementation, model monitoring and a model database. The Model Management function creates versions and controls the lifecycle of the models. The Model Serving and Implementation function translates model outputs into network configuration parameters. The Model Monitoring function compares model behaviour with real network observations and supports model update procedures. The Model Database stores generated, validated and updated models.

The Optimisation Control xApp complements the Modelling xApp by orchestrating the decision-making process. It receives high-level target scenarios from the user and translates them into optimisation objectives, such as improving energy efficiency, enhancing coverage or increasing throughput. Based on these objectives, it interacts with the Modelling xApp to select, evaluate and apply the most suitable model. This separation of responsibilities allows PoC#3 to distinguish between model lifecycle management and optimisation control, while keeping both functions connected through a closed-loop process.

The operation of PoC#3 is organised into three interrelated cycles: Model Creation, Model Actualisation and Validation, and Model Implementation, described in Figure 5-3 and in the following text [61], [62], [63].

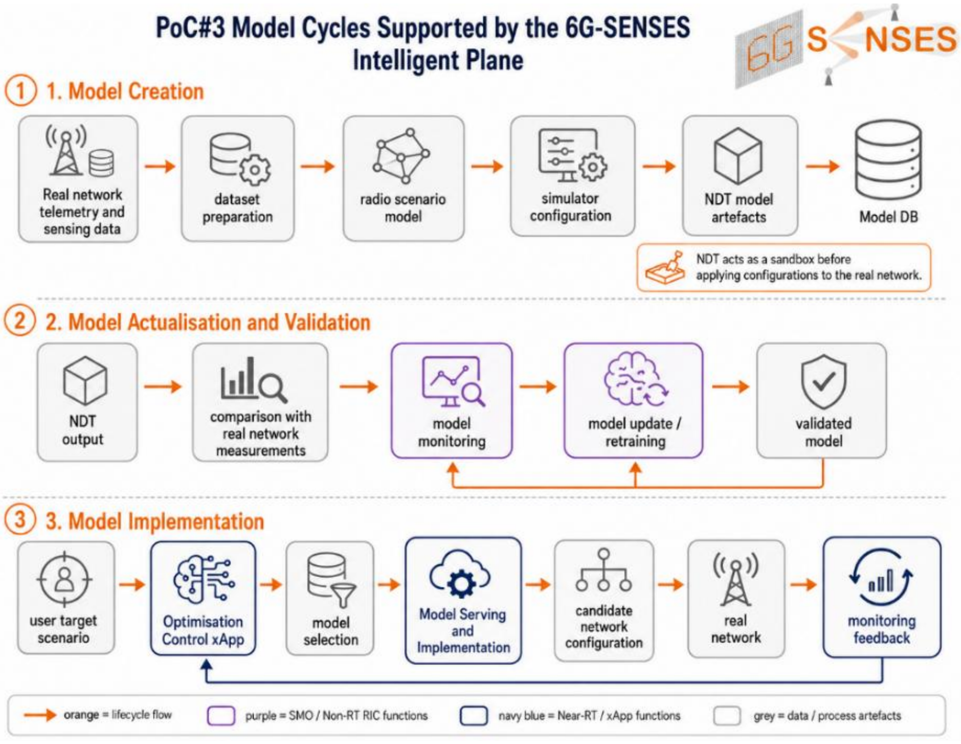


Figure 5-3 Model lifecycle in the Intelligent Plane

- The ***Model Creation cycle*** generates the initial NDT models from real network observations. It starts with the collection of telemetry data from the real network, including radio performance measurements, configuration parameters, contextual information and, when available, sensing-related inputs. The datasets may include metrics such as GPS position, latency, downlink bitrate, uplink bitrate, RSRP, RSRQ, RSSNR, Channel Quality Indicator (CQI) and Received Signal Strength Indicator (RSSI), together with network planning information such as gNB location, cell reselection events, handover information, cell identifiers and user location. This information is ingested by the Model Management function and used to create a structured representation of a specific radio environment scenario [64]. In this implementation, a model represents a radio scenario composed of known RU locations, known UE locations and associated radio metrics. This model representation is then translated into simulator-compatible configuration files that define the input conditions for the NDT environment. The NDT simulator executes the scenario and generates synthetic outputs corresponding to the defined radio environment. These model artefacts are stored in the Model Database. At this stage, the generated model is an initial approximation of the real network and must be refined through validation [64].
- The ***Model Actualisation and Validation cycle*** provides this refinement. The Model Monitoring function compares NDT simulation outputs with actual measurements from the real network. When discrepancies are identified, the model parameters or assumptions are adjusted to improve alignment with the observed behaviour. This process is similar to a CI/CD-like loop for models, where updates are continuously evaluated before being stored or used for decision-making. In the context of radio modelling, this may include updating propagation assumptions, refining environmental parameters, or modifying the mapping between measured KPIs and simulated behaviour [62].
- The ***Model Implementation cycle*** applies validated models to optimise the real network. The cycle is initiated when a user-defined target scenario is provided through the user interface. The Optimisation Control xApp translates this target into optimisation objectives and queries the Model Database for suitable models. The selected model is then processed by the Model Serving and Implementation function, which maps NDT-derived outputs into actionable network configuration parameters. These parameters can be evaluated in the simulated environment and, once validated, applied to the real network. After deployment, the network response is monitored and fed back into the optimisation process, enabling iterative refinement.

