Quality of perception prediction in 5G slices for e-Health services using user-perceived QoS

Yosra Ben Slimen[1], Joanna Balcerzak[1], Albert Pagès[2], Fernando Agraz[2], Salvatore Spadaro[2], Konstantinos Koutsopoulos [3], Mustafa Al-Bado[4], Thuy Truong [4], Pietro G. Giardina[5], Giacomo Bernini[5]

[1] *Orange Labs, Chatillon, France*

[2] *Universiat Politècnica de Catalunya (UPC), Barcelona, Spain*

[3] *Creative Systems Engineering, Athens, Greece*

[4] *Dell EMC, Cork, Ireland*

[5] *Nextworks, Pisa, Italy*

Quality of perception prediction in 5G slices for e-Health services using user-perceived QoS

Yosra Ben Slimen[1], Joanna Balcerzak[1], Albert Pagès[2], Fernando Agraz[2], Salvatore Spadaro[2], Konstantinos Koutsopoulos [3], Mustafa Al-Bado[4], Thuy Truong [4], Pietro G. Giardina[5], Giacomo Bernini[5]

[1] *Orange Labs, Chatillon, France*

[2] *Universiat Politècnica de Catalunya (UPC), Barcelona, Spain*

[3] *Creative Systems Engineering, Athens, Greece*

[4] *Dell EMC, Cork, Ireland*

[5] *Nextworks, Pisa, Italy*

Abstract

In order to compete for a prominent market share, network operators and service providers should retain and increase the verticals' subscription, catering to their needs in order to differentiate themselves from competitors. In this scenario, verticals' satisfaction arises of paramount importance. As such, user experience is becoming a reliable indicator for service providers and telecommunication operators to convey overall end-to-end system functioning. To properly estimate end user satisfaction, operators and service providers require efficient means for quality monitoring and estimation at all layers, in conjunction with mechanisms able to maintain said quality at optimum levels. Given these factors, this paper proposes a mechanism for Quality of Perception (QoP) estimation in e-Health services, enabling the QoP-aware management of network slices fulfilling the requirements of supported services. To this end, the paper proposes a cognitive-based architecture which allows for the collection and monitoring of verticals' data to estimate QoP and provides mechanisms to re-configure the underlying network slices according to the monitored quality levels. A machine learning (ML) model is introduced that aims to forecast any future degradation in the quality perceived by vertical users. In case of a predicted degradation, the proposed architecture reacts and triggers the necessary remedial actions, referred as actuations. In order to evaluate the developed ML model and to showcase the

interaction between the different components of the proposed architecture, an experimental study is presented with real data extracted from a roaming ambulance. In addition, a Proof of Concept of the actuation mechanism is demonstrated through an experimental testbed emulating e-Health services.

Keywords: 5G network slicing, quality of experience, machine learning, network cognitive management.

---

1. Introduction

5G subscriptions are predicted to reach 1.9 billion by 2025 where 35% of traffic will be carried by 5G networks [1]. The novelty of 5G technology is the concept of network slicing that should allow an increased speed, performance, scalability, and flexible service deployment. According to [2], a network slice (NS) consists of physical and/or virtual network functions (PNF/VNF) that can belong to access and core network part. Then, these network functions are interconnected by means of network resources (what is known as chaining), composing a synthetic infrastructure with specific characteristic, both in functionalities and resource capacities. The synthesis of a NS, then, serves a particular functional purpose and once instantiated, it is used to support certain communication services, which ultimately are deployed to support vertical services on top.

E-Health is identified as one of the key vertical sectors to which 5G infrastructures should focus in providing solutions for its effective materialization. In particular, 5G could provide support for stroke diagnosis through the provisioning of dedicated NSs. This would help a trained paramedic team to assist stroke patients with the cooperation of medical personnel (e.g. a doctor) present in remote facilities, such as hospitals.

To date, significant progress has been made on in-hospital stroke management, but a reliable pre-hospital in-ambulance solution has not yet been established. Hence, solving this problem could offer life-saving time gains and speed up treatment initiation by early activation of the in-hospital stroke response, thereby curtailing the risk of misdiagnosis and death. Generalised in-ambulance tele-medicine is a recently developed and promising approach, which is under the umbrella of the more generic e-Health vertical industry.

Following 5G service classification, e-Health services belong to the Enhanced Mobile Broadband (eMMB) type of services defined by 5G standards. These type of services require the provisioning of significant network capac-

3

ities in a highly mobile and distributed geographical area. E-Health has as added requirement strict Quality of Experience (QoE) service levels, which ultimately may have an impact on the well-being of stroke patients.

