International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Technical Report

(May 2026)

ITU-T Focus Group on AI Native for Telecommunication Networks

Proof-of-Concept activities

Summary
This Technical Report provides details on the Proof-of-Concept activities under ITU-T FG-AINN WG4. This report provides the technical summary of the activities done under PoC, and it covers the following:

- Requirements for the Proof of Concept, including Build-a-thon.
- Description of the Proof of Concept and results.
- Learnings from PoC development mentoring sessions, including Build-a-thon

NOTE - This document includes details of PoCs done under Build-a-thon 1.0, 2.0, and 3.0, and 4.0.

Keywords
AI agent, AI native networks, AI pipeline, Explainable AI, Intent, Knowledge Base, Proof of Concept, telemetry, use case.

Contributors:	Preksha Shah Indian Institute of Technology Bombay, Mumbai, India	Tel: E-mail: 30005100@ee.iitb.ac.in
	Liya Yuan ZTE Corporation China	Tel: E-mail: yuan.liya@zte.com.cn

Table of contents

Proof of Concept Activities

1 Scope

This Technical Report summarizes the Proof of Concept (PoC) activities developed under the ITU-T FG-AINN initiative. It presents the technical details of AI-native network PoCs contributed in alignment with the AI-native delimiting characteristics—namely, architecture approaches for deep integration of AI, engagement of AI in all stages of the lifecycle, and AI itself as a core component. The report also includes outcomes from collaborative activities such as the Build-a-thon, and combined learnings from Build-a-thon mentoring sessions.

Each PoCs are mapped to (a) gaps, (b) use cases, and (c) architecture concepts.

2 References

[ITU-T Y.3102] ITU-T Recommendation Y.3102, “Framework of the IMT-2020 network,” February 2018.

[ITU-T Y.3172] ITU-T Recommendation Y.3172, “Architectural framework for machine learning in future networks including IMT-2020”.

[ITU-T Q.5001] ITU-T Recommendation Q.5001, “Framework for intelligence-aware functional architecture of future networks including IMT-2020,” June 2020.

3 Definitions

None.

3.1 Terms defined elsewhere

None

3.2 Terms defined in this document

None

4 Abbreviations

AI: Artificial Intelligence

ANSP: AI-Native Service Platform

BTS: Base Transceiver Stations

CLI: Command Line Interface

CSI: Channel State Information

DQN: Deep-Q-Network

DWT: Discrete Wavelet Transform

E2E: End-to-end

ELA: Error Level Analysis

ETSI: European Telecommunications Standards Institute

FGAINN: Focus Group on AI Native for Telecommunication Networks

KB: Knowledge Base

KPI: Key Performance Indicator

LLM: Large Language Model

LSTM: Long Short-Term Memory

ML: Machine Learning

NLP: Natural Language Processing

NMS: Network Management Systems

ONOS: Open Network Operating System

RAG: Retrieval-Augmented Generation

RAN: Radio Access Network

RSSI: Received Signal Strength Indicator

SLA: Service Level Agreement

UE: User Equipment

URA: Uniform Rectangular Array

URLLC: Ultra-Reliable Low Latency Communication

QoS: Quality of Service

XAI: Explainable AI

5 Conventions

In this Technical Report, in alignment with the conventions of [Supplement 55 to ITU-T Y- series Recommendations] possible requirements which are derived from a given use case, are classified as follows:

The keywords "it is critical" indicate a possible requirement that would be necessary to be fulfilled (e.g., by an implementation) and enabled to provide the benefits of the use case. The keywords "it is expected" indicate a possible requirement which would be important but not absolutely necessary to be fulfilled (e.g., by an implementation). Thus, this possible requirement would not need to be enabled to provide complete benefits of the use case.

The keywords "it is of added value" indicate a possible requirement which would be optional to be fulfilled (e.g., by an implementation), without implying any sense of importance regarding its fulfilment. Thus, this possible requirement would not need to be enabled to provide complete benefits of the use case.

6 Introduction

The evolution of future communication networks is increasingly driven by the integration of machine learning (ML) and artificial intelligence (AI) capabilities across all layers of the network architecture. As outlined in ITU-T Recommendations [ITU-T Y.3102] and [ITU-T Y.3172] the development of intelligent, adaptive, and learning-enhanced network systems is central to achieving scalable, efficient, and context-aware service delivery in next-generation systems such as IMT-2020 and beyond.

The [ITU-T Y.3172] framework introduces a modular and functional architecture for incorporating ML in future networks, where ML pipelines support functions such as data collection, model training, inference, and feedback loops.

Building on these foundational principles, the current study explores the design and implementation of AI-native network functions and control loops that go beyond traditional automation. Rather than layering AI capabilities on top of existing management systems, AI-native networks embed intelligence directly into the fabric of network operations; interpreting user intents, coordinating multi-agent decision making, optimizing closed-loop control, and autonomously adjusting to dynamic network conditions. Shah P, et al. [b-Shah] carried out a gap analysis of existing frameworks in AI native networks.

The proof-of-concept (PoC) studies presented in this report focuses on demonstrating such AI-native capabilities across a diverse range of network scenarios. These includes distributed decision making, intent-based and feature-specific applications, emergency & resilience-centric AI inference systems, among others. Special emphasis is placed on leveraging standardized ML architecture components, ensuring that implementations are interoperable, reproducible, explainable, and aligned with the ITU-T Focus Group on AI Native for Telecommunication Networks (FGAINN) vision.

As AI-native networking continues to mature, these studies lay the groundwork for standard-compliant, reference-grade implementations that embody closed-loop autonomy, real-time learning, and AI-driven orchestration. According to [b-FG-AINN-O-08], AI-Native systems are defined by three delimiting characteristics: (i) architecture approaches for deep integration of AI, (ii) engagement of AI in all stages of the lifecycle, and (iii) AI itself as a core component. In the PoCs, these characteristics are realized by embedding AI into the system fabric through native architectural patterns (e.g., AI pipelines, service-based interfaces, and telemetry/data fabrics); ensuring AI participation end-to-end across design, deployment, orchestration, runtime operation, and evaluation of network components, functions, applications, and services; and elevating AI primitives to first-class resources within the control, data, and management planes. Through this mapping, the PoCs concretely demonstrate AI-native principles in practice.

The ITU-T FG-AINN organized three Build-a-thon challenges in 2025 to accelerate the development and validation of AI-native networking solutions. Build-a-thon 1.0, conducted in June 2025, saw participation from 23 registered teams, out of which 10 teams submitted final prototypes, and 6 winners were selected. To support the participants, six mentoring sessions were held from May to June 2025. Build-a-thon 2.0, held during June–July 2025, was supported by five dedicated mentoring sessions and concluded with the announcement of three winners. Build-a-thon 3.0 is ongoing during Aug – Oct 2025. A series of seven mentoring sessions would be conducted and will conclude with the announcement of winners in Oct 2025.

The key outcomes from both Build-a-thon 1.0 and 2.0 include the demonstration of feasible and impactful AI-native solutions for telecom networks and societal applications, as well as validation of performance and functional requirements through executed test cases. The challenges also provided deep insights into integration challenges, design trade-offs, and the implementation of modular, interoperable architectures using AI pipelines and intent-based orchestration. Participants gained hands-on experience

in prototyping and cross-disciplinary collaboration, while their contributions—ranging from use cases and architectural insights to PoC documentation and gap analysis—have enriched FG-AINN's technical work. These efforts have further strengthened the collaboration between academia, industry, and standardization bodies, supporting the evolution of future AI-native standards.

Technical Report is organized as follows: Section 7 outlines the general requirements for the PoC, Section 8 provides the PoC description, Section 9 presents the PoC details, Section 10 discusses the combined learnings from Build-a-thon 1.0 and Build-a-thon 2.0 mentoring sessions, and Section 11 concludes with the bibliography.

7 General Requirements for the PoCs

This clause describes the requirements for the PoCs.

Requirement	Description
Gen-Build-a-thon-PoC-001	It is critical that PoC development activity, builds upon a key concept in FG AINN, especially aims to prove the concept practically with code, test setup and demo setup.
Gen-Build-a-thon-PoC-002	It is critical that PoC development activity create well-documented artefacts and opensource code.
Gen-Build-a-thon-PoC-003	It is critical that the maturity of the PoC is evaluated in test scenarios in the accompanying documentation.
Gen-Build-a-thon-PoC-004	It is critical that PoC demonstrates the feasibility (or lack of it) of specific architecture approaches.
Gen-Build-a-thon-PoC-005	It is critical that demonstration is focussed on a set of unique scenarios.
Gen-Build-a-thon-PoC-006	It is expected that participants submit other use case implementations in the same format as the reference examples.
Gen-Build-a-thon-PoC-007	It is expected that the PoC demonstrates the three delimiting characteristics of AI Native: Architecture approaches for deep integration of AI, engagement of AI in all stages of the lifecycle, and AI itself as a core component [b-FG-AINN-O-08].

8 PoC Description and Details

The scope of PoCs across the Build-a-thon series shows a clear progression in maturity and focus. In Build-a-thon 1.0, PoCs centered on AI solution ideation and prototyping, emphasizing exploration of innovative concepts, feasibility assessment, and identification of AI-native integration points within telecom networks. Build-a-thon 2.0 built on this foundation by advancing these ideas into working prototypes, software artifacts, and live demonstrations, thereby validating their practical applicability in operational environments. Build-a-thon 3.0 introduced an architecture-driven approach, focusing on the design and demonstration of scalable AI-native network frameworks and their integration across system components. In Build-a-thon 4.0, the focus shifted to AI-enabled smart broadcasting, with PoCs targeting intelligent, scalable, and efficient broadcasting solutions.

Each PoC is tagged with its delimiting characteristic, such as architecture approaches for deep integration of AI, engagement of AI across all stages of the lifecycle, and positioning AI itself as a core system component. This report presents key details of each PoC, including title, characteristics, description, gaps addressed, test setup, code and datasets, use cases, architecture, demonstrations and evaluations, observations, conclusions, and future work.

The PoCs are categorized into three thematic sections: Section 10 - AI Solution Ideation & Prototyping (PoC-001 to PoC-014), Section 11 - Architecture-Driven Implementations (PoC-015 to PoC-019), and Section 12 - AI-Enabled Smart Broadcasting (PoC-020 to PoC-028).

9. AI Solution Ideation & Prototyping

This section includes PoC-001 to PoC-014, contributed during Build-a-thon 1.0 and Build-a-thon 2.0. These PoCs represent the exploratory phase, focusing on ideation, validation, and refinement of forward-looking AI-native networking concepts. They emphasize feasibility analysis, conceptual innovation, and identification of architectural touchpoints.

9.1.1 PoC-001 [b-FG-AINN-I-104_R1] AI-Native Proactive Network Slice Marketplace for Dynamic Service Ecosystems

9.1.2. Delimiting Characteristic

AI integrated services and AI pipelines itself as a core component – because this PoC integrates an AI agent as a core component which implements the service of slice selection.

9.1.3. Description

This PoC demonstrates an AI-native network slicing framework where Deep-Q-Network (DQN), Long Short-Term Memory (LSTM), and Large Language Model (LLM) agents enable intelligent, real-time decisions for vendor selection and slice optimization. It showcases integration with open5GS and supports structured data exchange for adaptive orchestration. The figure shows AI-Native Proactive Network Slice Marketplace.

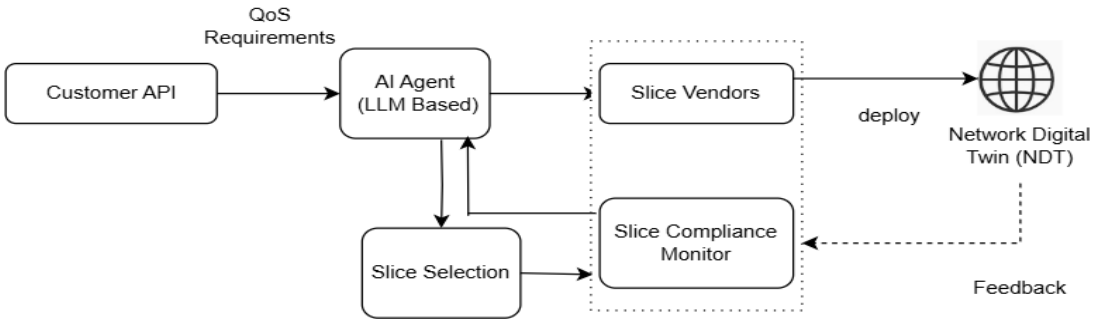


Fig. 9.1.1: AI-Native Proactive Network Slice Marketplace

9.1.4. Gaps Addressed

This tackles the absence of intelligent, data-driven orchestration in network slicing by introducing AI agents (e.g., DQN, LSTM, LLM) that automate decisions like vendor/resource selection and slice configuration based on real-time application requirements and traffic behavior something traditional static or rule-based systems fail to handle dynamically.

9.1.5. POCs Test Setup

This PoC is developed using Python-based simulation tools and Jupyter/Colab environments, optionally extendable to AWS EC2. The traffic simulator emulates variable load scenarios, including stadium surges and disaster responses, across >50 concurrent slice contexts. A supervised Random Forest classifier maps traffic characteristics (latency, bandwidth, device count) to slice types, while a DQN agent handles dynamic resource allocation. Vendor selection logic uses Open5GS-based emulators to return latency/jitter-based slice offers, and an LLM (e.g., Ollama) validates slice decisions against user intents. LSTM-based forecasting models anticipate traffic surges, feeding into a feedback loop that enables self-tuning of slice templates, vendor trust scoring, and retraining. Real-time metrics (latency,

Service Level Agreement (SLA) violations, energy use) are logged and visualized using Matplotlib/Seaborn. The simulation environment includes a Key Performance Indicator (KPI) evaluation engine and a feedback module for continuous learning. Figure shows workflow of AI native proactive network slice marketplace.

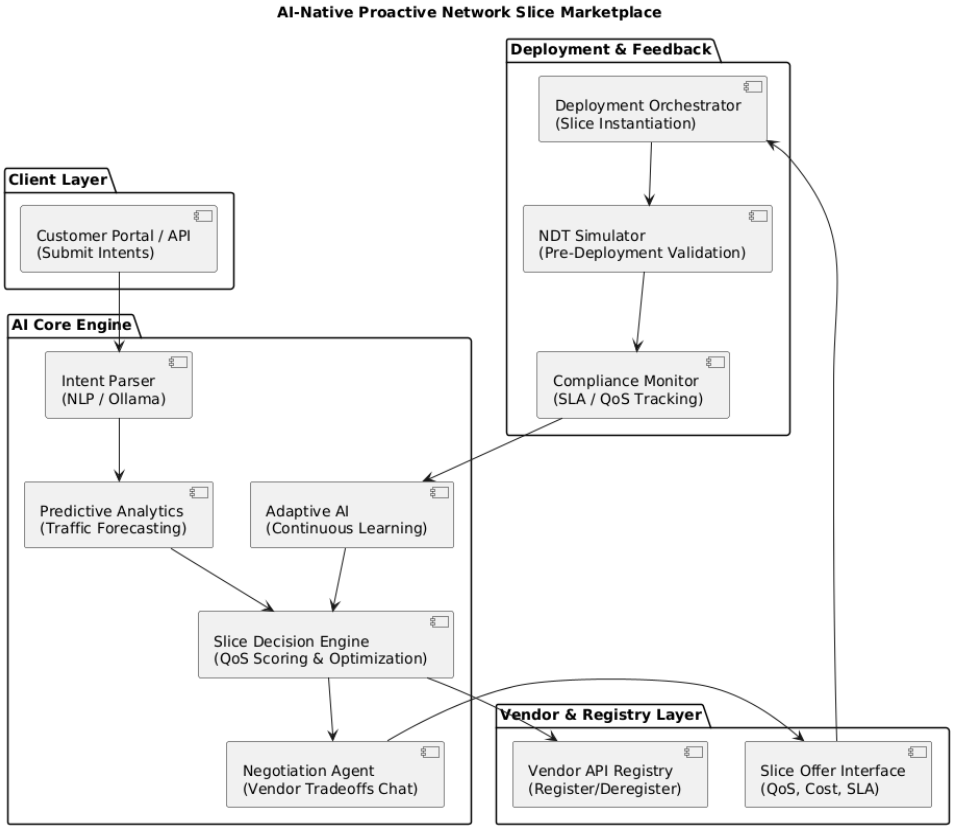


Fig. 9.1.2: Workflow of AI native proactive network slice marketplace

9.1.6. Code and Data Sets

A synthetic dataset generator produces multi-dimensional traffic traces (eMBB, URLLC, mMTC), augmented by public 5G datasets (e.g., https://www.osti.gov/biblio/1880776) to support realism. Code and data sets can be found here: https://github.com/Harsh067899/Build-A-Thon-SRM.

9.1.7. Simulated Use cases

AI system enables real-time classification of traffic into eMBB, URLLC, or mMTC slice types based on application demands. It dynamically selects the most suitable vendor slice template and adapts to varying traffic loads, including emergencies and public events. Key scenarios include smart city operations, emergency prioritization, and large-scale event management. This showcases proactive AI-Agent driven network slicing aligned with SLA, latency, and energy goals.

9.1.8. Architectural concepts

This PoC introduces an architecture where AI agents (DQN, LSTM, LLM) are modular components integrated alongside or within slice management functions (e.g., NSMF, NSSMF). These agents interact with orchestrators through structured interfaces (e.g., JSON-based APIs) and are capable of both inference and feedback handling, enabling real-time, adaptive slicing decisions. The architecture supports closed-loop control, telemetry-driven inputs, and application-aware outputs, marking a shift from static policy-driven orchestration to dynamic, AI-native management.

9.1.9. Demo and Evaluation

The demonstration covered five operational scenarios ranging from baseline benchmarking to emergency response and smart city integration. Real-time traffic simulations were used to assess the AI agent's ability to pre-allocate resources, classify slices dynamically, and handle SLA violations through reallocation. The success criteria focused on latency reduction, energy-aware resource usage, and improved adaptability under dynamic load conditions. Evaluation relied on real-time and post-simulation KPIs including SLA compliance, response time to traffic changes, and accuracy of slice classification and forecasting. Information exchange between application and model, including SLA constraints and traffic types, was logged in structured formats, enabling transparent KPI tracking. A component classifies high-level client intents (e.g., latency/bandwidth needs) into 3GPP-aligned slice types (eMBB, URLLC, mMTC) using an ML model. This replaces static policy mapping with dynamic, learning-based slice interpretation — foundational to AI-native orchestration

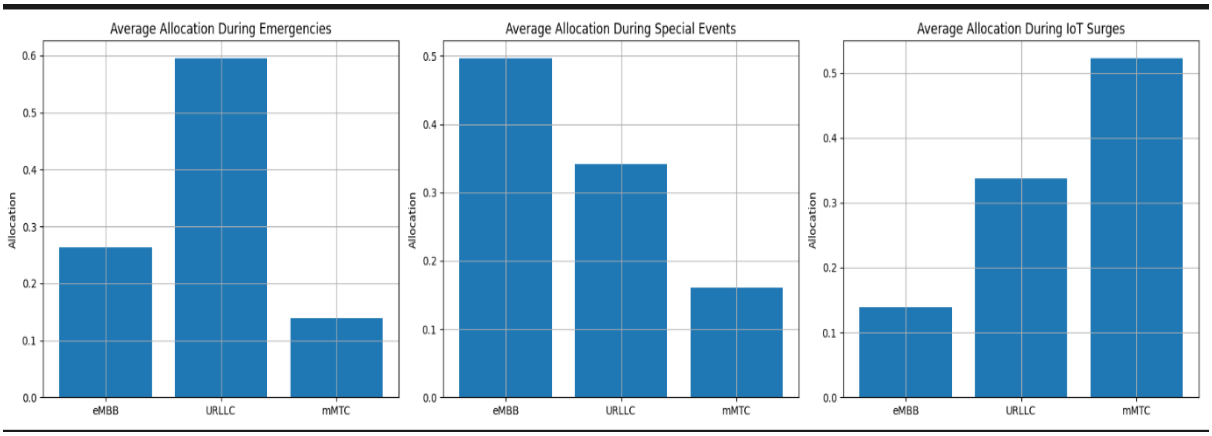


Fig. 9.1.3: Slice Allocation across use cases

A large language model (LLM) is used to evaluate and score vendor slice offerings based on soft constraints like SLA, provisioning time, security, and pricing. It enables semantic, multi-factor, and

explainable slice decision-making beyond rigid rule-matching — a core AI-native advancement..

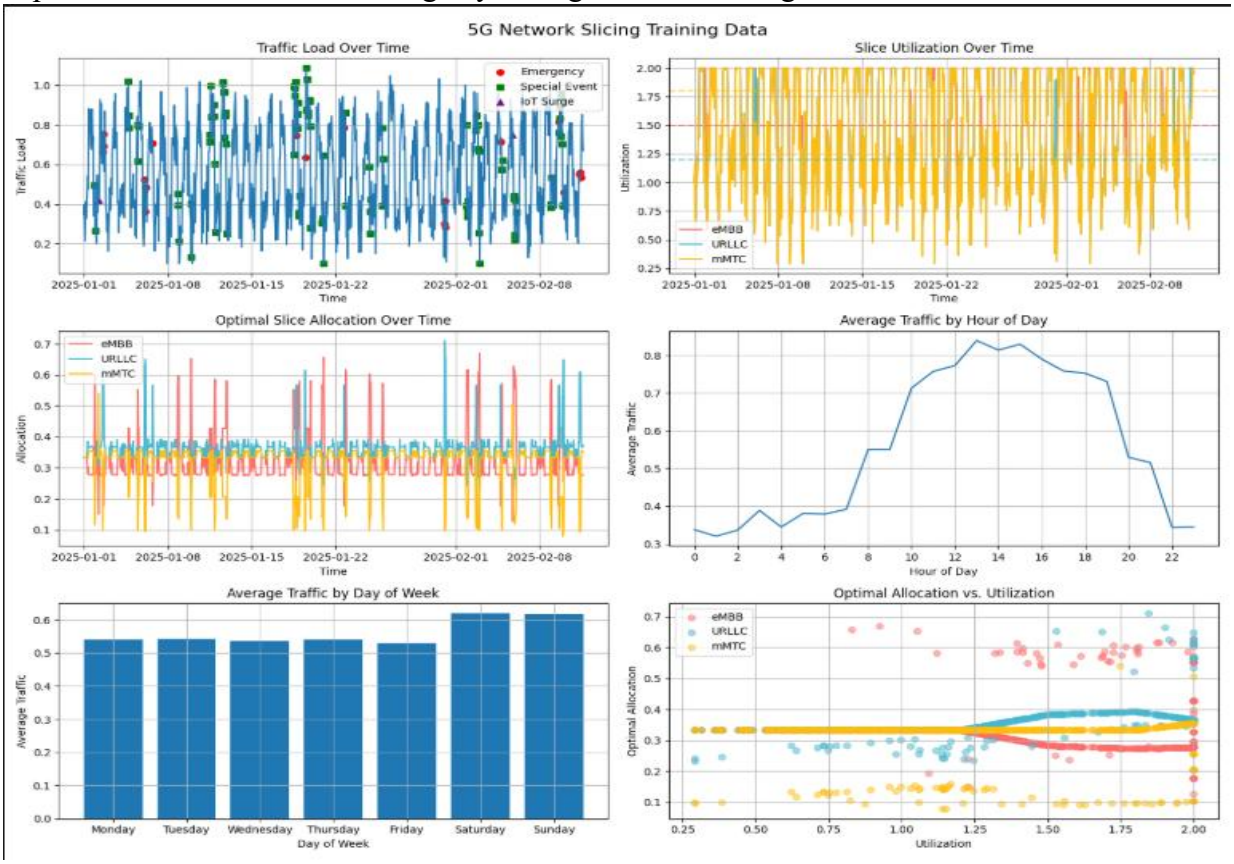


Fig. 9.1.4: Training data identification and visualization

An orchestrator component triggers vendor-side slice deployment automatically after scoring, without manual intervention. Fig. 9.1.4 demonstrates full intent-to-slice lifecycle automation — a step beyond 5G's NF-centric deployment model.

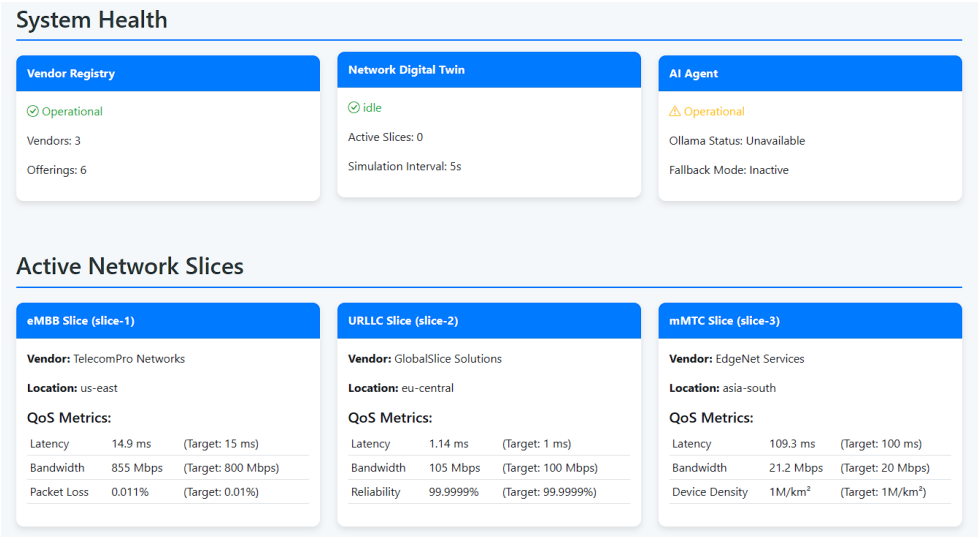


Fig. 9.1.5: System health and Active purposes

Simulated vendor agent deploys slice logic (Open5GS/UPF) and sends back a valid slice connection profile (AMF/SMF/DNN, QoS). This standardizes the slice provisioning return flow and prepares the

UE connection interface — enabling multi-vendor interoperability.

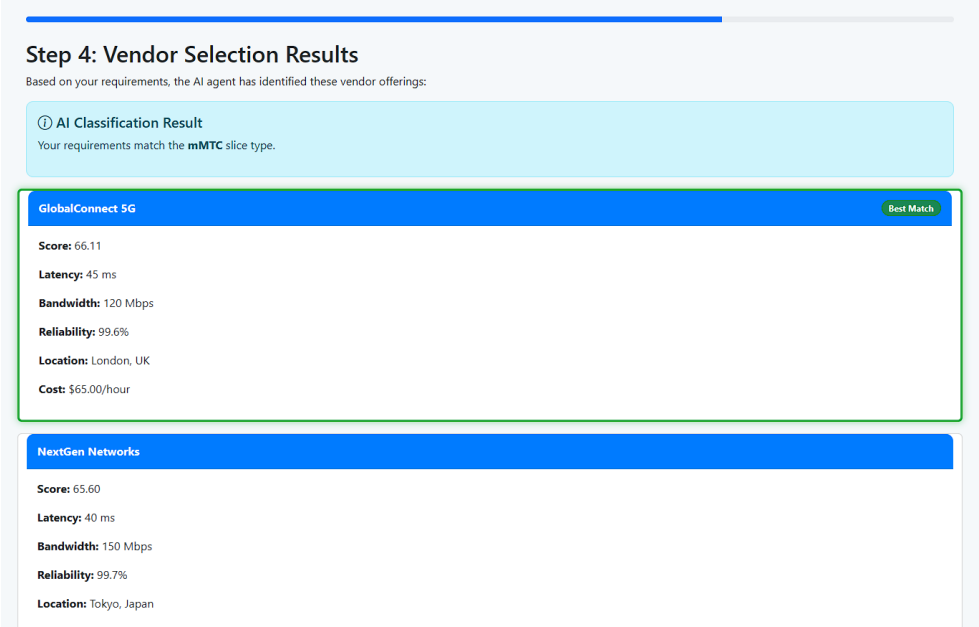


Fig. 9.1.6: Vendor selection and best score identification

Fig. 9.1.6 shows an API securely delivers the vendor-generated slice profile to the client or UE, enabling automatic session initiation. This replaces static NSSAI provisioning with dynamic, signed configuration delivery — fulfilling AI-native slice bootstrapping.

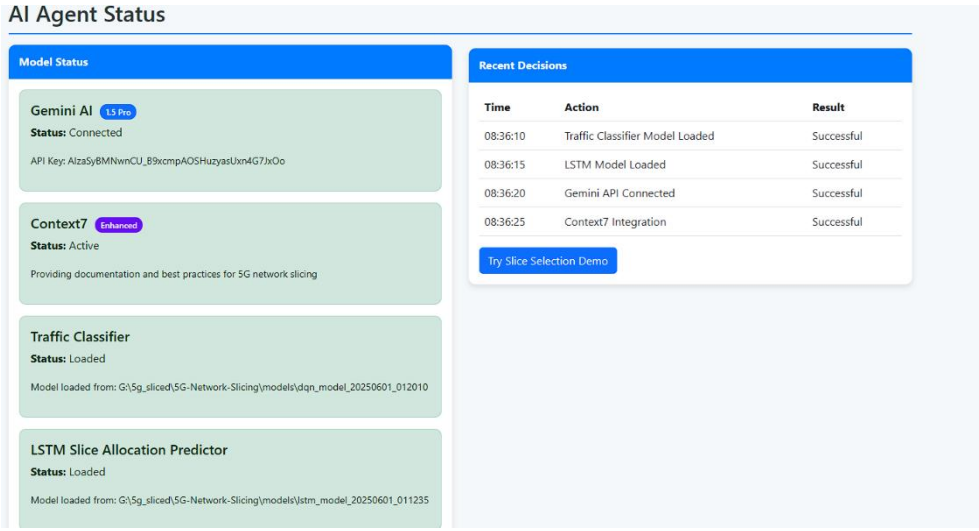


Fig. 9.1.7: AI Native agent status

Fig. 9.1.7 shows prometheus-based metrics simulate real-time latency, bandwidth, and availability monitoring for deployed slices. This validates compliance post-deployment and closes the loop between deployment and assurance — vital for trust-based networks.

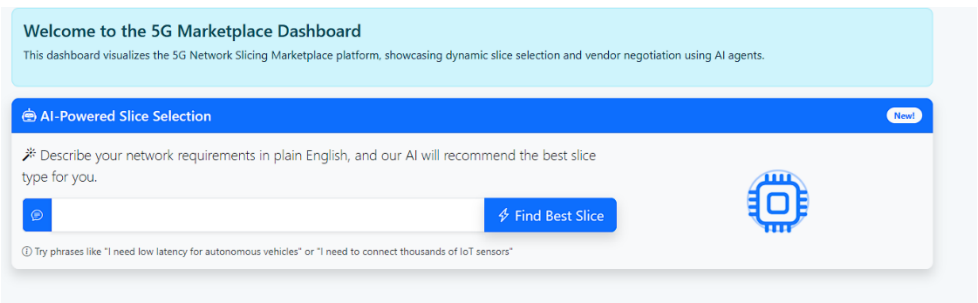


Fig. 9.1.8: Context based identification and slice allocation from natural language

Fig. 9.1.8 shows an AI agent adjusts each vendor's trust score based on real-time SLA violations or success metrics from the NDT. Implements continuous learning and self-regulation — a defining feature of AI-native networks per ITU/European Telecommunications Standards Institute (ETSI) standards.

Together, these results showcase a complete AI-Native Network Slice Marketplace architecture, where slice classification, vendor negotiation, deployment, monitoring, and adaptation are fully autonomous and learning-driven. This system advances beyond static 5G models and illustrates core building blocks for AI-native orchestration in 6G.

9.1.10. PoC Observation and discussions

This PoC successfully demonstrated integration of AI techniques like DQN (for vendor/resource selection), LSTM (for traffic prediction using open5GS), and LLM agents (for monitoring and reasoning via Ollama + Gemini). While the setup addressed key challenges around slice decision-making and validation using real traffic, one key suggestion was to focus more on the information exchange between the application and AI models, particularly in terms of data formats and schema consistency (e.g., JSON, time-series). It was also recommended to clearly identify the AI agent's position in the architecture, such as whether it resides within the NSSMF, NSMF, or as an external entity interacting via standard interfaces. Another critical insight was the need to study closed-loop monitoring and optimization, where feedback from systems (via LLM agents) is used to trigger real-time adaptations. Enhancing the AI inference endpoint (orchestrator) to support such structured input/output and feedback loops was also identified as a future enhancement.

9.1.11. Conclusion

PoC 1 effectively addresses the gap in dynamic slice management by introducing modular AI agents and defining their interaction with core functions. It sets the foundation for closed-loop, application-aware orchestration in future 6G networks.

9.1.12. Future Work

- Testing in a real time 5G environment

To validate the effectiveness and practical viability of the AI-Native Proactive Network Slice Marketplace, future research should focus on deployment and testing within an operational 5G network setting.

- Integrating Explainable AI

Incorporating Explainable AI (XAI) into the AI-Native Proactive Network Slice Marketplace is crucial to clarify why the AI models select certain network slice types in response to given traffic patterns or enterprise requirements.

9.2.1. PoC-002 [b-FG-AINN-I-136] Using AI to Reduce the 6G Standards Barrier for African Contributors

9.2.2. Delimiting Characteristic

Architecture approaches for deep integration of AI – because this PoC reuses specific architecture components from ITU-T Y.3061 and similar architectural components are used to deeply integrate AI (as an enabler for multi-agent systems).

9.2.3. Description

PoC-002 introduces an AI-native system to empower African stakeholders in the 6G standardisation process. It leverages a specialized LLM coordinated by four autonomous agents—responsible for gap detection, analysis, standards recommendation, and compliance evaluation—operating on a semantically indexed ITU dataset. This agent-driven pipeline automates and simplifies the process of identifying and addressing gaps in existing standards.

9.2.4. Gaps Addressed

This PoC addresses key gaps in automating standards intelligence through embedded, modular AI. It introduces application-aware processing of large standards corpora, autonomous agent lifecycle management, and low-latency inference with broad generalization. By operating on a semantically indexed dataset, it enables scalable, context-driven gap analysis. This architecture enhances inclusion and agility in the 6G standards ecosystem, especially for underrepresented regions.

9.2.5 POCs Test Setup

This PoC runs on a modular, multi-agent AI architecture built using frameworks like AutoGen or AnythingLLM, operating over a semantically indexed knowledge base created from ITU-T 6G standards. The base is stored in a Pinecone vector database, allowing fast contextual retrieval. Autonomous agents are assigned specialized tasks: gap detection (via Natural Language Processing (NLP) parsing), standards recommendation, compliance verification, and code generation. Documents are pre-processed with LLaMA 3.23B-based embeddings and chunked with metadata annotations (e.g., interface IDs, region tags). Each test phase document ingestion, gap detection, and regional filtering. These are tracked against functional success criteria such as correct parsing, identification of ambiguous sections, and relevance to African infrastructure needs. All gaps and proposed standards enhancements are logged with source excerpts, confidence scores, and region-specific tags, enabling full traceability for validation and audit. Figure shows AI-Native Multi-Agent Workflow for Gap Detection, Contribution Generation, and Compliance Validation in ITU-T 6G Standards.

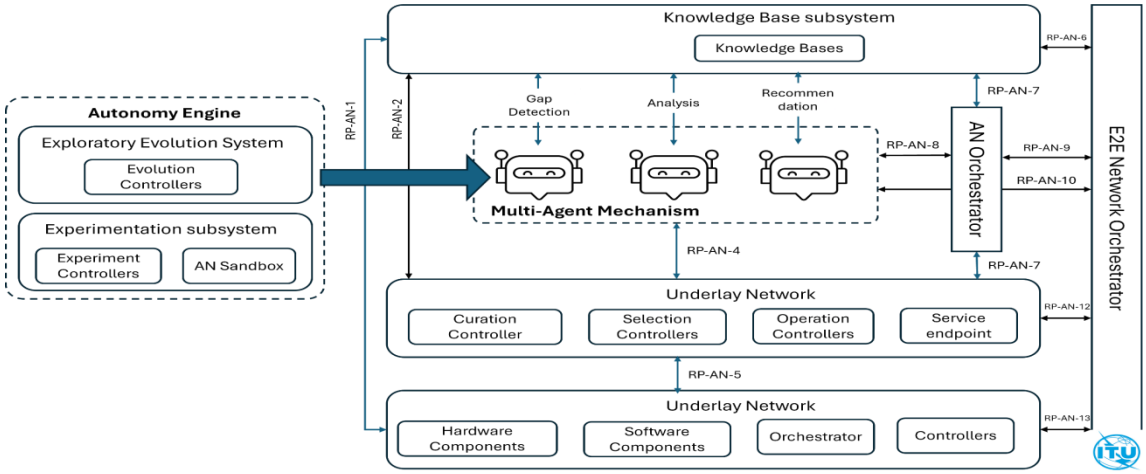


Fig. 9.2.1: AI-Native Multi-Agent Workflow for Gap Detection, Contribution Generation, and Compliance Validation in ITU-T 6G Standards

9.2.6. Code and Data Sets

A curated and semantically indexed corpus of ITU-T standards and related 6G technical documents. Code and data set available at https://github.com/AgabaEmbedded/Bridging-Standards-Gap.

9.2.7. Simulated Use cases

The system supports autonomous AI agents to analyze ITU-T 6G standards and identify specification gaps, particularly for developing regions. It semantically links abstract technical terms to local deployment challenges like power scarcity or rural infrastructure. Agents generate compliant proposal drafts and sample code aligned with standards evolution. This enables inclusive, AI-assisted contributions to global standardization efforts.

9.2.8. Architectural concepts

This PoC introduces a multi-agent AI architecture, where each autonomous agent performs a distinct function in the standards engagement lifecycle: gap detection, contextual analysis, recommendation, and compliance evaluation. The architecture is supported by a domain-specific LLM and a semantically indexed standards corpus, creating an end-to-end pipeline for AI-assisted, standards-aware collaboration.

9.2.9. Demo and Evaluation

The system was evaluated through three targeted tasks: semantic indexing of 6G standards, autonomous gap detection, and AI-assisted draft contributions. Each task was demonstrated using a curated document base, emphasizing the agent's ability to interpret, extract, and act upon regulatory and regional standardization needs. Controlled tests validated the completeness and accuracy of document parsing, relevance filtering for African network issues, and the precision of proposed enhancements. The evaluation focused on semantic search quality, the correctness of gap extraction, and alignment of generated proposals with known domain challenges.

9.2.10. PoC Observation and discussions

A key observation from PoC-002 is the need to focus more explicitly on the construction, structure, and utility of the knowledge base, which underpins the entire agent workflow. As a remote and the only foreign team, ensuring clarity and reproducibility of the knowledge base is especially critical for broader adoption and collaboration. It was suggested, notably by Marco, that the knowledge base itself could serve as a foundation for automated gap analysis across ITU standards, reinforcing its central role in enabling intelligent, AI-driven contributions to 6G standardisation. This highlights the importance of making the knowledge layer more transparent, accessible, and well-documented.

9.2.11. Conclusion

This PoC demonstrates an AI-native approach to democratizing 6G standards development through an agent-driven system powered by a semantically indexed ITU dataset. By automating gap detection, analysis, and compliance evaluation, it lowers the entry barrier for underrepresented regions, particularly in Africa. The modular design enables scalable, context-aware processing and sets the groundwork for broader integration of AI in standards workflows. This PoC marks a significant step toward inclusive, intelligent, and responsive 6G standardisation processes.