5.4.5 Deployment strategy and integration path

The implementation path for PoC#3 is organised in two main stages as presented in Figure 5-4. The first stage focuses on collecting network information to train and initialise the NDT. This requires a real laboratory setup connected to the dRAX telemetry ingestion framework, with an xApp or rApp collecting and processing the required measurements. The collected data may include measurements taken at known locations, potentially complemented by external sensing information such as radar-assisted UE location. This stage provides the datasets required to initialise the NDT and generate the first version of the radio environment model.

The second stage focuses on selecting, validating and applying trained models or NDT models for optimisation purposes. In this stage, the NDT is used as a controlled environment where alternative network configurations can be evaluated before being applied to the real system. This is particularly important for energy-efficient optimisation, since candidate configurations should be tested against coverage, throughput and reliability constraints before being enforced in the network.

The practical execution is organised through two deployment phases. The first phase is the pre-trial at the **OTE** testbed in Athens, where the DT rApp is pre-tested and network performance datasets are generated for estimating NDT performance. In this phase, an **ACC** O-RAN gNB is deployed and integrated with the **OTE** 5G CN. The generated datasets support the initial assessment and training of the PoC#3 DT rApp.

The second phase is the final trial at UC in Santander, where the Intelligent Plane is expected to support a more complete closed-loop scenario. This trial uses an open 5G Standalone Non-Public Network provided by **ACC** and extended with the 6G-SENSES Intelligent Plane. The environment hosts AI/ML-based PoC#3 applications and is enriched with sensing telemetry from **WP3** and PoC#1. In this phase, the rApp function models the radio environment and dynamically generates configuration files and simulation data for the NDT simulator, based on Sionna RT [65]. The expected outcome is a ray-tracing-based NDT system model that can support validation and optimisation of network configurations.

Overall, this implementation confirms that the 6G-SENSES Intelligent Plane is not only an architectural concept, but a concrete integration framework. The dRAX platform provides the RIC execution environment, telemetry ingestion, databases, rApps, xApps and near-real-time data access. MLRun, where deployed, provides the structured MLOps capability for model lifecycle management. The Modelling xApp manages NDT model creation, monitoring, serving and storage. The Optimisation Control xApp coordinates target-driven optimisation decisions. The real network and the NDT simulator provide the two complementary environments where models are trained, validated and applied. Together, these components realise the Intelligent Plane as an operational framework for AI/ML-driven, sensing-assisted and energy-aware network optimisation in 6G-SENSES.

5.5 Support towards WP5 Demonstrators and End-to-End Integration

The implementation approach described above provides a practical path for supporting **WP5** demonstrators and e2e integration. The main value of the Intelligent Plane is that it converts network telemetry into operational intelligence. This requires a complete loop from data collection to model training, model deployment, inference and feedback. The transcription confirms that the dRAX-based RIC environment already supports several of the necessary building blocks for this loop, while MLRun can be added to provide a more structured MLOps component.

For **WP5**, the first relevant contribution is dataset generation. The RIC collects radio, infrastructure, event and log information and stores it in databases that can be queried for historical analysis. This allows the demonstrators to create datasets from real or emulated network operation. These datasets can support training, validation and future DT creation. Since the data retention policy is configurable, the system can be adapted to store longer historical windows when needed by the demonstrator.

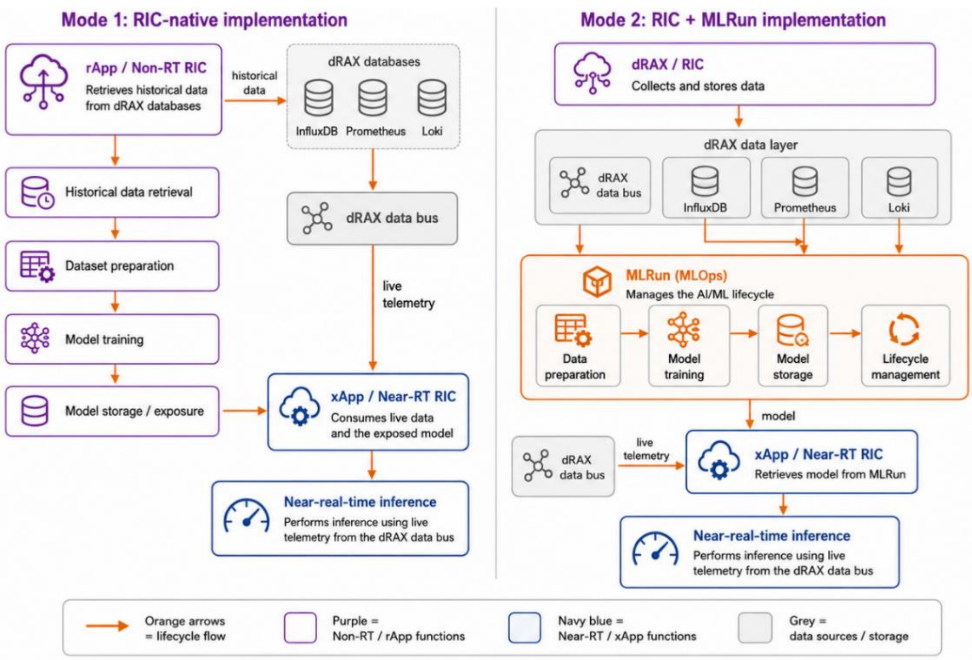


Figure 5-4 Implementation modes for the 6G-SENSES Intelligent Plane

The second contribution is the support for AI/ML training and retraining. **WP5** demonstrators involving AI/ML require not only a trained model, but also the ability to update the model when the network scenario changes. The implementation supports this either through rApp-based training or through MLRun-based MLOps. In the rApp-based mode, the workflow is implemented directly in code. In the MLRun-based mode, MLRun connects to the RIC databases, prepares the data, trains the model and stores it in a model database. Both options enable model creation and retraining, which are essential for realistic e2e demonstrations.