The targeted scenario in this paper begins with the continuous collection and streaming of patient data when the emergency ambulance paramedics arrive at the incident scene. Several Internet of Things (IoT) devices (e.g. wearables, cameras) enable the provision of enhanced patient insights. The goal is for all paramedics to have wearable clothing that can provide real-time video feeds as well as other sensor related to data pertaining to the immediate environment. Figure 1 summarizes the scenario in hand.
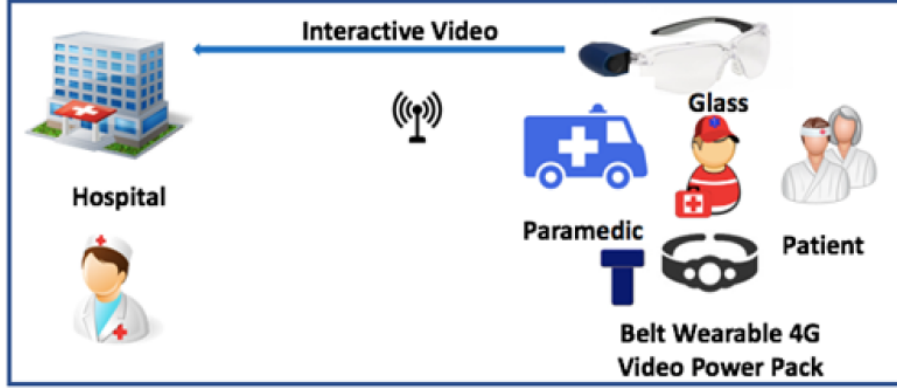

Figure 1: Summary of e-Health use case.

Under such scenario, the network should provide efficient communication capabilities that enable storing and potential real-time streaming of patient data to the medical team inside the ambulance. This can enable more intelligent decision support for the paramedics attending the patient. Therefore, detecting and resolving e-Health NSs' operational problems is of a paramount importance in order to minimise the downtime of slice operations and to provide a good QoE.

Quality of Experience has been defined by [3] as the degree of delight or annoyance of the user of an application or service. Nevertheless, a good quality of service does not guarantee the satisfaction of the customers. As defined in [4], the Quality of Service (QoS) is the characteristics of a telecommunications service that bears on its ability to satisfy stated and implied needs of the user of the service (e.g., packet loss, latency, etc.). Using these two definitions, it is clear that the customer experience matters the most in QoE.

In the context of network slicing, measuring customer experience (i.e. slice user) is a challenging problem for verticals because it is costly and complex due to the human involvement in the process. The challenging issue is mainly the subjectivity of this measurement. First, applications are of different importance to different users and they also have diversified network-level QoS requirements. Second, different application-level QoS performances bring different effects on user QoE. Third, many contributing factors of QoE change over time so they are highly time-variant. Therefore, it is inadvisable and far from reality to measure the QoE once forever.

In addition to that, although a good QoS does not imply a good QoE, the impact of network QoS parameters, such as delay and packet loss rate are considered as the contributing factors that influence the user perceived QoE. For example, user satisfaction can be expressed by reducing the traffic delay that can cause dropping of sessions. Thus, the network QoS parameters (i. e. bandwidth, delay, packet loss rate, etc.) are regulating the user perceived quality experience.

In this paper, we utilize a more objective user-centric approach that computes the User-perceived QoS also called 'the Quality of Perception' (QoP). We aim to provide an architecture that provides a top QoP to the vertical. A machine learning (ML) model is presented in order to analyse the vertical's QoP and to forecast the degradation of perceived signal strength quality that may arise in the future 5 minutes. In case of any predicted degradation, an alert is sent to an actuation framework that will react accordingly in order to correct the problem before it occurs.

The remaining of this paper is structured as follows: Section 2 provides a background review for the main topics targeted by the presented work, namely, cognitive management of NSs and quality estimation in network services, while highlighting the novel aspects introduced with the current proposal. Section 3 introduces the proposed architecture. Section 4 elaborates the e-Health use case and the data collection mechanism. Section 5 details the supervised forecasting ML model for QoP degradations. Section 6 introduces the actuation framework. Section 7 presents the testbed, along the numerical experiments that aim to evaluate the performance of the ML model as well as the details of the actuation mechanism in order to correct potential degradation of the measured slice QoP. Finally, Section 8 elaborates the conclusions and future works.