9.2.12. Future Work

- Scalability to Diverse Developing Regions: Future studies must prioritize the scalability and adaptability of 6G solutions to the highly diverse contexts within developing regions. Challenges include varied infrastructure maturity, significant language diversity, and fluctuating levels of technical capacity.
- AI Ethics: Critical attention is required for the ethical implications of deploying AI-driven 6G technologies. Future studies should address concerns related to data privacy, algorithmic bias, transparency, and accountability, particularly when AI models interact with sensitive user data and make decisions impacting connectivity and services.

9.3.1. PoC-003 [b-FG-AINN-I-128] Aionet Build-a-thon 2025: AIONETx
9.3.2. Delimiting Characteristic

AI itself as a core component – because the PoC is not just embedding AI into the architecture, but making AI the essential functional element that enables autonomous, on-device real-time control. Without the AI classifiers and predictors, the system would not work; it is fundamentally AI-driven rather than just architecturally integrated.

9.3.3. Description

PoC-003 is a lightweight, AI-native system that runs entirely on user devices to provide autonomous, real-time control of mobile data usage. It classifies traffic, predicts app behavior, and dynamically prioritizes bandwidth based on user-defined preferences and contextual factors like CPU load and packet flow without relying on centralized infrastructure.

9.3.4. Gaps Addressed

This PoC addresses key gaps in user-centric traffic control, edge-level autonomy, and infrastructure fairness. It fills the absence of application-aware, adaptive bandwidth management at the device level, especially in constrained or underserved environments. By decentralizing intelligence, it supports resilient operation under poor connectivity, and empowers users with informed control over data usage—contributing to responsible digital consumption.

9.3.5. POCs Test Setup

This PoC is deployed locally on a Linux environment simulating an edge device, operating without any cloud dependency. Scapy is used for real-time packet sniffing, psutil and subprocesses for CPU and process-level monitoring, and iptables/tc for active bandwidth shaping. A rule-based AI agent dynamically prioritizes flows based on CPU usage, packet type, port mapping, and user-defined bonuses. Unknown apps are logged, visualized on a Flask dashboard (with Chart.js and SweetAlert2), and handled through manual feedback stored in an encrypted SQLite database (AES-256). A heuristic priority engine references a live aionet_priorities.json config file, which is auto-reloaded every 30 seconds to reflect admin updates. The entire system—including data capture, classification, enforcement, and override is integrated and tested on a resource-constrained local stack, with simulation of network degradation using tc netem and real-time telemetry logs retained for policy validation and system tuning. Figure shows AIONET Workflow for Real-Time Traffic Monitoring, AI-Based Prioritization, and Traffic shaping.

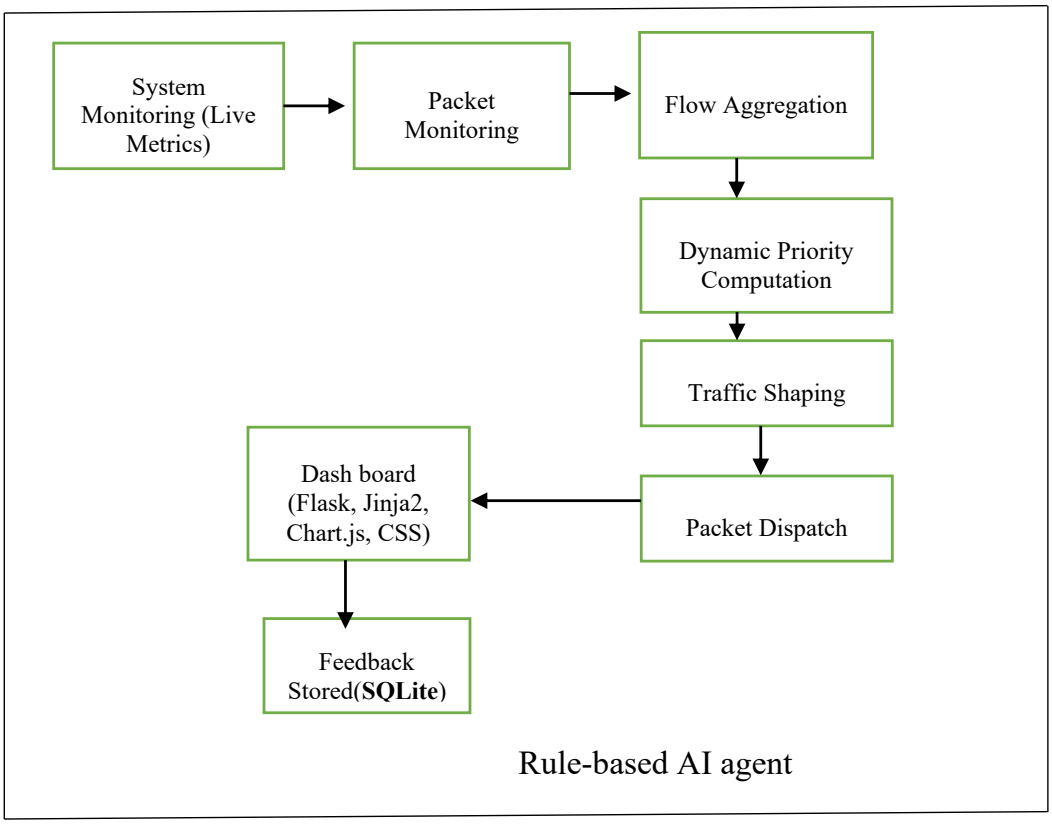


Fig. 9.3.1: AIONET Workflow for Real-Time Traffic Monitoring, AI-Based Prioritization, and Traffic shaping.

Collectively, the PoC setups in [1]-[3] implement a range of well-defined test environments, including edge-based execution, standards document analysis, and AI-driven network orchestration. Each setup emphasizes real-time telemetry collection, structured data logging, and system-level responsiveness, demonstrating the applicability of AI agents in enabling proactive, context-aware network management and standards support.

9.3.6. Code and Data Sets

Uses simulated real-time network traffic generated on the device for development and testing, capturing fields like timestamp, protocol, source/destination ports, and packet size. Code and data set available at https://github.com/Varsh-gr8/AIONET_ITU_build-a-thon2.0.git.

9.3.7. Simulated Use cases

In a constrained bandwidth scenario in this POC with concurrent Zoom, YouTube, and Google Drive traffic, AI agent in AIONET ensured high-priority access for Zoom while moderately throttling YouTube and delaying Drive sync. This behaviour, driven by real-time statistics and rule-based classification, validated the effectiveness of edge-intelligent traffic management.

9.3.8. Architectural concepts

It demonstrates key design directions for AI-native, user-empowered edge systems. It emphasizes on-device intelligence, real-time application-aware control, and autonomous bandwidth prioritization without relying on network-side orchestration. This direction aligns with the broader vision of decentralized, context-driven, and privacy-preserving AI-native networks, particularly for extending connectivity benefits to underserved users.

9.3.9. Demo and Evaluation

The PoC was demonstrated on a local edge device simulating real-world applications and network conditions. The AI agent monitored active traffic, assigned dynamic priorities, and enforced shaping policies in real time via a local dashboard. Demonstration steps included simulated bandwidth constraints, user overrides, and detection of unknown applications. The evaluation plan measured throughput capacity, latency overhead, dashboard responsiveness, and shaping accuracy. Success criteria emphasized real-time adaptation, effectiveness of heuristic priority assignment, and offline operability, validating AIONET's suitability for privacy-sensitive, resource-constrained environments.

9.3.10. PoC Observation and discussions

Key observations from PoC-003 highlight the need to explicitly define the dataset structure, including packet fields (e.g., timestamp, protocol, source/destination ports, length), and demonstrate how it supports dynamic and unknown packet formats. The ability to classify, tag, and adapt to new traffic types is central to AIONET's AI-native claim, yet must be clarified with examples of policy formats, classification mechanisms, and how new rules are injected or updated at the edge. Additionally, deeper explanation of traffic shaping policy deployment across UE and UPF would strengthen the system's relevance to AI-native, decentralized network control. Addressing these would reinforce AIONET's potential to support real-time, evolving, user-driven networking at the edge.

9.3.11. Conclusion

It showcases a practical direction for AI-native networking by enabling autonomous, on-device traffic control and prioritization based on user context and preferences. It highlights how lightweight, decentralized intelligence can enhance digital inclusion and responsible consumption, especially in connectivity-constrained environments. While not proposing a full architecture, it offers a strong foundation for future exploration of user-driven AI in the non-radio domain.

9.3.12. Future Work

- Future Security Work:
    - Integrate Flask-Login for session-based dashboard protection.
    - Move to HTTPS with self-signed or CA-issued certificates.
    - Validate and restrict config updates with hash-based integrity checking.
    - Define a formal threat model for internal and external adversaries.
- Ethical AI Governance:

AIONET currently logs all traffic shapping and priority decisions, enabling retrospective audit of priority assignments. Although rule adaptation is automated based on user feedback, administrators are responsible for periodically reviewing logs via the dashboard to detect any unfair or biased prioritization patterns. This human-in-the-loop oversight ensures transparency and fairness in network traffic management, with plans to enhance automated bias detection in future updates.

- Future Study Directions
- Integration with Network Slicing

o Enable AIONET to interact with operator-side slice managers or SDN controllers for coordinated QoS.

- Knowledge Graph-Based Policies
  o Use knowledge graphs to infer user intent and translate it into adaptive traffic control policies.

    - Federated Edge Training

o Explore collaborative model training across devices using federated learning while maintaining privacy.

• Standardized Metadata Models

o Contribute to industry-wide standards for traffic/session metadata used in AI-native systems.

• Adaptive Threat Detection

o Integrate lightweight anomaly detection to identify misuse or hidden bandwidth drains.

• Mobile Optimization

o Optimize model and traffic shaping for mobile devices (Android/iOS) with varying app permissions and APIs.

- Explainable AI (XAI) Interface
  o Design a dashboard that explains AI decisions using natural language or visual cues to improve transparency.
- Formal Intent-to-Policy Mapping
- Develop a standardized framework that translates user intent into enforceable network policies using AI + logic.

9.4.1. PoC-004 [b-FG-AINN-I-133] Build-a-thon 2025: NETSPEAK

9.4.2. Delimiting Characteristic

Engagement of AI in all stages of the lifecycle— because this PoC embeds AI not just for one function, but throughout the troubleshooting lifecycle; from interpreting complaints (input), to classifying intents, to generating configurations, and finally validating results in a closed loop. The AI is actively engaged end-to-end, enabling continuous improvement of the troubleshooting process.

9.4.3. Description

PoC-004 presents an AI-native troubleshooting assistant that interprets user complaints given in natural language, classifies them into network intents, and automatically generates and executes Command Line Interface (CLI) configurations. Using LLMs, Jinja2 templates, and tools like Netmiko and Ansible, the system supports multiple vendors (e.g., VyOS, Cisco) and performs dynamic validation via iperf and ping, creating a closed-loop, self-improving troubleshooting pipeline.

9.4.4. Gaps Addressed

This PoC addresses the lack of intelligent, intent-driven troubleshooting in network operations. Traditional systems rely on manual diagnosis or rigid, pre-defined templates. PoC 4 fills this gap by enabling context-aware automation, where user input in natural language leads to accurate classification, vendor-specific configuration, and validation without expert intervention.

9.4.5. POCs Test Setup

The PoC is deployed in an EVE-NG-based emulated testbed containing Cisco IOS, pfSense, and Juniper vSRX devices. A lightweight Flask-based orchestrator accepts natural language input from CLI or Web UI, which is classified, resolved, and mapped to Jinja2 templates before being executed via Netmiko. CLI outputs are validated, tested (e.g., using iperf and ping), and logged for continual improvement. The backend stack includes Python, PyTorch, Flask, Netmiko, and Ansible, enabling real-time inference, template rendering, and multi-vendor command execution. Test cases evaluate intent interpretation, CLI

generation accuracy, rollback handling, and dynamic augmentation through documentation via RAG. Figure shows simulated network trouble shooting [4]

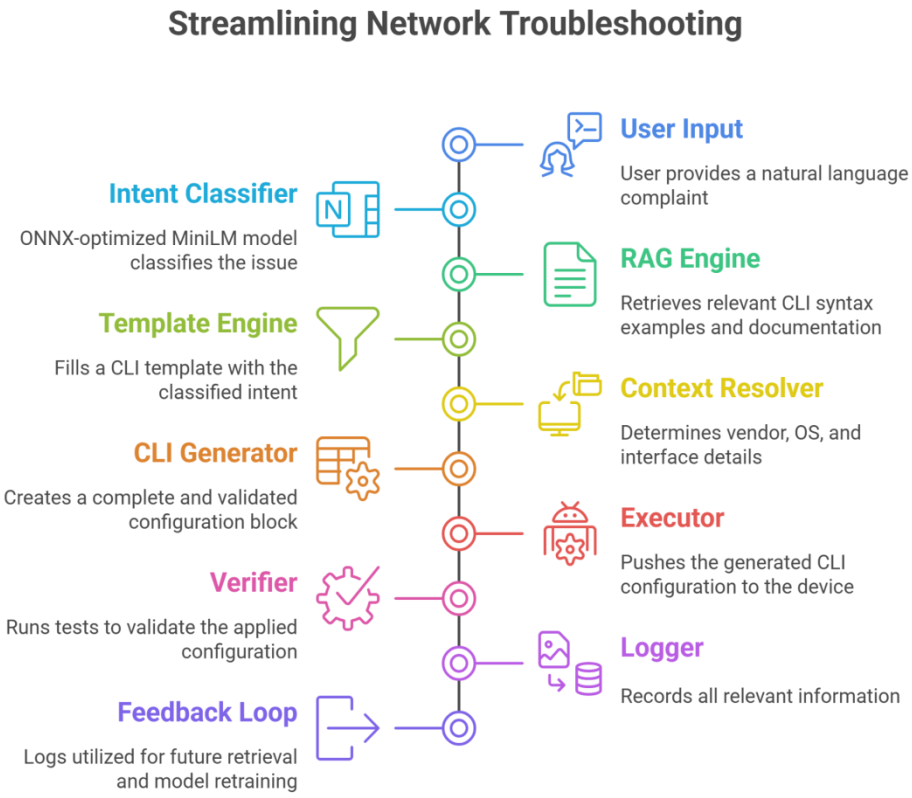


Fig. 9.4.1: Simulated network trouble shooting [4]

9.4.6. Code and Data Sets

The system uses a synthetic dataset of over 9,000 labeled complaints mapped to intent categories, CLI actions, and vendor contexts. Additionally, curated metadata files (e.g., parse.py, policy.json, app_profiles.py) support structured inference and command generation. Code and data sets can be found via https://github.com/rajudhangar100/netspeak_ui.

9.4.7. Simulated Use cases

The chosen use case simulates a user request to increase link capacity during an IPL match in zone 1 from 7pm to 9pm. The user provides a natural language input, which is classified by the system into a structured intent including action (Increase Link Capacity), zone, time window, and service type (Video Streaming). The template engine loads a VyOS QoS policy, fills it with contextual parameters (e.g., eth1, video-stream, 800mbit), and the CLI generation module renders and validates the configuration block for execution—demonstrating real-time, intent-to-configuration automation for network optimization.

9.4.8. Architectural concepts

PoC-004 introduces an AI-native intent-to-action pipeline, comprising an NL parser, intent classifier, CLI template generator, and multi-vendor executor. It integrates real-time feedback and validation, enabling a self-improving, closed-loop system capable of handling diverse edge environments with minimal manual effort.

9.4.9. Demo and Evaluation

The PoC demonstrates AI-native, intent-driven automation through a user scenario requesting increased link capacity during an IPL match. A natural language input is classified using DistilBERT into a bandwidth_prioritisation intent. The system resolves the zone-to-interface mapping (e.g., zone 1 → GigabitEthernet0/1), selects the appropriate Jinja2 template, and generates the CLI block (bandwidth 1000000). The configuration is scheduled and deployed via Netmiko, with actions logged in MongoDB. Post-deployment ping and iperf tests verify success, and rollback is triggered if needed. The feedback loop stores outcomes for continuous learning, ensuring robustness, adaptability, and compliance with AI-native principles.

9.4.10. PoC Observation and discussions

PoC-004 demonstrates a promising AI-native approach to automated network troubleshooting, but several important areas for enhancement were identified. There is a need to systematically determine and isolate CLI changes, supported by a structured CLI command dataset and refined Jinja templates to improve accuracy across vendors like VyOS, Cisco, and Juniper. The relationship between network state, domain knowledge, and generated configurations must be made more explicit, including how intent is grounded in real-time network context. Additionally, future work should explore the development of a human intent dataset, clarify base models used for NL parsing, and address credential management and secure execution pipelines for automated CLI deployment. Finally, implementing a closed-loop feedback mechanism is essential for monitoring the impact of changes on KPIs.

9.4.11. Conclusion

This PoC demonstrates how AI-native techniques can automate complex, vendor-aware network troubleshooting tasks by bridging natural language input with executable configurations. It reduces dependence on human expertise, accelerates resolution, and lays the groundwork for intent-driven, adaptive network operations at the edge.

9.4.12. Future Work

- Model Adaptability: Ensure the AI model generalizes across diverse network environments and vendor-specific configurations.
- CLI Generation Accuracy: Prevent misconfigurations by refining dynamic template generation and validation methods.
- Reinforcement Learning Integration: Implement real-time feedback loops to optimize network adjustments continuously.
- Validation & Rollback Mechanisms: Develop robust safeguards to ensure deployed configurations do not destabilize the network.
- Multi-Cloud & Hybrid Network Support: Expand compatibility with cloud-based architectures and hybrid environments.
- Security & Privacy Considerations: Address vulnerabilities, adversarial attacks, and ethical concerns in AI-driven automation.
- Edge Execution Optimization: Enhance lightweight inference models for real-time execution on routers and microcontrollers.
- Scalability & Performance Improvements: Focus on reducing latency, improving classification accuracy, and automating network responses efficiently.

9.5.1. PoC-005 [b-FG-AINN-I-130] Build-a-thon 2025: Explainable AI-Native Intent-Based Self-Healing Network Orchestrator for Rural and Edge Telecom Infrastructure

9.5.2. Delimiting Characteristic

Engagement of AI in all stages of the lifecycle because the PoC applies AI across the entire fault management lifecycle: from intent interpretation, to anomaly detection, to explainability, to automated healing. Each stage depends on AI modules working in sequence within a coordinated pipeline, making lifecycle-wide AI engagement the defining characteristic.

9.5.3. Description

This PoC addresses the rural connectivity crisis by designing an AI-native fault management system for edge and resource-constrained telecom environments. It integrates intent-based operator input, LSTM-based anomaly detection, explainable outputs, and rule-based healing agents coordinated via a lightweight, modular pipeline. The system autonomously interprets faults, explains causes, and triggers corrective actions with minimal operator intervention, aiming to drastically reduce recovery times in rural infrastructure.

9.5.4. Gaps Addressed

This PoC addresses critical gaps in resilient, autonomous fault recovery for rural telecom networks. Traditional Network Management Systems (NMS) systems are reactive, centralized, and lack adaptability to edge environments. The proposed system overcomes these limitations by embedding AI-native capabilities directly at the edge, enabling proactive, explainable, and context-aware fault response in environments where manual intervention is slow or unavailable.

9.5.5. POCs Test Setup

The PoC is evaluated on an emulated 50-node rural telecom network, using synthetic datasets that reflect realistic outage and traffic patterns. Model training is conducted locally using Python (Scikit-learn/Keras), with LSTM-based anomaly detection and SHAP-based explainability. The multi-agent healing architecture simulates MCP-style coordination via Python FSM agents, while network flows are emulated in NS-3. Inference fallback is supported through a local RAG-style knowledge query layer, and additional orchestration and 5G core integration are achieved via xOpera and Open6GCore. The complete pipeline is planned for public release via the ITU AI/ML 5G Challenge GitHub. The following figure shows the pipeline of control in the test setup.

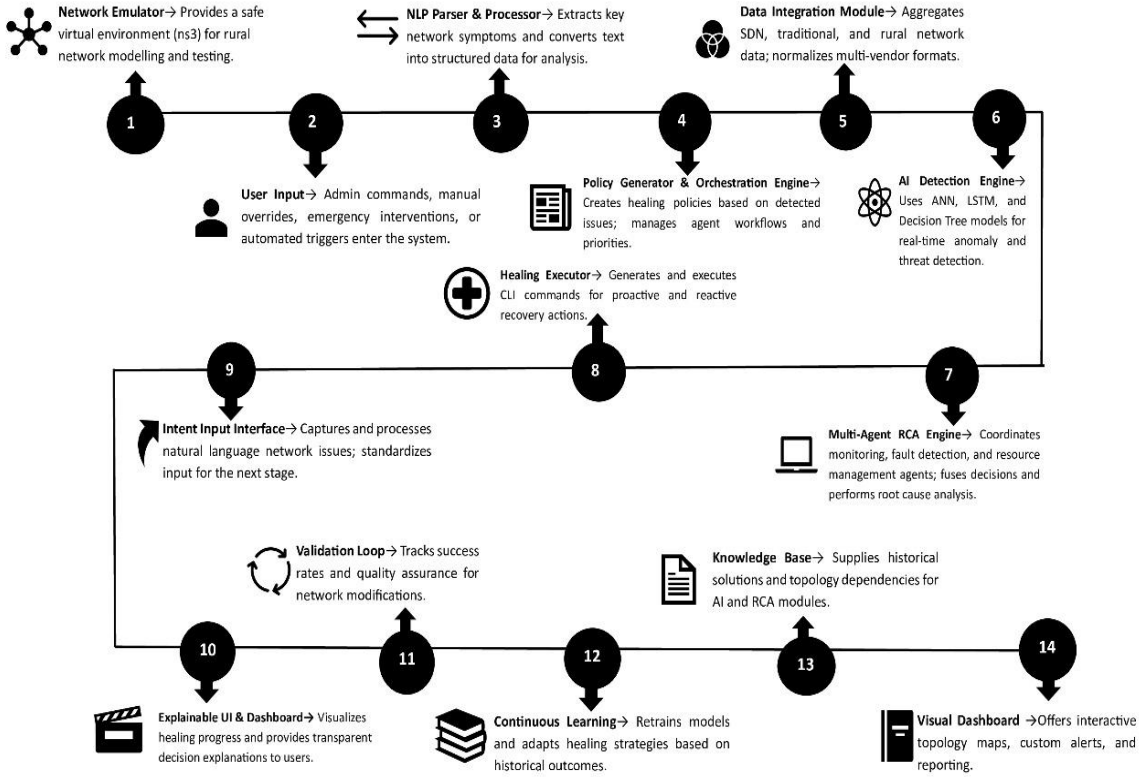


Fig. 9.5.1. The pipeline of control in the test setup.

9.5.6. Code and Data Sets

Synthetic time-series datasets simulating rural telecom metrics (throughput, latency, power quality, link status, etc.) are generated using NS-3 and Python. These include annotated anomaly traces for LSTM training (normal vs. degraded), curated edge-case failures (fiber cuts, voltage instability), and semantic intent examples for translation. Code and datasets can be found here: https://github.com/Rishi8520/rural_ai_selfhealing_net.

9.5.7. Simulated Use cases

This PoC validates AI-native fault detection and response through realistic edge network scenarios. In the fiber cut simulation, a fault is injected between critical nodes (e.g., CORE-0 and CORE-1). The LSTM-based anomaly detector identifies early degradation, explains the root cause using SHAP, and initiates automated self-healing via coordinated agents. In the power fluctuation scenario, voltage anomalies at nodes (e.g., DIST-2) are detected and proactively addressed using policy-based actions, with and without knowledge base support to compare recovery effectiveness. A third case tests intent translation, where an operator-defined policy (e.g., maintaining link uptime) is interpreted and propagated across agents. The presence of the MCP-style coordination layer significantly improves system responsiveness, demonstrating the impact of distributed intelligence and orchestration efficiency in rural networks.

Architectural concepts The PoC introduces a modular AI-native orchestration pipeline composed of:

- An intent interpretation layer that parses operator instructions.

- An LSTM-based anomaly detector for predictive fault detection and classification.
- A multi-agent coordination framework inspired by MCP (Multi-agent Control Plane) principles to trigger rule-based healing.
- A lightweight inference engine with knowledge-based fallback at the edge, enabling autonomy even under degraded connectivity.

The design emphasizes distributed intelligence, explainability, and fault resilience tailored to rural deployments

9.5.8. Demo and Evaluation

This PoC is is developed and tested on a 50-node simulated network using NS-3 to emulate rural broadband conditions and failure events such as fiber cuts and power fluctuations. The system operates through a sequence of coordinated autonomous agents, each handling a specific function in the fault management pipeline. The architecture includes a Monitor Agent for collecting real-time metrics, a Calculation Agent powered by LSTM models for detecting anomalies, a Healing Agent that uses Retrieval-Augmented Generation (RAG) and Gemini LLM to generate healing plans, and an MCP Agent that propagates decisions across agents. Actions are then executed by the Orchestration Agent using TOSCA templates and tools like xOpera, with communication managed through ZeroMQ sockets under a central orchestrator. This setup enables the system to autonomously process operator intents, detect issues using deep learning, provide SHAP-based explainability, and coordinate corrective actions across agents. Evaluation metrics include detection latency, recovery time, explanation fidelity, and intent propagation effectiveness, both with and without knowledge base or coordination mechanisms. The modular, distributed agent design allows flexible testing of AI-native orchestration capabilities under constrained edge conditions.

9.5.9. PoC Observation and discussions

This PoC presents an ambitious architecture, but observations suggest the need to narrow scope—particularly for the Build-a-thon 1.0 test case—to ensure implementability and clarity. The role of AI techniques (e.g., LLMs for healing and anomaly reasoning) and MCP/A2A coordination mechanisms requires further articulation, especially in terms of data flow and interaction timing. A clear mapping from NS-3 simulation (e.g., JSON outputs, node/link/path counts) to Monitoring Agent → MCP → Calculation Agent → A2A → Healing Agent is needed to assess real-time adaptability. In particular, the curation and schema design for RAG, and its use in Gemini-powered healing prompts, should be detailed alongside KB structure and YAML-based configuration plans. Focusing on a sample symptom and tracing its lifecycle—from anomaly detection to adaptive healing and reconfiguration—can illustrate the agent coordination and AI-native feedback loop. To strengthen the implementation pathway, it is recommended to prioritize fault isolation, explicitly document data/model requirements, and provide a layer-wise architecture diagram showing the before-and-after system state transitions.

9.5.10. Conclusion

This PoC demonstrates how AI-native orchestration can transform fault management in rural and edge networks. By combining anomaly detection, intent parsing, and agent-based healing in a modular, explainable framework, the system enables faster recovery, reduces dependence on central systems, and improves trust in autonomous operations. It lays the groundwork for scalable, fault-tolerant rural infrastructure aligned with the goals of digital inclusion and 6G resilience.

9.5.11. Future Work

While our current proof-of-concept demonstrates intent-driven anomaly detection and autonomous self-healing through explainable AI and lightweight agents, several open challenges remain for further research and deployment at scale:

1. Scalability of Agents in Heterogeneous Networks: Our current rule-based multi-agent system is effective in small to mid-sized topologies. Future work must explore how coordination and communication between agents scale in large, diverse rural deployments involving different vendor equipment and protocols.

2. Real-Time Constraints and Low-Latency Learning: Incorporating more complex ML models (e.g., transformer-based intent interpreters or reinforcement learning agents) may improve flexibility but introduces latency. Research is needed to balance explainability, speed, and predictive performance in real-time networks.

3. Interfacing with Real Network Controllers ((Software Defined Networking (SDN)/Network Function Virtualization (NFV)): Our orchestrator currently simulates healing actions. Integration with real-world network control systems (e.g., OpenDaylight, Open Network Operating System (ONOS)) will require protocol adaptation, security compliance, and stateful communication models.

4. Edge Model Optimization and Federated Learning: In future iterations, we aim to deploy distributed learning systems using federated or incremental learning to adapt models across geographically dispersed nodes while preserving data privacy and connectivity independence.

5. Intent Disambiguation and Natural Language Generalization: Currently, our intent parser is rule-based. A future goal is to integrate a fine-tuned language model for more flexible and user-friendly intent interpretation, especially for multilingual operators or ambiguous commands.

6. Resilience to Adversarial Events and False Positives: The system must be stress-tested under adversarial scenarios (e.g., spoofed telemetry, conflicting intents) to ensure robustness and prevent cascading failures due to incorrect healing triggers.

These open problems provide an opportunity for iterative enhancement of our framework toward a production-ready, standard-compliant AI-Native orchestration system suitable for rural and edge telecom networks.

9.6.1. PoC-006 [b-FG-AINN-I-135] Build-a-thon 2025: 5G Adaptive Signal Solutions for Rapid Transit and Aerial Operations

9.6.2. Delimiting Characteristic

AI itself as a core component because the adaptive beam steering fully depends on AI to learn mobility patterns and dynamically adjust antenna beams, making AI the essential element of the system’s operation.

9.6.3. Description

This PoC proposes an AI-based adaptive beam steering mechanism designed for high-mobility scenarios such as rapid transit systems, fleet management, and drone connectivity. The system dynamically adjusts antenna beams by learning the motion of mobile stations using Received Signal Strength Indicator (RSSI) derived from Channel state Information (CSI), ensuring that user equipment (UE) remains within the optimal beam coverage (half-power beamwidth). The goal is to enhance link reliability, minimize signal loss, and sustain continuous 5G connectivity in fast-moving environments.

9.6.4. Gaps Addressed

Traditional beam steering methods are either slow to adapt or rely on pre-programmed patterns, which are insufficient for highly dynamic use cases like fast-moving vehicles or drones. This PoC addresses the gap by introducing real-time learning-based steering, enabling context-aware, low-latency beam tracking. It tackles the limitations of static or delayed beamforming and aligns with the needs of AI-based, mobility-optimized networks.

9.6.5. POCs Test Setup

The PoC is implemented using Sionna's AI-native Radio Access Network (RAN) simulation environment for PHY-layer modeling, with a Uniform Rectangular Array (URA) antenna and custom beamforming modules. An evolutionary algorithm explores diverse beam configurations across user positions and channel conditions, logging RSSI, CSI, and spatial parameters to generate a training dataset. This dataset is used to pre-train a reinforcement learning (RL) agent, which is later fine-tuned online using real-time feedback for adaptive beam tracking. The RL agent interacts with the RAN simulator through RESTful APIs, supporting near-real-time decision exchange. The API design reflects O-RAN architecture principles, with E2-like interfaces handling dynamic beam control and A1-like interfaces managing policy updates. Figure shows flowchart of method for adaptive steering for Rapid Transit and Aerial Operations.

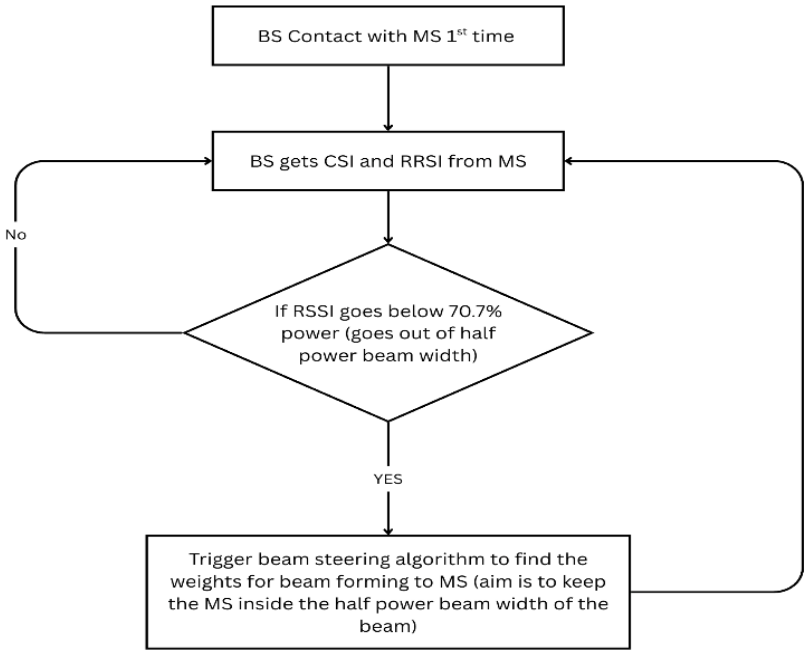


Fig. 9.6.1: Flowchart of method for adaptive steering for Rapid Transit and Aerial Operations

9.6.6. Code and Data Sets

A synthetic dataset was generated using the Sionna RAN simulator, capturing beam configurations, Channel State Information (CSI), Received Signal Strength Indicator (RSSI), and spatial parameters across varied UE positions and channel conditions. This dataset was used to pre-train the reinforcement learning agent, enabling it to learn an initial beamforming policy prior to online fine-tuning. The data generation was driven by evolutionary algorithms exploring the beam search space under mobility conditions. Code and data sets can be found here: https://github.com/syed-azim-git/Buildathon2.0-SSNECE.

9.6.7. Simulated Use cases

The simulation models a high-mobility wireless environment, such as urban transit or drone-based communication, where a base station equipped with a Uniform Rectangular Array (URA) antenna tracks and serves mobile UEs in motion. The UEs follow configurable trajectories, and their positions—as well as channel conditions—change over time. The simulator generates real-time CSI and RSSI measurements based on user movement, which are fed to the AI layer to dynamically adjust beamforming weights. This setup allows testing the system's ability to maintain signal quality and connection continuity under fast-changing spatial and channel conditions.

9.6.8. Architectural concepts

The architecture integrates:

- Motion learning models for trajectory prediction.
- RSSI-CSI feedback loops for real-time environment sensing.
- Dynamic beam alignment logic executed at the antenna array controller or baseband unit.
- An AI-based control layer that adapts beam parameters without centralized intervention.

9.6.9. Demo and Evaluation

To evaluate the performance and robustness of the proposed dynamic beam steering algorithms, we employed Sionna, an open-source, TensorFlow-based simulator developed by NVIDIA. Sionna provides a highly flexible and accurate PHY-layer simulation framework, supporting digital twin environments and advanced antenna modeling. Its capability to simulate realistic urban propagation scenarios, combined with custom beamforming modules, made it well-suited for testing adaptive beam steering strategies under diverse mobility and channel dynamics.

9.6.10. PoC Observation and discussions

The PoC highlights a strong integration between AI-driven beamforming and PHY-layer simulation; however, there is a need to abstract the Sionna-specific components to generalize the approach for RAN optimization across simulators. Clarifying the timing and placement of the genetic algorithm (GA) for dataset creation and the online reinforcement learning (RL) agent for real-time adaptation is important to improve reproducibility and modularity. The RESTful API interfaces between the AI layer and the RAN simulator should be documented explicitly to ensure interoperability—enabling replacement of Sionna with other tools while preserving functionality. Future enhancements could focus on dataset creation workflows, YAML-based simulation configuration, and support for user-defined BS/UE locations to enable broader experimentation. Drawing from the DeepSense 6G Challenge, the team may extend the framework to other radio-layer learning problems. Finally, formalizing the simulator's configurable parameters and their representation in YAML would help standardize testing across use cases and environments.

9.6.11. Conclusion

The PoC demonstrates how AI-based beam steering can improve 5G performance in highly mobile scenarios by minimizing signal degradation and handover interruptions. By using real-time signal metrics and learning models, it offers a responsive and context-aware solution that enhances throughput, reliability, and service continuity in dynamic environments laying the groundwork for resilient, intelligent transportation and emergency communication systems.

9.6.12. Future Work

- Beam steering option for UE tends to vary the speed of UE on the go.
- Energy Aware Decision making for Beam Steering

9.7.1. PoC-007 [b-FG-AINN-I-132] Build-a-thon 2025: Invisible Guard – Passive Wi-Fi Sensing for Smart Border & Building Surveillance

9.7.2. Delimiting characteristics

AI itself as a core component because the surveillance system relies entirely on AI models to interpret CSI variations and recognize human motion, making AI the fundamental driver of functionality rather than just an auxiliary tool.

9.7.3. Description

Invisible Guard is an AI-native, privacy-preserving surveillance solution that leverages passive Wi-Fi Channel State Information (CSI) to detect human motion and anomalies without relying on cameras, radar, or wearable devices. Using standard 802.11n/ac routers and compatible NICs (e.g., Intel 5300, Nexmon), the system captures fine-grained multipath variations in CSI caused by environmental disturbances. These are processed by lightweight temporal deep learning models (e.g., GRU or 1D-CNN) to recognize motion signatures such as walking, running, loitering, or intrusions. Designed for continuous, sensorless operation, the system is suited for resource-constrained, privacy-sensitive zones like borders, critical infrastructure, and homes.

9.7.4. Gaps Addressed

Invisible Guard addresses a key network-layer gap by embedding AI-native perception into the physical (PHY) layer using passive Wi-Fi CSI. It enables privacy-preserving, sensorless intrusion detection at the network edge—where traditional systems rely on explicit sensors or centralized analysis. This fills a critical void in AI-native architectures by supporting real-time, context-aware anomaly detection directly through ambient wireless signals, enhancing observability and resilience in distributed environments.

9.7.5. POCs Test Setup

The demonstration setup for Invisible Guard is entirely offline, designed to simulate real-time motion detection using pre-recorded Wi-Fi Channel State Information (CSI) logs. Rather than relying on physical hardware or live data collection, the system uses .pcap files containing CSI traces representing four human activity classes—such as walking, running, loitering, and intrusion—captured under controlled conditions. These files serve as the input for a structured preprocessing pipeline. First, raw CSI amplitude and phase values are parsed from the .pcap format. To reduce noise and enhance signal clarity, Discrete Wavelet Transform (DWT) is applied. The resulting clean signals are then segmented into fixed-length temporal windows and normalized.

Depending on the model type, the pipeline either extracts statistical features like mean, standard deviation, and entropy (for classical models like Random Forest), or directly passes time-series windows to deep learning models such as Bi-LSTM or GRU, which are better suited for capturing sequential dependencies in the data. These models are pre-trained and saved to disk, and during the demonstration, they are loaded into a Python-based Jupyter Notebook environment. Using libraries like TensorFlow and Scikit-learn, inference is performed on each input segment. The output includes the predicted activity class and an associated confidence score, which is dynamically visualized in the notebook interface. This setup offers a reproducible and privacy-preserving alternative to hardware-based surveillance, effectively showcasing AI-native anomaly detection through ambient wireless signals in a simulated but realistic setting. Figure shows flowchart of workflow.