The third contribution is near-real-time inference. Once a model has been trained, xApps can retrieve or receive the model and execute inference using live data from the dRAX data bus. This allows the demonstrators to show that AI/ML outputs can be used operationally, not only as offline analysis. The xApp can consume live measurements, provide them to the model and generate predictions or recommendations that can support intelligent network management. This is a key step towards demonstrating closed-loop operation.

The fourth contribution is the connection between non-real-time and near-real-time domains. The A1 interface provides a mechanism for exchanging AI-related information between rApps and xApps. This enables a complete lifecycle: the rApp or MLOps component can train or update a model; the xApp can use the model for inference; the xApp can expose model performance information; and the rApp can use this feedback to retrain or update the model. This interaction is highly relevant for **WP5** because it supports iterative validation and refinement of AI/ML behaviour during the demonstrations.

The implementation also supports different levels of demonstrator complexity. For a simple demonstrator or a tight implementation schedule, a custom rApp can be used to implement data extraction, preparation, model training and model storage. This can be combined with an xApp for inference. This approach is lightweight and avoids the need to deploy an additional MLOps platform. For a more complete demonstrator, MLRun can be deployed alongside dRAX to provide a structured MLOps workflow. In that case, dRAX and MLRun operate as complementary systems: dRAX provides data and RIC-based control, while MLRun provides model lifecycle management.

A proof-of-concept example was also mentioned in the transcription, where an rApp fetched RSRP data, performed a basic forecasting process, created a model and communicated it to an xApp, which then executed inference and provided predicted RSRP values. Although this example is described as an early and basic implementation, it is useful because it confirms the feasibility of the architectural pattern: rApp for training, xApp for inference, and RIC-based data access as the foundation for the loop.

Overall, the Intelligent Plane implementation supports **WP5** by providing a realistic, incremental and demonstrable path towards AI/ML-enabled network control. It starts from the existing RIC capabilities, including telemetry, databases, streaming data, rApps and xApps. It adds MLOps either through custom rApps or through MLRun. It uses xApps for near-real-time inference and the A1 interface for coordination between domains. This gives 6G-SENSES a practical implementation strategy for demonstrating how trained models can be created from network data, deployed into the RIC environment and used for intelligent network management.

6 Summary and Conclusions

6.1 Summary of Achievements

This deliverable has established the design and implementation direction of the 6G-SENSES Intelligent Plane, positioning it as the project framework for embedding AI/ML capabilities into the operational network architecture. Rather than defining the Intelligent Plane as an isolated component, the work presented in this document aligns it with the 6G-SENSES system architecture and with the evolution of O-RAN and 3GPP towards programmable, data-driven and AI-assisted network operation. This provides a coherent basis for supporting communication, sensing, computing and orchestration functions across multiple domains and time scales.

A first achievement is the consolidation of the architectural role of the Intelligent Plane within 6G-SENSES. The document clarifies how intelligence can be distributed across the SMO, Non-RT RIC, Near-RT RIC and, in future extensions, closer-to-RAN execution environments such as dApps or RT-RIC-like functions. This distribution allows long-term analytics, policy generation, model lifecycle management and orchestration to be handled in the non-real-time domain, while near-real-time inference and control can be executed through xApps. The deliverable also identifies the need for cross-domain observability, where RAN telemetry, sensing information, core-network analytics, edge/cloud metrics and external data sources can be collected and correlated to support AI/ML-driven decisions.

A second achievement is the review and positioning of relevant State-of-the-Art developments from 3GPP, O-RAN and SNS-JU projects. The document analyses how 3GPP functions such as NWDAF and ADRF can complement the O-RAN intelligence framework by exposing analytics from the core-network and service domains. It also reviews the O-RAN AI/ML framework, including the roles of the SMO, Non-RT RIC, Near-RT RIC, A1, E2, O1/O2, rApps and xApps. In addition, the BeGREEN Intelligent Plane has been considered as a relevant architectural precedent, especially for energy-efficient RAN operation, MLOps integration and the separation between model lifecycle management and control applications. These references support the 6G-SENSES decision to build on existing open-network intelligence concepts while extending them towards ISAC, multi-technology sensing and NDT support.

The identification of the technical enablers required by the Intelligent Plane, is presented as the third achievement. The deliverable summarises the AI/ML techniques investigated in **WP4**, including RL, supervised learning, FL and generative AI for intent-driven automation. These techniques impose different requirements in terms of data collection, training, inference, deployment location, feedback and retraining. The document therefore identifies MLOps as a core enabler, ensuring that AI/ML models can be trained, validated, stored, deployed, monitored and updated in a controlled way. In this context, MLRun is introduced as a candidate MLOps component, while the possibility of implementing lighter workflows directly through rApps is also maintained.

A fourth achievement is the definition of a practical implementation path based on the dRAX RIC platform. The deliverable identifies dRAX as the operational baseline for telemetry ingestion, data storage, rApp/xApp execution and near-real-time access to network data. The implementation approach makes use of live data streams for inference and historical data repositories for dataset creation, model training and retraining. The data foundation includes radio metrics, infrastructure KPIs, events and logs, supported through components such as the dRAX data bus, InfluxDB, Prometheus and Loki. This provides the basis for transforming raw network information into usable datasets and operational intelligence.