2. Background

Network slicing is a very hot topic in the research and industrial communities, due to the many benefits that it brings. In this regard, the cognitive management of network slices is seen as a promising candidate to the automated and autonomous fulfillment of vertical services requirements, such as e-Health. Thus, several works can be found tackling the multiple aspects of slice management in 5G shared infrastructures. In this section, we review some of the work found in the literature regarding network slicing in 5G, cognitive network management and quality estimation for vertical services to provide the reader with an understanding of the background in which the presented work is enclosed in.

2.1. Network slicing in 5G

Network slicing in the context of 5G is a concept that has been introduced by the Next Generation Mobile Network (NGMN) Alliance [5]. Network slicing consists on the partitioning of the underlying physical infrastructure, network elements (both physical and virtual) as well as functions to enable the deployment of self-contained networks, each one tailored to the specific needs of the services that will run on top.

Such a concept enables multi-tenancy in multi-domain/role/technology 5G infrastructures, in which all involved infrastructure segments are able to create isolated network slices towards the fulfillment of end-to-end (E2E) synthetic infrastructures, each one managed and controlled independently from the dynamic of the underlying substrate.

For instance, the concept of network slicing is described in [6]. Slice isolation is one of the important properties that should be guaranteed in performing network slicing and is one of the changes in relation to QoS. A network slice instance (slice activated from template) could be isolated from another network slice instance in several ways, e.g., full or partial isolation and logical or physical isolation.

As such, several works has tackled efficient ways to facilitate slicing in the multiple segments that constitute the 5G ecosystem. Starting from the Radio Access Network (RAN) segment, works have focused on the challenges posed by the wireless nature of the RAN medium, in order to effectively slice the RAN spectrum and provide slices with the requested resources. For instance, authors in [7] focus on the resource assignment aspect of RAN slices,

proposing allocation algorithms that coordinate both the resource assignment to slice instances as well as the sharing across provisioned slices.

Another critical segment is the Mobile Edge Computing (MEC), since it allows the deployment and execution of network functions close to the endpoints, reducing drastically the latency for communication services that are delay sensitive (e.g. e-Health). Thus, it is essential to devise architectures that facilitate the slicing and allocation of functions in MEC sites. To this end, authors in [8] present an architecture specially tailored to the challenges posed by network slicing in MEC segments, with the aim of minimizing communication latencies and signalling overheads for IoT services. These works, among others focused in other segments (e.g. Core, optical transport, multi-domain), provide the necessary foundations that enable the realization of 5G slices.

The multiple infrastructure segments involved onto the delivery of E2E services need to be sliced in a coordinated way in order to exploit all the technologies present at the data plane and satisfy the requirements posed by vertical applications. Such a task requires the translation of the service requirements into concrete E2E NS characteristics, to be further decomposed into specific slice instances deployed at the several segments. The deployment of such slices requires, in turn, the orchestration and configuration of multiple types of resources (e.g. networking, computing) [9]. As such, several works have targeted the development of architectures that allow for the efficient and automatic deployment of slices in multi-segment/domain scenarios.

On this regard, the work presented in this article is enclosed within an architectural framework proposed in the context of the 5GPPP Phase II SliceNet project (see Section 3 for further details). The novelty of the architectures relays on its multi-role structure, which enables network and infrastructures operators with the capacity to holistically provision E2E slices in a multi-domain environment. The modularity of the architecture allows for it to be instantiated at multiple scopes, thus, functionalities can be opt-in/out as required for the entity in charge of creating/managing the slices. Such an approach further progresses the network slicing concept, providing a framework to flexibly slice a multi-technological network infrastructure, encompassing RAN, MEC, core and transport resources.

As an illustration to standardized network management solutions devoted to the quality-aware runtime management of network slices, it is necessary to mention the Network Data Analytics Function (NWDAF) framework and Management Data Analytics Service (MDAS). As defined in [10] MDAS an-

alytics can be used for analysis e.g. resource usage prediction or network failure for a network slice. NWDAF, as described in [11], [12], can provide analytics functions related to the mobility, Slice load level, observed service experience but also including QoS sustainability which involves reporting and predicting QoS change for a specific period in future, in certain areas.

2.2. Cognitive network management

Aside from the aspects related to enable slicing in 5G networks from an infrastructure perspective, another important topic relates to the management of said slices. Due to the sheer scale of 5G systems, the dynamicity, heterogeneity and volatility of the supported traffic profiles, as well as the added complexity of softwarized/virtualized infrastructures, traditional management solutions are no longer suitable for the management of network slices. As such, novel management solutions should be capable to automatically and autonomously adapt to changes or unforeseen states on the underlying infrastructure and supported services.