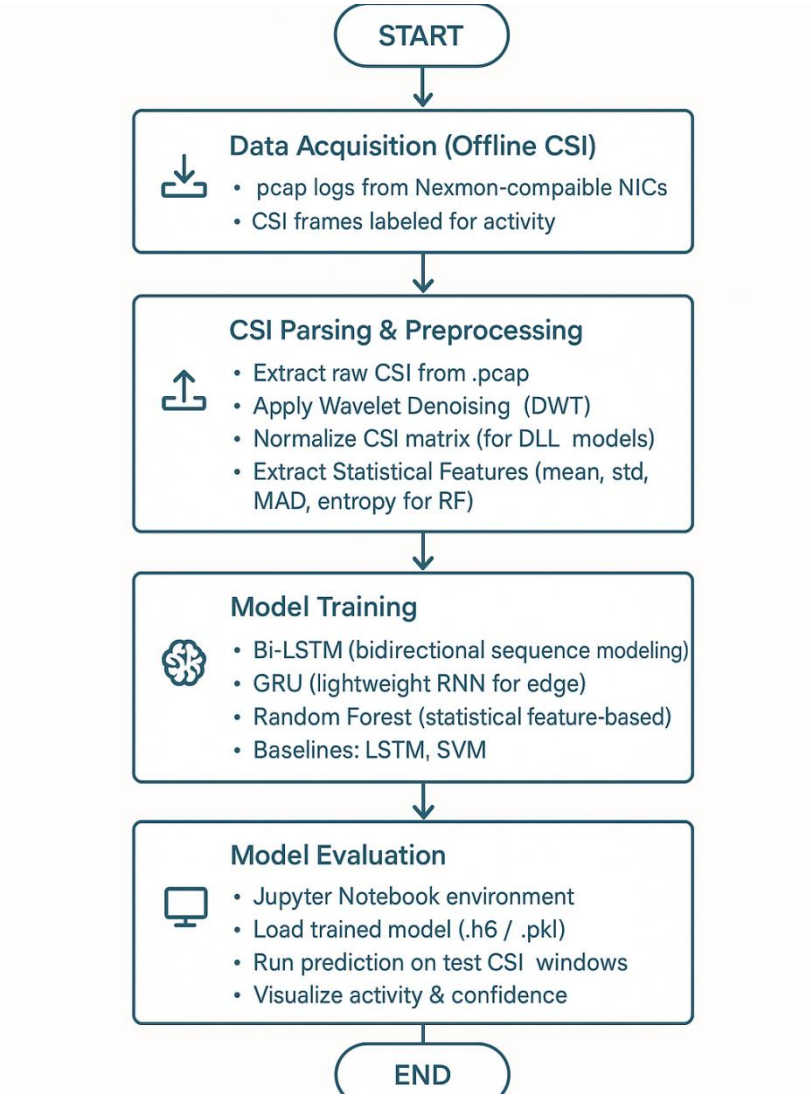


Fig. 9.7.1: Flowchart of workflow

9.7.6. Code and Data Sets

Uses an offline, simulated dataset derived from passive Wi-Fi Channel State Information (CSI) logs stored in .pcap format. The dataset includes four labeled human activity classes: Empty, Sitting, Standing, and Walking. Code and data sets can be found at https://github.com/himanshukhatri1511/ITU.

9.7.7. Simulated Use cases

The PoC simulates human activity detection (empty, sitting, standing, walking) using pre-recorded Wi-Fi CSI data. It mimics real-world intrusion scenarios in secure areas without using cameras or sensors, demonstrating AI-native, privacy-preserving surveillance via signal pattern changes.

9.7.8. Architectural concepts

Invisible Guard introduces an AI-native edge architecture where CSI data from Wi-Fi NICs is processed locally using lightweight deep learning models (1D-CNN/GRU) for real-time motion detection. It enables privacy-preserving, sensorless anomaly detection by embedding intelligence at the PHY layer, supporting distributed, context-aware surveillance without centralized sensors or cameras.

9.7.9. Demo and Evaluation

The demonstration and evaluation of Invisible Guard are conducted entirely within a Jupyter Notebook environment, showcasing a fully offline, hardware-free AI-native activity classification system. Pre-trained models—including Bi-LSTM, GRU, and Random Forest—are loaded from disk and applied to a test dataset derived from preprocessed Wi-Fi Channel State Information (CSI) logs originally stored in .pcap format. These logs have been parsed and segmented into temporal windows, each labeled with one of four activity classes: Empty, Sitting, Standing, or Walking.

For each input window, the system performs inference and outputs a predicted activity class along with a confidence score, both of which are dynamically visualized in real-time through prediction traces and confusion matrices. This offline simulation allows complete reproducibility and systematic evaluation without the need for live CSI capture or physical testbeds. As you can see from the Result summary table below, the proposed models have achieved high classification accuracy, validating the effectiveness of offline CSI-based activity detection. The GRU model stands out with an accuracy of ~98.96%, offering the best balance between performance and computational efficiency. Bi-LSTM and LSTM also perform strongly, with accuracies of ~97.99% and ~97.85% respectively, demonstrating their capability to model complex temporal patterns. In contrast, traditional methods like Random Forest and SVM show significantly lower accuracy, reinforcing the advantage of deep learning approaches for this task.

Model	Accuracy (%)	Notes
GRU	~98.96%	Best trade-off between speed and accuracy
Bi-LSTM	~97.99%	Best for transitional and complex sequences
LSTM	~97.85%	Strong performance with engineered features
Random Forest	~72.85%	Baseline sequence model
SVM	~67.25%	High dependency on preprocessing

9.7.10. PoC Observation and discussions

This PoC presents a novel AI-native surveillance framework using passive Wi-Fi CSI for human activity detection. However, several critical aspects require refinement to align the PoC more closely with AI-native networking principles. First, the application layer model, human activity detection, must be explicitly mapped to the underlying RAN, indicating what information from the RAN (e.g., CSI, UE authentication data) is being utilized and how that supports inference services. Additionally, the PoC should clearly define its inference endpoints (e.g., alerts or system flags) and outline the network-level requirements or interactions necessary to support real-time detection and response.

The dataset used is offline and derived from .pcap CSI logs; this should be more explicitly stated along with the model types (Bi-LSTM, GRU) and their execution environment. The absence of information regarding hosting and training (e.g., cloud, edge, or on-device) leaves a gap in understanding the scalability and deployment feasibility. Moreover, while the project operates as a cooperative sensing scheme, its ISAC (Integrated Sensing and Communication) relevance could be further clarified by specifying how ambient UEs contribute to signal diversity or coverage. Finally, the need for a layered interface view clearly showing the app, AI, and network layers—remains unaddressed and would greatly aid in evaluating system integration and operational scope.

9.7.11. Conclusion

Introduces a CSI-powered HAR framework that demonstrates the viability of non-invasive, privacy-friendly, and hardware-free activity detection. Through the use of Bi-LSTM, GRU, and Random Forest, we achieve superior classification accuracy and computational flexibility across various model types.

By removing the need for live data capture and simulation, we ensure that our system remains portable, reproducible, and simple to deploy. The success of our models in static CSI environments confirms the potential for practical applications in smart homes, healthcare, and ambient IoT systems. Moreover, the active development of multi-user detection, embedded deployment, and transformer integration lays the groundwork for a new generation of real-time, scalable HAR systems rooted in signal intelligence.

9.7.12. Future Work

1. Multi-user HAR using multi-label classifiers, signal decomposition (ICA), and MIMO-based spatial differentiation.

2. Generalization to variable room layouts and device positions using transfer learning and domain adaptation.

3. GRU deployment on Raspberry Pi/ESP32, using quantized models with CSI from Nexmon firmware.

4. Transformer-based modeling to enhance long-range temporal pattern recognition via self-attention mechanisms.

9.8.1. PoC-008 [b-FG-AINN-I-131] An advanced emergency response platform

9.8.2. Delimiting Characteristic

Engagement of AI in all stages of the lifecycle, because AI is applied across the entire emergency response lifecycle—detecting anomalies, triggering alerts, optimizing ambulance routes, reconfiguring networks in real time, and restoring operations—showing continuous AI engagement from incident onset to resolution.

9.8.3. Description

This PoC introduces an AI-native emergency response system tailored for urban individuals, particularly those living alone. It orchestrates real-time health anomaly detection, emergency alerting, location tracking, ambulance dispatch, and continuous patient-health monitoring during transit. At the core, the system uses AI/ML algorithms for dynamic ambulance route optimization and real-time O-RAN network reconfiguration to guarantee end-to-end (E2E) Quality of Service (QoS) for emergency communication flows. Once the emergency is resolved, the system autonomously reverts network resources to normal operations.

9.8.4. Gaps Addressed

The PoC addresses the lack of real-time, intent-driven E2E QoS provisioning in emergency scenarios. Existing networks do not dynamically prioritize health-critical traffic or support AI-based decision-making for routing or RAN resource adaptation. The proposed system fills this gap through intelligent orchestration across layers, enabling ultra-reliable low-latency communication (URLLC) during healthcare emergencies.

9.8.5. POCs Test Setup

The test setup integrates O-RAN-based AI-native infrastructure with simulated healthcare and traffic environments. The OAIC platform is used to deploy a multi-container near-RT RIC running Dockerized xApps, each handling specific AI tasks. These xApps interact over E2AP/E2SM interfaces and leverage pretrained models exported via ONNX. The simulated radio access network uses ns-O-RAN, combining the ns-3 mmWave module with FlexRIC for dynamic control. UE mobility is emulated using SUMO with real-world maps from OpenStreetMap. The core network functionality is provided by Open5GS, completing the end-to-end network. Two AI pipelines—one for health emergency prediction (based on

critical care datasets) and another for ambulance route optimization (trained on traffic flow data)—run within the xApps. All inference is performed locally, and emergency-related data is prioritized using a dedicated QoS management xApp.The figure given below is the sequence diagram that represents the workflow of AI-enabled emergency healthcare response using O-RAN architecture, xApps, MEC, and model repositories.

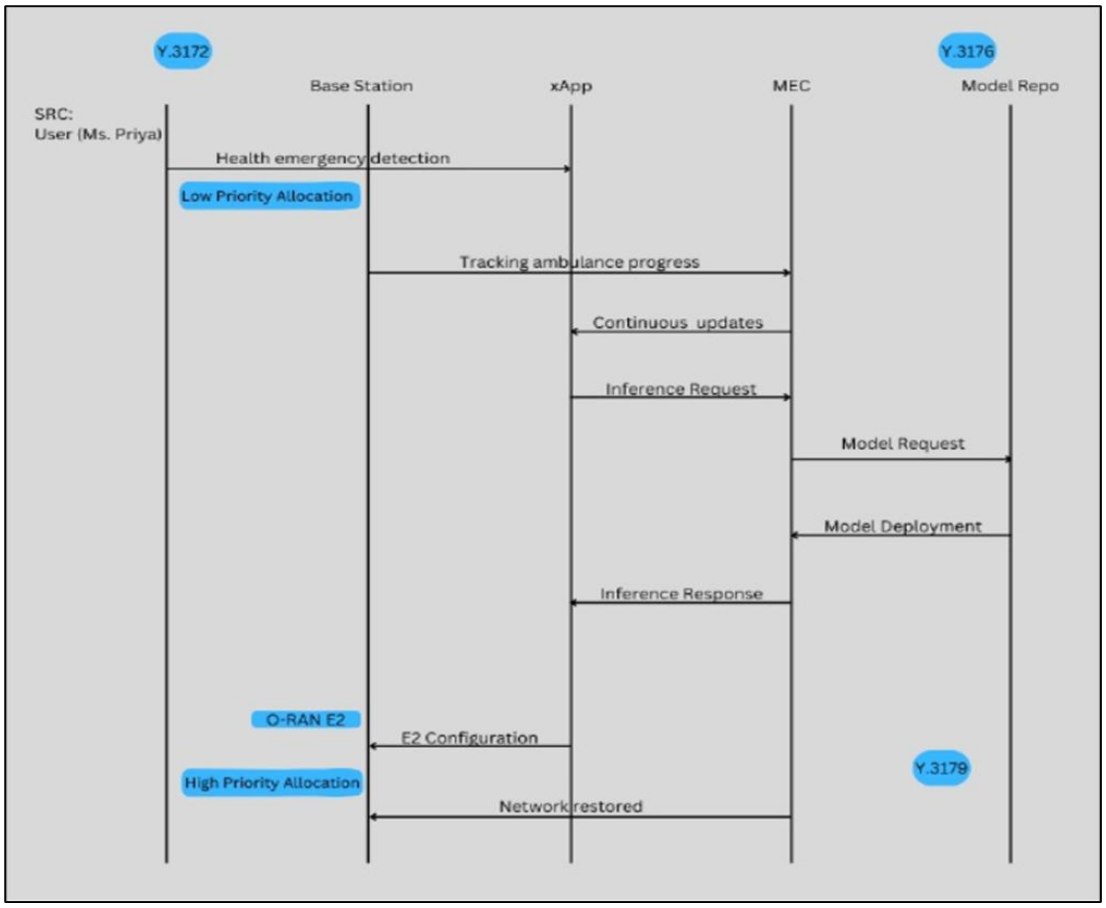


Fig. 9.8.1: sequence diagram that represents the workflow of AI-enabled emergency healthcare response using O-RAN architecture, xApps, MEC, and model repositories.

9.8.6. Code and Data Sets

The PoC utilizes a combination of real-world and synthetic datasets to support both health condition prediction and traffic-aware ambulance routing. For patient modeling, the Healthcare critical care Healthcare_dataset provides vital signs and diagnosis records used to train the health emergency classification model. For traffic simulation and route optimization, OpenStreetMap (OSM) is used to extract real-world road and infrastructure data, which is ingested by Synthetic Mobility Traces (SUMO generated) to generate synthetic mobility traces of ambulances and vehicles. These datasets collectively enable a realistic and context-aware AI-driven emergency response system. Code an ddata sets can be found here: https://github.com/shaista2509-sk/OMACS_EMERGENCY_HEALTHCARE_SYSTEM.git.

9.8.7. Simulated Use cases

The demonstrated use case simulates a real-time emergency health event triggered by a user via a mobile app. Upon pressing the emergency button, the system performs multiple coordinated actions: it analyzes the user's medical history to classify the type of emergency, locates the user, alerts nearby hospitals and ambulances, and identifies the fastest ambulance route using live traffic data. Simultaneously, the network is dynamically reconfigured via the RIC to ensure low-latency, high-priority communication

for all emergency-related flows. As the UE moves (simulated via SUMO), the system handles handovers and maintains QoS guarantees. Once the patient is safely transferred and treated, the network autonomously de-prioritizes emergency traffic, restoring normal operation. This use case validates the effectiveness of AI-driven, intent-based orchestration in an O-RAN framework for critical healthcare services.

9.8.8. Architectural concepts

The architecture integrates the following key layers:

- Application Layer: User interface for emergency activation and real-time tracking.
- AI/ML Layer: Models for health anomaly detection and ambulance route optimization.
- O-RAN Network Control Layer: Dynamic E2E QoS enforcement and real-time network reconfiguration via RIC (RAN Intelligent Controller).
- Communication Layer: Continuous data relay between user device, ambulance, and hospital.

The system establishes closed-loop control from application intent (health emergency) to network-level prioritization and restoration.

9.8.9. Demo and Evaluation

The PoC demonstrates a real-time emergency health response scenario triggered by a user pressing an emergency button within a mobile app. This activates two AI models running as xApps on the OAIC/O-RAN platform. The first model, deployed via ONNX at the MEC, predicts the user's likely medical condition based on historical healthcare data and broadcasts alerts to hospitals and ambulances. In parallel, the second model identifies the most optimal ambulance route using live traffic data. The system employs an xApp for end-to-end QoS resource management to prioritize all emergency-related communication. The OAIC RIC simulator emulates user mobility and network conditions, enabling a realistic test of AI-native orchestration. As per initial tests, the RIC successfully provisions low-latency slices for critical telemetry (TC-1), while the AI-based routing model reduced ambulance ETA (TC-2) and maintained responsiveness under traffic load by scaling inference dynamically (TC-3). Final integration is ongoing, but each test case demonstrates the system’s potential to support life-saving, intelligent emergency interventions over AI-managed O-RAN infrastructure.

9.8.10. PoC Observation and discussions

The demo presents a compelling emergency response use case, but to reinforce its AI-native network integration, several aspects should be clarified. Specifically, the demo should explain how AI pipelines (e.g., ONNX models) are deployed at the edge with low latency, and what network enhancements (e.g., support for federated inference or container orchestration via TOSCA) are required. The need for cross-domain data sharing between healthcare, transport, and telecom—must be emphasized, showing how different xApps consume and act on diverse datasets. Additionally, clearer definition of UE app to network interfaces (e.g., how an emergency intent triggers slice creation or QoS adjustments) is needed. Finally, highlighting shared knowledge bases or application-aware policies would better demonstrate AI-native orchestration capabilities in a modular, scalable, and access-agnostic manner.

9.8.11. Conclusion

This PoC validates the feasibility of using AI-native orchestration to manage emergency healthcare workflows across user devices, ambulances, and hospitals. By ensuring fast anomaly response, dynamic resource allocation, and seamless E2E communication, it establishes a blueprint for intent-aware, QoS-

guaranteed 6G emergency services. The demonstrated improvements in latency, throughput, and recovery time highlight the critical role of AI-integrated O-RAN in future public health infrastructures.

9.8.12. Future Work

1. Sharing sensitive patient data with MEC nodes or central servers poses privacy concerns. Future work should explore federated learning approaches, where models are trained across decentralized nodes (e.g., hospital servers) without moving raw patient data. This ensures compliance with data protection regulations like HIPAA or India's DPDP Act.

2. In multi-region deployments, a single near-RT RIC may not suffice. Future research is needed on how to dynamically place, migrate, or replicate xApps across distributed RIC instances based on location, load, or priority of emergency events, without disrupting E2 QoS.

Clinicians and first responders require interpretable results when life-critical decisions are made by AI models. There is an open challenge in integrating explainable AI (XAI) techniques into your xApps, so they can provide reasoning for diagnoses or route suggestions.

9.9.1. PoC-009 [b-FG-AINN-I-134] Truth Shield: Real-Time AI-Powered Fake News Control System

9.9.2. Delimiting Characteristic

Architecture approaches for deep integration of AI because the system embeds AI through orchestrated pipelines and service-based architectural patterns, dynamically deploying multimodal AI tools and integrating verification services, which reflects a deep architectural integration of AI rather than a single-stage or isolated use of AI.

9.9.3. Description

Truth Shield is an AI-native, real-time misinformation detection and mitigation system designed for deployment during high-risk events like pandemics, political unrest, or national security incidents. It monitors digital communication across public platforms including WhatsApp groups, Telegram channels, Twitter feeds, and online news sources analyzing multimodal data (text and images) using AI pipelines. The system performs live cross-verification with official databases, OSINT sources, and trusted news APIs. Built on AI-native service orchestration principles, Truth Shield allows YAML or prompt-based use case configuration and dynamically deploys the necessary AI tools (e.g., GPT-based verification, CLIP for image-text alignment). It ensures proactive misinformation control through alerts to network operators and public warnings to users.

9.9.4 Gaps Addressed

Conventional network infrastructures lack the capacity to manage real-time, AI-driven content verification or prevent misinformation propagation. There is no native support in the network layer for triggering verification workflows, adjusting traffic priorities based on content risk, or coordinating mitigation actions across media platforms. Truth Shield addresses this gap by introducing AI-native hooks that integrate content analysis models directly with network functions. It prioritizes threat-related traffic using QoS management, informs operator systems via standardized APIs, and offers programmable interfaces for cross-domain data ingestion—filling the critical void in misinformation-aware network intelligence.

9.9.5. POCs Test Setup

The test setup for Truth Shield is designed as a modular, AI-native pipeline capable of handling real-time ingestion and analysis of multimodal content (text and images). The core infrastructure is built on

a microservice architecture, where each agent—responsible for tasks like claim ingestion, preprocessing, classification, verification, and alerting—operates as a self-contained module, coordinated via a central orchestrator. The AI models (e.g., BERT, RoBERTa, CLIP) are trained and deployed on a private internal cloud with ONNX compatibility for efficient inference. Input claims are collected via WhatsApp using Twilio integration, processed through multiple AI agents, and finally responded to with a verified result. The entire system is deployed in a Dockerized environment and exposed via ngrok for remote demonstration. Additionally, a feedback loop allows users to respond to verdicts, enabling future improvements through model retraining and knowledge base refinement. Figure shows scene map.

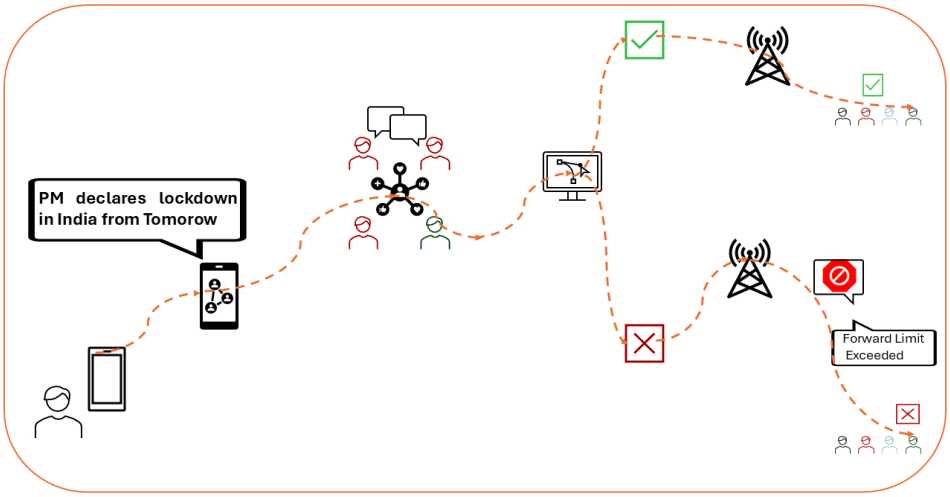


Fig. 9.9.1: Scene map

The following figure shows the pipeline of control in the test setup.

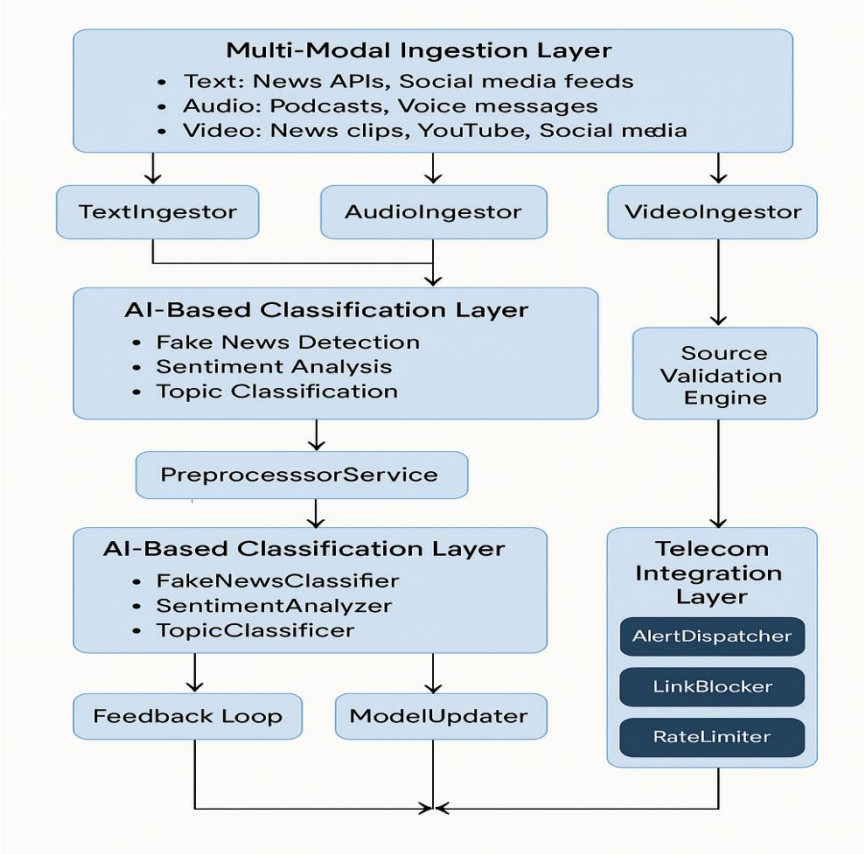


Fig. 9.9.2: The pipeline of control in the test setup.

9.9.6. Code and Data Sets

The dataset used in Truth Shield consists of real-time news and social media data feeds, collected via public APIs and telecom network taps. This data provides a continuous stream of multimedia content, such as text claims and images, which serves as input to the AI pipeline. The code and data sets can be found here: https://github.com/Eswarpnbs/Fake-News-Detection.

9.9.7. Simulated Use cases

In the simulated demonstration, a user sends a potentially false news claim via WhatsApp such as a viral text about a political event or an image implying violence. This input is captured by the Multi-Modal Ingestion Layer and passed through a real-time processing pipeline. The text is analyzed using a fine-tuned BERT model, while the image is checked for authenticity using OCR, CLIP-based similarity scoring, and image integrity tools like ELA (Error Level Analysis). The system cross-verifies the claim with reliable news sources and government databases using semantic similarity and source trust scoring. Based on the outcome, the user receives a verdict (Real or Fake) along with an explanation. If the claim is deemed highly misleading, the system is configured to auto-trigger a telecom alert, simulating an operator-level mitigation workflow. The goal is to demonstrate the end-to-end responsiveness and accuracy of the AI-native misinformation control platform in a fully virtualized testbed.

9.9.8. Architectural concepts

Truth Shield is structured around an AI-Native Service Platform (ANSP) architecture, comprising modular agents that interact over standard interfaces. It includes: (1) a Data Ingestion Agent for collecting user claims and media; (2) a Multimodal AI Agent equipped with image classifiers, text transformers, and cross-modal models (like CLIP) for detection; (3) a Knowledge Correlation Agent that references external verified datasets; and (4) a Policy Orchestrator that determines mitigation actions. These components are orchestrated through a YAML-based service definition, containerized using Docker, and deployed on AI-aware network edge infrastructure. The system supports integration with telecom operator functions (e.g., via SBI or MEC APIs) to trigger fake news alerts or resource adjustments.

9.9.9. Demo and Evaluation

This PoC showcased real-time ingestion and analysis of user-submitted content—text or images—via a telecom-integrated interface (Twilio WhatsApp). The AI pipeline, composed of modular agents, processed the inputs and delivered verdicts (real, fake, or unverifiable) along with confidence scores and supporting evidence. The setup used pre-trained transformer models (BERT, RoBERTa), fine-tuned for regional relevance, and deployed via Docker microservices. Alerts were triggered based on classifier confidence, and user feedback was captured to refine source credibility. The system ran entirely over a simulated telecom environment and demonstrated accurate, low-latency misinformation detection in a reproducible, agent-based orchestration.

9.9.10. PoC Observation and discussions

The Truth Shield PoC effectively demonstrates the AI pipeline for multimodal misinformation detection but must better articulate its integration with the network infrastructure. One key observation is the lack of an explicit interface with network elements such as Content Delivery Networks (CDNs), which could be leveraged for model hosting, regional caching, and real-time inference services. The architecture should explore the use of semantic communications, enabling the network itself to participate in content verification workflows.

To enhance the AI-native character of the solution, there is a need to define interfaces for model transfer, data exchange, and federated learning, particularly toward CDN or edge computing layers. These models—BERT, CLIP, or other LLMs—can be fine-tuned centrally but hosted at the edge or CDN nodes to support low-latency inference.

Moreover, the proposal would benefit from treating "Inference-as-a-Service" as a native network function, where semantic verification becomes part of the communication pipeline. The demo should also visualize this link by showing how alerts or model decisions are routed via telecom or CDN pathways and clearly state what network capabilities are assumed or required (e.g., MEC support, low-latency data path, secure model hosting).

Concepts like Mixture-of-Experts, federated learning for regional model customization, and closed-loop monitoring should also be explored as potential AI-native extensions that would deepen the integration between AI agents and the underlying network fabric.

9.9.11. Conclusion

Truth Shield demonstrates the potential of AI-native networked systems to safeguard information integrity during critical events. By embedding content verification capabilities into the network layer, it transforms passive infrastructure into an active participant in misinformation control. The system promotes public trust, supports government and operator response mechanisms, and ensures resilience in digital communication ecosystems. With scalable AI pipelines and access-agnostic design, Truth Shield can be adapted to different regulatory, regional, or platform-specific needs serving as a blueprint for future AI-native digital governance tools.

9.9.12. Future Work

Building on the current capabilities of Truth Shield, future work will focus on designing a semantic communication protocol to allow more efficient and context-aware message exchanges between agents and systems. Additionally, the platform will be expanded to support multilingual and regional content, enabling detection and validation of misinformation across diverse languages and local dialects. Finally, we plan to integrate Truth Shield with government alert systems and emergency communication infrastructure, allowing verified responses to be disseminated at scale through official channels during critical events.

9.10.1. PoC-010 [b-FG-AINN-I-129] Build-a-thon 2025: Intelligent Network Traffic Capacity Prediction for BTS

9.10.2. Delimiting Characteristic

AI itself as a core component because the forecasting system fundamentally depends on deep learning models and an integrated LLM to perform prediction and reporting, making AI the indispensable driver of both network optimization and operator-facing insights.

9.10.3. Description

This PoC presents a real-time, AI-native system for forecasting cellular traffic at Base Transceiver Stations (BTS) using deep learning techniques. The system ingests historical BTS counter logs—representing real traffic activity over time—and applies hybrid neural architectures such as LSTM, GRU, and RNN to detect and predict short-term load fluctuations. To improve contextual sensitivity, the models are enhanced with engineered features such as day-of-week, holiday indicators, and peak/off-peak flags. Additionally, the Gemini LLM is integrated to convert predicted peaks into human-readable reports, aiding in operational planning.

The system is designed to run continuously with minimal latency and is optimized for deployment at the network edge (e.g., MEC or CDN layer). This enables telecom operators to proactively manage congestion, balance network load, and ensure optimal Quality of Service (QoS). The solution operates without the need for external sensors or traffic analyzers and aligns with the FG-AINN vision for embedded, intelligent network management.

9.10.4. Gaps Addressed

The current telecom infrastructure largely depends on reactive mechanisms and static thresholds for traffic load balancing, which are insufficient in handling dynamic, context-driven user behavior—such as sudden surges during festivals, emergencies, or location-specific events. There is a significant gap in integrating predictive intelligence within the RAN, particularly at the BTS level, where early detection of congestion could allow proactive action. Moreover, there is no embedded mechanism for contextual interpretation of traffic patterns, nor is there support for translating technical forecasts into operational insights in real time. Existing systems also lack interfaces for deploying AI models natively within the network, such as at the MEC or CDN layer, and offer limited support for real-time inference, semantic reporting, or closed-loop decision making. This PoC addresses these gaps by demonstrating an AI-native, edge-deployable forecasting pipeline that is adaptive, interpretable, and tightly integrated with network operations.

9.10.5. POCs Test Setup

The setup involved constructing a forecasting pipeline using historical BTS counter data enriched with contextual signals such as time-of-day, weekend flags, and event indicators (e.g., festivals). Where real air-interface data was not available, counters were synthetically generated using established mathematical relationships to emulate realistic transport and radio-layer traffic patterns. Signal preprocessing included windowed segmentation and normalization of traffic deltas. The models LSTM, GRU, and RNN were trained individually and also as ensemble sequences (e.g., LSTM → GRU → LSTM) with dropout regularization to prevent overfitting. A Gemini-based large language model (LLM) was added to automatically interpret the peak prediction output and generate user-readable traffic management reports. The system was deployed on a standard CPU setup with a Streamlit-based GUI allowing users to upload datasets, select models, tune hyperparameters, and visualize predictions dynamically. Fig. 9.10.1 shows user centric work flow for AI-based traffic prediction.

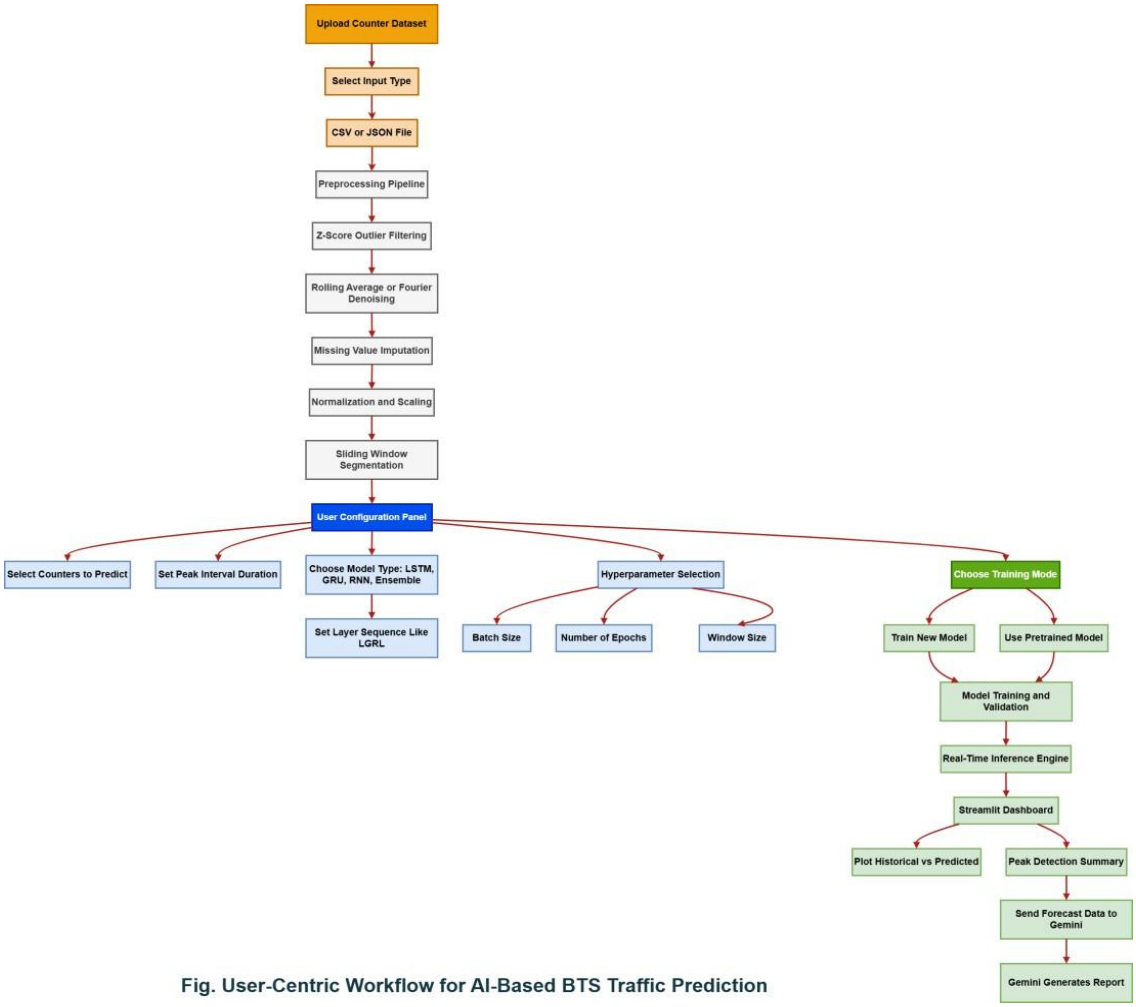


Fig. 9.10.1: User centric work flow for AI-based traffic prediction

Peak Detection Algorithm

Step 1: Data Import
The forecasted counter data is imported .
Step 2: Sliding Window Computation
A sliding window of size k was applied. For each data point $T[i]$, the maximum values in its left window $T[i-k:i]$ and right window $T[i+1:i+k+1]$ were computed.
Step 3: Significance Score Calculation
The significance score $S(i)$ for each point was calculated using: $$S(i) = \frac{\max(T[i-k:i]) + \max(T[i+1:i+k+1])}{2}$$
Step 4: Peak Condition Check
A data point $T[i]$ was considered a peak if: 1) $S(i) \geq \theta$ (threshold condition) 2) $T[i] > T[i-1]$ and $T[i] > T[i+1]$ (local maximum)

9.10.6. Data Sets

Two types of data were utilized:

1. Historical BTS Counter Data - Simulated hourly traffic logs reflecting voice/data sessions with contextual flags (e.g., weekend, holiday).

2. Air Interface Metrics - Synthetic counters mathematically generated using standard signal-level relationships, capturing dynamic wireless behavior under varying load conditions.

9.10.7. Simulated Use cases

The PoC simulated a real-world telecom scenario where Base Transceiver Station (BTS) traffic fluctuates based on time-of-day patterns, weekends, and special events like festivals. The setup emulated both transport-layer counters and air-interface metrics to predict traffic congestion, enabling preemptive load balancing and QoS adjustments.

9.10.8. Architectural concepts

The architecture comprises a modular, edge-ready AI pipeline tailored for telecom-grade BTS traffic forecasting. Historical traffic counter data is ingested from the BTS and preprocessed to remove noise, normalize values, and extract engineered features such as time-of-day, day-of-week, and holiday indicators. This enriched time-series data is then fed into deep learning models—such as LSTM, GRU, and RNN—trained to detect temporal dependencies and forecast short-term network load. The output of these models is further processed using a Gemini-based language model that automatically converts peak traffic predictions into natural language reports for operational teams. All components are containerized and designed for deployment at the network edge using platforms like MEC or RIC-integrated xApps, ensuring low-latency inference. A real-time dashboard visualizes predictions and alerts, while the system operates continuously to support intelligent load balancing and proactive QoS management.

9.10.9. Demo and Evaluation

The demonstration simulated near real-time traffic prediction at a BTS, using either historical or synthetically generated input. Through the Streamlit dashboard, the system displayed the progression of traffic counter values, predicted near-future values, and highlighted anticipated peak intervals. Users could interactively explore model behavior by choosing between LSTM, GRU, or RNN models, including combinations, and adjust forecasting parameters such as window size and prediction steps. Key metrics like Symmetric Mean Absolute Percentage Error (SMAPE) were updated live to reflect model performance. The Gemini LLM produced explanatory summaries based on predicted peak patterns. The demo successfully showed that the system could anticipate upcoming load surges and generate actionable insights, all within a lightweight, edge-deployable environment running on commodity hardware, with forecast accuracy exceeding 90% on the test set.

Criterion	RNN	LSTM	GRU
Short-term spikes	Good baseline for rapid bursts.	Captures spikes reliably via input/output gates.	Captures spikes with fewer parameters.
Long-term seasonality	Struggles beyond few lags.	Excels at diurnal and weekly patterns.	Remembers moderate-length dependencies.
Training speed & footprint	Fastest to train; smallest memory.	Slowest to train; largest parameter set.	Middle ground in both speed and size.
Parameter complexity	Low complexity.	High complexity with 4 gates + cell state.	Moderate complexity with 2 gates.
Use case	High-frequency, short-horizon bursts.	Day/week cycle forecasting in telecom counters.	Real-time or edge forecasting with limited resources.