The definition of two complementary AI/ML implementation modes is the next contribution. In the first mode, rApps can directly implement data extraction, dataset preparation, model training and model exposure, which is suitable for prototyping and narrowly scoped use cases. In the second mode, MLRun can be deployed alongside dRAX as a structured MLOps component, supporting more maintainable model lifecycle management when several datasets, models and retraining cycles are required. In both cases, trained models can be consumed by xApps in the Near-RT RIC for live inference and optimisation, creating a closed operational loop between non-real-time learning and near-real-time control.

A final achievement is the alignment of the Intelligent Plane with the 6G-SENSES demonstrator work, especially PoC#3 and WP5 integration. The deliverable describes how the Intelligent Plane can support the creation, validation and use of Network Digital Twin models through real network telemetry and sensing inputs. In this context, the Modelling xApp and Optimisation Control xApp provide a concrete direction for managing NDT model creation, actualisation, validation and implementation. This enables candidate configurations to be evaluated in a simulated or digital-twin environment before being applied to the real network, supporting safer and more informed optimisation. Overall, the work presented in this deliverable provides a first concrete realisation path for the 6G-SENSES Intelligent Plane as an AI/ML-driven, sensing-assisted and energy-aware framework for future 6G network operation.

6.2 Open Challenges and Future Enhancements

Although this deliverable defines an initial implementation direction for the 6G-SENSES Intelligent Plane, several challenges remain before the concept can be considered mature and fully validated. These challenges concern data harmonisation, model governance, interoperability, safe automation and the transition from proof-of-concept demonstrations to reusable 6G network intelligence capabilities.

One of the most fundamental challenges is the harmonisation of heterogeneous data sources. The Intelligent Plane relies on information from RAN, core, edge/cloud, transport, Wi-Fi, ISAC components and external sensing systems. These sources may differ in format, granularity, time resolution and reliability. Future work should therefore define common data models, metadata conventions, time-alignment mechanisms and feature preparation procedures to improve model reuse across testbeds and deployment scenarios.

Beyond data integration, ensuring the reliability and trustworthiness of AI/ML models remains a key concern. Models used for prediction, optimisation or control may degrade when traffic, mobility, radio propagation or sensing conditions change. Future enhancements should include stronger validation pipelines, testing through emulation or NDT environments, explicit acceptance criteria, rollback mechanisms and continuous monitoring. In parallel, the project should define how models are versioned, approved, deployed, updated and retired, including traceability of training datasets and operational assumptions.

As the number of intelligent functions increases, coordination between them becomes equally important. Different rApps, xApps or AI agents may optimise different objectives, such as energy efficiency, coverage, throughput, sensing accuracy, latency or resource utilisation. These objectives may conflict. Future work should therefore include policy conflict detection, prioritisation rules, multi-objective optimisation strategies and explainability mechanisms to support trustworthy closed-loop operation.

Another area requiring further attention is the integration of NDT capabilities. Open questions remain regarding NDT fidelity, update frequency, the mapping between real telemetry and simulation parameters, and the confidence level of simulated results. Future work should define how NDT accuracy is measured, how discrepancies between real and simulated behaviour are handled, and how NDT outputs are translated into safe operational actions.

Looking further ahead, the evolution towards more native AI/ML and AI-RAN concepts should also be assessed. Some sensing-aware or radio-level optimisation functions may require lower latency than current Near-RT RIC control loops can provide. Therefore, it will be necessary to determine which functions remain in rApps and xApps, which require edge inference, and which may need closer-to-RAN execution through dApps or similar mechanisms.

In parallel with these functional enhancements, the cost of intelligence itself must be considered. AI/ML workflows introduce additional computation, storage, signalling and energy consumption. Future work should therefore measure model training cost, inference latency, resource usage, data storage requirements and the energy impact of running the Intelligent Plane itself, ensuring that AI/ML-enabled optimisation provides a net operational benefit.

In summary, the next phase should move from architectural consolidation towards systematic validation. The main future enhancements are common data and feature models, stronger model governance, safer closed-loop control, improved NDT fidelity, better coordination between intelligent applications and measurable assessment of AI/ML overhead.

6.3 Contribution to the 6G-SENSES Vision

The work presented in this deliverable contributes directly to the 6G-SENSES vision of integrating communication, sensing, computing and intelligence into a common 6G system framework. By defining the Intelligent Plane as an operational layer built around O-RAN, 3GPP analytics, AI/ML lifecycle management and RIC-based control, the deliverable provides a practical path for transforming heterogeneous network and sensing data into actionable intelligence.

This contribution is particularly relevant for enabling sensing-aware and energy-aware network operation. The Intelligent Plane provides the mechanisms required to collect telemetry, prepare datasets, train and update AI/ML models, execute inference and connect model outputs to network optimisation functions. In this way, sensing information and communication KPIs can be jointly exploited to support more adaptive, efficient and context-aware network behaviour.

The deliverable also supports the project vision by linking architectural design with concrete implementation. The use of dRAX, r/xApps, MLOps workflows and NDT functions shows how the Intelligent Plane can move beyond a conceptual model and become a deployable framework for **WP5** demonstrations. This is especially important for PoC#3, where real network data and sensing inputs are expected to feed digital-twin models and support safer optimisation before changes are applied to the live network.