In this regard, cognition-based management, focused on Artificial Intelligence (AI)/ML procedures, is gaining momentum as a solution to overcome the challenges posed by network slice management in 5G environment. The importance of investigating AI/ML-based network management for 5G networks has been recognized by several large European Union projects under the 5G PPP in recent years. During Phase I of the 5GPPP programme, several projects addressed the broad concept of Self Organizing Networks (SONs), in which ML/cognition-aided network management is enclosed. In particular, three projects dealt specifically with network management, being Sesame, SelfNet and CogNet.

Starting with Sesame [13] this project has been devoted to the management of multi-tenant small radio cells. It defines a Cloud Enabled Small Cell (CESC) that combines the capabilities of traditional physical SCs with data centre (DC) capabilities to enhance the virtualization of the PNFs at the network edge. Then, a CESC Manager (CESCM) entity is introduced to allow for the management of the cells and its elements. To enable a flexible management, the CESCM integrates several SON capabilities, mainly Service Level Agreement (SLA) monitoring, to stimulate the multiple Element Management Systems (EMSs), which will operate autonomously the underlying network functions with the goal of SLA maintenance. Although Sesame follows the SON approach for self-operation of the network, it does not employ ML/cognition techniques to gain insights on the monitored SLAs

and help to make smart decisions. The presented work in this paper does follow the SON principles, in which several elements automate the operation of the underlying physical/virtual network. In addition, as it will be explained later, it incorporates several ML/cognition-based elements, i.e. the Cognition Sub-Plane, to enhance the SON capabilities, gaining better insights about the network state and providing proper context for the automated decisions.

Both projects SelfNet [14] and CogNet [15] incorporate ML capabilities to enhance the management of the underlying infrastructure. In particular, SelfNet is devoted to the management of key network technologies for 5G systems, including Software Defined Networking (SDN) and Network Function Virtualization (NFV)-enabled networks, with a special focus on SONs for self-monitoring, self-optimisation, self-protection and self-healing. Both its control and orchestration components are enriched with some analysis capabilities (based on ML) that allow for a better self-control/management of the network and data layers. Nevertheless, SelfNet is not fully dedicated to the cognitive management of networks. In this regard, CogNet is totally dedicated to the cognitive management of network infrastructures, with a special emphasis on the NFV aspect of the underlying substrate. To this end, CogNet defines a complete ML-based framework that provides monitoring, analysis and actuation capabilities to influence on the operation of an NFV infrastructure (NFVI). The main difference respect with the presented proposal in this work is that these projects focused on the generic case of cognitive management of SONs/NFVIs, while in our solution, we focus on the quality-aware management of 5G networks, which requires specialized analysis and actuation components to provide the necessary context. Beyond these aspects, the current work is devoted to the E2E cognitive management of NSs build on top of a 5G system, including both networking and NFV aspects. Not only that, the presented architecture takes special care to incorporate the slice/vertical context in its analysis and actuations so as to fully provide a quality-aware management of NSs.

Aside from research and industrial projects, cognitive and data-driven management of network infrastructures has gathered the interest of the main standardization bodies. As such several standardization initiatives have been developed, in order to set up the principles for which framework exploiting ML/AI for the management of complex network infrastructures should be guided. The ETSI Experiential Networked Intelligence (ENI) Industry Specification Group (ISG) focuses on the specification of a cognitive network management system for the optimisation and adjustment of network operator

9

systems over the time based on metrics extracted from the underlying network system. Given the current trends on both research and industry, where technologies such as SDN, NFV and slicing are of paramount importance, the ENI ISG proposes an architecture that adopts a "observe-orient-decide-act" approach to guide an assisted system and influence in its behaviours by providing recommendations, for example, in the form of policies [16]. ENI's approach advocates for a declarative approach, in which the recommendation block suggests changes on policies or system states, which are then left to the discretion of the assisted system. On the other hand, the presented approach in this paper, engages with the assisted system in a more imperative way. More specifically, the QoE Optimizer module (see Section 3), which is responsible for determining the remedial actions once quality violations occurs or undesired situations are detected, contacts the assisted system (the Orchestration Plane) with the exact action that needs to be carried out, leaving the specific details on how to achieve this to the logic of the assisted system. This difference, although subtle, allows our proposal to be more focused on its goals, which is the quality-aware management of NSs as offered to vertical clients, and, for the particular scenario of this paper, is the QoP-based management of eHealth slices/services.