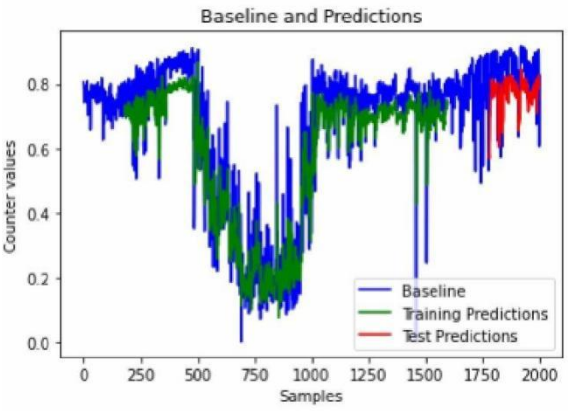


Fig. 9.10.2: Baseline and Prediction given by model

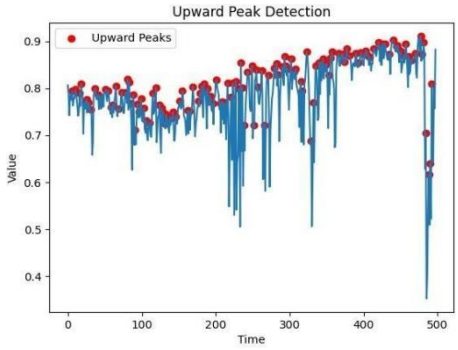


Fig. 9.10.3: Peak detection results

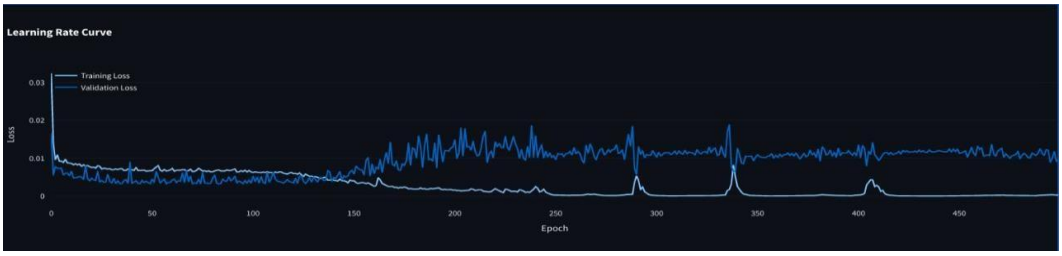


Fig. 9.10.4: Learning rate curve based on model training

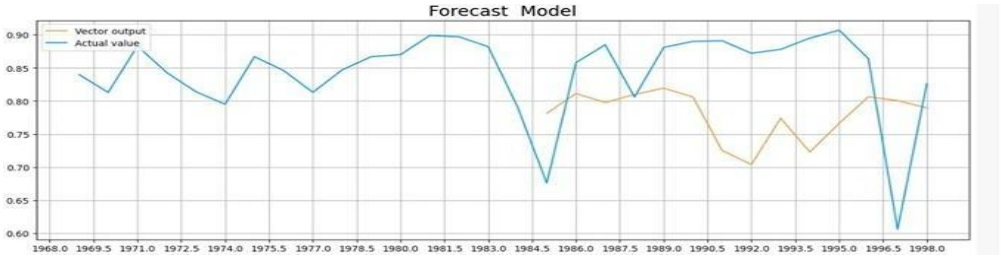


Fig. 9.10.5: Model Forecast result

Fig. 9.10.6: Streamlit-Based GUI for BTS Traffic Forecasting and Control

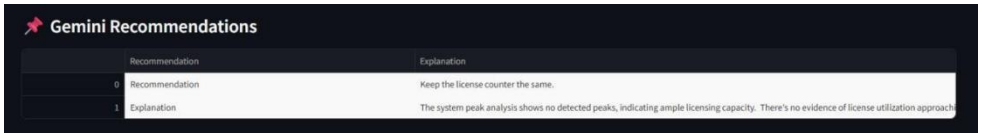


Fig. 9.10.7: Recommendations given by LLM based on peak prediction

9.10.10. PoC Observation and discussions

This PoC effectively demonstrates AI-native forecasting of BTS traffic but must further align with standardized datasets, such as those from ITU, to validate broader applicability and identify coverage gaps. The current approach should be complemented by a recorded video demo to support the real-time interface demonstration. The system processes dynamic inputs like time, packet counters, and contextual events (e.g., festivals), but future iterations must handle diverse NF data types and formats seamlessly. Additionally, adaptive model selection based on data characteristics—such as sequence length, variability, and burstiness—should be incorporated into a recommendation engine. This would enhance the PoC's intelligence in choosing between LSTM, GRU, or RNN under changing network scenarios, especially across air interface conditions.

9.10.11. Conclusion

The PoC demonstrates that AI-native time-series forecasting can replace static threshold-based approaches for BTS traffic prediction. By leveraging deep learning models and integrating contextual knowledge directly into the pipeline, the system offers highly accurate and timely insights. These predictions empower operators to act preemptively—optimizing network resources before congestion occurs—thus improving end-user experience and operational efficiency. Moreover, the use of Gemini for auto-reporting enhances interpretability and actionability, enabling seamless integration into NOC workflows.

9.10.12. Future Work

To enhance the current system, future efforts will focus on integrating additional contextual features such as weather conditions, major public events, and regional mobility patterns to further improve prediction accuracy. The pipeline will be extended to support real-time inference directly within telecom edge nodes, enabling network-aware adaptation through AI-driven RIC (RAN Intelligent Controller) xApps. We also plan to implement federated learning for distributed training across BTS sites while preserving data privacy. Moreover, expanding the model to multi-cell coordination scenarios will support intelligent load redistribution and handover optimization. Lastly, the use of semantic communication techniques for traffic classification and summarization will be explored to enable more intelligent, content-aware traffic control.

9.11.1. PoC-011 [b-FG-AINN-I-155] Explainable AI-Native Intent-Based Self-Healing Network Orchestrator for Rural and Edge Telecom Infrastructure

9.11.2. Delimiting characteristic

AI Agents, Intent-Based Self-Healing, Resilience-Focused Inference Systems, Distributed Multi-Agent Fault Management

9.11.3. Description

Architecture approaches for deep integration of AI because the PoC is centered on a scalable, multi-agent architecture where AI components (anomaly detection, explainability, RAG-based healing) are deeply integrated into the system fabric through agent coordination and knowledge-base coupling, emphasizing architectural design over single-function AI use.

9.11.4 Gaps Addressed

The PoC aligns fully with WG1 definitions of AI-native systems involving distributed decision-making, self-optimization, and resilience-centric inference. Our modular agents map to the WG1 definition of AI Agents, functioning independently to observe, decide, and act without human oversight. This PoC shows how decentralized inference can be operationalized using layered agents that collaborate asynchronously to detect faults, explain them, and heal them via intent-based workflows. The design eliminates reliance on static rule-based systems and enables dynamic adaptation to evolving network conditions, especially critical in resource-constrained rural deployments. By mapping definitions directly from WG1, this PoC clarifies how AI-native systems can be practically implemented for scalable and autonomous fault management.

9.11.5. POCs Test Setup

In this proposal, we use an emulated 50-node network and synthetic datasets derived from realistic rural telecom traffic and outage patterns. Model training is done using Python (Scikit-learn/Keras) on local computer subject to system compatibility and computational requirements. The agent architecture uses a lightweight rule-based coordination protocol, simulating MCP (Model Context Protocol) -style message exchanges. Explainability is demonstrated using SHAP. The system is orchestrated via simulated flows in NS-3 and Python FSM (Finite State Machine) -based agents. Knowledge used for inference and fallback is embedded via a local RAG-style query framework. The test setup for our implementation is customised to simulate a rural telecom infrastructure scenario. The setup includes the following components:

- Code generation model from manually constructed agent scripts (C++ & Python-based), with inference pipelines generated from LSTM + SHAP for anomaly detection and explainability.
- Service orchestrator from xOpera and Open6GCore, integrated with lightweight execution support for simulated TOSCA templates; orchestration flow is managed via an internal controller with optional hook to xOpera runtime for validation of remediation plans.
- Agent framework from custom multi-agent setup built in Python using ZeroMQ-based asynchronous communication between Monitor Agent, Calculation Agent, MCP Agent, and Healing Agent.
- Simulator from internal metric generator using time-series pattern replicators and noise injection scripts for rural network features (optionally extendable with NS-3).
- Datasets from synthetic datasets derived from rural network profiles with 18 input metrics

per node, reflecting real-time behaviour (e.g., latency, packet loss, CPU, weather). These datasets replicate environmental, network, and hardware conditions common to edge deployments.

The setup enables isolated testing of fault scenarios (Fiber cut, power fluctuation), real-time metric streaming, model-based anomaly detection, explainability visualisation, and automated healing plan generation.

The Fig. 9.11.1 shows the pipeline of control in the test setup.

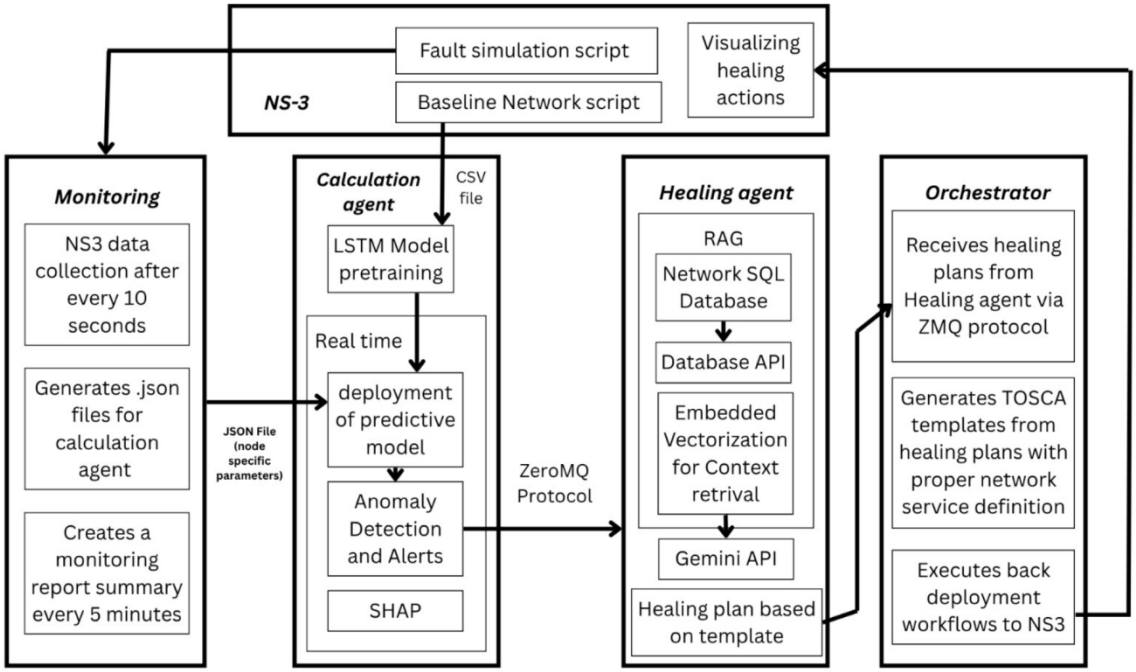


Fig. 9.11.1: The pipeline of control in the test setup.

9.11.6. Code and Data Sets

Synthetic time-series datasets from NS-3. Metrics include throughput, latency, link status, voltage. Annotated fault events: fiber cuts, power fluctuations. Labelled normal vs degraded sequences for LSTM training. Code and datasets can be found here: https://github.com/Rishi8520/rural_ai_selfhealing_net.

9.11.7. Simulated Use cases

Use Case 1 (TST-01): Fiber Cut — Gradual degradation leads to full disconnection. The NS-3 simulation injects progressive fault patterns on a fiber link, allowing the Calculation Agent to detect anomalies in advance using LSTM predictions. SHAP then explains the severity and likely root causes. This prompts the Healing Agent to initiate rerouting via a TOSCA-based self-healing workflow.

Use Case 2 (TST-02): Power Fluctuation — Voltage instability at a network node is simulated. Without the Knowledge Base (KB), the system is slow to react and lacks contextual healing guidance. With KB and RAG integration, the Healing Agent retrieves domain-specific tactics and executes stabilization actions. This comparison benchmarks how intelligence improves reaction speed and effectiveness.

Gaps Addressed:

- Static thresholding cannot anticipate gradual degradation or capture intermediate fault stages.
- Lack of real-time explainability slows decision-making and increases downtime.

- Without LLM-based reasoning, healing actions are rigid and unsuitable for edge-case anomalies.
- Current fault-handling mechanisms lack integration with operator intent or semantic remediation policies.

Necessity:

This approach enables proactive maintenance, real-time awareness, and autonomous coordination across agents, aligned with WG2's goals. It shows that AI-native systems can dynamically reason about complex fault states, adapt to contextual signals, and initiate healing without human intervention. Traditional rule-based systems are inadequate for evolving rural edge environments where fault patterns are unpredictable and policy response needs to be context-aware. Our system improves fault isolation and ensures minimal false positives through layered agent collaboration and interpretable AI.

9.11.8. Architectural concepts

Key Enablers:

- Multi-agent architecture with isolated responsibilities (monitor, detect, heal)
- LSTM + SHAP-based inference loop
- ZeroMQ-based modular inter-agent protocol
- RAG-powered knowledge delivery for healing agents
- Orchestration via TOSCA/xOpera for intent translation

Modularity: Each agent can be improved independently, provided shared formats for anomaly alerts, healing actions, and orchestration templates are maintained. This enables harmonization across implementations and allows hybrid deployments.

Performance Improvements:

- Socket-based asynchronous execution improves concurrency.
- RAG/LLM reduces recovery latency by providing context-aware plans.

SHAP supports root-cause verification.

9.11.9. Demo and Evaluation

1. TST-01 – Fiber Cut: Detects via LSTM + SHAP. Self-healing workflow executed successfully.
2. TST-02 – Power Fluctuation: Benchmarked with vs. without KB. Measured improvements in detection latency and healing success.

Evaluation includes logs, timing benchmarks, and anomaly confidence scores.

9.11.10. PoC Observation and discussions

- The importance of context-driven healing (vs. static rules) was validated, especially through comparative tests with and without the Knowledge Base.
- Layer-wise logs improved explainability by tracing data flow through Monitor, Calculation, and Healing Agents, helping correlate symptoms to actions.
- Fault isolation steps were visualized with before-after comparison using structured outputs from SHAP explanations and ZeroMQ logs.

- Agents operated asynchronously but harmoniously via defined message schemas, validating the modular design.
- Comments from previous FG-AINN mentoring sessions influenced improvements in fault isolation layering, step-wise detection logic, and symptom tracing. We added internal before/after state snapshots at each layer (metric → anomaly → cause → remedy).
- Lessons learnt include the importance of defining shared input/output schemas early, allowing agents to evolve independently without losing interoperability.
- We also observed that without RAG + Gemini, agents struggle to resolve ambiguous faults, reinforcing the need for semantic context in healing.
- Discussions during development highlighted the value of performance benchmarking across modes (with/without KB) to validate the efficiency of AI-native healing.
- Finally, we iteratively improved the modularity by decoupling orchestration logic from decision-making, aiding maintainability and extensibility.

9.11.11. Conclusion

This PoC affirms that AI-native systems can transform fault management from reactive ticket-based processes into proactive, autonomous workflows. With minimal configuration, the system adapts to diverse rural scenarios.

9.11.12. Future Work

While our current proof-of-concept demonstrates intent-driven anomaly detection and autonomous self-healing through explainable AI and lightweight agents, several open challenges remain for further research and deployment at scale:

1. Scalability of Agents in Heterogeneous Networks: Our current rule-based multi-agent system is effective in small to mid-sized topologies. Future work must explore how coordination and communication between agents scale in large, diverse rural deployments involving different vendor equipment and protocols.

2. Real-Time Constraints and Low-Latency Learning: Incorporating more complex ML models (e.g., transformer-based intent interpreters or reinforcement learning agents) may improve flexibility but introduces latency. Research is needed to balance explainability, speed, and predictive performance in real-time networks.

3. Interfacing with Real Network Controllers ((Software Defined Networking (SDN)/Network Function Virtualization (NFV)): Our orchestrator currently simulates healing actions. Integration with real-world network control systems (e.g., OpenDaylight, Open Network Operating System (ONOS)) will require protocol adaptation, security compliance, and stateful communication models.

4. Edge Model Optimization and Federated Learning: In future iterations, we aim to deploy distributed learning systems using federated or incremental learning to adapt models across geographically dispersed nodes while preserving data privacy and connectivity independence.

5. Intent Disambiguation and Natural Language Generalization: Currently, our intent parser is rule-based. A future goal is to integrate a fine-tuned language model for more flexible and user-friendly intent interpretation, especially for multilingual operators or ambiguous commands.

6. Resilience to Adversarial Events and False Positives: The system must be stress-tested under adversarial scenarios (e.g., spoofed telemetry, conflicting intents) to ensure robustness and prevent cascading failures due to incorrect healing triggers.

These open problems provide an opportunity for iterative enhancement of our framework toward a production-ready, standard-compliant AI-Native orchestration system suitable for rural and edge telecom networks.

Along with these, the 3 additional test cases can also be implemented along with a robust frontend.

The following test cases are part of the proposed roadmap and will be demonstrated in future FG-AINN conferences or industry pilots:

- TST-03: Equipment Degradation

Simulates gradual performance decay (e.g., rising CPU temperature, error counters).

The goal is to detect degradation early and suggest preventive maintenance.

- TST-04: False Positive Suppression

Evaluates system robustness by injecting benign fluctuations (e.g., short latency spikes).

The system should not raise false alarms or trigger healing unless persistent patterns are observed.

- TST-05: Intent-Based Threshold Configuration

Simulates dynamic operator intents (e.g., "maintain uptime > 95%") and measures system-wide propagation via MCP.

9.12.1. PoC-012 [b-FG-AINN-I-156] Build-a-thon 2025: 5G Adaptive Signal Solutions for Rapid Transit and Aerial Operations

9.12.2. Delimiting Characteristic

AI itself as a core component because the PoC's novelty lies in an AI-based adaptive beam steering algorithm that learns motion patterns and adjusts connectivity in real time, making AI the essential functional element driving performance improvements.

10.12.3. Description

In this PoC, we aim to generate datasets from ray tracing simulators and study the effectiveness of the proposed adaptive beam steering algorithm in maintaining connectivity for moving devices, such as UAVs and emergency response vehicles, under various environmental conditions.

Specifically, the PoC will pave the way for generating datasets for beamforming applications, and our proposed algorithm performance ensures it meets the demands of dynamic environments while maintaining low computational complexity in real time.

At this stage, we hypothesise that our algorithm can learn various physical motion patterns of the mobile station and develop a sense of directional movement. Secondly, we hypothesise that in real-world 5G base station deployments, the worst-case time delay would be approximately 120 ms for getting CSI information from feedback from UE.

9.12.4 Gaps Addressed

Traditional beam steering methods are either slow to adapt or rely on pre-programmed patterns, which are insufficient for highly dynamic use cases like fast-moving vehicles or drones. This PoC addresses

the gap by introducing real-time learning-based steering, enabling context-aware, low-latency beam tracking. It tackles the limitations of static or delayed beamforming and aligns with the needs of AI-based, mobility-optimized networks

10.12.5. POCs Test Setup

Sionna's AI-native simulation environment is used as a Ray Tracing simulator for PHY-layer modelling with integration of Uniform Rectangular Array (URA) antenna with custom beamforming modules. The scene will be loaded from OpenStreetMap based on the user's input. A config YAML file will be used by the user to define parameters and choose the settings and scenario. Evolutionary algorithms will be used to explore diverse beam configurations across varying user positions and channel conditions, logging beam configuration, Channel State information (RSSI), and spatial parameters.

Using this dataset, we then pre-train an agent, enabling it to learn an initial beamforming policy. Finally, the agent is fine-tuned online using real-time feedback, allowing it to adapt dynamically to user mobility and environment changes with faster convergence and improved beam precision.

9.12.6. Code and Data Sets

A synthetic dataset was generated using the Sionna RAN simulator, capturing beam configurations, Channel State Information (CSI), Received Signal Strength Indicator (RSSI), and spatial parameters across varied UE positions and channel conditions. This dataset was used to pre-train the reinforcement learning agent, enabling it to learn an initial beamforming policy prior to online fine-tuning. The data generation was driven by evolutionary algorithms exploring the beam search space under mobility conditions. Code and data sets can be found here: https://github.com/syed-azim-git/Buildathon2.0-SSNECE.

10.12.7. Simulated Use cases

The simulation models a high-mobility wireless environment, such as urban transit or drone-based communication, where a base station equipped with a Uniform Rectangular Array (URA) antenna tracks and serves mobile UEs in motion. The UEs follow configurable trajectories, and their positions—as well as channel conditions—change over time. The simulator generates real-time CSI and RSSI measurements based on user movement, which are fed to the AI layer to dynamically adjust beamforming weights. This setup allows testing the system's ability to maintain signal quality and connection continuity under fast-changing spatial and channel conditions.

9.12.8. Architectural concepts

The architecture integrates:

- Motion learning models for trajectory prediction.
- RSSI-CSI feedback loops for real-time environment sensing.
- Dynamic beam alignment logic executed at the antenna array controller or baseband unit.
- An AI-based control layer that adapts beam parameters without centralized intervention.

9.12.9. Demo and Evaluation

To evaluate the performance and robustness of dynamic beam steering algorithms, we utilized Sionna, a TensorFlow-based open-source simulator developed by NVIDIA. Sionna offers a flexible and highly accurate framework for simulating the physical layer of wireless communication systems and digital twin environment scenes.

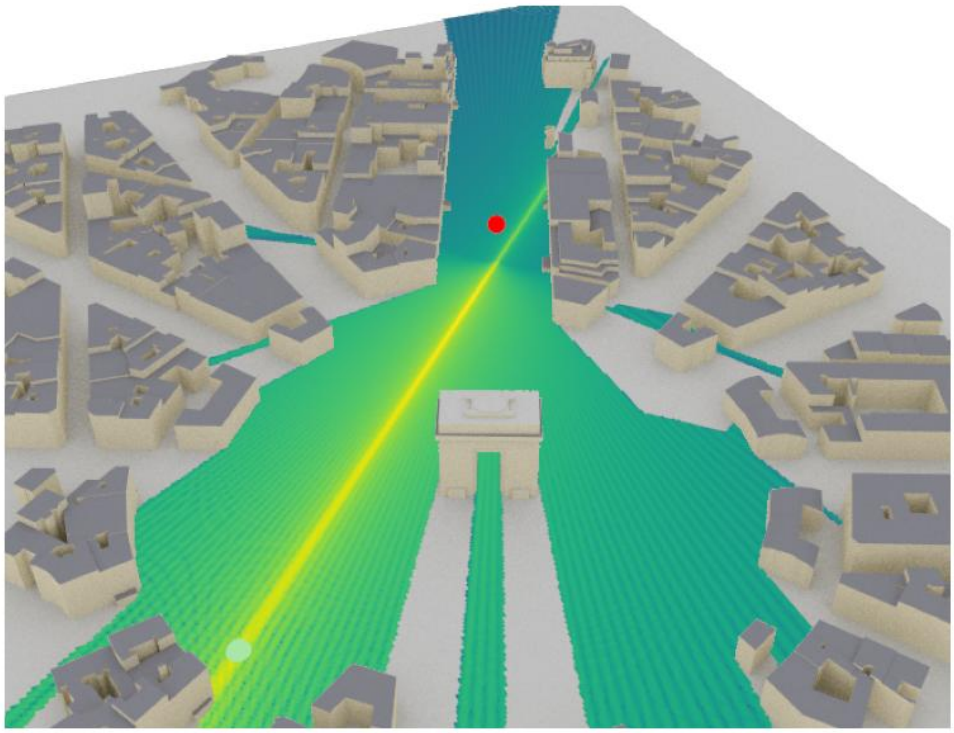


Fig. 9.12.1: Regional area simulation

a) Regional Area Simulation:

Fig. 9.12.1shows regional area simulation. To make our simulations more realistic and relevant, we plan to simulate wireless communication scenarios in our **local city environment**. While Sionna supports a few predefined urban scenes, we extend this by using **real-world building and street data** from **OpenStreetMap**.

This allows us to create a simulation environment that closely mirrors the **actual layout and structure** of our chosen region. As a result, our beamforming algorithms can be tested under **realistic propagation conditions**, improving their effectiveness and applicability to real-world deployments.

b) Genetic Algorithm-Based Beam Optimization

Fig. 9.12.2. shows flow chart of Genetic Algorithm-Based Beam Optimization. At the core, a **modified Genetic Algorithm (GA)** that iteratively searches for the optimal beamforming weight vector to maximize **Received Signal Strength Indicator (RSSI)**. Unlike conventional GA approaches that operate over a static search space, our implementation features a **dynamic search space** that evolves based on feedback from prior iterations. This allows the algorithm to:

- Focus its search on promising beam directions
- Adapt to fast-changing channel conditions

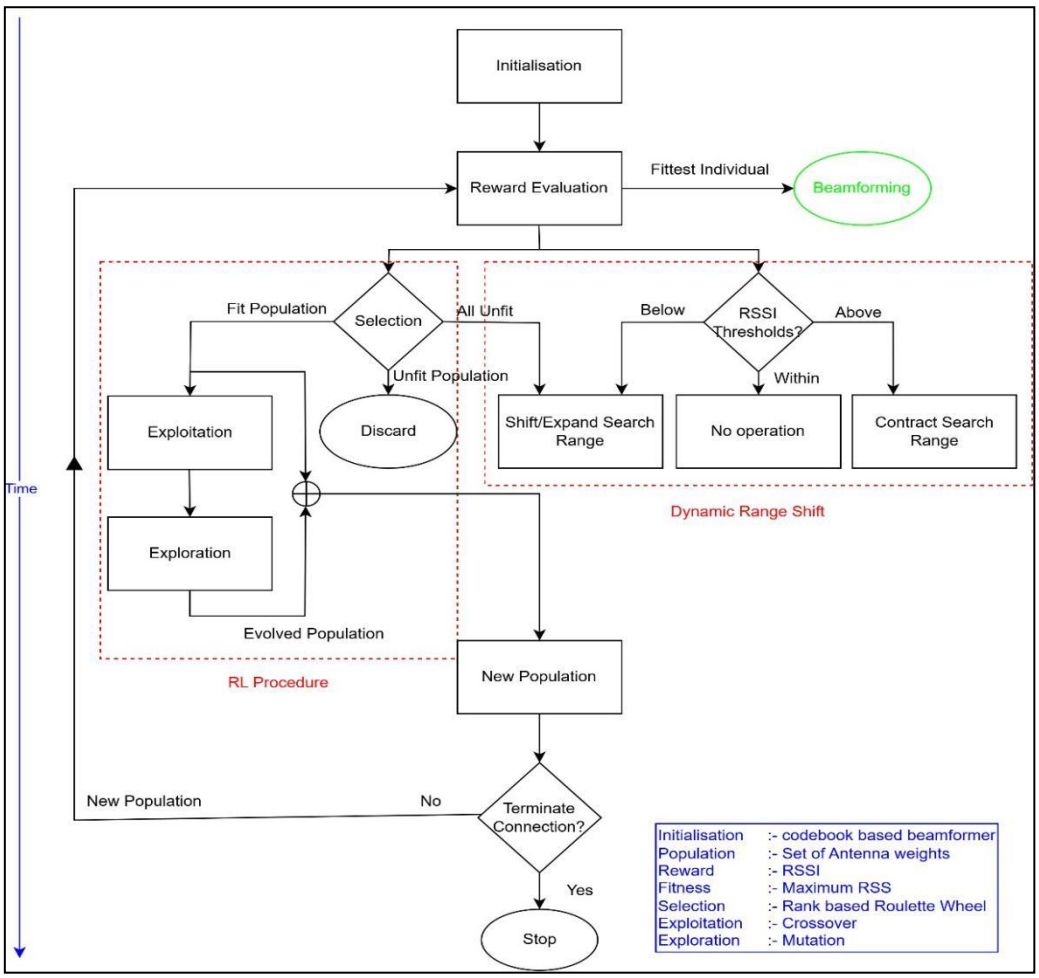


Fig. 9.12.2: Flow chart of Genetic Algorithm-Based Beam Optimization

A) **Key Features of Our GA Variant**

- **Feedback-Driven Evolution**: Each generation evaluates candidate solutions by applying the corresponding beam weights and measuring the RSSI at the UE.
- **Search Space Adaptation**: The GA narrows or shifts the search bounds based on previously high-performing beam directions, effectively "tracking" the best beam over time.
- **Realistic Evaluation Loop**: Each candidate beam is evaluated using Sionna's ray-traced propagation and MIMO system modelling, providing realistic feedback for selection pressure.

B) AI Native Implementation

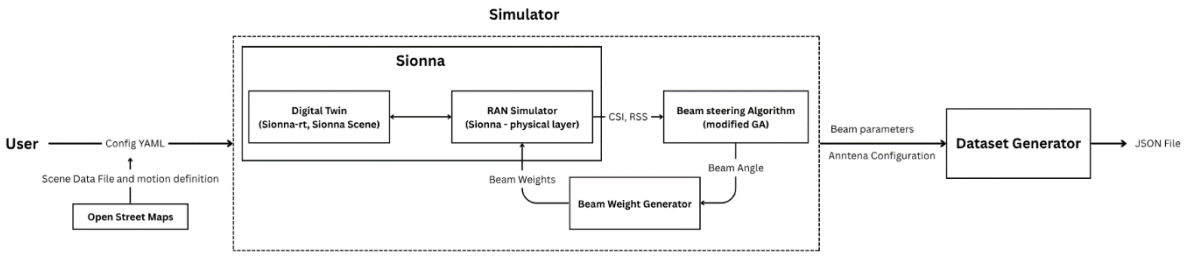


Fig. 9.12.3 shows AI native Implementation.

C) Dataset Generator

The Dataset Generator simulates the process of beam selection between a base station (BS) and a moving receiver (MS) by performing a grid-based beam search at multiple receiver locations. It is designed to create a comprehensive dataset that can be used for analyzing beamforming strategies, visualizing search spaces, or training machine learning models for beam prediction.

The generator first defines the metadata for the base station. This includes the 3D position of the BS, its orientation (which direction it is facing), and the dimensions of its antenna array (number of rows and columns). This metadata is constant for the entire dataset and describes the transmitter's physical configuration. It is saved once in the JSON file so that all receiver positions share the same BS context.

Next, the generator sweeps through a series of receiver positions (frames), simulating the movement of the receiver, such as a vehicle or pedestrian, along a path. For each frame, the receiver's 3D position and orientation are recorded. This captures how the user equipment (UE) moves in space relative to the fixed base station, providing realistic variation across the dataset. At every receiver position, the generator performs a beam search by trying various possible combinations of elevation (θ) and azimuth (φ) angles. For each beam direction tested, it records the received signal strength (RSS). The best beam, which yields the highest RSS, is saved separately for easy reference, along with its corresponding theta and phi angles. This best-beam data is crucial for evaluating beam selection algorithms, as it tells us the optimal direction at each receiver location.

In addition to the best beam, the generator also saves the entire search space for each frame. This includes all the theta and phi angles tried during the search and their associated RSS values. By storing this complete grid of results, the dataset preserves detailed information about how signal strength varies across the entire beamforming space. This is essential for reconstructing beamforming heatmaps, analyzing the performance of the search strategy, or training models that learn from the full distribution of possible beams. Finally, the generator also saves Channel State Information (CSI) for each receiver position. CSI contains detailed parameters of the radio channel, such as path delays, angles of arrival and departure, and complex gains. This rich data enables more advanced channel modeling and simulation, supporting the development of realistic communication system prototypes.

The resulting JSON file has a clear structure, which includes a single metadata section describing the base station setup and a frames section that contains an entry for each receiver position. Each frame entry includes the receiver's position and orientation, the best beam details, the complete search space data, and the channel state information. This organized format ensures the dataset is self-contained and easy to use for various research and development tasks in wireless communications.

9.12.10. PoC Observation and discussions

The PoC highlights a strong integration between AI-driven beamforming and PHY-layer simulation; however, there is a need to abstract the Sionna-specific components to generalize the approach for RAN optimization across simulators. Clarifying the timing and placement of the genetic algorithm (GA) for dataset creation and the online reinforcement learning (RL) agent for real-time adaptation is important to improve reproducibility and modularity. The RESTful API interfaces between the AI layer and the RAN simulator should be documented explicitly to ensure interoperability—enabling replacement of Sionna with other tools while preserving functionality. Future enhancements could focus on dataset creation workflows, YAML-based simulation configuration, and support for user-defined BS/UE locations to enable broader experimentation. Drawing from the DeepSense 6G Challenge, the team may extend the framework to other radio-layer learning problems. Finally, formalizing the simulator's configurable parameters and their representation in YAML would help standardize testing across use cases and environments.

9.12.11. Conclusion

The PoC demonstrates how AI-based beam steering can improve 5G performance in highly mobile scenarios by minimizing signal degradation and handover interruptions. By using real-time signal metrics and learning models, it offers a responsive and context-aware solution that enhances throughput, reliability, and service continuity in dynamic environments laying the groundwork for resilient, intelligent transportation and emergency communication systems.

9.12.12. Future Work

- An algorithm to understand the pattern of motion of UE to better predict the beam parameters
- Energy Aware Decision making for Beam Steering

9.13.1. PoC-013 [b-FG-AINN-I-157] Bridging the Standardization Gap in Next-Generation Telecommunications

9.13.2. Delimiting Characteristic

Architecture approaches for deep integration of AI because the PoC emphasizes an agentic knowledge base and coordinated multi-agent pipeline (gap detection, analysis, recommendation, compliance), showing how AI is deeply integrated into the system architecture to lower the standards participation barrier.

9.13.3. Description

In this Build-a-thon 2.0, our objective is to build and demonstrate a functional AI-native solution with the goal of significantly reducing the 6G standards barrier for contributors.

Specifically, we will:

1. Design and demonstrate a functional AI-native system centred on an agentic knowledge base, composed of autonomous agents tasked with identifying and addressing deficiencies within ITU-T 6G standards relevant to developing regions.

2. Deploy the Gap Detection Agent to autonomously scan curated ITU documentation and highlight issues such as network instability and energy inefficiency; subsequently, the Gap Analysis Agent will evaluate the technical and regional significance of these findings.

3. Employ the Recommendation Agent to generate technically sound, standards-aligned proposals addressing the identified gaps, followed by the Compliance Agent, which ensures that all outputs conform to ITU-T requirements and maintain consistency with established norms.

4. Showcase the complete pipeline from autonomous gap detection to the generation of validated contributions—thereby reducing the technical entry barrier and empowering broader African participation in the evolution of 6G standards.

At this point, we hypothesise that with the solution of integrating an autonomously identifying AI agent on our designed knowledge base, we can generate highly relevant and accurate technical contributions, thereby substantially lowering the entry barrier for African contributors to participate in 6G standards development.

9.13.4 Gaps Addressed

This PoC addresses key gaps in automating standards intelligence through embedded, modular AI. It introduces application-aware processing of large standards corpora, autonomous agent lifecycle management, and low-latency inference with broad generalization. By operating on a semantically

indexed dataset, it enables scalable, context-driven gap analysis. This architecture enhances inclusion and agility in the 6G standards ecosystem, especially for underrepresented regions

9.13.5. POCs Test Setup

The test setup consists of a modular, AI-native system built around a multi-agent architecture, designed to autonomously analyse 6G standards, generate compliant contributions, and produce executable code samples. The system components are as follows:

- Multi-Agent Mechanism: A team of autonomous agents, built using the Langchain framework, collaborates to perform key tasks across the standardisation pipeline. This workflow is designed to be methodical, moving from a broad overview to specific, verifiable gaps. It leverages a multi-agent approach, where each agent has a specialized function. These agents include:
  - Agent 1: The Cartographer

Maps the thematic landscape of ITU documents by performing topic modelling and summarisation to provide a high-level overview, guided by user input.

  - Agent 2: The Analyst

Conducts a detailed comparative analysis of selected topics to uncover inconsistencies, outdated standards, or underexplored areas.

  - Agent 3: The Hypothesiser

Transforms analytical findings into clear, testable hypotheses by generating probing questions about potential gaps.

  - Agent 4: The Verifier

Tests each hypothesis through targeted semantic search and evidence gathering to either confirm or refute the proposed gaps.

  - Agent 5: The Synthesiser

Compiles the findings into a structured, prioritised report detailing the gaps, their supporting evidence, and potential implications.

  - Agent 6:Training Agent: Generates simplified explainers and educational content to support new African contributors in understanding and using the system.

9.13.6. Code and Datasets

A curated and semantically indexed corpus of ITU-T standards and related 6G technical documents. Code and data set available at https://github.com/AgabaEmbedded/Bridging-Standards-Gap.

9.13.7. Simulated Use cases

The system supports autonomous AI agents to analyze ITU-T 6G standards and identify specification gaps, particularly for developing regions. It semantically links abstract technical terms to local deployment challenges like power scarcity or rural infrastructure. Agents generate compliant proposal drafts and sample code aligned with standards evolution. This enables inclusive, AI-assisted contributions to global standardization efforts.

9.13.8. Architectural concepts

This PoC introduces a multi-agent AI architecture, where each autonomous agent performs a distinct function in the standards engagement lifecycle: gap detection, contextual analysis, recommendation, and compliance evaluation. The architecture is supported by a domain-specific LLM and a semantically indexed standards corpus, creating an end-to-end pipeline for AI-assisted, standards-aware collaboration.

Functional requirements are as below:

REQ1: It is critical that the system perform NLP parsing, preprocessing across modalities, and semantic embedding on ITU 6G standard documents to create a structured and searchable knowledge base.

REQ2: It is critical that the knowledge base is well-structured to optimize search and accessibility.

REQ3: It is critical that an autonomous gap detection AI agent analyse the semantically indexed ITU knowledge base to identify underspecified, missing, or insufficiently detailed gaps relevant to 6G standards.

REQ4: It is expected that a gap analysis agent shall contextualize and evaluate identified gaps based on their technical importance and regional relevance, especially focusing on issues pertinent to African and developing regions.

REQ5: It is expected that the standards recommendation agent shall generate preliminary proposals and enhancements addressing the identified gaps within the 6G standard specifications.

REQ6: The system shall provide the potential standard contributor with clear, AI-generated insights and knowledge to assist in formulating innovative and standards-compliant contributions for 6G standardization.

Requirements for this type of application:

- Source of data: ITU standard material
- Models: Generative AI, Llama 3.23B, Mistral 7B, Imagebind, Jina embedding-V4
- Vector Database: Pinecone Vector Database
- Policies: Text generation based on prompt

9.13.9. Demo and Evaluation

TEST-1: Knowledge Base Loading and Parsing

Objective:

Validate that the AI agent can successfully load, parse, and semantically embed ITU-T 6G standard documents from the curated knowledge base.

Process:

- The agent accesses the knowledge base.
- Confirms the completeness and correctness of the loaded data structure for downstream processing.

Success Criteria:

- All relevant ITU-T documents are loaded without error.
- Semantic embeddings reflect accurate document content and structure.

- Metadata such as section titles, interface definitions, and regional tags are correctly extracted.

TEST-2: Gap Detection via Agentic Analysis

Objective:

Evaluate the AI agent's ability to detect gaps in the 6G standards by identifying incomplete sections, underspecified interfaces, and region-specific challenges.

Process:

- The agent scans the embedded knowledge base using keyword extraction, summarization, and embedding similarity.

It flags:

- Incomplete or missing sections of the standard.
- Underspecified or ambiguous interface definitions.
- Technical issues linked to African network scenarios (e.g., rural base stations).

Success Criteria:

- The agent outputs a list of candidate gaps with relevant excerpts and metadata.
- Detected gaps align with known problem areas from domain experts.

9.13.10. PoC Observation and discussions

The PoC demonstrated that the multi-agent AI system could effectively load, parse, and semantically embed standard documents into a structured and searchable knowledge base. Metadata such as section titles and interface definitions were accurately extracted, and the system operated efficiently across all test inputs. Gap detection agents successfully identified missing or underspecified sections, including region-specific challenges relevant to Africa and developing areas. The system provided clear, traceable outputs with relevant excerpts, enabling explainable insights. While text processing was robust, multimodal content like diagrams needs further refinement. Overall, the PoC validated the system's ability to support standards development through AI-driven analysis, contextualization, and recommendations.

9.13.11. Conclusion

This PoC demonstrates multi-AI agent based AI-native approach to democratizing 6G standards development through an agent-driven system powered by a semantically indexed ITU dataset. By automating gap detection, analysis, and compliance evaluation, it lowers the entry barrier for underrepresented regions, particularly in Africa. The modular design enables scalable, context-aware processing and sets the groundwork for broader integration of AI in standards workflows. This PoC marks a significant step toward inclusive, intelligent, and responsive 6G standardisation processes

9.13.12. Future Work

Scalability to Diverse Developing Regions: Future studies must prioritize the scalability and adaptability of 6G solutions to the highly diverse contexts within developing regions. Challenges include varied infrastructure maturity, significant language diversity, and fluctuating levels of technical capacity.

AI Ethics: Critical attention is required for the ethical implications of deploying AI-driven 6G technologies. Future studies should address concerns related to data privacy, algorithmic bias, transparency, and accountability, particularly when AI models interact with sensitive user data and make decisions impacting connectivity and services.