Overall, the Intelligent Plane strengthens the 6G-SENSES objective of moving towards AI-native, programmable and sensing-assisted 6G networks. It provides the basis for future extensions in cross-domain automation, native AI/ML in the RAN, AI-RAN integration and trustworthy closed-loop control, while maintaining alignment with open and standards-oriented network architectures.

7 References

[1] 6G-SENSES deliverable D2.2, "System architecture and preliminary evaluations", June 2025.

[2] Pavlos Doanis, Federico Trombetti, Salvatore Pontarelli, Neco Villegas, Luis Diez, Ramón Agüero, Shahid Mumtaz, Ioanna Mesogiti, Elina, Sark, Jesús Gutiérrez, '6G-SENSES - D2.3 Final report on 6G-SENSES network architecture evaluation'. Feb. 01, 2026.

[3] BubbleRAN, 'MX-AI: AI-RAN Platform & DevKit'. 2026. [Online]. Available: https://bubbleran.com/products/mx-ai/

[4] ChenChiehChun, *BubbleRAN MX-PDK: Enabling 5G ISAC Spatial Intelligence with O-RAN Technologies*, (Jun. 2026). [Online Video]. Available: https://www.youtube.com/watch?v=v9vyofl_-9I

[5] I. Chatzistefanidis *et al.*, 'MX-AI: Agentic Observability and Control Platform for Open and AI-RAN'. 2025. [Online]. Available: https://arxiv.org/abs/2508.09197

[6] BubbleRAN, 'BubbleRAN Multi-X Product Family Datasheet', BubbleRAN, Product datasheet, 2026. [Online]. Available: https://bubbleran.com/resources/files/BubbleRAN-MX-Product-Family-Datasheet.pdf

[7] BubbleRAN, 'Lab 10: Lower-Layer Control ORAN SM'. 2026. [Online]. Available: https://bubbleran.com/docs/user-guide/xapp-training/lab10

[8] M. Qu *et al.*, 'An Overview of Enabling Artificial Intelligence in 3GPP 5G-Advanced Networks', in *2023 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, Danzhou, China: IEEE, Dec. 2023, pp. 65–70. doi: 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics60724.2023.00035.

[9] 3GPP, 'Management and Orchestration; Management Data Analytics (MDA)', 3rd Generation Partnership Project, Technical Specification TS 28.104 V19.3.0, Oct. 2025.

[10] 3GPP, 'Management and Orchestration; Artificial Intelligence/Machine Learning (AI/ML) Management', 3rd Generation Partnership Project, Technical Specification TS 28.105 V19.5.0, Apr. 2026.

[11] W. Yi, Y. Fu, J. Cao, L. Gan, L. Xiong, and H. Li, 'Toward Seamless 6G and AI/ML Convergence: Architectural Enhancements and Security Challenges', *IEEE Network*, vol. 39, no. 4, pp. 157–165, Jul. 2025, doi: 10.1109/MNET.2025.3537608.

[12] W. Lee, H. Chae, I. Bang, E. Kim, T. Kim, and G. Noh, 'AI/ML-Driven Predictive Mobility Management in 3GPP: Technical Insights and Standardization Trends', in *2025 16th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, Korea, Republic of: IEEE, Oct. 2025, pp. 1313–1318. doi: 10.1109/ICTC66702.2025.11388879.

[13] L. Dong and N. Hu, 'AI-Driven NG-RAN Assisted Positioning in 3GPP 5G-Advanced', in *2025 IEEE International Conference on Communications Workshops (ICC Workshops)*, Montreal, QC, Canada: IEEE, Jun. 2025, pp. 208–213. doi: 10.1109/ICCWorkshops67674.2025.11162409.

[14] 3GPP, 'Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services', 3rd Generation Partnership Project, Technical Specification TS 23.288 V19.6.0, Mar. 2026.

[15] O-RAN Alliance, 'O-RAN Architecture Description', O-RAN Alliance, Technical Specification O-RAN.WG1.OAD.

[16] O-RAN Alliance, 'O-RAN AI/ML Workflow Description and Requirements', O-RAN Alliance, Technical Specification O-RAN.WG2.AIML.

[17] ETSI, 'O-RAN; A1 Interface: General Aspects and Principles', European Telecommunications Standards Institute, Technical Specification ETSI TS 103 983 V3.1.0, Jan. 2024.

[18] O-RAN Alliance, 'Near-RT RIC and E2 Interface Use Cases and Requirements', O-RAN Alliance, Technical Specification O-RAN.WG3.

[19] O-RAN Alliance nGRG, 'Research Report on Native and Cross-domain AI', O-RAN Alliance, Research Report RR-2023-03, 2023.

[20] O-RAN Alliance nGRG, 'Native AI Architecture Description', O-RAN Alliance, Research Report RR-2023-02, 2023.

[21] O-RAN Alliance nGRG, 'dApp Use Cases and Requirements', O-RAN Alliance, Research Report RR-2024-10, 2024.

[22] A. Elyasi and others, 'O-RAN xApps: Survey and Research Challenges', *Computer Networks*, 2025.

[23] O-RAN Alliance nGRG, 'Principles and Methodologies for AI/ML Testing in Next Generation Networks', O-RAN Alliance, Research Report, 2024.

[24] 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*, vol. 25, no. 2, pp. 1376–1411, 2023, doi: 10.1109/COMST.2023.3239220.