The NGMN also defines as one of the key requirements for 5G systems the concept of Autonomic Networking (AuN). Given the complexity of 5G systems, in which multiple network technologies and service requirements coexist, it is essential to push the concept of AuN from an E2E perspective for automating the configuration, management, operation and awareness of 5G systems. For this, in [17], NGMN introduces the presence of Autonomic Management and Control (AMC) in the overall E2E 5G network architecture. Similar to ENI, it proposes a framework in which an underlying network infrastructure layer is governed by another entity that provides insights about the overall system and then enforces actions to the infrastructure layer. In this regard, NGMN divides the said entity into two different roles. First, an Automated Management layer, in which reasoning, learning and adaption procedures take place. Then, the created insights are distributed to an AMC layer, which focuses on the efficient implementation and automation of configuration and monitoring workflows. The key elements of the AMC layer are the network Decision Entities (DEs), which, given the inputs disseminated from the Automated Management layer (e.g. policies), autonomously take decisions for self-configuration, self-healing, self-diagnosis, etc. of the network layer.

In addition, there are several works in literature that target different facets of cognitive management in softwarized network environments. Focusing on the aspect of media delivery in 5G networks, authors in [18] propose a system that leverages on ML for a network resource allocation system as the main contribution which enables autonomous network management with quality-awareness. As a summary of the application of ML/AI techniques in SDN/NFV environments, authors in [19] review the challenges and opportunities of ML/AI in softwarized network environments, with a special focus on data-driven decision making for management and control of SDN/NFV-based infrastructures. To this end, the authors propose to enhance the functional primitives of monitoring, composition and control with ML modules. All these works, among many others, pave the path towards cognitive-aided management of networks, specially in the presence of SDN/NFV technologies.

Given all these bases, the present work expands the architectural proposals of past projects, current standards and works found on the literature by going beyond a simple network infrastructure but encompassing the E2E aspect of the service delivery, by tackling the cognitive management of NSs. Moreover, a key aspect of the developed solution is the "verticals-in-the-loop" philosophy adopted, which makes the verticals an integral part of the full system. Due to that, the multiple analysis functions of the Cognition Sub-Plane are specialized for the quality monitoring/analysis of verticals' slices, with properly curated monitored data to reflect the requirements of the managed slices.

2.3. Quality of Experience/Service estimation

Due to the more user oriented nature of 5G networks, a concept that has been named as "verticals in the loop", quality assurance becomes primordial for an optimal delivery of services towards the vertical customers. While network level QoS is quite easy to monitor/estimate, with a plethora of mechanism and strategies to guaranty good QoS levels, the quality perceived by end users is more challenging to determine and optimize, since it involves the subjective perception of human actors about the performance of a delivered services.

As such, several works in literature propose strategies in support of quality estimation and optimization at all levels, ranging from systems that aid on decision taking for optimal service provisioning, to QoE estimation in media/network services, passing through works related to network-level quality

monitoring and estimation.

For instance, authors in [20] propose a novel NFV orchestration solution that employs ML techniques to enhance the Quality of Decision (QoD) of an NFV-based system when orchestrating and provisioning services. The objective is to enhance NFV systems to achieve a near-optimal placement of virtualized network functions at minimum monitoring costs.

Focused on the estimation of QoS in cloud environments, authors in [21] review the main open questions about QoS attributes modeling and prediction in cloud services and showcase the limitations posed by statistics-based techniques in their forecasting. In this regard, authors discuss how ML-based techniques arise as a superior alternative in the presence of dynamic workloads for proper QoS estimation.

In the aspect of QoE estimation, the main line of research tackled in the literature is QoE estimation/prediction in video delivery services. Indeed, video delivery is a primordial service in nowadays telecom/cloud ecosystem and economic activity, encompassing both consumer grade services (e.g. video on demand, video chat) or business grade services, as it would be the case of e-Health. Thus, several works explore the main challenges of this aspect and propose several solutions to overcome them. The main common denominator is the use of ML/cognitive-aided solution, since the own nature of QoE requires of mechanisms that are able to learn what is the perceived quality from a user perspective. As an example, the work presented in [22] employs ML in order to enhance traditional RAN bases stations with QoE awareness, improving the delivery of video streams in current and future Long Term Evolution (LTE) radio networks. In general, QoE management for video delivery services is an E2E process, which requires not only QoS management at network level but also real time monitoring and control of end user QoE. Therefore, traditional strategies relying on statistical or reactive measures fall short in addressing properly the posed challenges. In this regard, the work presented in [23] reviews the most significant QoE management methods in video services, pinpointing to the most effective ML methods employed in the literature in this regard.