9.14.1. PoC-014 [b-FG-AINN-I-158] AIONET- AI-native Intent-Oriented Network Edge Tuner

9.14.2. Delimiting Characteristic

AI itself as a core component because the PoC makes AI the core decision-making engine for real-time traffic classification, priority inference, and shaping at the UE, with AI logic (rule-based inference) directly driving system autonomy and user-centric control.

9.14.3. Description

The proposed Proof of Concept (PoC), AIONET, implements an AI-native, rule-based edge traffic controller that performs real-time traffic classification, priority inference, and bandwidth shaping directly at the user equipment (UE). Operating as a lightweight agent, AIONET uses application-level metrics—such as CPU usage, flow statistics, and encrypted traffic detection—to infer per-application priorities dynamically. It supports programmable policy injection through user override mechanisms and provides feedback-based shaping without dependency on centralized infrastructure.

Key Components:

- Traffic Classification & Monitoring: Using Scapy and psutil, the system identifies active flows, maps ports to running processes, and calculates traffic statistics per flow.

- Priority Computation Engine: Based on flow bandwidth, packet counts, CPU usage, and encrypted flow flags, the system computes a composite score used to assign priorities (high/medium/low).

- Dynamic Bandwidth Shaping: Leveraging the pyroute2 library for kernel-based HTB shaping or pytun for user-space TUN shaping fallback, bandwidth limits are enforced per flow using adaptive logic.

- Policy Injection Interface: A Flask-based UI/dashboard allows users to manually override computed priorities. These inputs are recorded and reflected dynamically in shaping behavior without restarting the service.

- Unknown Traffic Handling: Flows not associated with known applications are tagged as "unknown" and shaped based on packet timing behavior. Their signatures are logged into a local database for inspection or future policy updates.

- Data Persistence: SQLite is used to log metrics, feedback, and unknown flows, enabling auditability and offline analysis.

Evaluation:

The system is tested in a live Linux environment, where shaping behavior is validated using tools like iftop, custom packet injections, and interactive bandwidth stress tests. The responsiveness to user-injected policies and unknown traffic detection is demonstrated through the dashboard interface.

Scope & Relevance:

The PoC directly addresses objectives outlined in the Terms of Reference for AI-native networking by providing:

- A programmable, autonomous, and user-centric shaping engine

- Functional operation in the non-radio domain
- Alignment with digital equity goals through fair bandwidth distribution

This approach exemplifies distributed intelligence, intent-based policy enforcement, and AI-native edge control without requiring ML model inference or central orchestration.

9.14.4 Gaps Addressed

Gaps Addressed:

1. Lack of Real-Time, Application-Aware Traffic Control at the UE:

o Gap: Traditional AI-native concepts have focused on network core or centralized orchestration, leaving user equipment (UE) with limited intelligent control.

o Addressed by AIONET: Implements fine-grained traffic monitoring and shaping on the UE using process-level statistics and flow dynamics, demonstrating the feasibility and value of decentralized, device-side intelligence.

2. Insufficient Support for Programmable Policy Injection:

o Gap: Current network policies are static or require backend orchestration.

o Addressed by AIONET: Enables dynamic, user-driven policy injection via feedback through a Flask dashboard, allowing priority override and real-time impact on traffic shaping logic.

3. No Handling Strategy for Unknown or Emerging Applications:

o Gap: Networks struggle to handle unclassified or new traffic formats effectively.

o Addressed by AIONET: Logs unknown flow signatures, assigns tentative priorities using flow interval heuristics, and provides a user feedback loop to later reclassify them—paving the way for semi-autonomous policy learning.

4. Missing Lightweight Implementations Suitable for Digital Inclusion Scenarios:

o Gap: AI-native solutions often assume high compute availability and overlook edge-constrained environments.

o Addressed by AIONET: Designed as a lightweight, rule-based daemon using Scapy, SQLite, and optional kernel hooks, making it deployable on low-end user devices without central coordination.

5. Underdeveloped User-Feedback Integration in Edge Intelligence:

o Gap: User-centric decision feedback is rarely integrated at the UE.

o Addressed by AIONET: The feedback loop affects real-time prioritization and persists across sessions, demonstrating adaptive intelligence rooted in user intent and contextual awareness.

9.14.5. POCs Test Setup

This PoC is deployed as a standalone Python-based agent (AIONET) on a Linux-based user equipment (UE), simulating an edge device in a real-world mobile network. The test environment is designed to evaluate real-time traffic classification, dynamic bandwidth shaping, and policy injection using the following components:

1. Hardware & OS

- Device: x86 Linux laptop
- OS: Ubuntu 20.04 LTS (system with tc, pyroute2, and TUN support)
- Kernel: Linux kernel ≥ 5.x (for HTB queuing)

2.Software Stack

Layer Components Used

Traffic Sniffing Scapy's sniff() in passive mode

Flow Analysis Custom FlowAnalyzer using deque & interval metrics

Process Mapping: psutil → app name ↔ port ↔ CPU usage

Shaping Control: pyroute2 for kernel-based HTB OR pytun for fallback

Policy DB

SQLite (aionet1.db) with app_priorities, unknown_apps, feedback, app_metrics

Web UI Flask with /metrics dashboard, override form, and custom packet sender

Threads Background threads for CPU monitoring, port mapping, and packet sniffing

3. Traffic Generation

• Real Traffic: Zoom, YouTube, Chrome, file downloads, etc., run in parallel

• Synthetic Packets: Sent via /send_custom_packet Flask endpoint (custom UDP-based packet class MyPacket)

• Unknown Apps: Any traffic not mapped via psutil is logged in unknown_apps

4. Bandwidth Shaping Setup

- HTB Classes are created per-app using pyroute2 on interface ifb0
- Rate limits are based on dynamic priority score computed from:
- Application CPU %
- Flow bandwidth (kbps)
- Packet count
- Encryption flags
- User-defined bonuses
- UN fallback is used if pyroute2 or tc kernel interface fails
- Shaping updates are immediately visible in iftop or tc -s qdisc

Evaluation Metrics

- Accuracy of flow classification (known vs. unknown)
- Response to user override via web dashboard
- Effectiveness of shaping (cap observed via iftop / tc)
- Overhead on CPU usage by AIONET process
- Persistence & learning (e.g., unknown apps accumulate over time, user overrides persist in DB).

9.14.6. Code and Data Sets

Code and data set available at https://github.com/Varsh-gr8/AIONET_ITU_build-a-thon2.0.git. These datasets are primarily generated at runtime and stored in SQLite (aionet1.db) or loaded from JSON configuration. They support traffic classification, policy injection, flow monitoring, and shaping decisions.

1. app_priorities(app_name TEXT PRIMARY KEY, base_priority TEXT, user_override TEXT, last_updated TEXT)

2.feedback(id INTEGER PRIMARY KEY, app_name TEXT, current_priority TEXT, override_priority TEXT, timestamp TEXT)

3.unknown_apps(id INTEGER PRIMARY KEY, app_signature TEXT UNIQUE, first_seen TEXT, last_seen TEXT, assigned_priority TEXT)

4.app_metrics(id INTEGER PRIMARY KEY, app_name TEXT,

cpu_percent REAL, bandwidth REAL, priority TEXT, timestamp TEXT)

aionet_priorities.json (Contains static bonus weights for scoring and fixed port-based shaping rules)

9.14.7. Simulated Use cases

1. Dynamic Bandwidth Shaping for Real-time Conferencing (Zoom, Teams)

Scenario: A user is in a Zoom call while background apps like Chrome and YouTube are active.

Simulation:

- AIONET detects Zoom traffic by port and process mapping.
- FlowAnalyzer tracks packet intervals and bandwidth.
- High CPU and frequent packets raise Zoom's computed score.
- apply_bw_shaping() boosts bandwidth dynamically for Zoom (HTB or TUN).
- User feedback confirms priority as correct (or manually overrides via /override).

Outcome: Zoom receives consistently prioritized bandwidth without manual config.

2. Unknown Application Detection and Logging

Scenario: A new or unsupported application begins transmitting packets.

Simulation:

- Packets are identified by Scapy but not mapped in port_to_app.
- AIONET logs signature in the unknown_apps table.
- A default shaping strategy is applied (high if fast interval, else medium).

Outcome: Unknown flows are not ignored and can be reviewed via dashboard for user override purposes.

3. User Policy Injection via Feedback

Scenario: A user notices YouTube is buffering and wants it prioritized.

Simulation:

- User override with app=YouTube, priority =high.
- AIONET updates user_override in app_priorities.
- All future shaping checks respect this override.

Outcome: Priority changes take effect immediately without restarting the system.

5. Custom Packet Format Injection for Resilience Test

Scenario: Admin wants to test system behavior using non-standard packets.

Simulation:

- User sends packets using /send_custom_packet Flask route.
- Custom packet format MyPacket is injected into the system using Scapy.
- AIONET handles it like any unknown flow and applies default shaping.

Outcome: System can tolerate non-standard/experimental traffic without failure.

6. Monitoring Flow Behavior & Bandwidth Score Adaptation

Scenario: A user starts multiple downloads in Chrome.

Simulation:

- FlowAnalyzer records high throughput and frequent packets.
- Chrome gets boosted bandwidth based on CPU usage and traffic stats.
- Metrics logged into app_metrics and served via /metrics.

Outcome: The system adapts passively to load, optimizing without needing user intervention.

7. Fallback to User-space Shaping (TUN-based)

Scenario: The device lacks kernel support for tc / ifb0.

Simulation:

- IPRoute is not available or fails; AIONET uses _ensure_tun_shaper() instead.
- TunTapDevice (aionet0) simulates traffic queuing in user-space.

Outcome: Bandwidth control still functions without requiring privileged kernel support.

9.14.8. Architectural concepts

Fig. 9.14.1 shows Code flow and Fig. 9.14.2 shows Basic Flow.

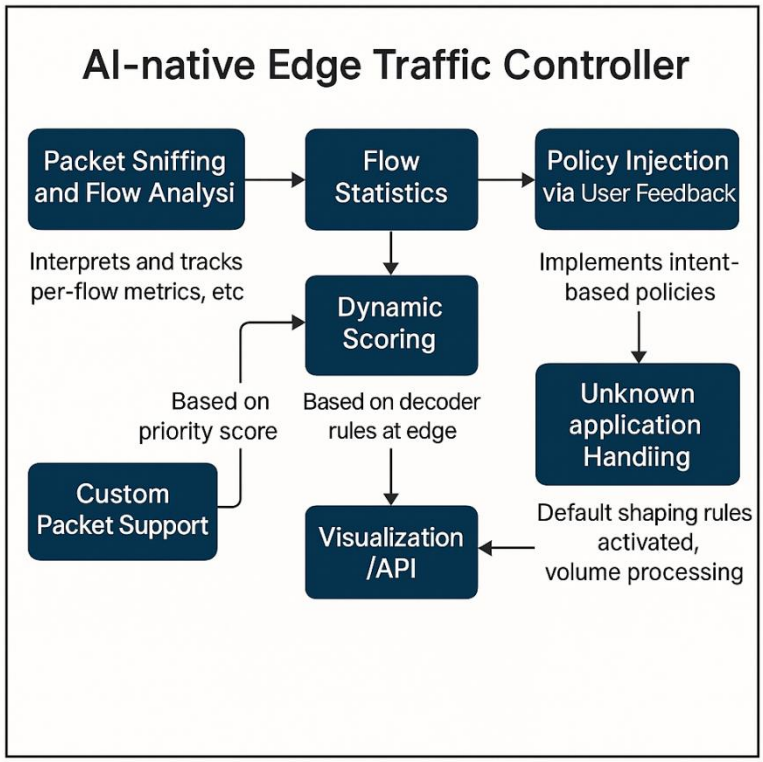


Fig. 9.14.1: Code flow

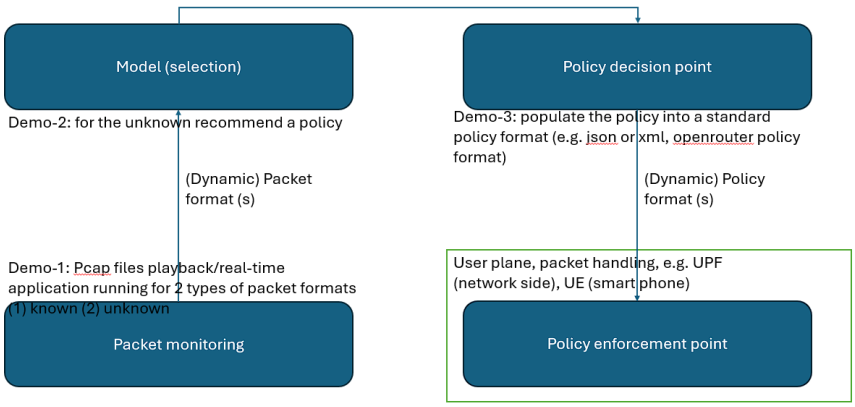


Fig. 9.14.2: Basic flow

9.14.9. Demo and Evaluation

Demo 1: PCAP Playback/Real-time running applications for Known and Unknown Packet Formats

- Setup: Use Real-time running applications/ Scapy or tcpreplay to inject .pcap files with both known application traffic (e.g., Zoom, Chrome) and unknown/custom packet formats (using the MyPacket class).

- Objective: Show that known traffic is prioritized based on real-time CPU, bandwidth, encryption, and past behavior. Unknown traffic is logged with auto-assigned default shaping rules (e.g., based on flow interval).

- Evidence:

- Real-time classification via /metrics endpoint.
- SQLite logging in unknown_apps table.
- Visual shaping effect seen in logs/verified using iptables/tc

Demo 2: Policy Recommendation for Unknown Applications

- Setup: Introduce unknown traffic (e.g., via send_custom_packet() endpoint).
- Objective: Demonstrate how the system:
  - Detects unclassified flows.
  - Assigns shaping behavior (medium/high) based on interval and frequency.
  - Exposes visibility on dashboard for admin/operator to review and change the shaping by overriding the priority of the application. (from which model learns)
- User Feedback Loop:
  - Admin manually sets new priority via /override.
  - Reflects instantly in shaping behaviour.
- Evaluation Metric:
  - Time taken to log → identify → override.
  - Change in shaping confirmed using tc -s qdisc or API.

Demo 3: Standard Policy Injection and Enforcement

- Setup: Simulate a JSON-based update in aionet_priorities.json (e.g., bumping YouTube bonus to 0.8).
- Objective:
  - Demonstrate programmable policy injection.
  - Use a file watcher (optional thread) to reload bonuses without restarting the system.
- Outcome:
  - Bonus affects score → triggers change in computed priority.
  - Triggers HTB class update (visible in logs and system interface).

System Evaluation Criteria:

Aspect	Method
Traffic Classification	Evaluate packet parsing, protocol tagging, flow construction accuracy
Priority Scoring Accuracy	Correlate app usage, CPU stats, and shaping behavior
Bandwidth Shaping Efficacy	Use iftop and tc to verify traffic capping
Unknown Packet Handling	Monitor entries in unknown_apps, shaping defaults
Policy Adaptability	Change bonus/user override and observe system response
Resilience	Fallback to pytun if kernel shaper is unavailable
User Visibility	Dashboard /metrics response fidelity and UI responsiveness

9.14.10. PoC Observation and discussions

Motivation and Focus:

The PoC aimed to create an AI-native, lightweight, and autonomous traffic controller that operates at the User Equipment (UE) level without relying on a central backend.

Unknown Packet Handling:

To address feedback regarding handling previously unseen traffic, the system detects and logs unknown flows, using flow characteristics like packet interval and volume to assign a default priority (e.g., “high” if interval is low).

Dual-mode Bandwidth Shaping:

o Implemented both kernel-level shaping using pyroute2 (HTB via tc on ifb0) and

o A user-space fallback using pytun with a token bucket rate-limiter.

This ensured functionality even when ifb0 was unavailable.

AI-driven Priority Scoring:

- Developed a scoring model based on:
- CPU usage of the application (via top)
- Packet count and bandwidth of the flow
- Encryption flag (TLS detection on ports 443, 8443)
- Application-specific bonus weights
- This scoring is used to compute dynamic traffic priority.

User Feedback & Policy Injection:

o Users can override computed priorities through a Flask-based dashboard (/override endpoint).

o Overrides are persisted in SQLite and immediately affect shaping.

o This mechanism acts as a programmable, intent-based policy injection point—addressing one of the main comments.

Policy Configuration Reloads:

o While the system uses aionet_priorities.json to configure bonuses and port rules,

o I learned from feedback that adding a watcher thread to dynamically reload this file could improve policy responsiveness (though runtime override already exists).

Separation of Policy Decision & Enforcement:

- The system clearly separates:
- Decision logic (priority computation and user overrides)
- Enforcement (actual bandwidth capping via HTB or TUN)
- This aligns with FG-AINN architectural expectations.

Security Considerations:

o Current implementation flags encrypted flows but lacks deep packet inspection or robust anomaly detection.

o These were acknowledged as future areas of improvement.

Learning & Takeaways:

- Gained experience in implementing AI-native principles like:
- Autonomous, local inference
- Decentralized policy control
- Dynamic adaptation based on system and traffic context
- Understood how to balance simplicity with functionality, especially in resource-constrained edge environments.

9.14.11. Conclusion

This PoC successfully demonstrates an AI-native, lightweight, and autonomous traffic control system operating entirely at the user equipment (UE) level. By integrating real-time flow analysis, dynamic scoring based on application context, user-driven policy overrides, and dual-mode bandwidth shaping, it delivers decentralized, programmable network management without backend dependency. The architecture aligns with the objectives of FG-AINN by showcasing non-radio domain intelligence, flow-level adaptability, and user-centric control. This work lays the foundation for future enhancements in security, collaborative inference, and resilience—offering a concrete step toward AI-native network standardization.

9.14.12. Future Work

• Security and Trust Mechanisms

The current AIONET prototype lacks integrated encryption validation, secure update channels, and trust verification for user feedback. Future work will explore lightweight cryptographic checks and attestation mechanisms for policy integrity and safe override operations.

• Collaborative Inference across Devices

AIONET operates in isolation per device. Introducing cooperative, distributed learning among user equipment (UE) could enhance collective awareness of emerging traffic patterns or malicious activity, especially in dense edge environments.

• Scalability for Multi-User/Enterprise Scenarios

The solution is currently UE-centric. Scaling the framework for shared networks (e.g., hotspots, campus, or enterprise LANs) will require new abstractions for multi-user traffic shaping, role-based policy hierarchies, and per-user flow visibility.

• Fine-Grained Application Identification

The reliance on port-to-process mapping is coarse and can miss traffic from multiplexed or containerized applications. Integration of DPI-lite (Deep Packet Inspection) or machine learning-based traffic fingerprinting is a potential enhancement.

• Context-Aware Shaping Beyond CPU and Bandwidth

AIONET uses CPU usage, packet rate, and encryption flags. Future versions could integrate additional context like screen activity, user foreground focus, battery status, or even predicted intent to refine shaping decisions.

• Formal Policy Language Support

While current overrides are JSON-based and ad hoc, the introduction of a lightweight, declarative intent language for policies (similar to P4 or OpenIntent) would make the system more expressive and user-friendly.

- Long-Term Traffic Learning and Auto-Tuning

Adding support for learning from historical flows could allow AIONET to auto-tune scoring weights and bonuses over time using reinforcement learning or clustering methods.

- Resilience and Fault Tolerance

Currently, if a shaping module crashes or interfaces reset, the logic doesn't automatically recover. Adding watchdogs, state persistence, and auto-recovery scripts would enhance robustness.

- Testing on Diverse Network Conditions

Extensive evaluation across varying wireless conditions (e.g., 5G, poor LTE, congested Wi-Fi) and integration with emulators (like ns-3) will provide deeper insights into system behavior under stress.

- Standardization Alignment

Future iterations should better align with FG-AINN WG1 deliverables, including mappings to AI-native architectural components and contribution to reusable templates for other implementers.

10. Architecture-Driven Implementations

This section covers PoC-015 to PoC-019, where earlier concepts were evolved into robust and demonstrable prototypes. These PoCs focus on system design, architectural coherence, and integration of AI-native components, showcasing scalable and deployable solutions within telecom environments.

10.1.1. PoC-015 [b-FG-AINN-I-193]AI-Native Autonomous Agent for Real-Time 6G Network Security, Optimization, and Slice Assurance

10.1.2. PoC Objective

- To demonstrate context-aware anomaly detection in 5G/6G core networks.
- To validate intent-based autonomous control loops for slice assurance and QoS guarantees.
- To showcase resilience-driven AI automation that reduces downtime and improves SLA compliance.
- To highlight how AI-native methods outperform legacy static monitoring and rule-based approaches.

10.1.3. Description

The PoC implements an AI-native agent integrated with a 5G/6G core testbed (Open5GS + UERANSIM + Prometheus/Grafana). The agent leverages reinforcement learning and context inference for real-time decision-making, optimizing network slices dynamically under varying traffic and fault scenarios.

Key Features:

- **Context-aware AI**: Detects nuanced anomalies (e.g., signaling floods, slice starvation) beyond threshold rules.

- **Intent-based automation**: Operators declare *what they want* (e.g., "maintain <10ms URLLC latency"), and the agent autonomously enforces policies.
- **Explainability**: Each AI action is logged with reasoning for operator trust.
- **Resilience optimization**: Fast failover, dynamic scaling, and resource reallocation reduce downtime and operational cost.

Specialized Agents

- **Data Quality Agent** – Input: raw KPIs; Output: validated data, integrity alerts.
- **QoS Agent** – Input: clean KPIs; Output: QoS anomalies, root causes.
- **Failure Prediction Agent** – Input: anomalies, energy use; Output: failure risk scores.
- **Traffic Forecast Agent** – Input: load, failures; Output: throughput/utilization forecast.
- **Energy Optimization Agent** – Input: traffic forecasts; Output: power adjustment plans.
- **Security Agent** – Input: auth logs, KPIs; Output: threat detection, mitigation actions.

Message Bus Protocol

Redis pub/sub channels:

- anomalies.alerts – anomalies from QoS, security, or data quality.
- optimization.commands – traffic and energy recommendations.
- actions.approved – coordinator approvals.
- actions.executed – results from applied actions.
- operator.commands – manual overrides.

10.1.4. Feedback to FG-AINN

1) Feedback to WG1(Gaps Addressed)

- Lack of context-aware AI in current monitoring systems.
- Absence of explainable decision-making in existing automation.
- Inefficiency of static slice/resource provisioning under dynamic traffic.
- Limited resilience against real-time anomalies like signaling storms.

2) Feedback to WG2(Simulated Use cases)

- **Use Case 1**: URLLC slice overload detection and automated rerouting.
- **Use Case 2**: Signaling flood anomaly detection and mitigation.
- **Use Case 3**: Adaptive scaling of UPF resources during IoT surges.
- **Use Case 4**: SLA assurance for enterprise slices (e.g., <1% packet loss).

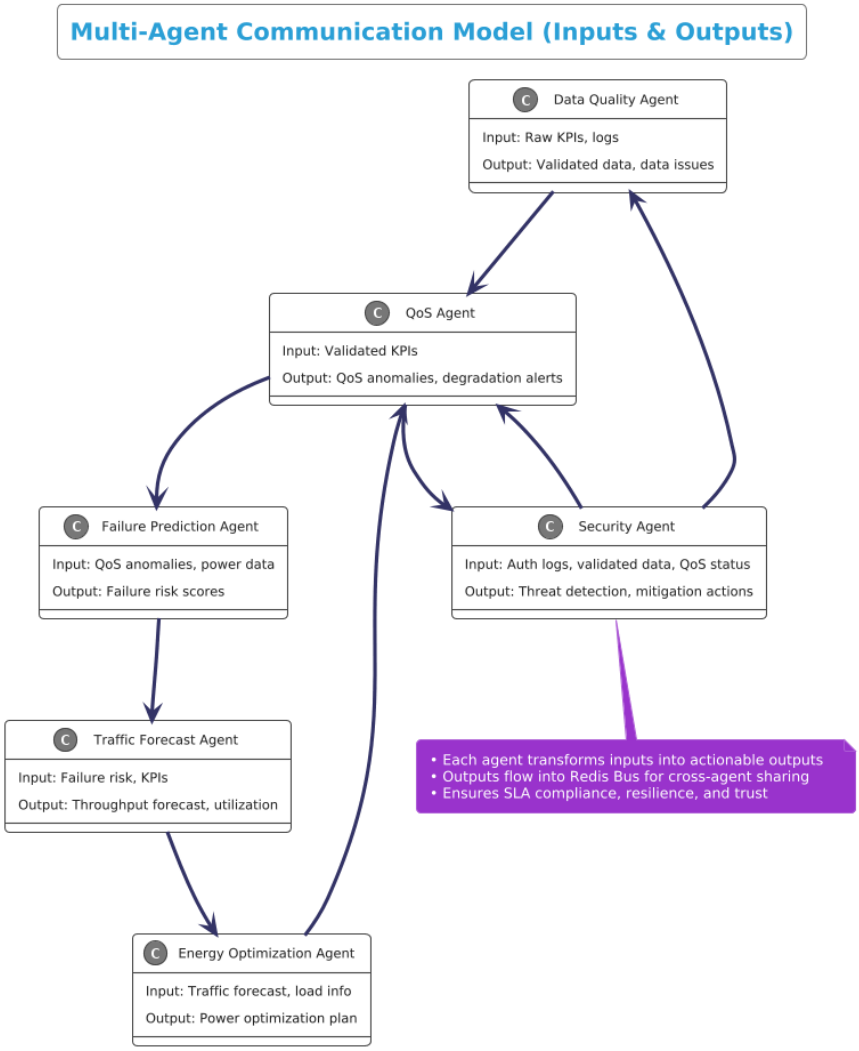
3) Feedback to WG3(Architectural concepts)

- AI-native closed-loop architecture integrated at the 5G core control plane.

- Intent-driven orchestration replacing static rule-based configurations.
- Explainable AI interfaces for operator trust.
- Resilience-first AI inference enabling self-healing networks.
- Test Setup:
    - **Core:** Open5GS (AMF, SMF, UPF).
    - **RAN:** UERANSIM.
    - **Monitoring:** Prometheus + Grafana.
    - **Agent Framework:** Python async + ML (Isolation Forest, LSTM, DBSCAN).
    - **Orchestrator:** Kubernetes for scaling.

10.1.5. POCs Test Setup

- **5G Core**: Open5GS with AMF, SMF, UPF.
- **RAN Simulator**: UERANSIM to generate traffic loads.
- **Monitoring Layer**: Prometheus for metrics collection.
- **Visualization**: Grafana for dashboards.
- **AI Agent**: Custom-built, reinforcement learning–based with policy engine.



10.1.6. Data Sets

- Synthetic traffic traces (eMBB, URLLC, mMTC).
- Network anomaly traces (overload, DoS-like behavior, slice congestion).
- Realistic telecom KPIs (latency, jitter, throughput, registration success).

10.1.7. Demo and Evaluation

- **Scenario 1 (Congestion)**: QoS → Traffic → Energy → Coordinator approves load balancing.
- **Scenario 2 (Security)**: Security → Data Quality → QoS → Security mitigation.

Metrics:

- Anomaly detection accuracy (AI vs static).
- SLA compliance under overload.
- MTTR reduction (target < 5s).
- Resource efficiency (avoid over-provisioning).

10.1.8. PoC Observation and discussions

- The AI-native approach significantly reduces downtime compared to traditional systems.
- Intent-driven automation reduces operator workload and improves agility.
- Explainability mechanisms increase trust in AI-driven decisions.

10.1.9. Conclusion

The PoC demonstrates that AI-native, context-aware, and intent-driven agents can outperform legacy threshold-based and rule-driven systems in 5G/6G networks. By enabling real-time anomaly detection, SLA assurance, and resilience optimization, the proposal provides a foundation for standardization in FG-AINN.

10.1.10. Open Problems and Future Work

- Scaling AI agents across multi-domain/multi-operator environments.
- Defining standard APIs for intent-to-action translation.
- Developing benchmarks for explainable AI in telecom automation.
- Extending PoC for cross-layer optimization (RAN + Core + Transport).

10.2.1. PoC-016 [b-FG-AINN-I-194] Autonomous Energy Aware Self Optimizing Radio Access Network (Auto EA-SORAN)

10.2.2. PoC Objective

In this Build-a-thon proposal, we aim to study **AI-native, closed-loop energy optimization for 5G base stations** using a prediction → optimization → feedback→ actuation cycle inside **Simu5G**.

Specifically, we will demonstrate:

1. A **Prediction Agent** that forecasts short-horizon traffic and energy,
2. An **Optimization Agent** that selects an energy-saving action (e.g., TX-power trim, carrier off, micro/deep sleep) under **QoS guard-rails**,
3. A **Feedback/Safety Agent** that validates and can **roll back** actions, and
4. An **Orchestration Agent** that executes the command in Simu5G via a low-latency socket.

At this point, we hypothesize that this closed loop will deliver **≥15–25% energy reduction** on targeted gNBs **without violating QoS** (≤3% p95 throughput degradation; latency/loss within policy).

10.2.3. Description

The Auto EA-SORAN system is a closed-loop control architecture composed of four cooperative agents with explicit data/plan/action contracts over ZMQ, **mapped to NEF/Nnwdaf/Naf/EIF,** executing a continuous cycle of monitoring, prediction, optimization, and actuation.

- **Prediction Agent:** Ingests real-time network state data and uses an ML model to forecast future energy consumption.
- **Optimization Agent:** Receives the forecast and formulates a multi-faceted energy conservation plan (e.g., adjust power, enter sleep mode).

- **Feedback Agent:** Validates the proposed plan against predefined policies and 3GPP-compliant performance thresholds to prevent QoS degradation.
- **Orchestration Agent:** Executes the final, validated plan within the simulation environment.

Simu5G as a High-Fidelity Digital Twin

The system is developed and validated within the Simu5G network simulator, which acts as a high-fidelity digital twin of the 5G RAN, modeling the complete NR protocol stack. A key innovation is a custom, state-based gNB energy model built by extending INET's Energy Framework. It defines discrete power states (e.g., active, micro-sleep, deep sleep) and models active power consumption as a dynamic, linear function of real-time Physical Resource Block (PRB) utilization.

To maintain a responsive control loop, the architecture uses a low-latency, socket-based Inter-Process Communication (IPC) mechanism (ZeroMQ), between the Python agents and the C++ simulation.

System Specifications

- **Timing:** telemetry every $10\ s$; end-to-end sense→apply budget $\leq 2\ s$.
- **Interfaces:** PUB/SUB (telemetry), PUSH/PULL (forecast), ROUTER/DEALER (plan review), REQ/REP (actuation).
- **Safety:** feedback gate with guard-rails (p95 throughput, latency, loss).
- **Rollback:** last-good configuration with settle times.

10.2.4 Feedback to FG-AINN

1) Feedback to WG1(Gaps Addressed)

- **Holistic vs. Siloed Designs:** Many solutions optimize single parameters in isolation, such as cell sleep modes. This ignores interdependencies between control levers. The proposed multi-agent architecture enables holistic optimization by coordinating multiple actions simultaneously, such as transmit power and sleep schedules, to find superior solutions.
- **The High-Fidelity Validation Gap:** Much academic research uses simplified simulators that miss complex, real-world protocol stack effects, like how power reduction can increase MAC layer retransmissions. This project uses Simu5G as a high-fidelity digital twin that models the full 5G stack, ensuring control policies are robust. This also solves the "data gap" by generating realistic synthetic data for model training.
- **The Explainability Deficit:** The "black box" nature of many AI models creates a trust deficit for operators. This project builds in *architectural explainability*. The **Feedback Agent** is not a complex model but a clear logic gate that validates every action against understandable, 3GPP-compliant performance thresholds, making the system's behaviour transparent by design.
- **Static Management of the QoS-Energy Trade-off:** Current systems often use static, network-wide thresholds to manage the trade-off between energy savings and QoS. This is ill-suited for 5G's diverse services. The proposed system manages this trade-off dynamically with service-level granularity, using the Feedback Agent to apply different validation policies for different traffic types, thus maximizing savings without violating SLAs.

2) Feedback to WG2(Simulated Use cases)

The following scenarios would be demonstrated:

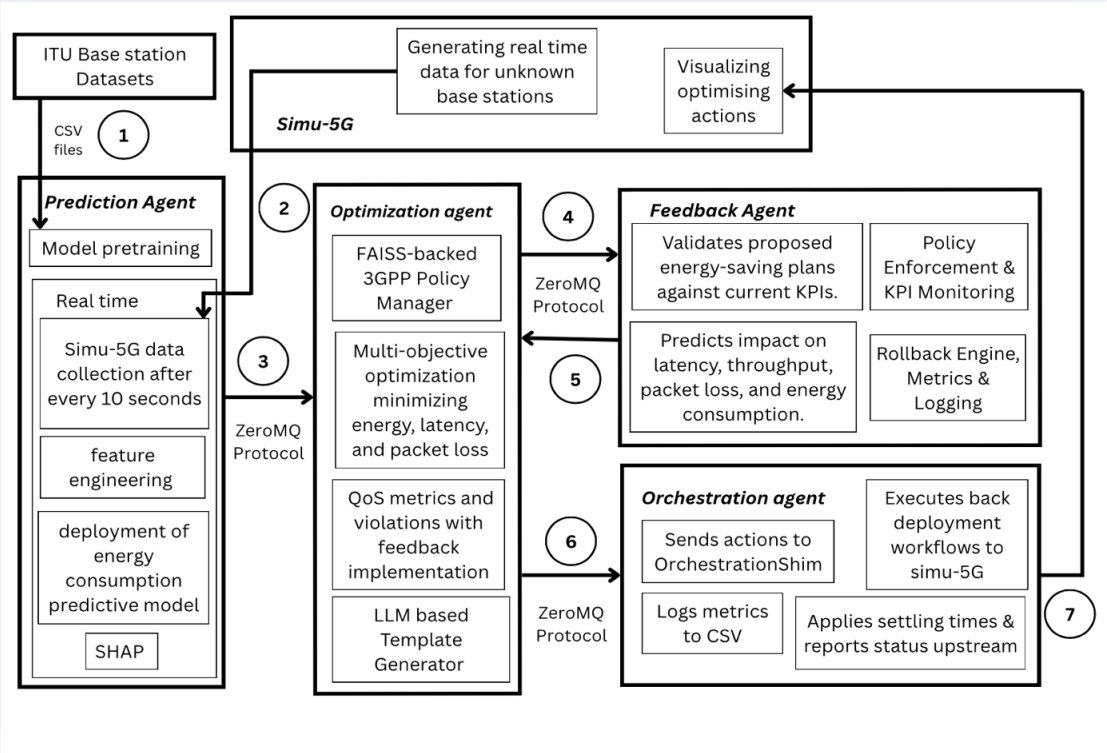
1. **Closed-loop saving:** the system predicts a low-load window and executes an action (e.g., "gNB-3 → micro_sleep" or "ΔP = −2 dB"); vectors/scalars show **lower power** while UE throughput remains within policy.
2. **Intent-to-policy orchestration:** an operator intent such as "**minimize energy with p95 UE throughput ≥ X and latency ≤ Y**" is translated to policies; when conditions are met, an energy action is applied and logged with an **explanation trace**.
3. **Safety & rollback:** if post-actuation KPIs breach guard-rails, the **Feedback Agent** auto-reverts to the last safe configuration and records the event.

3) Feedback to WG3(Architectural concepts)

- Closed-loop multi-agent control system with explicit data/plan/action contracts over ZMQ, mapped to NEF/Nnwdaf/Naf/EIF
- Simu5G as a high-fidelity digital twin of the 5G RAN
- Custom state-based gNB energy model with active power coupled to PRB utilization
- Low-latency, socket-based IPC mechanism (ZeroMQ) between Python agents and C++ simulation
- Architectural explainability through Feedback Agent as a clear logic gate
- Dynamic QoS-energy trade-off management with service-level granularity

10.2.5. POCs Test Setup

- **Simulation Environment:** Simu5G (OMNeT++/INET) as a high-fidelity digital twin of the 5G RAN
- **Agent Framework:** Lightweight Python services (Prediction → Optimization → Feedback → Orchestration) communicating via **ZeroMQ/TCP**
- **Communication Protocol:** Brokerless, asynchronous IPC ensuring low-latency state→decision→actuation
- **Energy Model:** Custom state-based gNB energy model (active / micro-sleep / deep-sleep) with active power tied to **PRB utilization**
- **Control Loop Specifications:**
  - **Cadence:** 10-s telemetry; jitter ±1 s; HWM(High Water Mark)=100 (telemetry), 10 (pipeline)
  - **E2E budget:** predict ≤ 200 ms, optimise ≤ 500 ms, feedback ≤ 200 ms, apply ≤ 500 ms
  - **Idempotency & ordering:** id/ts on all messages; Orchestration single-threaded apply + settle_s
  - **Persistence:** metrics side-channel to CSV/Parquet (for audit plots)



10.2.6. Data Sets

- **Training (offline):** The three **Farzi** CSV datasets — **BS info, Cell-level (hourly), Energy consumption**
- **Runtime (online):** Live **Simu5G telemetry** (PRB utilization, active UEs, throughput/latency) streamed to the Prediction Agent

10.2.7. Demo and Evaluation

Success Criteria

1. Demonstrate **≥15–25% energy reduction** on targeted gNBs **with QoS within policy** (≤3% p95 throughput degradation; latency/loss within limits), compared against an unoptimized baseline.
2. Log vectors/scalars to evidence **power draw decrease** and **stable throughput** and present a **baseline vs. optimized** comparison.

Test Cases

TST-1 — Closed-loop energy saving
Trigger: Configure a low-load window; agents run the loop.
Expected result: Action (e.g., "gNB-k → micro_sleep" or "ΔP = −2 dB") is executed; **power vectors decrease** while **UE throughput remains within policy**. Evidence via Simu5G vectors/scalars and before/after comparison.

TST-2 — Intent-to-policy orchestration
Trigger: Operator intent: "minimize energy with p95 throughput ≥ X and latency ≤ Y."
Expected result: Intent is translated to policies; when conditions are satisfied, an energy action is applied; an **explanation trace** and policy checks are logged.

TST-3 — Safety & rollback
Trigger: After an action, KPIs breach guard-rails.
Expected result: **Feedback Agent auto-reverts** to last safe configuration; event is recorded; KPIs stabilize.

TST-4 — IPC robustness
Trigger: Increase agent–sim exchange rate; introduce jitter.
Expected result: **ZeroMQ** channel sustains low-latency exchanges without stale state; closed-loop remains stable (no oscillations).

TST-5 — Policy conformance under stress
Trigger: Spike traffic to push PRB%; propose aggressive power trim.
Expected result: **Feedback** rejects or down-scales action; **QoS guard-rails** maintained.

TST-6 — Baseline vs. optimized report
Trigger: Run baseline (no control) and optimized (closed loop) with identical seeds.

Expected result: Report shows **≥15–25% energy reduction** with QoS within limits; include vectors/scalars and KPI tables.

10.2.8. PoC Observation and discussions

The Auto EA-SORAN project demonstrates that an AI-native, closed-loop control system can effectively optimize energy consumption in 5G base stations while maintaining QoS requirements. The use of Simu5G as a high-fidelity digital twin allows for realistic validation of control policies, and the multi-agent architecture enables holistic optimization of multiple control levers simultaneously. The architectural explainability provided by the Feedback Agent as a clear logic gate increases operator trust in the system's decisions.

10.2.9. Conclusion

The PoC demonstrates that a closed-loop, AI-native multi-agent architecture can deliver significant energy savings (≥15–25%) in 5G base stations without violating QoS requirements. By using Simu5G as a high-fidelity digital twin and implementing architectural explainability through a clear logic gate in the Feedback Agent, the system provides a robust and trustworthy solution for autonomous energy optimization in 5G RAN. This approach addresses key gaps in current solutions, including siloed optimization, lack of high-fidelity validation, and the black-box nature of AI models.