[25] B. Brik *et al.*, 'Explainable AI in 6G O-RAN: A Tutorial and Survey on Architecture, Use Cases, Challenges, and Future Research', *IEEE Communications Surveys & Tutorials*, 2024, doi: 10.1109/COMST.2024.3510543.

[26] A. Lacava *et al.*, 'dApps: Enabling Real-Time AI-Based Open RAN Control', *Computer Networks*, vol. 269, p. 111342, Sep. 2025, doi: 10.1016/j.comnet.2025.111342.

[27] SNS JU Technology Board, 'AI/ML as a Key Enabler of 6G Networks: Methodology, Approach and AI-Mechanisms in the SNS JU', Smart Networks and Services Joint Undertaking, Jan. 2025. [Online]. Available: https://smart-networks.europa.eu/wp-content/uploads/2025/02/ai_ml_white-paper-sns_tb_v1.0.pdf

[28] Smart Networks and Services Joint Undertaking (SNS JU) Reliable Software Networks Working Group (SNWG), 'The AI/ML Landscape for Smart Networks and Services: Taxonomy, Standards & Innovation Pathways', Smart Networks and Services Joint Undertaking (SNS JU), White Paper, Jan. 2026. doi: 10.5281/zenodo.17708459.

[29] 'BeGREEN', BeGREEN. Accessed: Jun. 25, 2026. [Online]. Available: https://www.sns-begreen.com/

[30] BeGREEN Consortium, 'Final Architecture and Evaluation of BeGREEN O-RAN Intelligence Plane, and AI/ML Algorithms for NFV User-Plane and Edge Service Control Energy Efficiency Optimisation', BeGREEN Project, Deliverable D4.3, Sep. 2025. Accessed: Jun. 25, 2026. [Online]. Available: https://content.webfactorysite.co.uk/7f995ac3-5287-431b-a334-b9fbadbb8553.pdf?t=1758869242

[31] BeGREEN Consortium, 'Initial Evaluation of BeGREEN O-RAN Intelligence Plane, and AI/ML Algorithms for NFV User-Plane and Edge Service Control Energy Efficiency Optimization', BeGREEN Project, Deliverable D4.2, Sep. 2024. Accessed: Jun. 25, 2026. [Online]. Available: https://content.webfactorysite.co.uk/1b69db87-63ae-4ba4-ac4d-4d54e4d88c16.pdf?t=1728664602

[32] M. Catalan-Cid, G. Castellanos, D. Reiss, and J. Armstrong, 'Demo: BeGREEN Intelligence Plane for AI-driven Energy Efficient O-RAN Management', presented at the IEEE INFOCOM 2025 Demo, 2025.,

[33] J. X. Salvat *et al.*, 'BeGREEN Use Case Demonstration and KPI Assessment', BeGREEN Project, Deliverable D5.3 D5.3, Sep. 2025. Accessed: Jun. 25, 2026. [Online]. Available: https://www.sns-begreen.com/deliverables/d53---begreen-use-case-demonstration-and-kpi-assessment/?Page_ID=147531&id=1151411

[34] Pavlos Doanis, Markos Anastasopoulos, Anna Tzanakaki, Ramón Agüero, Luis Diez, Novella Bartolini, Federico Trombetti, Salvatore Pontarelli, Germán Castellanos, Valerio Frascolla, Jessica Sanson, Muhammad Saadi, Shahid Mumtaz, '6G-SENSES - D4.3 AI/ML techniques for enhanced network performance.' May 13, 2026.

[35] D. Kreuzberger, N. Kuehl, and S. Hirschl, 'Machine Learning Operations (MLOps): Overview, Definition, and Architecture', *IEEE Access*, vol. 11, pp. 31866–31879, 2023, doi: 10.1109/ACCESS.2023.3262138.

[36] Google Cloud, 'MLOps: Continuous Delivery and Automation Pipelines in Machine Learning'. [Online]. Available: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

[37] MLRun, 'MLRun Documentation'. [Online]. Available: https://docs.mlrun.org/

[38] MLRun, 'MLRun Architecture'. [Online]. Available: https://docs.mlrun.org/en/latest/concepts/architecture.html

[39] MLRun, 'Model Monitoring'. [Online]. Available: https://docs.mlrun.org/en/latest/model-monitoring/

[40] MLRun, 'Real-Time Serving'. [Online]. Available: https://docs.mlrun.org/en/latest/serving/

[41] Cisco, 'What Is AIOps? Artificial Intelligence for IT Operations'. [Online]. Available: https://www.cisco.com/site/us/en/learn/topics/artificial-intelligence/what-is-aiops.html

[42] IBM, 'AIOps vs. MLOps: Harnessing Big Data for Smarter IT Operations'. [Online]. Available: https://www.ibm.com/think/topics/aiops-vs-mlops

[43] 3GPP, '3GPP TS 23.288 V18.5.0; Architecture enhancements for 5G System (5GS) to support network data analytics services (Release 18)', Mar. 2024.

[44] 3GPP, 'Management and orchestration; Artificial Intelligence/ Machine Learning (AI/ML) management', 3rd Generation Partnership Project (3GPP), Technical Specification TS 28.105, Apr. 2026. [Online]. Available: https://www.3gpp.org/DynaReport/28105.htm

[45] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, 'Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges', *IEEE Commun. Surv. Tutorials*, vol. 25, no. 2, pp. 1376–1411, 2023, doi: 10.1109/COMST.2023.3239220.