Few works have focused on the QoE estimation and management-based in network slices, due to its challenging aspects. Quality estimation has to be done considering the context of the slice as a whole, which may span several resources distributed across far apart infrastructure segments. Thus, multiple data sources have to be considered, which usually are owned by different entities, thus adding the complexity of data sharing/anonymization. Thus,

proper solutions need to be derived. For example, authors in [24] provided a Proof of Concept (PoC) for optimal NS placement for latency stringent gaming use cases. The presented framework considers user location as a primary constraint to ensure QoS and QoE. Following similar principles, the present work employs several radio channel quality indicators in order to derive the QoP for an eHealth E2E slice, with the aim to predict quality degradations ahead of time and apply corrective measures. The novel aspect of the work resides on the fact that the ML/cognitive artifacts are made an integral part of the whole NS infrastructure. That is, each of the deployed NSs is equipped with a QoP prediction ML model by definition, which acts as an advanced sensor for the slices, which then guide the whole orchestration/management loops of the overall system in keeping up a healthy status for deployed slices. Such an approach runs away from proposals in which quality estimation is performed as another aspect of the control/management loops. In the presented proposal, during the definition of slices, based on the service requirements, the system also derives and defines a set of cognitive functions particularized to the type of slice and the service supported on top, which span from ML models, as the currently presented QoP estimation one, to actuation policies.

3. Overview of the cognitive management architecture

In this section, we present a management architecture which aims on enabling 5G systems with cognitive-based management of NSs in support of advanced e-Health vertical services, with quality guarantees assurance being the main pilar in which the architecture is built. The presented architecture rests within the framework of the 5GPPP Phase II SliceNet project [25]. SliceNet aims to provision NSs with QoE guarantees to support vertical industries. As such, it proposes a management/control architecture devoted to QoE-aware life-cycle management of 5G slices. Figure 2 depicts a schematic of the architecture, putting emphasis on the functionalities and roles that are detailed in the remaining of this section.

The depicted architecture considers a multi-role environment, in which several stakeholders with clear responsibilities engage for the delivery of 5G NSs. The following sub-sections provide more details about the key components and entities involved.
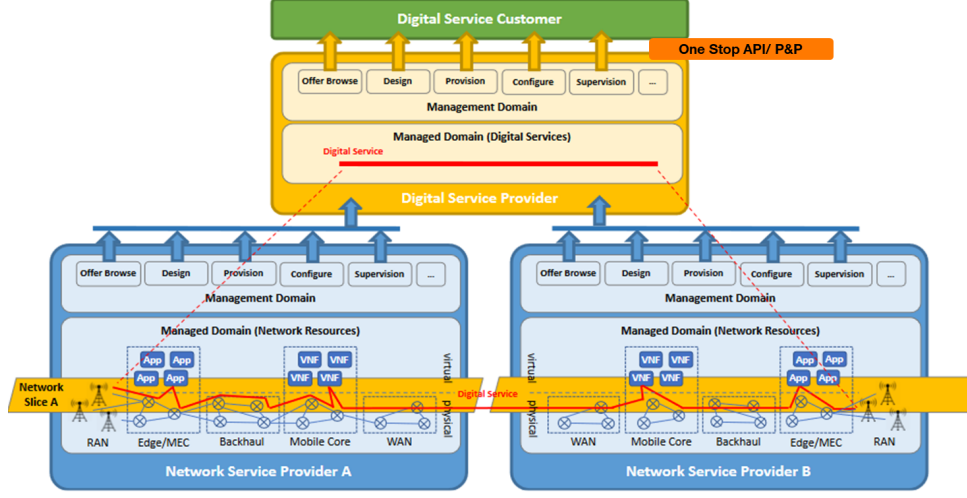
Figure 2: Overview of SliceNet architecture and stakeholders relationships.

3.1. Network Service Providers

A network service provider (NSP) is the owner of the physical infrastructure. Multiple NSPs operate under a "network slice as a service" (NSaaS) business model, offering their infrastructure capabilities to an upper layer (i.e the Digital Service Provider). In this regard, the NSPs are responsible for the maintenance of the physical network infrastructure, which may encompass several segments, for example, RAN, MEC, core and NFV Points of Presence (such as data-centres).

In this regard, NSs deployed at NSP level serve the networking requirements that are needed within a single domain towards the composition of E2E services, i.e. a NS offered to vertical customers. To this end, the NSPs have in place control and management elements that enable the provisioning, configuration and monitoring of physical/virtual resources that constitute NSs. These NSs are bound by specific SLAs that guarantee the properties of the deployed slices towards the fulfilment of specific needs requested by upper layer entities.