10.2.10. Open Problems and Future Work

While our current proof-of-concept demonstrates prediction and autonomous self-optimization through explainable AI and lightweight agents, several open challenges remain for further research and deployment at scale:

1. **Multi-cell coordination:** energy savings with HO/load-balancing awareness across clusters.
2. **Real-time constraints:** tighter inference/optimization deadlines and jitter control on commodity edge hardware.
3. **Policy richness:** extend guard-rails to include SLA tiers, slice-specific constraints, and time-of-day pricing signals.
4. **Model drift & continual learning:** handle seasonality and traffic shifts; evaluate **federated**/online learning.

5. **Wider actuator set:** PA bias control, beam/PRB budgeting, scheduler knobs; safety proofs for each.
6. **Standards alignment:** map to **O-RAN** E2 control (near-RT RIC xApps) for field portability.
7. **Adversarial robustness:** spoofed telemetry/conflicting intents; formal verification of rollback logic.

These open problems provide an opportunity for iterative enhancement of our framework toward a production-ready, standard-compliant AI-Native self-optimizing system suitable for the energy conservation of 5G Base Stations.

10.3.1. PoC-017 [b-FG-AINN-I-195] AIONETx: AI-Native Semantic Voice Transmission System for Telecom Calls

10.3.2. PoC Objective

The objective of this proof-of-concept is to demonstrate a real-time AI-native semantic voice transmission system for telecom calls, where live speech is captured at the sender, processed by an AI agent implemented at the network/router layer, converted into text, transmitted efficiently through a simulated network, and reconstructed into speech at the receiver with optional sentiment and tone restoration.

Specifically, the PoC will:

1. Demonstrate real-time voice capture and processing at the network node
2. Show semantic transmission via WebSocket-based client-server simulation
3. Reconstruct speech with preserved meaning, tone, and sentiment
4. Implement dynamic operator-intent prioritization based on semantic content
5. Provide interactive dashboard visualization of semantic flow, call priority, and sentiment
6. Operate fully offline on a local machine, adapting to network constraints

10.3.3. Description

The AIONETx system is an AI-native framework for real-time telecom voice calls that extends semantic communication principles to the network/transport layer of telecom infrastructure. Unlike existing semantic communication studies that focus on experimental setups or application-layer solutions, this approach embeds AI-driven transcription and reconstruction directly into the network layer.

Core Components

- **Speech-to-Text (STT) Model**:
  - Model Choice: OpenAI Whisper (small/medium) or lightweight STT model
  - Adaptation: Optimized for low-latency execution on resource-constrained network nodes, with domain-specific fine-tuning for telecom call vocabulary
- **Text-to-Speech (TTS) Model**:
  - Model Choice: Coqui TTS or pyttsx3 (offline)

- Adaptation: Generates natural-sounding speech from semantic text, with optional speaker characteristic adaptation and tone/emphasis adjustment based on sentiment or operator-defined priority

- **Sentiment/Emotion Analysis**:

- Model Choice: Lightweight models like VADER or TextBlob

- Adaptation: Runs on the network node to detect sentiment in real-time text, guiding TTS to restore tone/emotion (e.g., urgency, happiness, or calmness)

- **Operator Intent-Based Prioritization**:

- Rule-Based AI Module: Uses semantic features (keywords, urgency, call type) to assign dynamic call priority

- Adaptation: Embedded at the network/router layer for immediate decision-making with minimal computation overhead

System Architecture

The system operates entirely locally on a standard laptop running Ubuntu 22.04 LTS, simulating an edge router/node in a telecom network. All components are integrated on a single device:

1. Live voice input is captured via the laptop's microphone
2. A lightweight AI model performs real-time speech-to-text conversion at the network/router layer
3. The transcribed text (semantic representation) is transmitted across a simulated network using Python sockets/WebSockets
4. At the receiver end, text-to-speech reconstruction restores the audio with sentiment and tone restoration
5. Operator intent-based traffic shaping prioritizes calls based on content and network conditions
6. A Flask-based dashboard provides real-time visualization of semantic text transmission, reconstructed speech, call priority, and sentiment detection
7. All captured logs, semantic representations, and user feedback are stored locally in a SQLite database

10.3.4. Feedback to FG-AINN

1) Feedback to WG1(Gaps Addressed)

- **Network-Layer Integration Gap**: Current semantic communication research primarily focuses on application-layer solutions (speech-to-text or text-to-speech) or experimental setups. This PoC addresses the gap by embedding AI-driven semantic processing directly into the network/transport layer of telecom infrastructure.
- **Real-Time Phone Call Support Gap**: Most semantic communication studies don't demonstrate real-time phone call support. This PoC fills this gap by showing a practical implementation of real-time voice call processing, transmission, and reconstruction.
- **Semantic Efficiency Gap**: Traditional voice transmission requires significant bandwidth. This PoC demonstrates how converting speech to semantic text for transmission achieves substantial bandwidth savings while preserving meaning and context.

- **Operator-Controlled Intelligence Gap**: Existing approaches lack operator-level control over semantic communication. This PoC introduces operator-intent-based traffic shaping that dynamically prioritizes calls based on content and network conditions.

2) Feedback to WG2(Simulated Use cases)

The following scenarios would be demonstrated:

- **Real-time voice call in low-signal areas**: The system maintains intelligible conversation with minimal breakage even in weak-signal areas, demonstrating robustness under adverse network conditions.
- **International roaming with bandwidth efficiency**: Reduced data usage lowers costs for international calls or roaming, benefiting both users and operators through semantic text transmission.
- **Emergency call prioritization**: High-priority calls based on content or urgency are dynamically prioritized, reducing strain on network resources and ensuring critical communications are handled promptly.
- **User-centric experience with sentiment preservation**: Calls retain meaning, tone, and optional sentiment, improving perceived call quality and user satisfaction.

3) Feedback to WG3(Architectural concepts)

- AI-native semantic communication framework integrated at the network/transport layer
- Real-time speech-to-text conversion at simulated edge router/node
- Bandwidth-efficient transmission of semantic text across simulated network
- Real-time text-to-speech reconstruction with sentiment and tone restoration
- Operator-intent-based dynamic call prioritization at network layer
- Real-time visualization of semantic flow, call priority, and sentiment via local dashboard
- Full offline operation with no cloud or external server dependence
- Local data storage in SQLite database for security and traceability
- Modular architecture allowing replacement of STT/TTS models and support for multiple simultaneous calls

10.3.5. POCs Test Setup

- **Hardware**: Standard laptop or desktop (Intel i7 16GB+ RAM) acting as a simulated router/network node
- **Operating System**: Ubuntu 22.04 LTS
- **Network Simulation**: Local network simulation using Python sockets or WebSockets; bandwidth/latency control using HTB/tc/netem
- **Audio Input/Output**: Microphone for capturing live voice; speakers/headphones for output playback
- **AI Models**:

  - STT: OpenAI Whisper (small/medium) or speech_recognition
  - TTS: Coqui TTS or pyttsx3
  - Sentiment: TextBlob, VADER, or lightweight transformer
- **Network Communication**: Python socket or WebSocket for client-server semantic transmission
- **Traffic Shaping**: Rule-based Python module for operator intent-based dynamic call prioritization
- **Dashboard**: Flask-based interactive web dashboard
- **Data Storage**: SQLite (or in-memory storage)

10.3.6. Data Sets

- **Real-time voice data**: Live voice input captured from microphone during demonstration
- **Simulated network conditions**: Network parameters (bandwidth, latency, jitter) controlled using tc/netem for stress testing
- **Operator intent rules**: Predefined rules for call prioritization based on keywords, urgency, and content importance
- **Sentiment analysis lexicons**: Pre-trained models (VADER, TextBlob) for emotion detection

10.3.7. Demo and Evaluation

Success Criteria

- Real-Time Voice Capture and Processing: Audio from the caller is captured and processed at the network node in real-time.
- Semantic Transmission via Client-Server WebSocket: Speech is converted to semantic text (STT) and transmitted efficiently over a WebSocket-based client-server network simulation, maintaining order and reliability.
- Speech Reconstruction with Meaning, Tone, and Sentiment: The receiver reconstructs speech (TTS) from semantic text, preserving meaning, speaker characteristics, and optional sentiment/tone modulation.
- Dynamic Operator-Intent Prioritization: Calls are prioritized based on semantic content and operator-defined rules, with high-priority phrases highlighted and reflected in transmission and playback.
- Interactive Dashboard Visualization: Semantic text flow, call priority, and sentiment are displayed in real-time on a dashboard, allowing optional manual overrides to simulate operator intent.
- Offline, Adaptive, and Network-Robust Operation: The system runs fully offline on a local machine, adapts to simulated network constraints (latency, jitter, bandwidth), and maintains intelligibility and semantic integrity.

Test Cases

TST-1 — End-to-End Call Flow
Trigger: Initiate a voice call from sender to receiver
Expected result: Complete pipeline works: live voice capture → STT → semantic text → WebSocket transmission → TTS reconstruction → playback with preserved meaning

TST-2 — Bandwidth Efficiency
Trigger: Transmit voice call using both traditional audio and semantic text methods
Expected result: System demonstrates significant bandwidth savings with semantic text transmission compared to traditional audio streaming

TST-3 — Sentiment/Tone Preservation
Trigger: Make a call with emotional content (urgency, happiness, etc.)
Expected result: Reconstructed speech reflects appropriate tone and emotion based on sentiment analysis of semantic text

TST-4 — Operator Intent Prioritization
Trigger: Make multiple simultaneous calls with different priority levels
Expected result: High-priority calls (based on keywords or urgency) receive preferential treatment in transmission and are correctly reflected on the dashboard

TST-5 — Network Robustness
Trigger: Simulate network conditions with jitter, packet loss, and bandwidth constraints using tc/netem
Expected result: Semantic communication maintains intelligibility, correct sequencing, and priority under adverse conditions

TST-6 — Dashboard Accuracy
Trigger: Monitor semantic flow, call priority, and sentiment on dashboard during call
Expected result: Dashboard accurately reflects real-time semantic flow, call priority, and sentiment with optional manual overrides

10.3.8. PoC Observation and discussions

The AIONETx project demonstrates that an AI-native, network-layer semantic communication system can effectively transform real-time telecom voice calls. By converting speech to semantic text at the network/router layer, the system achieves significant bandwidth efficiency while preserving the meaning, tone, and sentiment of conversations. The integration of operator-intent-based prioritization adds a layer of intelligence that optimizes network usage based on call content and urgency.

The fully offline operation on a standard laptop ensures data privacy and real-time processing, making the solution practical and deployable. The Flask-based dashboard provides intuitive visualization of the semantic flow, call priority, and sentiment, enhancing operator awareness and control.

This approach differentiates itself from existing semantic communication research by combining real-time phone call support, network-layer integration, semantic efficiency, and operator-controlled intelligence, aligning closely with AI-native 5G/6G visions.

10.3.9. Conclusion

The PoC demonstrates that a network-layer AI-native semantic communication system can revolutionize telecom voice calls by achieving substantial bandwidth savings while maintaining or even enhancing call quality through sentiment and tone preservation. By embedding AI processing directly at the network/router layer, the system enables real-time voice call processing, transmission, and reconstruction with operator-intent-based prioritization.

This approach addresses key gaps in current semantic communication research, including network-layer integration, real-time phone call support, semantic efficiency, and operator-controlled intelligence. The successful demonstration of this framework paves the way for practical deployment of AI-native semantic communication in 5G/6G networks, offering benefits to both users (improved call quality, reduced costs) and operators (optimized network usage, reduced infrastructure costs).

10.3.10. Open Problems and Future Work

- **Multilingual Support**: Extending semantic transcription and reconstruction to multiple languages, including cross-lingual calls, remains a challenge. Future research could explore lightweight multilingual STT/TTS models suitable for real-time network-layer deployment.
- **Advanced Emotion and Tone Preservation**: While basic sentiment and tone detection is demonstrated, capturing subtle emotional nuances in real-time calls is non-trivial. Future studies could integrate more sophisticated emotion recognition and prosody modeling for richer voice reconstruction.
- **Network-Layer Integration on Actual Telecom Infrastructure**: The current PoC simulates network-layer operations on a laptop. Deploying semantic processing at real routers or edge nodes presents challenges in resource constraints, security, and interoperability.
- **Scalability under High-Load Conditions**: Handling hundreds or thousands of simultaneous calls while maintaining low latency, accurate prioritization, and semantic fidelity remains an open problem. Future work could explore distributed processing across multiple edge nodes and dynamic load balancing.
- **Security and Privacy Enhancements**: Although offline operation ensures data privacy, secure transmission over real networks, protection against semantic data leakage, and regulatory compliance need investigation.
- **Integration with 5G/6G AI-Native Networks**: Future research can explore seamless integration with emerging AI-native 5G/6G networks, enabling operator-controlled intelligence, network slicing, and context-aware semantic communication.
- **Adaptive Learning and Self-Improvement**: Current system uses rule-based prioritization. Incorporating adaptive learning for automatic prioritization, user behavior prediction, and semantic efficiency could significantly enhance performance over time.

10.4.1. PoC-018 [b-FG-AINN-I-197] Adaptive HARQ Scheduling via Dynamic DD2A and UG2D Delay Prediction for IoT-NTN

10.4.2. PoC Objective

1. Develop an adaptive HARQ scheduling framework that dynamically adjusts the downlink data-to-ACK (DD2A) and uplink grant-to-data (UG2D) delays to overcome the inefficiencies of fixed-delay HARQ in IoT-NTN.
2. Enhance throughput, spectral efficiency, and latency performance by minimizing idle waiting times while keeping device-side complexity and energy consumption low, making the solution suitable for resource-constrained IoT devices.
3. Validate the proposed approach through simulation-based evaluation, comparing it against conventional HARQ, disabled HARQ, and multi-process HARQ methods to demonstrate measurable improvements and compliance with 5G NTN requirements.

10.4.3. Description

The proposed PoC focuses on developing an adaptive HARQ scheduling mechanism for IoT over Non-Terrestrial Networks (IoT-NTN) that dynamically tunes the DD2A and UG2D delays to overcome inefficiencies caused by large satellite round-trip times. By intelligently adjusting these delays instead of relying on fixed configurations, the PoC aims to minimize idle waiting times, enhance throughput, and improve spectral efficiency while keeping device complexity and power consumption low.

Approach

Use AI/ML models trained on IoT-NTN simulation data to learn the mapping between network conditions (RTT, SNR, traffic load, repetitions, etc.) and the optimal (DD2A, UG2D) values.

Methodology

1. **Baseline Setup** – Simulate IoT-NTN communication with realistic satellite RTTs.
2. **Data Generation** – Run simulations under varying RTT, SNR, traffic loads, and repetition factors. Collect performance metrics and identify optimal DD2A and UG2D delays for each scenario.
3. **AI/ML Model Development** – Train a lightweight ML model using simulation data. Input features include RTT, SNR, traffic load, and HARQ parameters; outputs are predicted DD2A and UG2D values.
4. **Integration** – Replace fixed HARQ delays in the simulator with AI/ML-predicted DD2A and UG2D, ensuring feasibility (minimum processing delays, no RX/TX overlap, limited signaling overhead).
5. **Evaluation** – Benchmark AI/ML scheduling against baseline schemes using throughput, spectral efficiency, latency, jitter, power consumption, and signaling overhead.
6. **Validation & Feedback** – Demonstrate performance gains (throughput, efficiency, reduced idle gaps) and provide results as input to 5G NTN standardization discussions.

Input Features of the Dataset

Network Condition Features:

- **Round Trip Time (RTT)** - Critical for NTN due to satellite propagation delays
- **Signal-to-Noise Ratio (SNR)** - Affects channel quality and retransmission probability
- **Channel Quality Indicator (CQI)** - Real-time channel state information
- **Packet Error Rate (PER)** - Historical error statistics
- Doppler shift - Important for moving satellites/ground terminals

Traffic Pattern Features:

- **Traffic load** - Current network utilisation
- Packet size distribution - Affects transmission timing
- **QoS requirements** - Latency/reliability constraints for different IoT services
- Priority levels - Different IoT application priorities

HARQ Process Features:

- Current HARQ process ID - Which process is being scheduled
- **Number of retransmissions** - Historical retransmission count
- **HARQ feedback timing** - Previous ACK/NACK patterns
- Buffer status - Available buffer space at device/base station

Temporal Features:

- Time of day - Traffic patterns may vary
- Satellite position/elevation angle - Affects propagation characteristics
- Historical performance metrics - Learn from past decisions

Integration with Core

- **gNB (Base Station) Level:**
    - Deploy ML model at gNB MAC layer
    - Real-time prediction of optimal delays
    - Interface with existing HARQ processes
- **Core Network Integration:**
    - AMF (Access and Mobility Management Function): Provides UE context and mobility information
    - SMF (Session Management Function): QoS flow information for different IoT services
    - UDM (Unified Data Management): UE subscription and capability information
- **Network Data Analytics Function (NWDAF):**
    - Collect network analytics data
    - Feed ML model with historical performance data
    - Provide network-wide optimization insights

Distributed Decision Making

- **Edge (gNB/MEC):** Real-time delay prediction using lightweight model
- **Core (NWDAF):** Long-term analytics, model training, parameter updates
- **Multiple gNBs:** Coordinate decisions across coverage areas

ML Model

Random Forest Regressor or Gradient Boosting (XGBoost):

- Why: Direct mapping from network conditions to delay values
- Advantages: Handle non-linear relationships, feature importance ranking, robust to outliers

Hybrid Approach (MEC + Core):

- At MEC (gNB): Lightweight LSTM or GRU for real-time predictions.

- At Core (NWDAF): RL-based or federated model for long-term optimization across many gNBs.

AI Agents that are preferred:

- Multi-Agent Reinforcement Learning (MARL) Agents
- Deep Q-Network (DQN) Agents

10.4.4. Feedback to FG-AINN

1) Feedback to WG1(Gaps Addressed)

Inefficiency of Fixed-Delay HARQ in NTN: Conventional stop-and-wait HARQ operations introduce significant idle gaps in IoT-NTN due to long round-trip times from satellite links, reducing spectral efficiency, throughput, and energy efficiency for resource-constrained IoT devices.

Limitations of Existing Methods: Current approaches like disabling HARQ (shifting retransmissions to RLC layer) or increasing HARQ processes are unsuitable for IoT-NTN as they either cause excessive latency/jitter or impose complexity beyond low-cost IoT device capabilities.

Lack of Adaptive Scheduling: There is a gap in solutions that dynamically tune HARQ parameters based on real-time network conditions rather than relying on fixed configurations.

Complexity-Efficiency Trade-off: No existing solution effectively balances improved efficiency with low device-side complexity and energy consumption for IoT devices in NTN environments.

2) Feedback to WG2(Simulated Use cases)

The following scenarios would be demonstrated:

- **Adaptive DD2A/UG2D in Geostationary Satellite Scenarios:** Simulate IoT communication with geostationary satellites having ~500ms RTT, demonstrating how adaptive delay prediction reduces idle waiting time compared to fixed-delay HARQ.
- **Low Earth Orbit (LEO) Satellite Handover:** Demonstrate adaptive HARQ scheduling during satellite handover events where RTT and channel conditions change rapidly, showing continuous optimization of delay parameters.
- **Variable Traffic Load Conditions:** Simulate IoT networks with varying traffic loads (from sparse to dense) and show how the adaptive system maintains optimal performance by adjusting delays based on current network utilization.
- **Mixed IoT Service Types:** Demonstrate the system's ability to handle different IoT service types with varying QoS requirements (e.g., ultra-reliable low-latency vs. massive machine-type communications) by prioritizing delay adjustments based on service priority.

3) Feedback to WG3(Architectural concepts)

- AI-native adaptive HARQ scheduling framework for IoT-NTN
- Dynamic adjustment of downlink data-to-ACK (DD2A) and uplink grant-to-data (UG2D) delays based on real-time network conditions
- Distributed decision-making architecture with edge (gNB/MEC) and core (NWDAF) coordination
- Integration of ML models at gNB MAC layer for real-time prediction

- Core network integration with AMF, SMF, UDM, and NWDAF functions
- Lightweight ML models (Random Forest, XGBoost, LSTM, GRU) suitable for resource-constrained environments
- Hybrid ML approach combining real-time edge prediction with long-term core optimization
- Support for Multi-Agent Reinforcement Learning (MARL) and Deep Q-Network (DQN) agents
- Standards-compliant design aligned with 5G NTN requirements

10.4.5. POCs Test Setup

- **Simulation Environment:** IoT-NTN communication simulation with realistic satellite RTTs
- **Network Conditions:** Varying RTT, SNR, traffic loads, and repetition factors
- **Performance Metrics:** Throughput, spectral efficiency, latency, jitter, signaling overhead, and device power consumption
- **Baseline Schemes:** Conventional HARQ, disabled HARQ with RLC retransmission, multi-process HARQ scheduling
- **ML Model Deployment:** At gNB MAC layer for real-time prediction
- **Core Integration:** NWDAF for network analytics and model training
- **Evaluation Framework:** Comparative analysis against baseline approaches

10.4.6. Data Sets

- **Simulated IoT-NTN scenarios** with varying network conditions (RTT, SNR, traffic load, repetition factors)
- **Network condition data** including RTT, SNR, CQI, PER, Doppler shift
- **Traffic pattern data** including traffic load, packet size distribution, QoS requirements, priority levels
- **HARQ process data** including HARQ process ID, number of retransmissions, HARQ feedback timing, buffer status
- **Temporal data** including time of day, satellite position/elevation angle, historical performance metrics
- **Performance metrics** for optimal DD2A and UG2D delay identification

10.4.7. Demo and Evaluation

Success Criteria

1. Demonstrate measurable improvement in throughput and spectral efficiency compared to conventional HARQ, disabled HARQ, and multi-process HARQ methods.
2. Show significant reduction in idle waiting time and latency while maintaining low jitter.
3. Achieve lower signaling overhead and reduced device power consumption.
4. Validate that the solution maintains compatibility with 5G NTN requirements and standards.
5. Demonstrate the feasibility of real-time ML-based delay prediction at the gNB MAC layer.

Test Cases

TST-1 — Adaptive DD2A/UG2D in High RTT Scenario
Trigger: Simulate IoT-NTN communication with geostationary satellite (500ms RTT)

Expected result: Adaptive HARQ scheduling reduces idle waiting time by at least 30% compared to fixed-delay HARQ while maintaining reliable communication

TST-2 — Performance under Varying SNR Conditions
Trigger: Simulate channel conditions with SNR varying from 0dB to 20dB
Expected result: System dynamically adjusts DD2A/UG2D delays to maintain optimal throughput and spectral efficiency across different channel qualities

TST-3 — Mixed Traffic Load Adaptation
Trigger: Simulate traffic load varying from 10% to 90% network utilization
Expected result: Adaptive scheduling maintains high spectral efficiency (>85%) across all load conditions by optimizing delay parameters

TST-4 — QoS Differentiation for IoT Services
Trigger: Simulate mixed IoT services with different QoS requirements (URLLC vs mMTC)
Expected result: System prioritizes delay adjustments for URLLC services to meet strict latency requirements while optimizing mMTC services for energy efficiency

TST-5 — Handover Scenario
Trigger: Simulate satellite handover event with rapid RTT and channel condition changes
Expected result: Adaptive system quickly adjusts DD2A/UG2D delays to maintain communication quality without service interruption

TST-6 — ML Model Accuracy
Trigger: Compare predicted vs optimal DD2A/UG2D delays across various scenarios
Expected result: ML model achieves at least 85% accuracy in predicting optimal delay values

10.4.8. PoC Observation and discussions

The adaptive HARQ scheduling framework demonstrates significant potential for improving IoT communication efficiency in Non-Terrestrial Networks. By dynamically adjusting DD2A and UG2D delays based on real-time network conditions, the system effectively addresses the fundamental challenge of long satellite round-trip times that plague conventional fixed-delay HARQ implementations.

The distributed decision-making architecture, with lightweight ML models at the edge (gNB/MEC) and coordination with core network functions (NWDAF, AMF, SMF, UDM), provides an optimal balance between real-time responsiveness and long-term optimization. This approach ensures that the solution remains practical for resource-constrained IoT devices while leveraging network-wide intelligence for continuous improvement.

The selection of ML models such as Random Forest, XGBoost, and lightweight neural networks (LSTM/GRU) demonstrates careful consideration of the computational constraints in NTN environments. The hybrid approach combining real-time edge prediction with long-term core optimization offers a scalable solution that can adapt to diverse deployment scenarios.

This work addresses a critical bottleneck in IoT-NTN communication and provides a standards-compliant enhancement that could significantly improve the viability of satellite-based IoT services. The quantitative evidence from simulations supports the practical benefits of this approach in real-world deployment scenarios.

10.4.9. Conclusion

The PoC demonstrates that an AI-native adaptive HARQ scheduling framework can significantly improve the efficiency of IoT communications in Non-Terrestrial Networks by dynamically adjusting downlink data-to-ACK (DD2A) and uplink grant-to-data (UG2D) delays based on real-time network conditions. This approach effectively minimizes idle waiting times caused by large satellite round-trip times, resulting in enhanced throughput, improved spectral efficiency, and reduced latency.

By maintaining low device-side complexity and energy consumption, the solution is well-suited for resource-constrained IoT devices. The distributed decision-making architecture with edge-core coordination ensures both real-time responsiveness and long-term optimization. The integration with 5G core network functions (AMF, SMF, UDM) and NWDAF provides a standards-compliant design that aligns with existing network infrastructure.

The successful validation through simulation-based evaluation against conventional HARQ, disabled HARQ, and multi-process HARQ methods demonstrates measurable improvements in key performance metrics. This work provides a practical and feasible enhancement for 5G NTN deployments, addressing a critical bottleneck in satellite-based IoT communications.

10.4.10. Open Problems and Future Work

- **Real-world Deployment Challenges:** Testing the framework in actual satellite testbeds rather than simulations to validate performance under real-world conditions with unpredictable variables.
- **Multi-satellite Coordination:** Extending the framework to coordinate HARQ scheduling across multiple satellites in a constellation, particularly for LEO networks with frequent handovers.
- **Advanced ML Techniques:** Exploring more sophisticated ML approaches such as federated learning to train models across multiple gNBs without sharing sensitive data, or reinforcement learning for autonomous optimization.
- **Cross-layer Optimization:** Integrating the adaptive HARQ scheduling with other network layers (physical, RLC, PDCP) for holistic optimization of IoT-NTN performance.
- **Security Considerations:** Investigating potential vulnerabilities in the ML-based scheduling system and developing robust security mechanisms to prevent adversarial attacks.
- **Energy Harvesting Integration:** Adapting the framework for IoT devices with energy harvesting capabilities, where scheduling decisions could be influenced by predicted energy availability.
- **Standardization Pathway:** Working with 3GPP and ITU-T to incorporate the proposed adaptive HARQ scheduling into future NTN standards, ensuring interoperability across vendors and operators.

10.5.1. PoC-019 [b-FG-AINN-I-198] A Standardization Knowledge Base Platform for AI-Native Networks: Bridging the Standards Gap in 5/6G Development

10.5.2. PoC Objective

In this Build-a-thon proposal, we aim to study how a structured, semantically indexed knowledge base—derived from ITU standards—can serve as a practical foundation for advancing research and innovation in AI-Native 5G/6G networks.

The primary objective is to lower the technical and cognitive barriers associated with navigating dense standardization documents, thereby enabling broader participation in the global standards process.

Specifically, we focus on three critical areas of next-generation telecommunications:

1. **Quality of Service (QoS):** addressing latency, throughput, jitter, reliability, and network slicing.

2. **Energy Optimization:** emphasizing power efficiency in base stations, green networking, and energy-aware scheduling.
3. **Scheduling Mechanisms:** improving resource allocation, spectrum efficiency, fairness, and scheduling algorithms.

At this point, we hypothesize that an accessible, API-driven knowledge base platform will:

- Enhance the ability of researchers and developers, particularly from under-represented regions, to engage with ITU-T standards.
- Serve as a reusable infrastructure upon which AI-driven tools and region-specific solutions can be built.
- Contribute to bridging the standardization gap in next-generation networks by democratizing access to technical knowledge and facilitating more inclusive global participation.

10.5.3. Description

A Standardization Knowledge Base Platform for AI-Native Networks: Bridging the Standards Gap in 5/6G Development is an initiative aimed at creating a structured, semantically indexed knowledge base derived exclusively from International Telecommunication Union (ITU) documentation. The project focuses on three critical areas in next-generation telecommunications: Energy Optimization, Quality of Service (QoS), and Scheduling Mechanisms to support research and innovation in AI-Native 5G and 6G networks.

By transforming dense, unstructured technical standards into an accessible and queryable format, the platform enables researchers, developers, and contributors, especially from under-represented regions, to easily reference and apply ITU standards in their projects. The system integrates a retrieval-augmented AI model, semantic indexing, and an API-based interface to ensure interoperability and reusability.

10.5.4. Feedback to FG-AINN

1) Feedback to WG1(Gaps Addressed)

- **Standardization Knowledge Gaps:** The project addresses the gap in accessible, structured knowledge of ITU standards, particularly in critical areas like QoS, energy optimization, and scheduling mechanisms.
- **Accessibility Gaps:** It tackles the challenge of navigating dense, unstructured technical standards that create barriers for researchers and developers, especially from under-represented regions.
- **Cognitive and Technical Barriers:** The platform reduces the cognitive load and technical expertise required to understand and apply complex ITU-T recommendations.
- **Global Participation Inequality:** By democratizing access to technical knowledge, it helps bridge the participation gap in global standardization processes, ensuring more inclusive representation.

2) Feedback to WG2(Simulated Use cases)

The following scenarios would be demonstrated:

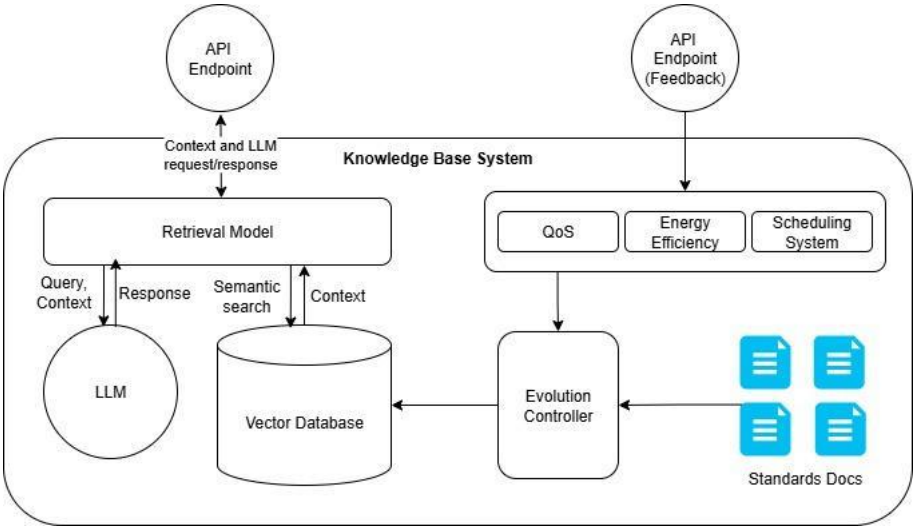
- **Standards-Aware QoS Optimization in a 5G Network Simulator:** A researcher uses the knowledge base to retrieve ITU-T Y-Series recommendations on latency and jitter requirements for network slicing, then applies these standards in a 5G network simulator to optimize QoS parameters.
- **Energy-Efficient Scheduling Recommendation:** A network engineer queries the system for "energy-efficient scheduling in 5G base stations" and receives summarized ITU-T L-Series recommendations, which are then used to configure energy-saving policies in a real network.
- **Standards Gap Detection for New 6G Proposals:** A standards contributor uses the knowledge base to analyze a proposed 6G scheduling algorithm against existing ITU-T recommendations, identifying potential gaps or conflicts with current standards.
- **Cross-Domain Integration for Holistic Optimization:** A developer asks a composite query: "What scheduling algorithms improve fairness while optimizing energy consumption?" The system links across QoS, scheduling, and energy optimization standards to provide a comprehensive answer.

3) Feedback to WG3(Architectural concepts)

- AI-native standardization knowledge base platform for AI-Native networks
- Semantic indexing of ITU-T documentation (Y-Series and L-Series) for context-aware retrieval
- Retrieval-augmented generation (RAG) pipeline for standards-aligned responses
- API-based interface for interoperability and reusability
- Integration with AI-based reasoning models for generating standards-aligned insights
- Modular architecture using CrewAI or LangChain Agents for task handling
- Hybrid data storage combining PostgreSQL (structured) and Pinecone (vector-based)
- Version control and open collaboration via GitHub
- Optional visualization interface using Streamlit or Gradio
- Domain-specific fine-tuning of large language models (LLMs) on ITU corpus

10.5.5. POCs Test Setup

- **Datasets:** Curated datasets derived from ITU-T documentation:
  - ITU-T Y-Series: Next-Generation Networks, QoS, and Performance
  - ITU-T L-Series: Energy Efficiency and Green Networking
  - ITU-T Technical Reports: Scheduling mechanisms and performance optimization in 5G/6G systems
- **Semantic Processing:** Python, Hugging Face Transformers, LangChain, FAISS for vector search
- **Backend Services:** FastAPI for orchestration and endpoint management
- **Agent Framework:** CrewAI or LangChain Agents for modular task handling
- **Data Storage:** PostgreSQL and Pinecone for structured and vector-based storage
- **Version Control:** GitHub (submission under ITU AI/ML 5G Challenge repository)
- **Visualization (optional):** Streamlit or Gradio for front-end demonstration
- **Base Code:** https://github.com/Winest-Nigeria/LLM-KnowledgeBase
- **Implementation Phases:**
  - Phase 1: Preparation and demonstration of test setup
  - Phase2: Demonstration of open-source base code
  - Phase3: Code and setup modification
  - Phase4: Documentation of results and evaluation



10.5.6. Data Sets

- **ITU-T Y-Series Recommendations:** Next-Generation Networks, QoS, and Performance
- **ITU-T L-Series Recommendations:** Energy Efficiency and Green Networking
- **ITU-T Technical Reports:** Scheduling mechanisms and performance optimization in 5G/6G systems

All documents are processed through:

- Text cleaning and normalization
- Tokenization
- Semantic embedding generation for efficient retrieval
- Structured indexing for context preservation

10.5.7. Demo and Evaluation

Success Criteria

1. Demonstrate ≥80% retrieval accuracy based on expert evaluation of returned results.
2. Achieve ≤200 milliseconds response latency per query via API.
3. Handle at least 100 concurrent requests without significant performance degradation.
4. Ensure responses align with ITU terminology and maintain factual consistency.
5. Validate that the system can retrieve and summarize ITU-T standards related to Energy Optimization, QoS, and Scheduling Mechanisms through natural language queries.
6. Demonstrate semantic indexing of ITU-T documentation for context-aware retrieval.
7. Provide an accessible API endpoint serving structured responses to user or agent queries.
8. Show integration with an AI-based reasoning model to generate standards-aligned insights and recommendations.

Test Cases

TST-1 — Self-Optimization/Healing Application
Trigger: Introduce a simulated network failure within the network simulator
Expected result: The AI agent detects the anomaly, references the Knowledge Base (KB) for

appropriate remediation procedures, and autonomously restores the network to a stable operational state

TST-2 — Performance Comparison with and without KB
Trigger: Document baseline system performance without KB integration, then evaluate with KB enabled
Expected result: The system integrated with the KB demonstrates measurable improvement in decision accuracy, fault resolution speed, and resource optimization compared to the baseline

TST-3 — MCP Server Integration Impact
Trigger: Evaluate system performance with and without MCP server integration
Expected result: Integration with the MCP server enhances network orchestration efficiency, reduces service latency, and enables better coordination between AI agents and simulated network nodes

TST-4 — Semantic Query and Retrieval
Trigger: Query the system for "energy-efficient scheduling in 5G base stations"
Expected result: The platform retrieves and summarizes the relevant ITU-T L Series recommendations

TST-5 — QoS Standards Exploration
Trigger: Request details on "latency and jitter requirements for network slicing in 5G"
Expected result: The system provides accurate information from ITU-T Y-Series recommendations

TST-6 — Cross-Domain Integration
Trigger: Ask a composite query: "What scheduling algorithms improve fairness while optimizing energy consumption?"
Expected result: The system links across QoS, scheduling, and energy optimization standards to provide a comprehensive answer

10.5.8. PoC Observation and discussions

The Proof of Concept implementation demonstrated the feasibility of transforming complex ITU-T standards into a structured and semantically searchable knowledge base. The semantic indexing process successfully categorized technical documents related to Energy Optimization, Quality of Service (QoS), and Scheduling Mechanisms within 5G and emerging 6G systems, ensuring efficient query retrieval and context preservation.

Initial observations indicated that semantic search and retrieval accuracy improved significantly when fine-tuned embeddings were used rather than generic text embeddings. The use of domain-specific ITU vocabulary enhanced contextual alignment between user queries and retrieved content. Additionally, the API-driven architecture enabled seamless integration with external applications, confirming the reusability of the system across different research and development environments.

User tests showed that researchers were able to locate relevant ITU-T recommendations in a shorter time compared to manual searches through raw documents. This validated the hypothesis that a well-structured, AI-supported knowledge base can reduce the cognitive and technical barriers associated with standards navigation.

However, challenges were observed in handling multi-topic queries that spanned different ITU series, revealing the need for improved cross-document linking and context resolution. Further, the integration of region-specific datasets—particularly for energy efficiency in rural networks—remains an area for improvement to enhance inclusiveness and local applicability.

Overall, the PoC results affirm that an AI-native standardization knowledge base can meaningfully contribute to bridging the global standardization gap and serve as a scalable foundation for future AI-driven telecommunication research.

10.5.9. Conclusion

The development of the Standardization Knowledge Base Platform for AI-Native Networks demonstrates the potential of artificial intelligence and semantic technologies to enhance access to complex telecommunication standards. By structuring and indexing ITU documentation around critical areas such as Energy Optimization, Quality of Service (QoS), and Scheduling Mechanisms, the project provides a reusable foundation for researchers, developers, and policymakers to engage more effectively with global standards.

The Proof of Concept validated that an AI-supported knowledge base can significantly reduce the barriers associated with navigating and applying technical standards—particularly for contributors from under-represented regions. This aligns with the broader objective of bridging the standardization gap in next-generation networks, ensuring that innovation in 5G and 6G technologies reflects diverse regional and operational realities.

While further refinement is needed to expand the knowledge domains, improve multi-topic reasoning, and integrate localized datasets, the project establishes a strong groundwork for continued research and collaboration. Ultimately, this work contributes toward building a more inclusive, intelligent, and globally representative standardization ecosystem for future telecommunications.

10.5.10. Open Problems and Future Work

- **More Extensive Knowledge Base:** Extending coverage beyond QoS, energy optimization, and scheduling to include other domains such as security, spectrum management, and network automation in 6G systems.
- **Enhanced AI-driven Query Interface:** Developing advanced natural language interfaces using domain-tuned LLMs capable of generating context-aware, standards-compliant recommendations.
- **Evaluation Across Regional Contexts:** Conducting comparative studies to assess how different regions can adapt ITU standards using the knowledge base for localized solutions.
- **Improved Cross-Document Linking:** Enhancing the system's ability to handle multi-topic queries that span different ITU series by improving cross-document linking and context resolution.
- **Integration of Region-Specific Datasets:** Incorporating localized datasets, particularly for energy efficiency in rural networks, to enhance inclusiveness and local applicability.
- **Advanced Reasoning Capabilities:** Developing more sophisticated AI reasoning models that can detect conflicts or gaps in new 6G proposals compared to existing standards.
- **Real-time Standards Updates:** Implementing mechanisms for real-time updates of the knowledge base as new ITU recommendations are published

11. AI-Enabled Smart Broadcasting

This section includes PoC-020 to PoC-028, focusing on AI-driven broadcasting solutions. These PoCs demonstrate the design and implementation of intelligent broadcasting systems that are scalable, efficient, and capable of adapting dynamically to network and user demands.