[46] 'nGRG-RR-2024-02-O-RAN Cross-domain AI v1.6'. Accessed: Jun. 11, 2026. [Online]. Available: https://mediastorage.o-ran.org/ngrg-rr/nGRG-RR-2024-02-O-RAN%20Cross-domain%20AI%20v1.6.pdf?utm_source=chatgpt.com

[47] A. Lacava *et al.*, 'dApps: Enabling Real-Time AI-Based Open RAN Control', *Computer Networks*, 2025, doi: 10.1016/j.comnet.2025.111342.

[48] C. Zhou *et al.*, 'Network Digital Twin: Concepts and Reference Architecture', Internet Engineering Task Force, Internet-Draft draft-irtf-nmrg-network-digital-twin-arch-05, Mar. 2024. [Online]. Available: https://datatracker.ietf.org/doc/draft-irtf-nmrg-network-digital-twin-arch/05/

[49] D. Chen, H. Yang, and C. Zhou, 'Requirements for Interfaces of Network Digital Twin', Internet Engineering Task Force, Internet-Draft draft-chen-nmrg-dtn-interface-00, 2022. [Online]. Available: https://datatracker.ietf.org/doc/draft-chen-nmrg-dtn-interface/00/

[50] 3GPP, 'Management and orchestration; Management aspects of Network Digital Twins', 3rd Generation Partnership Project, Technical Specification TS 28.561 V19.2.0, Apr. 2026.

[51] T. Gong, C. Dai, Y. Shi, H. Cao, and S. Mumtaz, 'Fairness-Oriented Decoupled User Association and Resource Allocation in Fully-Decoupled RAN: A Two-Layer MAB Approach', in *ICASSP 2026–2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2026, pp. 21301–21305. doi: 10.1109/ICASSP55912.2026.11464659.

[52] N. A. Khan and S. Schmid, 'AI-RAN in 6G Networks: State-of-the-Art and Challenges', *IEEE Open Journal of the Communications Society*, vol. 5, pp. 294–311, 2024, doi: 10.1109/OJCOMS.2023.3343069.

[53] D. Abouelmaati, A. Esfahani, S. Goudarzi, and S. Mumtaz, 'Empowering Next-Gen Networks: AI-Driven Autonomy in O-RAN and SON Architectures', *IEEE Access*, vol. 13, pp. 186903–186936, 2025, doi: 10.1109/ACCESS.2025.3626452.

[54] L. Kundu, X. Lin, R. Gadiyar, J.-F. Lacasse, and S. Chowdhury, 'AI-RAN: Transforming RAN with AI-Driven Computing Infrastructure', *IEEE Communications Magazine*, vol. 64, no. 1, pp. 168–174, Jan. 2026, doi: 10.1109/MCOM.001.2500018.

[55] 3GPP, 'AI/ML Management for 5G Systems'. Accessed: May 22, 2024. [Online]. Available: https://www.3gpp.org/technologies/ai-ml-management

[56] 3rd Generation Partnership Project, 'Management and orchestration; Artificial Intelligence/Machine Learning (AI/ML) management', 3GPP, Technical Specification 3GPP TS 28.105, V19.5.0, Release 19, Apr. 2026. [Online]. Available: https://www.3gpp.org/DynaReport/28105.htm

[57] G. M. Almeida *et al.*, 'RIC-O: Efficient Placement of a Disaggregated and Distributed RAN Intelligent Controller With Dynamic Clustering of Radio Nodes', *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 2, pp. 446–459, 2024, doi: 10.1109/JSAC.2023.3336159.

[58] Accelleran, *dRAX xApp Dev Guide*. [Online]. Available: https://accelleran.github.io/drax-xapp-dev-guide/v6.0/

[59] MLRun, 'MLRun Documentation'. Accessed: Jun. 25, 2026. [Online]. Available: https://docs.mlrun.org/

[60] srsRAN, *O-RAN NearRT-RIC and xApp. e@ Interface.* [Online]. Available: https://docs.srsran.com/projects/project/en/latest/tutorials/source/near-rt-ric/source/index.html#overview

[61] J. Mirzaei, I. Abualhaol, and G. Poitau, 'Network Digital Twin for Open RAN: The Key Enablers, Standardization, and Use Cases', Aug. 04, 2023, *arXiv*: arXiv:2308.02644. doi: 10.48550/arXiv.2308.02644.

[62] J. Mirzaei, I. Abualhaol, and G. Poitau, 'Network Digital Twin for Open RAN: The Key Enablers, Standardization, and Use Cases', 2023, *arXiv*. doi: 10.48550/ARXIV.2308.02644.

[63] O-RAN Alliance next Generation Research Group, 'Research Report on Digital Twin RAN Use Cases', O-RAN Alliance, Contributed Research Report RR-2024-07, May 2024. [Online]. Available: https://mediastorage.o-ran.org/ngrg-rr/nGRG%20RS01%20Digital%20Twin%20Use%20Case%20Research%20Report%20v1.0.pdf

[64] X. Lin, L. Kundu, C. Dick, E. Obiodu, T. Mostak, and M. Flaxman, '6G Digital Twin Networks: From Theory to Practice', *IEEE Commun. Mag.*, vol. 61, no. 11, pp. 72–78, Nov. 2023, doi: 10.1109/MCOM.001.2200830.

[65] J. Hoydis *et al.*, 'Sionna: An Open-Source Library for Next-Generation Physical Layer Research', *IEEE Communications Magazine*, vol. 60, no. 4, pp. 22–28, 2022, doi: 10.1109/MCOM.001.2100932.