3.2. Digital Service Providers

A Digital Service Provider (DSP) is responsible for the creation and management of E2E NSs to be offered in a NSaaS model towards vertical customers to satisfy the needs for supporting their associated services. Therefore, SliceNet architecture divides its functionalities between the multiple

roles (NSP/DSP). While single-domain control functionalities are basically assumed within the NSP role, the E2E multi-domain management functionalities are integrated within the DSP role.

This latter aims to translate the vertical service requirements towards specific network/infrastructure requirements, which then are translated onto a specific E2E NS to be fulfilled. The E2E NSs are then broken down into single domain NS which are materialized thanks to the capabilities offered by the multiple NSPs that have a business relationship with the DSP.

3.3. Access and FCAPS management

SliceNet adopts a vertical (i.e. slice user) centric approach that aims at abstracting technology and domain related enablers towards the provision of services tailored to the needs of the topmost user of the infrastructure. In this respect a Plug and Play (P&P) framework has been designed with the aim of provisioning vertical oriented exposure of the delivered slice.This requires that specially crafted control plugs are developed and available to be selected by verticals when requesting the provisioning of their slices. Each of the control plugs provides a specific customized northbound slice control view that fits the needs of the vertical. A dedicated slice P&P control instance (composed by one or more plugs) is automatically created as part of the slice provisioning process, and then exposed to the vertical. Each P&P control instance is able to process, at its southbound, the platform specific information as this is maintained via the layered and multi-domain Fault, Configuration, Accounting, Performance and Security (FCAPS) enablers. The FCAPS framework is also provided in an abstraction fashion starting from the pillar technologies per NSP domain and spanning up to inter-domain aggregations as administered by the DSP domains. The overall governance of the several domain enablers and abstractions is applied through the One Stop Application Programming Interface (OSA) that spans vertically from NSP to vertical. Besides, OSA provides role based views of the artifacts that have to be administered under a producer consumer relationship between overlayed roles.

In practice FCAPS allows NSPs to abstract their domain specific technologies with respect to actuation and sensing capabilities and expose these as offerings through the OSA towards DSPs. DSPs can consume the monitoring information as this is arranged per slice via the NSPs Data Lakes as well as invoke actuation information through the layered orchestration entities. DSPs maintain higher level slice counters and metrics, in their own

Data Lakes, as these have been made available through exposure functions instantiated per slice. At the level of DSP, Data Lake information is subject to be processed by Cognitive, QoE and P&P functionalities that have been activated per slice to deliver higher level and novel FCAPS strategies.

Beyond the bottom-up approach for the delivery of slice specific metrics and due to the vertical centric approach, verticals can contribute to the Data Lake information for the exploitation of service metrics that the user endpoints are able to produce. In this respect the vertical tailored user interface (UI) view provided by P&P and OSA is augmented to create slice specific endpoints that expose the existing P&P options towards more automated (UI less) procedures to contribute to the slice segment of the DSP Data Lake and trigger cognitive and QoE functions via direct quantitative vertical feedback. OSA allows therefore slice owners to indicate which P&P functions should be exposed as Representational State Transfer (REST) endpoints that can be invoked by metric and counter agents deployed in vertical user equipments (UEs) and collecting valuable metrics from the user domain.

3.4. Cognition sub-plane

It is precisely in the DSP role in which SliceNet defines a novel Cognition Sub-Plane, incorporating AI/ML capabilities, for the QoE-aware management of E2E NSs. The role of the Cognition Sub-Plane is dual. On one hand, it monitors the multiple metrics and events that are relevant to the quality of deployed slices. Given this information, it derives elaborated data that allows to predict the quality of slices thanks to specialized ML models tailored to the characteristics of NSs and vertical services. On the other hand, given the alerts generated by the models, the Cognition Sub-Plane is responsible to apply remedial actions so as to maintain optimal quality levels. This is done through a policy-driven actuation framework, which requests for specific (re-)configurations to selected NSP. It aims to modify the characteristics of the underlying single-domain NSs in order to have an optimal E2E NS able to provide the topmost quality levels towards vertical customers.

4. e-Health vertical use case

Following the architecture previously depicted in Figure 2, the E2E slice customer is a national/regional health service organization (in our case National Ambulance Service in Ireland), which operates multiple (static) hospitals, dispatch centres, and (moving) ambulances. The E2E e-Health slice

offered to the customer by the DSP consists initially of a "base" NS Instance (NSI) containing the minimal set of network functions and services. The base slice is fairly static and is centered around a geographical area in the vicinity of the hospital. The hospital/dispatch hosts experts who provide real-time support to the paramedics.