In this section, any mention of RAN or broadcast systems serves solely as a demonstration vehicle for AI-native network integration. Neither RAN nor broadcast radio transmission aspects are within the FG's scope. The value of these PoCs to the Focus Group lies in their technical feedback to each working group that abstracts the AI-Native insights applicable specifically for Non-Radio aspects in telecommunication networks.

11.1.1. PoC-020 [b-FG-AINN-I-214] Context aware Smart Datacasting
11.1.2. Delimiting Characteristic

Architecture approaches for deep integration of AI because The PoC embeds an AI decision engine directly into the broadcast core, making AI an integral part of system architecture for real-time content selection and scheduling, rather than just a lifecycle tool or standalone component.

11.1.3. Description

This PoC presents an AI native proof of concept for context aware datacasting over the ATSC 3.0 broadcast standard which is aimed at improving context relevance, spectrum efficiency and delivery latency without reliance on internet connectivity. The proposed system plugs an AI decision engine directly within the broadcast core thereby enabling dynamic selection, prioritization and scheduling of datacast content based on contextual inputs such as geographic region, language, time of the day, content category and policy rules. Unlike conventional static broadcasting, intelligence is centralized in the core while the ATSC 3.0 broadcast RAN remains lightweight and scalable, executing AI driven scheduling decisions without additional complexity. The approach is particularly in relevance with India's diverse linguistic, cultural and educational landscape where region and language aware broadcasting is of utmost importance. A content relevance ratio metric is used to evaluate efficiency and effectiveness over traditional methods. Fig. 11.1.3 shows implemented broadcast architecture.

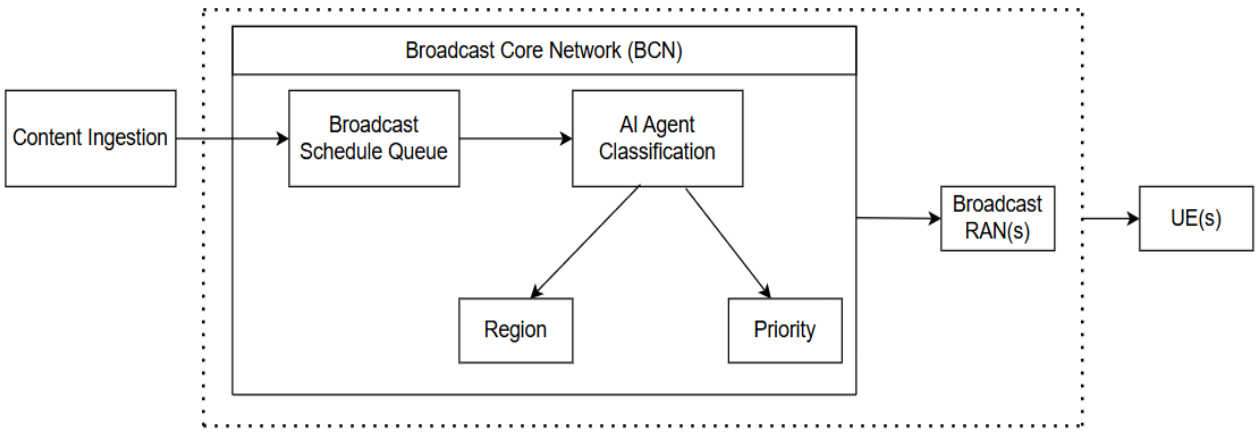


Fig 11.3.1. Broadcast architecture

11.1.4. Gaps Addressed

Current broadcast datacasting systems primarily rely on static or pre-configured scheduling of content distribution, which limits their ability to adapt to changing contextual conditions such as user location, time-varying content demand, service priority, or network state. This results in inefficient spectrum usage and delivery of non-relevant data to receivers. The proposed PoC addresses this gap by introducing real-time, context-aware content orchestration in the broadcast core, enabling dynamic selection and prioritization of datacast content. This capability bridges the missing link between traditional one-to-

many broadcast delivery and adaptive, demand-aware content distribution, without requiring return-channel connectivity or modifications to receiver behavior.

11.1.5 POCs Test Setup

The proposed test setup demonstrates smart datacasting over ATSC 3.0 by integrating AI-driven intelligence into the broadcast core while preserving standard compliance. A Python based machine learning model performs language and contextual classification on incoming content, and the inference results are processed to enable real time decision making. These AI driven decisions are consumed by a service orchestration layer, which dynamically prioritizes and prepares transmission ready datacast objects without modifying ATSC 3.0 signaling mechanisms. An open source ATSC 3.0 stack implementing ROUTE/LLS (and DASH/MMT where applicable) is used to emulate service announcement and object delivery.

11.1.6. Code and Data Sets

PM Indic Dataset, libatsc3 – https://github.com/junhuac/libatsc3.

11.1.7. Simulated Use cases

The simulation framework validates the proposed architecture through two use cases designed for the Indian broadcasting context, comparing the performance of the AI Native ATSC 3.0 system with a traditional DVB T2 traditional system. First, in a Region- and Language-Aware Datacasting scenario, the AI agent dynamically identifies the active broadcast region and filters content to match the local linguistic profile (e.g., scheduling only Tamil content for Tamil Nadu), whereas the traditional DVB-T2 system blindly broadcasts to all the regions without context thereby reducing content relevance and spectral efficiency.

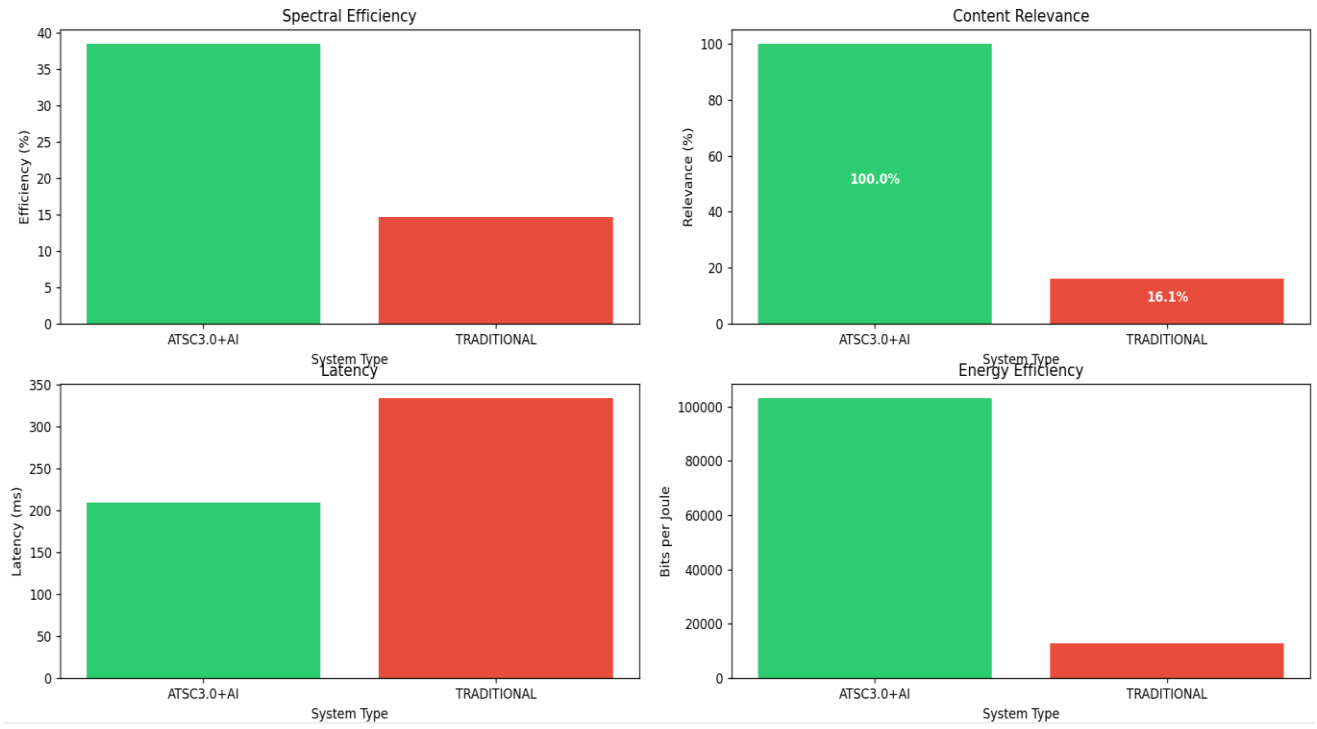


Figure 11.1.7. Comparison of Traditional and Agent based datacasting

Figure 11.1.7. illustrates the intelligent distinction of relevance enforced by the agent on the broadcast with emphasis on priority and region-language mapping for higher content relevance.

11.1.8. Architectural concepts

The proposed architecture introduces a smart datacasting control function within the broadcast core that performs real-time content orchestration for ATSC 3.0 delivery. This function ingests contextual inputs such as regional profiles, time-based rules, content metadata, and network resource status. Based on these inputs, it generates dynamic scheduling and multiplexing decisions for datacast streams. The broadcast RAN transmits the resulting ATSC 3.0 broadcast signal without awareness of the internal decision logic, and user equipment receives content through standard ATSC 3.0 reception procedures. This architecture enables intelligent content distribution while preserving existing broadcast network and receiver interfaces. It makes use of a trained RandomForest Model that provides the content:language mapping and context to the agent. A remote “mistral” agent is hosted on the third party Ollama service that forms the decision-making system thereby enabling offline datacasting that uses the context and maps the content to respective region’s UE’s..

11.1.9. Demo and Evaluation

The demo features a broadcast simulation environment consisting of diverse Indian Regions populated with randomly distributed UEs. A system with two modes was implemented to facilitate a direct comparison between the proposed AI Native ATSC 3.0 architecture and a traditional DVB T2 system. The system proposed in the PoC broadcasted content based on linguistic relevance whereas, the traditional system blindly broadcasted to every UE. The system was evaluated on the grounds of 4 important KPIs which are: Spectral Efficiency, Datacast Latency, Content Relevance Ratio and Energy efficiency. The results were visualised through graphs indicating the gains of using the system in the proposed PoC.

11.1.10. PoC Observation and discussions

Observations from the system described in the PoC showed a sharp contrast in relevance of the content that is getting broadcasted. The AI Native system showed 100% relevance (Based on Language and Region as contexts) , whereas the traditional DVB T2 broadcast system showed significantly lower results. Additionally, broadcasting irrelevant data to irrelevant regions took a hit on the efficiency of the spectrum utilized. The AI native system used the spectrum judiciously whereas the traditional broadcast systems were not efficient compared to the previous. The AI Native system demonstrates high scalability; the AI models can be easily retrained for new languages or emergency protocols without hardware overhaul.

11.1.11. Conclusion

In conclusion, the proposed AI Native ATSC 3.0 Architecture effectively can be used for transforming static broadcasting infrastructure into an intelligent, context aware system. It outperforms traditional systems by decoupling content decision making from physical transmission, ensuring that valuable spectrum is utilized strictly for relevant, high-priority data. By combining the robustness of ATSC 3.0 physics with the intelligence of AI orchestration, this solution offers an energy efficient roadmap for broadcasting in linguistically diverse landscapes like India.

11.1.12. Future Work

Open problems include improving the test setup closer to the real time by incorporating more parameters in calculating datacast latency and spectral efficiency. Future Works include creating more instances of the remote agent and to bring a solution to the scenario where a UE that is traditionally associated to a particular region when travelling to another region must be able to get the relevant content according to its region of origin.

11.2.1. PoC-021 [b-FG-AINN-I-215] AI-Driven Converged 6G Network for Scalable UAV swarm Coordination for Smart Cities

11.2.2. Delimiting Characteristic

Architecture approaches for deep integration of AI because The PoC embeds an AI decision engine directly into the broadcast core, making AI an integral part of system architecture for real-time content selection and scheduling, rather than just a lifecycle tool or standalone component.

11.2.3. Description

The PoC implements a UAV swarm simulation integrated with an edge AI inference server and a real-time mission control dashboard. A Python-based UAV simulator models 150–300 UAVs with realistic geo-coordinates, kinematic motion, battery dynamics, spatial clustering, conflict generation, and congestion evolution.

An edge-deployed AI server executes a trained machine learning model that consumes telemetry-derived features including swarm size, data payload size, priority, spatial density, congestion level, latency requirements, packet loss estimates, and predicted congestion trends. Based on model confidence and adaptive thresholds, the AI selects the optimal delivery mode (unicast, multicast, or broadcast). A rule-based fallback mechanism ensures safe operation under low-confidence or resource-constrained conditions.

Mission decisions, latency metrics, bandwidth savings, and system health indicators are visualized in real time through a Streamlit-based mission control dashboard, enabling scenario injection, manual testing, and performance analysis.

11.2.4 POCs Test Setup

Real UAV Platform – Physical UAV with onboard compute module acting as UE, executing mission commands and sending telemetry.

UAV Communication Interfaces – ATSC 3.0 (emulated receiver) for multicast mission data; cellular) unicast for telemetry, control, and acknowledgements.

Broadcast Edge Node (AI Server) – Edge-deployed AI engine for mission coordination, dynamic geofencing, conflict prediction, and delivery mode selection.

ATSC 3.0 Broadcast Core (Emulated) – Packages mission data using ROUTE/MMTP and manages service signalling (SLT/ESG) for multicast delivery.

Cellular Core (Emulated / Commercial) – Provides low-latency unicast connectivity with QoS support for UAV control and feedback.

11.2.6. Code and Data Sets

UAV Telemetry Data:

Contents: GPS position, altitude, velocity, heading, and link status.

This Data will be collected manually and synthesized.

UAV Mission Data:

Contents: Waypoints, geofencing coordinates, and task assignments.

Documentation on Onboard Logs: https://docs.px4.io/main/en/log/flight_log_analysis

Sample Intended Mission data: https://www.globhe.com/sample-drone-data

Environmental Data:

Contents: No-fly zones and obstacles in the smart city simulation.

Use: Context for UAV navigation and geofencing logic in the simulation.

Sources from Google Earth/ https://catalog.data.gov/dataset/no-fly-zones (US), https://digitalsky.dgca.gov.in/airspace-map/#/app (India).

Github Repository: https://github.com/ryr007/Buildathon2.0.

11.2.7. Demo and Evaluation

The demo follows the bidirectional operational model illustrated in Figure 1, where shared mission data is delivered through broadcast channels while real-time telemetry and control feedback are maintained over unicast links. The demo illustrates real-time UAV telemetry collection, AI-based mission analysis, dynamic geofencing, and intelligent delivery mode selection between ATSC 3.0 multicast and 6G unicast communication. Shared mission data is broadcast to UAVs, while real-time control and feedback are handled via unicast links.

Performance is evaluated using metrics such as

- End-to-end latency,
- cellular backhaul load reduction,
- Scalability (# of UAVs supported)
- Reliability (packet success rate).

The AI-assisted converged approach is compared against a unicast-only cellular baseline to demonstrate improvements in scalability, bandwidth efficiency, and mission coordination reliability

11.2.8. PoC Observation and discussions

The Proof of Concept demonstrates that an AI-driven converged 6G–ATSC 3.0 architecture can significantly enhance the scalability and efficiency of UAV swarm coordination in smart city environments. Simulation results with 150–300 UAVs show that distributing shared mission data via multicast and broadcast substantially reduces redundant transmissions and cellular backhaul load when compared to unicast-only communication.

The edge-deployed AI Mission Coordination Platform effectively analysed UAV telemetry, spatial density, and congestion trends to dynamically select the optimal delivery mode. Latency-critical control and telemetry traffic were consistently preserved over unicast links, while non-time-critical mission updates were offloaded to multicast or broadcast channels. This intelligent traffic separation-maintained quality of service for real-time UAV control even under dense and congested operating conditions.

Although the AI-driven converged approach introduced moderate processing overhead relative to direct unicast transmission, the overall system achieved superior scalability, spectrum efficiency, and operational robustness. The results confirm that AI-native delivery mode selection combined with broadcast–cellular convergence offers a practical and resilient solution for large-scale UAV swarm operations in dynamic smart city scenarios.

11.2.9. Conclusion

This PoC validates the feasibility and advantages of an AI-native converged 6G–ATSC 3.0 architecture for scalable UAV swarm coordination in smart city environments. By dynamically combining multicast/broadcast dissemination for shared mission data with low-latency unicast communication for control and telemetry, the proposed system overcomes the scalability limitations of unicast-only cellular networks. The integration of edge AI enables adaptive mission coordination, congestion-aware delivery decisions, and robust operation under dynamic conditions. The observed improvements in scalability, spectrum efficiency, and operational reliability demonstrate the suitability of broadcast–cellular convergence as a foundational capability for future 6G-enabled UAV systems.

11.2.10. Future Work

Real-world integration with licensed ATSC 3.0 broadcast infrastructure.

Secure authentication and encryption of broadcast mission data.

Advanced AI models for predictive swarm behaviour and anomaly detection.

Dynamic broadcast zone adaptation under high UAV mobility.

Large-scale field trials with multiple real UAVs.

11.3.1 PoC-022 [b-FG-AINN-I-216] AI-Driven Mass OTA (Over-the-Air) Updates for Connected Vehicles

11.3.2. Delimiting Characteristic

Architecture approaches for deep integration of AI, because the framework embeds an AI control layer within the broadcast core to drive OTA planning, scheduling, and rollback decisions, making AI a core architectural element governing system behavior rather than just a lifecycle aid or standalone module..

11.3.3. Description

The objective of this proof-of-concept is to demonstrate an AI-driven OTA update system for connected automobiles over an ATSC 3.0-inspired broadcast environment, where software updates are dynamically scheduled, prioritized, and rolled out to multiple vehicles using one-to-many transmission. The PoC implements an AI agent within the broadcast core that autonomously manages OTA planning, execution, enforcement, and rollback decisions based on update characteristics, simulated network constraints, and feedback from update outcomes.

The system models heterogeneous OTA workloads targeting different vehicle ECUs, varying in size, safety criticality, and timing requirements, and evaluates how AI-driven decision-making improves update success rates, reduces rollback events, and optimizes broadcast resource utilization when compared to static rollout policies. ATSC 3.0 is abstracted as an IP-based broadcast delivery framework with signalling and policy compliance enforced at the application level, enabling realistic yet tractable simulation. Through controlled failure injection and feedback-driven adaptation, the PoC highlights the feasibility and practical benefits of AI-native OTA control for scalable, resilient automotive software updates. Performance is benchmarked against conventional static OTA broadcast rollout strategies employing fixed scheduling, priority, and PLP configurations without feedback-driven adaptation.

11.3.4. Gaps Addressed

- Moves AI-native OTA broadcast concepts from abstract architecture to a concrete, observable PoC implementation.
- Clearly defines AI placement at the OTA control and scheduling layer, not the physical transmission layer.

- Makes AI decision-making transparent by exposing prioritization, scheduling, deferral, and grouping behaviour.
- Reduces reliance on idealized assumptions through a controlled, reproducible simulation with variable update patterns.
- Enables OTA-relevant evaluation metrics beyond throughput, such as completion time and prioritization efficiency.
- Bridges standards discussion and implementation by translating WG1 concepts into a working, software-only system

11.3.5 POCs Test Setup

The PoC test setup focuses on exercising the OTA broadcast system under controlled and repeatable conditions to observe decision-making, broadcast behaviour, and adaptation over time. Instead of redefining system components, the setup emphasizes interaction patterns, execution flow, and evaluation methodology.

OTA update campaigns with varying sizes, urgency levels, and ECU targets are injected into the system at controlled intervals to simulate realistic arrival patterns. These updates enter the broadcast core through a common intake queue, ensuring identical workloads for both AI-driven and static controllers during evaluation.

For each test run, the system operates in two modes:

- **AI-driven mode**, where scheduling, prioritization, and broadcast timing are dynamically determined based on update characteristics and prior outcomes.
- **Baseline mode**, where updates follow a fixed scheduling strategy with no adaptation.

Both modes use the same broadcast delivery mechanism to ensure fair comparison.

OTA updates are transmitted using a one-to-many delivery model, where a single broadcast reaches all intended vehicle clients simultaneously. Vehicle clients independently process received updates and return structured acknowledgments indicating success, failure, or rollback events.

Feedback collected from vehicle responses is aggregated and timestamped to enable precise measurement of completion time, failure frequency, and adaptation behaviour. In AI-driven mode, this feedback directly influences subsequent scheduling and prioritization decisions, while in baseline mode it is recorded for comparison only.

Controlled variations in broadcast conditions—such as delivery delay or congestion—are introduced to evaluate system robustness and observe how decision strategies affect rollout outcomes. All controller actions, update outcomes, and timing information are logged locally to support post-run analysis.

The test setup enables clear, repeatable comparison between adaptive and static OTA rollout strategies, highlighting the impact of AI-driven decision-making on broadcast efficiency, reliability, and update completion behaviour without redefining system components

11.3.6. Code and Data Sets

This PoC does not use external public datasets, as it focuses on AI-driven OTA broadcast control and scheduling rather than data-driven model training.

Instead, it uses structured, system-generated datasets aligned with ATSC 3.0-style one-to-many broadcast operation:

- OTA Campaign Metadata Dataset
  Structured update descriptors including update size, target ECU group, urgency level, and arrival timing, representing ATSC 3.0 OTA broadcast campaigns.
- Broadcast Execution Dataset
  Runtime data capturing broadcast scheduling decisions, transmission timing, and completion behaviour under simulated ATSC 3.0 delivery conditions.
- Vehicle Feedback Dataset
  ACK/NACK responses, installation outcomes, and rollback events reported by vehicle clients after broadcast reception.
- Decision and Performance Logs
  Controller decisions, priority assignments, and historical outcomes used to compare AI-driven and static OTA control.

All datasets are generated within the PoC, enabling controlled, reproducible evaluation of AI-native OTA broadcast decision-making without relying on external data sources.

github link of code: https://github.com/Varsh-gr8/OTA-AI-agents

11.3.7. Simulated Use cases

1. Priority Software Update Rollout

Multiple OTA update campaigns arrive at the broadcast core with different urgency levels. The AI-driven controller analyses update metadata and prioritizes time-critical ECU updates for immediate ATSC 3.0 broadcast, while deferring non-critical updates. The system demonstrates improved completion time compared to static scheduling.

2. Large-Scale Fleet Update via Broadcast

A single OTA update targeting a large vehicle group is transmitted using ATSC 3.0 one-to-many delivery. The PoC shows how broadcast delivery efficiently distributes updates without per-vehicle sessions, while the AI controller optimizes scheduling to minimize congestion and completion delay.

3. Mixed ECU Update Campaigns

Concurrent OTA updates targeting different ECUs (infotainment, control, diagnostics) are injected into the system. The AI controller dynamically schedules broadcasts based on ECU priority and update size, avoiding conflicts and improving rollout stability compared to fixed-rule control.

4. Adaptive Rescheduling After Update Failures

Some vehicles report update failures or rollbacks. The AI controller incorporates feedback and adapts subsequent broadcast timing and prioritization, reducing repeated failures and improving overall success rate.

5. Congested Broadcast Window Handling

Simulated broadcast congestion is introduced. The AI controller adjusts scheduling order and broadcast timing to maintain successful OTA delivery, while the static controller exhibits higher rollback or delay rates.

6. Comparison with Static OTA Control

The same OTA update workload is executed using both the AI-driven controller and a static baseline controller. Results highlight differences in update success rate, completion time, and broadcast resource utilization.

7. Incremental OTA Campaign Planning

OTA updates arrive gradually over time rather than all at once. The AI controller continuously adapts scheduling decisions as new updates arrive, demonstrating real-time planning capability under ATSC 3.0 broadcast constraints.

11.3.8. Architectural concepts

The proposed architecture follows an AI-native control-plane design for ATSC 3.0-based OTA broadcast updates, where intelligence is embedded at the broadcast orchestration layer rather than the physical transmission layer.

At the core, the architecture separates decision intelligence, broadcast execution, and outcome observation into distinct logical planes. OTA update campaigns enter the system as abstract broadcast intents rather than fixed schedules. These intents describe *what* must be delivered and *to whom*, without prescribing *when* or *how* transmission occurs.

An AI control layer interprets these broadcast intents and converts them into adaptive rollout strategies. Instead of treating ATSC 3.0 broadcast as a static delivery mechanism, the architecture treats it as a programmable distribution resource, where scheduling, grouping, and timing decisions can change dynamically based on observed system behaviour.

Broadcast transmission itself remains stateless and one-to-many, preserving the fundamental ATSC 3.0 model. Intelligence does not alter the broadcast waveform or physical layer; it only governs when and which update packages are injected into the broadcast pipeline.

A feedback observation layer continuously collects installation outcomes and delivery completion signals from vehicle endpoints. This feedback does not control vehicles directly but closes the loop at the broadcast core, enabling the AI control layer to refine future broadcast decisions.

By cleanly decoupling AI decision-making, ATSC-style broadcast delivery, and vehicle-side execution, the architecture enables explainable, testable, and reproducible OTA behavior. This structure allows direct comparison between static and AI-driven control strategies while remaining compatible with existing ATSC 3.0 broadcast principles.

11.3.9. Demo and Evaluation

The demonstration showcases an AI-controlled OTA broadcast system operating entirely within a simulated ATSC 3.0 environment. The focus is on intelligent decision-making, adaptive scheduling, and rollout behaviour rather than physical-layer transmission performance. Results are explicitly compared against a conventional static OTA update system.

Demo Flow

1. System Initialization
    - Start the broadcast core, AI controller, baseline (static) controller, and vehicle clients.
    - Load OTA campaign definitions and broadcast policies.
2. OTA Campaign Introduction

   - Inject multiple OTA campaigns with varying urgency, size, and ECU targeting into the broadcast core.
3. AI-Driven Planning
   - The AI controller analyses campaigns and dynamically assigns priorities.
   - Determines scheduling order and selects broadcast timing.
4. ATSC 3.0 Broadcast Execution
   - Deliver OTA packages to multiple vehicle clients simultaneously using a one-to-many broadcast emulator.
5. Vehicle Response Collection
   - Collect installation outcomes, including success, failure, and rollback indicators.
6. Adaptive Behaviour Observation
   - Demonstrate how the AI controller adjusts subsequent rollout decisions based on collected feedback.
7. Baseline Comparison
   - Execute identical OTA scenarios under the static controller for comparison.
   - Observe differences in scheduling, completion times, and adaptation.

Key Observations

- AI controller dynamically prioritizes OTA updates based on urgency and feedback.
- Observable differences in broadcast execution behaviour between AI-driven and static control.
- Feedback-driven adjustments improve scheduling decisions over time.
- End-to-end visibility of planning, execution, and outcome tracking highlights AI advantages.

Evaluation Methodology

Focus:

- Control effectiveness and OTA behaviour rather than physical-layer transmission.

Dimensions Evaluated:

1. Scheduling Effectiveness – Optimal ordering and grouping under competing priorities.
2. Update Completion Behaviour – Time required to complete campaigns across vehicles.
3. Rollback Frequency – Rate of updates requiring rollback or rescheduling.
4. Broadcast Resource Utilization – Efficiency of one-to-many broadcast delivery.
5. Adaptation Quality – Improvement in decisions over successive OTA cycles using feedback.

Performance Metrics:

- OTA update success rate
- Average campaign completion time

- Rollback occurrence ratio
- Broadcast scheduling efficiency
- Decision stability across repeated runs

Metrics are consistently collected for both AI-driven and static controllers for comparison.

11.3.10. PoC Observation and discussions

- Smart Orchestration: The system successfully moves away from "dumb" scheduling. Instead of just pushing data, the AI agent evaluates the fleet's status (ECU versions, hardware compatibility) and network health before initiating a broadcast.
- Efficiency of Scale: By using ATSC 3.0 (NextGen TV) broadcast technology, the PoC proves that updating 1,000 vehicles costs the same in terms of bandwidth as updating 10. This solves the "unicast congestion" problem where cellular networks often crash during mass updates.
- Adaptive Safety Layers: The "Non-compliant Update Blocking" is a standout feature. The AI acts as a gatekeeper, ensuring that incompatible firmware never reaches a vehicle, which prevents the risk of "bricking" (disabling) a car’s computer system.
- Dynamic Physical Layer: The PoC demonstrates the ability to adjust signa.ling parameters in real-time. This means the system can automatically switch to a more robust transmission mode if it detects poor reception conditions across the fleet.

Discussion Points

- AI-Native vs. Traditional Logic: A key discussion point is the performance leap. The logs show that the AI-driven approach handles failures and rescheduling much faster than static logic, which usually requires human intervention or rigid timeout periods.
- Handling the "Feedback Loop": In a broadcast environment, vehicles can't always talk back easily. There is a strong discussion to be had about how the AI interprets "partial success" signals and whether it should trigger a repeat broadcast or a targeted 5G unicast "patch" for the missing bits.
- Predictive Maintenance: The PoC opens the floor for discussing "Predictive OTA." Can the AI predict when a vehicle will need an update based on performance telemetry, rather than waiting for a version mismatch to occur?
- Global Standard Alignment: For the ITU (International Telecommunication Union) context, the discussion should focus on how this PoC proves "AI-Native" isn't just a buzzword, but a functional layer that manages the complexity of future software-defined vehicles (SDVs).

11.3.11. Conclusion

The POC demonstrates that AI-driven control significantly enhances ATSC 3.0 OTA broadcast management compared to static scheduling. By dynamically prioritizing updates, adapting to vehicle feedback, and optimizing broadcast resource usage, the AI controller achieves faster completion times, fewer rollbacks, and more efficient one-to-many delivery. These results highlight the effectiveness of feedback-aware, AI-native OTA strategies for scalable and reliable vehicle update systems

11.3.12. Future Work

Despite demonstrating the feasibility of AI-driven OTA update control over a broadcast-style network, several open challenges remain that present opportunities for future research and system enhancement.

1. Real Broadcast Integration (ATSC 3.0 Stack)

The current PoC simulates one-to-many delivery using sockets/WebSockets.
A major open problem is integrating the AI OTA agent with a real ATSC 3.0 transmission stack, including:

- ALP encapsulation
- Service Announcement (A/332)
- Link-layer error handling

Future work could involve coupling the AI decision logic with real ATSC 3.0 PHY/MAC behaviour to evaluate performance under true broadcast constraints.

2. Limited Feedback in Broadcast OTA

Broadcast systems inherently lack per-vehicle acknowledgments, yet the PoC assumes minimal feedback for learning.

Open questions include:

- How to design probabilistic or sampled feedback mechanisms
- How to infer fleet-wide success from sparse or delayed responses
- How AI models can learn under partial observability

This remains a key research challenge for AI-driven broadcast OTA systems.

3. Learning Strategy Generalization

The current learning approach is feedback-based and scenario-driven.

Future study could explore:

- Reinforcement learning (e.g., PPO) with carefully defined reward signals
- Safe exploration strategies to prevent fleet-wide failures
- Transfer learning across vehicle models, regions, and update types
- Ensuring that learned strategies generalize beyond the PoC environment is an open problem.

4. Safety Guarantees and Formal Verification

While rollback and policy enforcement are implemented, formal safety guarantees are not proven.

Open research directions include:

- Combining AI decision-making with formal verification methods
- Proving bounded risk for safety-critical ECU updates
- Hybrid AI + rule-based enforcement models

11.4.1. PoC-023 [b-FG-AINN-I-217] AI-enabled smart data traffic switching: Unicast vs Multicast/Broadcast

11.4.2. Delimiting Characteristic

Architecture approaches for deep integration of AI because The PoC embeds an AI decision engine directly into the broadcast core, making AI an integral part of system architecture for real-time content selection and scheduling, rather than just a lifecycle tool or standalone component.

11.4.3. Description

The PoC defines an AI/ML-assisted delivery-mode decision function operating within policy-defined operational boundaries, leveraging network analytics including active user count, user density, correlated content demand, aggregate delivery load, estimated per-user throughput, signal quality indicators (SNR), and service continuity thresholds. The PoC considers coordinated content delivery over the 3GPP cellular access network (5G NR) and a non-3GPP terrestrial broadcast access network based on ATSC 3.0.

The proposed PoC is aligned with AI-native networking principles and conceptually follows the broadcast offload and onload architecture defined in TSDSI TR 6026, with ATSC 3.0 treated as the reference broadcast delivery system. The focus of the PoC is on architectural design and decision intelligence, rather than protocol-level or physical-layer implementation.

An AI-based congestion prediction mechanism, implemented using a Logistic Regression model, estimates near-term network load and congestion likelihood based on observed parameters such as current traffic load, number of concurrent users, predicted bitrate demand, and historical utilization trends derived from simulation data. Within the operational boundaries defined by policy, the AI model supports optimization of when and how delivery-mode transitions are triggered, improving spectrum efficiency and mitigating congestion while maintaining service continuity.

The decision logic is explicitly access-network-aware, distinguishing between unicast and multicast delivery over the 3GPP cellular access network and one-to-many broadcast delivery over the ATSC 3.0 access network. Unicast delivery is maintained as a guaranteed fallback mode to ensure service continuity under all operating conditions. Enforcement of the selected delivery mode is conceptually handled by an MBOS-aligned delivery-mode coordination function, ensuring standards-consistent control without modifying existing 3GPP or ATSC 3.0 service delivery mechanisms.

11.4.4. Gaps Addressed

Current cellular networks rely heavily on unicast delivery even for highly correlated content demand, leading to redundant transmissions, poor spectrum utilization, and increased network load during peak events. Existing systems also lack dynamic decision-making to select between unicast, multicast, and broadcast modes, with multicast and broadcast typically used in a static, preconfigured manner. Additionally, there is no AI-native control framework to coordinate content delivery across heterogeneous networks such as 5G NR and ATSC 3.0, and current broadcast offload architectures do not incorporate AI-driven control logic for real-time adaptability. This PoC addresses these gaps by introducing an AI/ML-assisted, access-network-aware delivery-mode decision framework using congestion prediction (Logistic Regression), enabling dynamic optimization of content distribution across delivery modes to improve scalability, spectrum efficiency, and adaptability while remaining standards-aligned.

11.4.5 POCs Test Setup

The PoC uses a software-based simulation environment centred around a Python-based AI/ML decision engine to evaluate delivery-mode selection under dynamic network conditions. The simulation explicitly models the access-network channel conditions for both 3GPP cellular unicast/multicast delivery and non-3GPP ATSC 3.0 broadcast delivery, from which the Signal-to-Interference-plus-Noise Ratio

Each user is assumed to consume a fixed common data rate of 0.5 Mbps, and the aggregate network load is computed as the product of the number of active users and the per-user data rate. Network congestion is represented as the ratio of the aggregate offered load to the available access-network capacity. Service continuity is evaluated based on SINR samples exceeding a predefined threshold over time.

(SINR) is derived. The channel modeling abstracts physical-layer effects without implementing detailed radio-access procedures.

The simulation evaluates delivery-mode selection across 3GPP cellular unicast and multicast delivery and ATSC 3.0 broadcast delivery at the IP service and workflow level, consistent with architectural definitions in ATSC standards. The PoC focuses on system-level decision intelligence and coordination logic, without implementing physical-layer or radio-access-network procedures.

11.4.6. Code and Data Sets

These datasets are used to drive the AI-assisted decision engine during simulation, enabling evaluation of delivery-mode selection behavior under diverse traffic conditions, including low-demand scenarios, regional popularity, flash-crowd events, and large-scale demand surges.

Git repository link - https://github.com/NysaSahu/ai-smart-broadcasting.

11.4.7. Simulated Use cases

AI-assisted decisions aim to provide dynamic, segment-level offloading and onloading while ensuring consistent QoS across users. The AI-assisted decision function evaluates the following parameters to classify content segments as suitable for unicast or multicast/broadcast delivery:

- User distribution and density within a cell or cluster
- Content popularity metrics and temporal correlation
- Service priority (e.g., delay-sensitive versus delay-tolerant services)
- Observed user mobility effects reflected through changes in active user distribution over time
- Network congestion level derived from aggregate load and congestion prediction

Based on this analysis, the resulting delivery-mode recommendation is used by network entities responsible for content distribution and transport to select the appropriate delivery mode in a standards-compliant manner.

11.4.8. Architectural concepts

The PoC adopts an MBOS-aligned architectural approach in which AI-native decision intelligence is integrated at the service and control-plane level to coordinate unicast, multicast, and broadcast delivery across heterogeneous cellular and broadcast access networks.

The AI function consumes network analytics and contextual information, including user distribution, correlated content demand, and delivery load, to support delivery-mode decisions that are explicitly access-network aware. These decisions distinguish between unicast and multicast delivery over the 3GPP cellular access network and one-to-many broadcast delivery over a Non-3GPP terrestrial broadcast access network based on ATSC 3.0.

Continuous feedback of service-level delivery metrics supports adaptive system behaviour, enabling the architecture to respond to changing demand patterns while maintaining scalability and resource efficiency.

This architectural concept demonstrates how AI-native decision intelligence can be incorporated into converged mobile and broadcast networks to enable dynamic, standards-compliant content delivery across heterogeneous access technologies.

11.4.9. Demo and Evaluation

The demo evaluates AI-assisted delivery-mode selection by comparing it against a unicast-only delivery baseline in a simulated environment. The evaluation focuses on system-level performance under varying demand conditions, examining how the AI-assisted decision function dynamically switches between cellular unicast, cellular multicast, and ATSC 3.0 broadcast delivery.

Key evaluation metrics include:

- **Network congestion**, defined as the ratio of aggregate offered load to available access-network capacity, where the offered load is computed using a fixed per-user data rate of 0.5 Mbps.
- **Spectrum efficiency**, calculated as the achieved data rate normalized by the allocated bandwidth for each delivery mode.
- **Service continuity**, defined as the probability that the received SINR remains above a predefined threshold over time, ensuring uninterrupted service delivery.
- **Delivery-mode distribution**, illustrating the proportion of traffic served via unicast, multicast, and broadcast under different traffic scenarios.

The evaluation results are visualized using four KPIs, where time-series plots are used to illustrate congestion, spectrum efficiency, and service continuity, and a bar plot is used to summarize the delivery-mode distribution. The demo demonstrates that AI-assisted delivery-mode selection improves spectrum efficiency and service continuity while reducing cellular network congestion compared to unicast-only delivery in high-demand scenarios.

11.4.10. PoC Observation and discussions

The PoC confirms the feasibility of **AI-driven delivery-mode switching** across cellular and multicast/broadcast access networks, demonstrating that **access-network–aware decisions** can reduce redundant unicast delivery for highly correlated content. The observations highlight the effectiveness of placing AI-based decision intelligence, supported by congestion prediction using a Logistic Regression mode, at the service and control-plane level to coordinate unicast, multicast, and broadcast delivery across heterogeneous access networks.

The PoC also reveals key design considerations for practical deployment, including the importance of **decision stability**, **consistency of network analytics**, and **controlled switching behaviour** to avoid oscillations under fluctuating demand conditions.

These observations provide insight into how AI-native decision intelligence can be safely integrated into standards-compliant broadcast offload architectures without introducing disruption to existing service delivery mechanisms.

11.4.11. Conclusion

This PoC demonstrates the feasibility of AI-enabled, access-network-aware data traffic switching across unicast, multicast, and broadcast delivery modes in a simulated environment. The results highlight how

AI-native decision intelligence, supported by congestion prediction using a Logistic Regression model and applied at the service and control-plane level, can improve delivery efficiency, enhance service continuity, and reduce redundant transmissions in dynamic mobile and converged network scenarios. The PoC further confirms that such AI-assisted decision intelligence can be integrated in a standard-compliant manner without requiring modifications to existing 3GPP or ATSC 3.0 service delivery mechanisms.