8 Acronyms

Acronym	Definition
3GPP	3rd Generation Partnership Project
5G	Fifth Generation
5GC	5G Core
5GS	5G System
6G	Sixth Generation
A2A	Agent-to-Agent
ADRF	Analytics Data Repository Function
AF	Application Function
AI	Artificial Intelligence
AIA	AI Engine Assist
AIMLsys	Artificial Intelligence / Machine Learning for 5G System
AIOps	Artificial Intelligence for IT Operations
AI-RAN	Artificial Intelligence Radio Access Network
AMF	Access and Mobility Management Function
AnLF	Analytics Logical Function
AP	Access Point
API	Application Programming Interface
AR	Augmented Reality
BAT	BubbleRAN Agentic Toolkit
BS	Base Station
CDN	Content Delivery Network
CF	Cell-Free
CF-MIMO	Cell-Free Multiple-Input Multiple-Output
CN	Core Network
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CQI	Channel Quality Indicator
CSI	Channel State Information
CU	Central Unit
CU-CP	Central Unit Control Plane
CU-UP	Central Unit User Plane
DNN	Deep Neural Network

DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
DU	Distributed Unit
e2e	End-to-End
EMF	Electromagnetic Field
ETSI	European Telecommunications Standards Institute
FCAPS	Fault, Configuration, Accounting, Performance and Security
FIFO	First In- First Out
FL	Federated Learning
GNN	Graph Neural Network
GPS	Global Positioning System
HFL	Horizontal Federated Learning
HRLLC	Hyper Reliable Low Latency Communication
I/Q	In-Phase and Quadrature
IASA	Institute of Accelerating Systems and Applications
ICN	Information-Centric Networking
ICT	Information and Communication Technology
IHP	Innovations for High Performance Microelectronics
IMT	International Mobile Telecommunications
IoT	Internet of Things
ISAC	Integrated Sensing and Communication
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
LCM	Life Cycle Management
LFU	Least Frequently Used
LLM	Large Language Model
LLMOps	Large Language Model Operations
LOC	Location
Loki	Log and Event Database used in dRAX
LRU	Least Recently Used
MAC	Medium Access Control
MADDPG	Multi-Agent Deep Deterministic Policy Gradient
MCP	Model Context Protocol
MDA	Management Data Analytics

MDAF	Management Data Analytics Function
MEC	Multi-access Edge Computing
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MLKTL	ML-Knowledge-Based Transfer Learning
MLOps	Machine Learning Operations
MLRun	Machine Learning Run / MLOps Framework
mMIMO	massive Multiple-Input Multiple-Output
mmWave	Millimetre Wave
MRU	Most Recently Used
MTLF	Model Training Logical Function
N3GPP	Non-3GPP
N3IWF	Non-3GPP InterWorking Function
NBI	Northbound Interface
NDT	Network Digital Twin
Near-RT	Near-Real-Time
NEF	Network Exposure Function
NGMN	Next Generation Mobile Networks
nGRG	O-RAN next Generation Research Group
NG-RAN	Next Generation Radio Access Network
NLoS	Non-Line-of-Sight
Non-RT	Non-Real-Time
Non-RT RIC	Non-Real-Time RAN Intelligent Controller
NOP	Network Operator
NR	New Radio
NSSF	Network Slice Selection Function
NTU	Nottingham Trent University
NWDAF	Network Data Analytics Function
O1	O-RAN Management Interface
O2	O-RAN Cloud Interface
O-Cloud	O-RAN Cloud
O-CU	O-RAN Central Unit
O-CU-CP	O-RAN Central Unit Control Plane
O-CU-UP	O-RAN Central Unit User Plane

O-DU	O-RAN Distributed Unit
O-RAN	Open Radio Access Network
O-RU	O-RAN Radio Unit
PCF	Policy Control Function
PDK	Platform Development Kit
PHY	Physical Layer
PoC	Proof of Concept
PRB	Physical Resource Block
QoS	Quality of Service
R1	O-RAN R1 Interface
RAM	Random Access Memory
RAN	Radio Access Network
rApp	
RAT	Radio Access Technology
RIC	RAN Intelligent Controller
RIS	Reconfigurable Intelligent Surface
RL	Reinforcement Learning
RLC	Radio Link Control
RNN	Recurrent Neural Network
ROS	Robot Operating System
RR	Research Report
RRC	Radio Resource Control
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSSI	Received Signal Strength Indicator
RSSNR	Reference Signal Signal-to-Noise Ratio
RT	Real-Time
RT-RIC	Real-Time RAN Intelligent Controller
RU	Radio Unit
SBA	Service-Based Architecture
SDN	Software-Defined Networking
Sionna RT	Sionna Ray Tracing
SLA	Service Level Agreement
SMF	Session Management Function

SMO	Service Management and Orchestration
SNPN	Standalone Non-Public Network
SNS JU	Smart Networks and Services Joint Undertaking
SotA	State of the Art
SP	Service Provider
SRS	Sounding Reference Signal
Sub-6	Sub-6 GHz
TB	Technology Board
TS	Technical Specification
UC	Universidad de Cantabria
UE	User Equipment
UPF	User Plane Function
V2X	Vehicle-to-Everything
VFL	Vertical Federated Learning
VNF	Virtual Network Function
VR	Virtual Reality
WAT	Wireless Access Technology
Wi-Fi	Wireless Fidelity
WP	Work Package
XAI	Explainable Artificial Intelligence
xApp	Extended Application
XR	eXtended Reality
Y1	O-RAN Y1 Interface