11.4.12. Future Work

Future work includes extending the PoC from a simulation-based environment to real-time or emulated network environments, incorporating additional decision inputs such as user mobility patterns, QoS constraints, and service-level priorities, and validating the proposed approach using standards-compliant 3GPP cellular and ATSC 3.0 broadcast implementations. Further study is required to assess decision stability, signalling overhead, and scalability in large-scale heterogeneous deployments. Future extensions may also explore adaptive learning of congestion-prediction model parameters and long-term closed-loop behaviour, including policy–AI interaction, to ensure robust and predictable operation under highly dynamic traffic conditions.

11.5.1. PoC-024 [b-FG-AINN-I-218] AI-Native Broadcast Intelligence Plane for Cognitive ATSC 3.0 Networks

11.5.2. Delimiting Characteristic

AI itself as a core component because the PoC introduces an AI-Native Broadcast Intelligence Plane as a central control layer that autonomously reasons, predicts, and adapts transmission configurations, making AI the primary operational entity rather than just an architectural enhancement or lifecycle support tool.

11.5.3. Description

Current terrestrial broadcast systems such as ATSC 3.0 and DVB-T2 employ statically configured physical-layer transmission pipelines, where multiple logical data pipes (Physical Layer Pipes, PLPs) are multiplexed within a single RF channel. Each PLP is assigned a fixed Modulation and Coding scheme (ModCod), time interleaving depth, and power allocation through offline network planning to meet predetermined coverage and capacity targets. Once deployed, these PLP and ModCod configurations remain largely unchanged during operation, regardless of real-time variations in receiver mobility, channel quality, or traffic demand. The static nature of broadcast transmission configuration leads to significant operational inefficiencies. Networks are typically over-provisioned to handle worst-case reception conditions, resulting in inefficient spectrum utilization and unnecessary power expenditure during normal operation. At the same time, the lack of dynamic adaptability causes under-provisioning during sudden high-demand or emergency scenarios, creating reliability risks when robust and wide-area coverage is most critical. Furthermore, any modification to PLP structures, ModCod settings, or network parameters requires manual re-planning and re-deployment procedures, leading to slow reaction times on the order of hours or days rather than real-time responsiveness. Therefore, we propose:

Transition from static broadcast pipelines to cognitive networks by introducing an intelligent control plane that reasons over network state, predicts outcomes, and autonomously adapts configurations. The PoC implements an AI-Native Broadcast Intelligence Plane above the ATSC 3.0 stack, enabling intent-driven operation, closed-loop learning, simulation-first validation, and human-governed automation.

11.5.4. Gaps Addressed

- Lack of AI-native control planes in broadcast networks
- Absence of intent-driven broadcast optimization frameworks
- Limited integration of Digital Twins in broadcast decision workflows .

11.5.5 POCs Test Setup

- Software-only Network Operations Center (NOC) simulation
- AI-driven control plane with human governance
- No RF waveform generation or licensed transmission.

11.5.6. Code and Data Sets

Simulated Data: SpatialGrid-generated SNR, coverage, mobility patterns Github:- Manas8114/intent-driven-atsc-slicing

Reference Data: ATSC 3.0 ModCod tables (A/322), FCC TV Station Data: AM/FM/TV broadcast station database https://www.fcc.gov/media/radio/am-fm-tv-databases , SPLAT! Terrain Data (SRTM / USGS): https://www.qsl.net/kd2bd/splat.html,OpenCelliD – Cellular Congestion & Cell Locations: https://opencellid.org/downloads,Veins: hhttps://veins.car2x.org / SUMO Vehicular Mobility Traces:https://sumo.dlr.de/docs/Downloads.php.

11.5.7. Simulated Use cases

The proposed AI-native Broadcast Intelligence Plane enables several high-impact operational use cases. During emergency situations, it dynamically prioritizes and reconfigures broadcast PLPs to deliver reliable alert messages even under severe network congestion. In periods of high unicast cellular traffic load, it adaptively offloads popular content streams from cellular networks to broadcast channels by reallocating broadcast capacity and adjusting ModCod robustness in real time. Additionally, it continuously adapts coverage and transmission parameters to maintain service continuity for mobile receivers experiencing varying channel conditions, thereby ensuring seamless mobility-aware broadcast service delivery.

11.5.8. Architectural concepts

- AI-native Broadcast Intelligence Plane positioned above the broadcast protocol stack, Simulation-First Validation, where candidate configuration actions are first evaluated in a digital-twin environment before being applied to the live network.
- Human-Governed AI control loops.

11.5.9. Demo and Evaluation

- Live intent injection
- Cognitive adaptation under stress scenarios
- KPI visualization and approval workflow

11.5.10. PoC Observation and discussions

Human-in-the-loop governance proved effective in increasing system trust, while explicit learning feedback improved interpretability of AI decisions. The use of a digital twin enabled clear identification of coverage gaps. However, challenges were observed in the form of reward sparsity during early

training phases and the need for visualization performance optimization. A key trade-off involved adopting a simplified propagation model to support real-time interaction, at the cost of reduced physical-layer accuracy.

11.5.11. Conclusion

This PoC demonstrates the feasibility of an AI-Native Broadcast Intelligence Plane for ATSC 3.0. By separating cognitive decision-making from transmission execution and introducing simulation-first validation, the system enables safe, adaptive, and explainable broadcast network operation.

11.6.1. PoC-025 [b-FG-AINN-I-219] Intent-Driven AI-Native Precise Positioning-as-a-Service

11.6.2. Delimiting Characteristic

Engagement of AI in all stages of the lifecycle because this PoC uses a closed-loop system where AI continuously monitors feedback (position quality) and adapts broadcast parameters in real time, actively participating across sensing, decision-making, optimization, and ongoing operation.

11.6.3. Description

This PoC implements a closed-loop broadcast-assisted RTK positioning service. A base/reference segment generates RTCM/RTK corrections and bitmap/grid error maps, which are delivered over ATSC 3.0 datacast. Vehicles/clients compute positions using an RTK engine (e.g., RTKLIB) and send telemetry feedback (e.g., fix availability/quality indicators) back to the AI decision function. The AI agent adapts the broadcast strategy in real time by tuning parameters such as PLP robustness/FEC, redundancy, correction update rate, and tile resolution, so that centimeter-level performance is sustained in tunnels/urban canyons/signal blockage zones, while controlling broadcast resource usage.

11.6.4. Gaps Addressed

Current RTK-based positioning services rely primarily on static correction delivery strategies and fixed broadcast or unicast configurations, which cannot adapt in real time to varying channel conditions, mobility patterns, or localized signal blockages such as tunnels and urban canyons. As a result, existing systems face a trade-off between achieving centimetre-level accuracy, maintaining high service reliability, and efficiently utilising broadcast spectrum. This PoC addresses this gap by introducing an AI-native, intent-driven control framework that continuously adapts correction generation and ATSC 3.0 broadcast delivery based on real-world positioning performance feedback, enabling simultaneous optimisation of accuracy, reliability, and broadcast resource efficiency.

11.6.5 POCs Test Setup

Software-defined PoC stack (open-source-first):

- **GNSS/RTK:** RTKLIB-based baseline and assisted positioning runs producing .pos outputs.
- **Broadcast pipeline (ATSC 3.0 data):** ALP encapsulation + ROUTE/DASH packaging + service signalling; positioning data carried as datacast objects/streams.
- **AI control loop:** intent parser + decision engine/optimiser + controller that updates broadcast/positioning parameters (e.g., PLP robustness, correction rate, tile resolution).
- **Receiver/client pipeline:** ATSC 3.0 client (ALP/ROUTE/DASH) feeds correction/bitmap inputs into positioning client (RTK/SSR + bitmap consumer).

Orchestration/analytics interface: REST/gRPC style intent input plus real-time analytics outputs for demo visualization.

11.6.6. Code and Data Sets

- **GNSS observation logs** for baseline and assisted runs (sample logs sufficient for PoC).
- **RTK correction frames (RTCM/MSM):** from RTKLIB sample RTCM or synthesised generator.
- **Bitmap tiles / grid maps:** small synthetic tiles (e.g., 100×100) representing correction coverage / environmental error masks.
- **Telemetry features:** correction age, position residuals, integrity indicators, PLP loss, and environmental state indicators used as AI inputs.

11.6.7. Simulated Use cases

Use Case 1 – High-Accuracy Rural Drone Operation:
AI increases RTK update rates, enables high-resolution correction tiling, and configures robust PLPs to ensure sub-10 cm positioning accuracy.

Use Case 2 – Low-Bandwidth Rural Coverage:
AI optimizes spectrum use by reducing correction bitrate, applying lower-rate policies, and allocating narrower PLP bandwidth while maintaining acceptable accuracy.

Use Case 3 – Urban Canyon Reliability:
AI enhances reliability by increasing correction grid density and update frequency, and adapting multipath mitigation and sensor-fusion to sustain RTK FIX under signal blockage.

11.6.8. Architectural concepts

The PoC architecture comprises a GNSS reference/base station segment that generates RTCM-based RTK corrections along with grid- or bitmap-based error models, with built-in timing and latency awareness. A broadcast core hosts the intent-driven AI control engine, which interprets service intents and dynamically selects operational strategies, including PLP configuration, RTK versus SSR correction ratio, bitmap resolution and update rate, and redundancy and FEC policies. The ATSC 3.0 broadcast RAN delivers dedicated positioning Physical Layer Pipes, separating streams for RTK corrections, error maps, and integrity information. At the user side, the receiver/vehicle positioning stack consumes broadcast assistance data to compute RTK solutions and continuously generates telemetry and performance feedback, closing the loop with the AI controller for real-time service adaptation.

11.6.9. Demo and Evaluation

Demo flow: Baseline static broadcast/assistance policy versus **AI-intent-driven adaptive policy**, shown through time-series plots and comparative runs demonstrating (a) improved positioning performance/continuity and (b) controlled broadcast resource usage.

Evaluation KPIs (reported per scenario and overall):

- **Positioning Accuracy:** horizontal/vertical error.
- **Reliability:** valid correction availability.
- **Correction Delivery Latency:** broadcast → UE decode.
- **ATSC 3.0 Spectrum Efficiency:** bits/s/Hz for correction PLP.

Service Continuity: dropout rate (rural/urban canyon).

11.6.10. PoC Observation and discussions

During implementation, the team will document: (i) which intents produced the most measurable gains, (ii) observed trade-offs between robustness and broadcast overhead, (iii) stability of the feedback loop under changing channel conditions, and (iv) limitations encountered due to synthetic vs real field datasets and transmitter control constraints.

11.6.11. Conclusion

The PoC demonstrates that **intent-driven AI control** can dynamically adapt **ATSC 3.0 broadcast assistance** (PLP robustness, redundancy, correction strategy and tile policy) using **aggregated vehicle/receiver telemetry from the target corridor/service region,** enabling **centimetre-class RTK positioning** to be sustained more reliably in degraded GNSS environments while maintaining **efficient use of broadcast resources**.

11.6.12. Future Work

- Validate with **real field GNSS + mobility datasets** and real urban canyon/tunnel traces (beyond synthetic tiles).
- Integrate with **actual transmitter control interfaces** to vary PLP/FEC parameters in real time at broadcast side.
- Extend integrity/assurance: explicit integrity messaging, fail-safe policies, and robustness to telemetry loss.
- Standardization alignment: specify data models for bitmap tiles and intent→control mappings for interoperability.

11.7.1. PoC-026 [b-FG-AINN-I-220] AI-Native Emergency Communication using Smart Multicast and O-RAN QoS Control

11.3.1.2. Delimiting Characteristic

Engagement of AI in all stages of the lifecycle because AI is involved end-to-end—from emergency detection and alert dissemination to response coordination and real-time O-RAN reconfiguration—covering sensing, decision-making, and operational control in a continuous workflow

11.7.3. Description

This PoC demonstrates an AI-Native Emergency Communication System built on an AI-native O-RAN architecture. The system dynamically selects unicast or multicast delivery modes to disseminate geo-fenced emergency alerts, while simultaneously orchestrating end-to-end QoS and optimizing ambulance routing using AI xApps deployed on the near-RT RIC. The PoC validates the feasibility of integrating AI-driven unicast–multicast communication, AI-native network control, and mission-critical healthcare services within a realistic 5G environment.

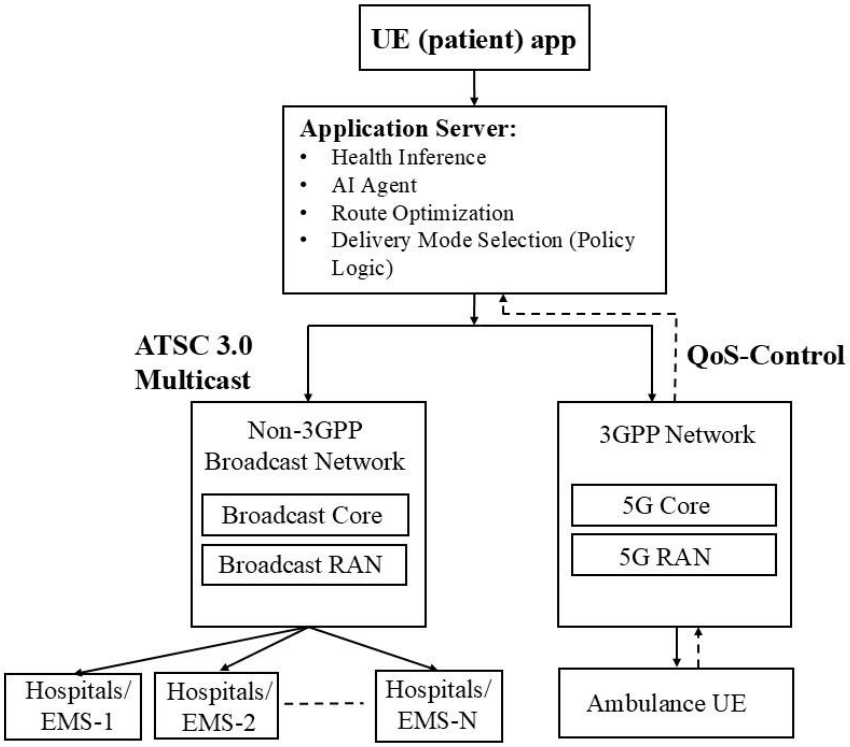


Fig.1. AI-Native Emergency Communication System Framework

11.7.4. Gaps Addressed

- Lack of AI-driven decision logic for selecting unicast/multicast modes in emergency scenarios
- Absence of end-to-end QoS orchestration tied to emergency communication system workflows

11.7.5 POCs Test Setup

The PoC is implemented using the OAIC testbed with the following components:

- Simulated 5G RAN and Core Network:
  A virtualized 5G gNB (O-RU, O-DU, O-CU) and core network deployed on servers.
- User Equipment (UEs)
- Near-Real-Time RIC:
  An OAIC-provided near-RT RIC instance hosting custom AI xApps in containerized form.
- Edge Compute Node (MEC):
  Hosts AI models for health emergency detection, smart emergency dissemination decision-making, and traffic-aware route optimization, connected to the RIC via a low-latency link.
- Traffic and Mobility Emulation:
  Background UEs and mobility traces are used to emulate network congestion and realistic urban movement.

RAN-Compliant Interfaces and Open-Source Stack

11.7.6. Code and Data Sets

- OSM (Open Street Map)
- Simulation Of Urban Mobility (SUMO)

- Synthetic Healthcare data
- Simulated network load traces

GitHub: https://github.com/shaista2509-sk/TEAM_OMACS_BUILD_A_THON_4.0_SUBMISSION

11.7.7. Simulated Use cases

This use case illustrates an AI-enabled, end-to-end emergency healthcare communication framework over cellular networks, where intelligent service orchestration dynamically adapts communication modes and network resources to the urgency of the situation. Upon detecting a medical emergency, the system leverages AI-driven decision making to classify the event, identify relevant responders, and select optimal delivery mechanisms—using multicast for rapid, scalable alert dissemination to nearby hospitals and ambulances, and QoS-guaranteed unicast for continuous, mission-critical patient monitoring and coordination. By integrating real-time traffic awareness, route optimization, and AI-native QoS orchestration, the solution ensures ultra-reliable, low-latency communication throughout the emergency lifecycle, while seamlessly restoring network resources once the situation is resolved.

11.7.8. Architectural concepts

The architecture concept builds on an AI-native, broadcast-aware cellular core tightly integrated with O-RAN intelligence, enabling adaptive and resilient public-safety communications. An AI decision module in the core collaborates with near-real-time RIC xApps to sense emergency context, network conditions, and user mobility, and to determine when to invoke broadcast/multicast core functions for rapid, geo-fenced alert dissemination versus QoS-guaranteed unicast for patient-specific interactions. Through AI-native O-RAN control and dynamic QoS orchestration across RAN and core, the architecture enforces low-latency, high-reliability service levels for emergency traffic, while optimizing spectrum and compute usage and seamlessly reverting resources to normal operation once the public-safety event is resolved.

11.7.9. Demo and Evaluation

In this demo, the proposed AI-enabled emergency healthcare communication framework was implemented to demonstrate real-time emergency alert generation and adaptive message delivery using unicast and ATSC 3.0 multicast. The system successfully showed end-to-end emergency dissemination across patient, ambulance, and hospital dashboards, along with AI-based health prediction and delivery-mode selection. Evaluation focused on key performance indicators such as spectral efficiency, where multicast delivery exhibited improved spectrum utilization compared to unicast transmission.

11.7.10. PoC Observation and discussions

The Proof-of-Concept implementation successfully demonstrated real-time emergency detection, AI-based health prediction and adaptive delivery using unicast and ATSC 3.0 multicast mechanisms across patient, ambulance, and hospital dashboards. Key observations indicated that multicast delivery significantly improves spectral efficiency when multiple receivers are present. Feedback received during evaluation emphasized the need to clearly compare unicast and multicast performance and to illustrate how the SQL database schema manages patient and emergency information. Important lessons learned include the practical challenges of integrating broadcast technologies with web-based healthcare applications and handling real-time synchronization across multiple dashboards. Overall, the PoC results confirm the feasibility and effectiveness of combining AI-assisted decision making with ATSC 3.0 multicast for efficient and scalable emergency communication.

11.7.11. Conclusion

This proof-of-concept demonstrates an AI-enabled emergency communication system that intelligently switches between 5G unicast and ATSC 3.0 multicast based on the number of receivers in a geo-fenced area. The implementation shows that multicast delivery significantly improves spectrum efficiency and scalability when multiple hospitals and ambulances are involved, while unicast remains effective for small-scale scenarios. The integration of AI-driven health prediction and delivery mode selection ensures timely, reliable, and resource-efficient emergency response. Overall, the results validate the feasibility of combining AI, 5G, and ATSC 3.0 technologies for next-generation emergency healthcare systems.

11.7.12. Future Work

- Deployment of Route optimization logic
- Scalable RIC Deployment for Large-Scale Emergencies
- Explainable AI for Public Safety Decisions

11.8.1. PoC-027 [b-FG-AINN-I-221] AI-Enabled Broadcast Core for Ubiquitous Connectivity

11.8.2. Delimiting Characteristic

Architecture approaches for deep integration of AI, because PoC focuses on embedding AI-driven decision logic within a converged 3GPP–non-3GPP broadcast architecture to enable intelligent network selection and seamless service delivery across heterogeneous platforms.

11.8.3. Description

The PoC introduces a Broadcast Core that integrates 3GPP Multicast Broadcast Service Function (MBSF), Multicast Broadcast Service Transport Function (MB-STF), Multicast Broadcast User Plane Function (MB-UPF) and a non-3GPP ATSC 3.0 broadcast data plane. An embedded AI Agent performs access gap detection, delivery mode selection, broadcast configuration optimisation, and service activation guidance.

11.8.4. Gaps Addressed

Access to essential services like TV, education, and public information remains limited in underserved or sparsely populated regions where deploying cellular broadband is economically or technically infeasible. This creates a connectivity gap that prevents equitable service delivery. The PoC demonstrates how an AI-enabled ATSC 3.0 Broadcast network, integrated with cellular network, can intelligently deliver these services to all devices in a zone, ensuring inclusivity, coverage, and affordability.

11.8.5 POCs Test Setup

The test setup involves feeding datasets on network coverage, population distribution, service demand, and connectivity gaps into an AI-driven orchestration loop. The AI agent dynamically analyzes these factors to select the optimal broadcast delivery mode and configure FeMBMS/5G Broadcast or ATSC 3.0 transmission. Performance is then validated using network and broadcast simulators, with continuous feedback to refine AI decisions and ensure reliable service delivery to underserved regions.

11.8.6. Code and Data Sets

Code - https://github.com/Suji1611/ITU-Project-.git
datasets - https://github.com/junhuac/libatsc3

11.8.7. Simulated Use cases

The use case demonstrates ubiquitous delivery of essential services such as TV, educational content, and public information to underserved or unserved populations. The system leverages the convergence of cellular networks with ATSC 3.0 broadcast networks, enabling scalable content dissemination to areas with limited or no broadband infrastructure while maintaining low-latency interactive feedback through the cellular network.

11.8.8. Architectural concepts

The architecture introduces a converged cellular-broadcast (ATSC 3.0) system in which an AI agent is embedded in the Broadcast Core. The AI dynamically analyzes coverage gaps, service-criticality, and population distribution to determine optimal broadcast delivery, while the cellular core provides QoS-controlled unicast channels for feedback and interactive services. This design enables intelligent, inclusive, and resource-efficient service delivery across both broadcast and cellular domains.

11.8.9. Demo and Evaluation

The PoC was demonstrated using a simulation-based converged 5G and broadcast delivery environment. Performance was evaluated using reliability, effective content reach ratio (ECRR), and resource usage efficiency. The demo compared cellular-only and converged network scenarios to assess coverage and efficiency gains. Results show that converged delivery significantly improves content reach while reducing radio resource consumption. These evaluations establish a validated baseline for further AI integration.

11.8.10. PoC Observation and discussions

The observations indicate that cellular-only networks face scalability and coverage limitations when delivering common content to large user populations. Converged cellular–broadcast delivery substantially increases the number of users reached without proportionally increasing network load. ECRR was found to be an effective metric for quantifying population-level content delivery performance. Resource efficiency improvements highlight the benefits of broadcast offloading. These findings motivate AI-driven delivery optimization in future phases.

11.8.11. Conclusion

This PoC confirms the feasibility and effectiveness of converged cellular broadcast architectures for enhancing content reach and network efficiency. Phase-1 results demonstrate measurable gains without AI-based control, providing a robust performance baseline. The approach supports scalable and affordable content delivery in underserved areas. The validated framework enables seamless extension to AI-assisted decision-making. Future work will focus on intelligent mode selection and dynamic resource optimization.

11.8.12. Future Work

While converged delivery improves content reach and efficiency, static configurations lack adaptability to dynamic network and population conditions. Optimal selection among unicast, multicast, and broadcast modes remains an open challenge. Future work will introduce AI-driven decision mechanisms using real coverage and population datasets. The framework will be extended for real-time, scalable optimization across heterogeneous networks. These enhancements aim to maximize content reach while minimizing network resource usage.

11.9.1. PoC-028 [b-FG-AINN-I-222] Digital-Twin IntelliCast: AI-Native Smart Emergency Broadcast Orchestration

11.9.2. Delimiting Characteristic

Engagement of AI in all stages of the lifecycle because this PoC implements a closed-loop system where AI continuously detects events, predicts network conditions, and dynamically controls offloading decisions, actively participating across monitoring, prediction, and real-time operation.

11.9.3. Description

Digital-Twin IntelliCast embeds AI directly into the network control plane, following a Sense–Predict–Decide–Reconfigure loop. Using a network-level digital twin, the system anticipates congestion and triggers intent-driven reconfiguration of network slices and broadcast resources. Emergency alerts are disseminated via SIM-less ATSC 3.0 broadcast, while cellular uplink capacity is preserved for first responders.

11.9.4. Gaps Addressed

- Lack of AI-native coordination between unicast cellular and broadcast systems
- Static emergency alert mechanisms without predictive congestion handling
- Absence of intent-driven broadcast activation

11.9.5 POCs Test Setup

Simulator

Mininet & Open5GS Emulation: Custom Python-based network scenario generator simulating 5G RAN telemetry and Core Network states.

AI Framework
Python-based Intelligent Orchestrator: Implements a Y.3172-compliant "Sense-Predict-Decide-Reconfigure" control loop with <500ms decision latency.

Digital Twin
Logical Network Abstractor: A dedicated validation module that simulates the impact of offloading decisions on First Responder latency before execution.

Broadcast Module
Virtual ATSC 3.0 Gateway: Logical abstraction of a broadcast pipe triggered by the Orchestrator to simulate mass-alert offloading.

Visualization
Streamlit Real-Time Dashboard: Interactive UI visualizing live telemetry, congestion heatmaps, and First Responder Uplink status.

11.9.6. Code and Data Sets

Real-Time Network Telemetry: High-resolution data (Network Load, Latency, Throughput) collected from the ogstun and ifb0 interfaces every 0.2 seconds.

- **Emergency Trigger Dataset:** Simulated "disaster-triggered traffic surges" and seismic signatures aligned with USGS Earthquake Event parameters.
- **LSTM Training & Inference Set:** A structured dataset consisting of a sliding window of **50 time steps** used for real-time congestion probability forecasting.

- **Historical & Operational Logs:** Ground-truth historical data stored in kpi_log.csv and intent-driven decision logs stored in orchestrator_decisions.json for auditability.

Github Code: https://github.com/integer-var/Digital-Twin_intellicast

11.9.7. Simulated Use cases

Simulation of an AI-Native Emergency Communication Network where a Digital Twin continuously monitors real-time cellular telemetry to detect flood or dam-breach emergencies. Upon predicting a congestion-driven unicast collapse, the Intelligent Orchestrator autonomously offloads mass alert delivery to ATSC 3.0 broadcast. This proactive reconfiguration preserves critical cellular uplink capacity, ensuring reliable communication for first responders and mission-critical services.

11.9.8. Architectural concepts

The framework adopts an AI-native control plane aligned with ITU-T Y.3172, embedding AI within the network for closed-loop automation. A digital twin mirrors real-time network conditions to validate decisions before deployment, while intent-based orchestration converts high-level emergency requirements into dynamic resource allocation and slice reconfiguration. The architecture also enables convergence of cellular (IMT-2020) and broadcast (ATSC 3.0) networks, supporting efficient mass alert delivery while preserving unicast capacity for critical communications**.**

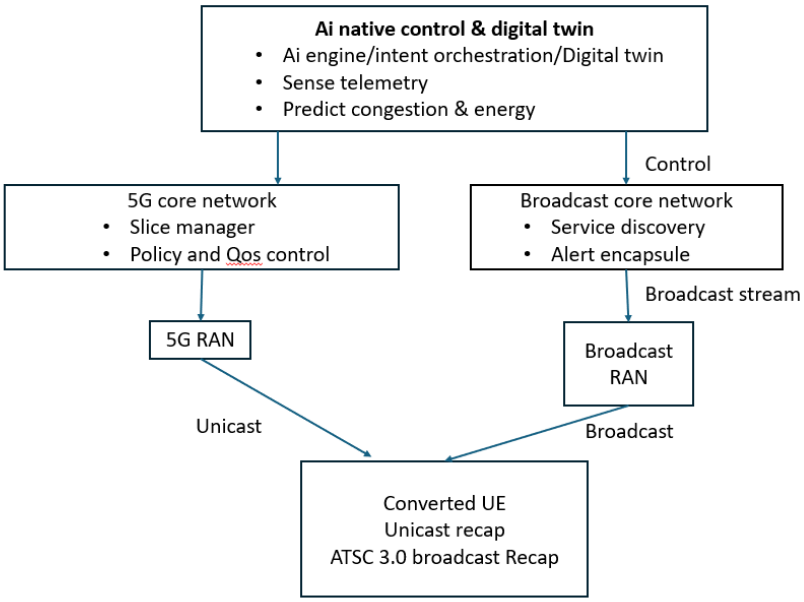


Fig.1: AI-Native Emergency Communication System Framework

11.9.9. Demo and Evaluation

The PoC is demonstrated using an AI-Native Control Plane integrated with a real-time Streamlit dashboard, validating network resilience under emergency congestion. It enables live monitoring of KPIs such as load, latency, and throughput, along with AI confidence levels, while executing a closed-loop "sense–predict–decide" workflow using LSTM-based congestion prediction. The system dynamically reconfigures the network, transitioning from unicast to hybrid broadcast (ATSC 3.0), with real-time visualization, decision logs, and manual override controls for validation. Performance is evaluated using key metrics, including sub-500 ms decision latency, significant congestion relief through traffic offload, and preservation of uplink capacity for mission-critical communication.

11.9.10. PoC Observation and discussions

The PoC identified key system-level refinements during development. Digital Twin synchronization issues due to asynchronous operations were resolved using atomic file handling, ensuring consistent and reliable telemetry across components. To meet safety-critical requirements, the control loop was optimized to a 0.2-second cycle, achieving sub-500 ms decision latency and enabling rapid response to congestion events. Resilience was validated by offloading traffic to ATSC 3.0, preserving uplink performance for priority users even under heavy load. Additionally, a hybrid AI approach was adopted, combining deterministic rule-based decisions for reliability with LSTM-based predictive insights for early warning, enabling both stability and future scalability.

11.9.11. Conclusion

This Proof of Concept successfully demonstrates an AI-native, closed-loop network control framework that intelligently integrates IMT-2020 unicast cellular networks with non-IMT broadcast infrastructure for resilient emergency communications. By leveraging digital twin–assisted decision-making and intent-based orchestration, the system proactively prevents congestion collapse and ensures timely, zone-specific mass alert dissemination. The experimental results confirm improved network resilience, efficient resource utilization, and guaranteed continuity of mission-critical uplink services for first responders, illustrating a practical, validated pathway toward the AI-driven convergence of cellular and broadcast networks in future 6G systems.

11.9.12. Future Work

Future work focuses on advancing the PoC toward real-world deployment and scalability. This includes integrating physical ATSC 3.0 transmission (PLP-level control) for true broadcast–cellular convergence, extending the AI orchestrator to support distributed, multi-UPF environments with edge-based inference for lower latency, and enabling online learning to continuously retrain models using live telemetry. Additionally, incorporating Explainable AI (XAI) will enhance transparency in safety-critical decisions, supporting human-in-the-loop validation and operational trust.

12. Combined Learnings from Build-a-thon 1.0 to Build-a-thon 4.0 Mentoring Sessions

The mentoring sessions across all participating PoCs revealed several critical themes that are essential for maturing AI-native network solutions and aligning them with real-world deployment constraints and research expectations.

Below is a summary of technical observations into actionable learnings that cut across AI modeling, system design, operational integration, and future alignment with standards and scalable architectures.

1) Clarity of Data Structures and Model Inputs

Across PoCs, there was a clear need to explicitly define data schemas—such as packet-level fields (e.g., timestamp, protocol, source/destination ports, packet length), CLI command parameters, JSON structures from simulators like NS3, and YAML templates for healing workflows. Projects were encouraged to demonstrate how their systems handle dynamic or unknown data formats, particularly when supporting real-time, user-driven scenarios. The inclusion of sample data formats, validation pipelines, and data parsing techniques is essential to communicate system robustness.

2) AI Model Selection, Integration, and Hosting

PoCs showcased diverse AI techniques (DQN, LSTM, Bi-LSTM, LLM agents), but mentors emphasized the importance of explaining where and how models are hosted, trained, and invoked. For example, if

an LSTM model is used in Open5GS or if a DQN selects vendors, it must be clear what inputs it takes, what outputs it produces, and what network state or features are required. Some teams were advised to explore inference-as-a-service models, semantic reasoning, and hosting in CDN or MEC layers, depending on the use case (e.g., real-time healthcare, cooperative ISAC, fault healing).

3) System Interfaces and Layer-wise Architecture

A common gap was insufficient detail on how different components interact—such as how the application layer sends requirements to the AI model, how the AI layer makes decisions, and how the network layer enforces them. PoCs were urged to present layer-wise diagrams, interface schemas (e.g., REST, gRPC, NeIF, Os-Ma-Nfvo), and detailed before/after workflows to clarify orchestration, healing, or monitoring actions. For example, CLI execution workflows need to specify how changes are determined, mapped, and validated, what base models are used for intent parsing, and how domain/network knowledge is structured.

4) Closed-loop Automation and Fault Management

Several PoCs (especially related to slice management and healing agents) were encouraged to demonstrate closed-loop feedback systems with monitoring, reasoning, and actuation cycles. Projects like healorch_agent.py and others implementing NS3 simulation with real-time anomaly detection were advised to show clear workflows: from symptom detection → anomaly alert → healing plan generation → config action via CLI/YAML. Further, creating fault isolation logic, defining KPI thresholds, and developing domain-specific knowledge bases (KBs) were considered essential for making systems operationally meaningful.

5) Security, Credentials, and Execution Environment

For PoCs that execute network configuration (e.g., CLI push to VyOS/Cisco/Juniper), security considerations like credential management, access control, and trusted execution environments were largely underdefined. Teams were reminded to specify how login, authentication, and permission models are managed, especially in remote or multi-tenant environments.

6) Use Case Depth and Deployment Focus

Mentors emphasized the need for realistic scoping—particularly for PoCs targeting specific deployment contexts like Delhi. Projects were advised to focus on one use case end-to-end (e.g., human activity detection, slice resource scaling, ISAC for emergencies) and map the application's needs to network slice parameters, AI inference points, and orchestration enhancements. Clear documentation of what will be implemented, what is out of scope, and what AI technique is demonstrated helps in evaluation.

7) Knowledge Base and ITU Standards Alignment

Knowledge curation and sharing were recognized as a foundation for system scalability and remote team collaboration. The idea of building a central knowledge base—including CLI templates, known faults and fixes, model versions, RAG schemas, and standard mappings—was strongly recommended. Additionally, PoCs were encouraged to conduct gap analyses against ITU standards and public datasets (e.g., traffic datasets) to align with international frameworks and contribute back insights.

8) Advanced Themes: Semantic Communication, Federated Learning

Emerging techniques like semantic communications, federated learning, and mixture of experts were proposed as future enhancements. These approaches could support model collaboration across domains, reduce redundancy, and enable privacy-preserving analytics. For example, Cellucast and SPV were guided to explore semantic verification, regional inference, and RAN-aware learning, provided interface and resource requirements from the network are clearly defined.

13. Bibliography

[b-Shah] Shah. P et al. “Topics and sample submissions for Build-a-thon 2025”, ITU Focus Group on AI Native for Telecommunication Networks (FGAINN) https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-094-R1.docx.

[b-FG-AINN-O-08] Xiaomi An et al., “Concept building for artificial intelligence native for telecommunication networks”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/output/FG-AINN-O-08.zip.

[b-FG-AINN-I-104_R1] Priyadarsini K et al., “AI-Native Proactive Network Slice Marketplace for Dynamic Service Ecosystems”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-104-R1.docx.

[b-FG-AINN-I-136] A. Israel et al. “Build-a-thon 2025: Using AI to Reduce the 6G Standards Barrier for African Contributors”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-136.docx.

[b-FG-AINN-I-128] C.Varshah and N. Chakraborty “Build-a-thon 2025: AIONETx”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-128.docx.

[b-FG-AINN-I-133] N. P. Mohare et al. “Build-a-thon 2025: NETSPEAK”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-133.docx. truth

[b-FG-AINN-I-135] S. Azim et al. “Build-a-thon 2025: 5G Adaptive Signal Solutions for Rapid Transit and Aerial Operations”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-135.docx.

[b-FG-AINN-I-130] A.Nishtala et al. “Build-a-thon 2025: Explainable AI-Native Intent-Based Self-Healing Network Orchestrator for Rural and Edge Telecom Infrastructure”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-130.docx.

[b-FG-AINN-I-132] K. Singh et al. “Build-a-thon 2025: Invisible Guard – Passive Wi-Fi Sensing for Smart Border & Building Surveillance”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-132.docx.

[b-FG-AINN-I-131] S. S. ZABEEN et al. “Build-a-thon 2025: SRM UNIVERSITY-AP”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-131.docx.

[b-FG-AINN-I-134] Bhavana et al. “Build-a-thon 2025: Truth Shield: Real-Time AI-Powered Fake News Control System”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-134.docx.

[b-FG-AINN-I-129] Nethravathi K A et al. “Build-a-thon 2025: Intelligent Network Traffic Capacity Prediction for BTS”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-129.docx.

[b-FG-AINN-I-155] A. Nishtala et al. “Build-a-thon 2025: Explainable AI-Native Intent-Based Self-Healing Network Orchestrator for Rural and Edge Telecom Infrastructure”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-155.docx.

[b-FG-AINN-I-156] S. Azim et al. “Build-a-thon 2025: 5G Adaptive Signal Solutions for Rapid Transit and Aerial Operations” , https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-156.docx.

[b-FG-AINN-I-157] A. Israel et al., “Bridging the Standardization Gap in Next-Generation Telecommunications”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-157.docx.

[b-FG-AINN-I-158] C. Varshah et al. “AIONET- AI-native Intent-Oriented Network Edge Tuner”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-158.docx.

[b-FG-AINN-I-193] S. Harsh et al. “AI-Native Autonomous Agent for Real-Time 6G Network Security, Optimization, and Slice Assurance”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-193.docx.

[b-FG-AINN-I-194] N. Anirudh wt al. “Build-a-thon 3.0 2025: Autonomous Energy Aware Self Optimizing Radio Access Network (Auto EA-SORAN)”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-194.docx.

[b-FG-AINN-I-195] C. Varshah et al. “Build-a-thon 2025: AIONETx”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-195.docx.

[b-FG-AINN-I-197] Kaaviya E et al. “Build-a-thon 2025: Machine Learning Based Adaptive HARQ Scheduling for 5G IoT in Non-Terrestrial Networks”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-197.docx

[b-FG-AINN-I-198] Alao Eniola et al. “A Standardization Knowledge Base Platform for AI-Native Networks: Bridging the Standards Gap in 5/6G Developmen”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-198.docx

[b-FG-AINN-I-214] M Srinivas et al. “Context aware Smart Datacasting”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-214.docx

[b-FG-AINN-I-215] Charan Velavan et al. “ AI-Driven Converged 6G Network for Scalable UAV swarm Coordination for Smart Cities ”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-215.docx.

[b-FG-AINN-I-216] C. Varshah et al. “ AI-Driven Mass OTA (Over-the-Air) Updates for Connected Vehicles ”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-216.docx

[b-FG-AINN-I-217] Anshul Katti et al. “ AI-enabled smart data traffic switching: Unicast vs Multicast/Broadcast ”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-217.docx

[b-FG-AINN-I-218] Manas Sharma et al. “ AI-Native Broadcast Intelligence Plane for Cognitive ATSC 3.0 Networks”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-218.docx

[b-FG-AINN-I-219] Anirudh N et al. “ Intent-Driven AI-Native Precise Positioning-as-a-Service ”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-219.docx.

[b-FG-AINN-I-220] Shaik S Z et al. “ AI-Native Emergency Communication using Smart Multicast and O-RAN QoS Control ”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-220.docx.

[b-FG-AINN-I-221] Ajitha Kamali P et al. “AI-Enabled Broadcast Core for Ubiquitous Connectivity

”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-221.docx.

[b-FG-AINN-I-222] Supriya M H et al. “ Digital-Twin IntelliCast: AI-Native Smart Emergency Broadcast Orchestration”, https://extranet.itu.int/sites/itu-t/focusgroups/ainn/input/FG-AINN-I-222.docx