As ambulances are dispatched, additional Network Sub-Slice Instances (NSSIs) may be instantiated in order to increase the geographical coverage of the slice or to guarantee the latency and availability requirements of the slice. For the latter, additional processing functions may be dynamically instantiated at suitable MEC locations. Dispatch may trigger a handover of a paramedic's communication stream to a different hospital. Handover between domains might be needed while the ambulance is moving. The vertical customers only see services provided by their DSP.

For data collection, we installed G-Net Pro application in a Samsung phone and configured the application to capture the statistics for down-link traffic within 1s. Following the ambulance paths, the application collected the network statistic information including the timestamp, longitude, latitude, mobility speed, technology (e.g., 4G, 3G, WiFi), technology mode, bit-rate, cellID, serving cell, operator ID, etc. in multiple files. We then applied a data pre-processing program to extract the interesting fields for data training which is described in Table 1. In total, we have 12 data-sets, each data-set contains 2-3 hours statistic information.

Table 1: Collected QoP metrics from the ambulance

Feature	Description
Perceived Reference Signal Receive Power (RSRP)	Average power received from a reference signal
Perceived Reference Signal Receive Quality (RSRQ)	It indicates the quality of the received signal
Perceived Signal to Noise Ratio (SNR)	It indicates the signal strength relative to background noise
Perceived Channel Quality Indicator (CQI)	It indicates how good/bad the communication channel quality is
Perceived Received Signal Strength Indicator (RSSI)	It indicates the signal strength measured from all base stations
Perceived DL-bit-rate	Bit-rate of the download link
Perceived UL-bit-rate	Bit-rate of the upload link

As shown in Figure 3, the information collected from the UEs at the ambulance are fed towards the monitoring systems at the NSPs, since UEs are connected to a specific NS within the E2E NS, for which a concrete NSP is responsible. The collected information is then transferred to the DSP, more specifically, to the Anomaly Detection ML model that resides in the Analyser.
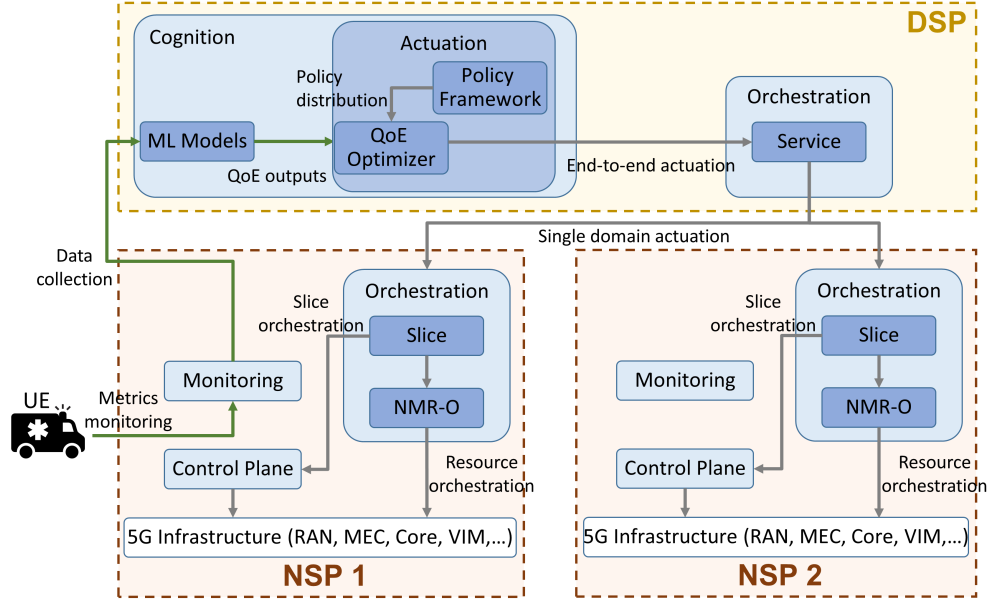

Figure 3: SliceNet e-Health use case actuation workflow.

This model acts as an intelligent sensor that observes the last 5 minutes of the QoP metrics of the NS supporting the e-Health services. The sensor will predict every 5 seconds if a degradation in the network signal strength in the RAN segment may be perceived by the vertical in the future 5 minutes as depicted in Figure 4. Based on the alerts generated by the model, the actuation framework reacts and triggers the necessary remedial actions.

Due to the mobility nature of ambulances, which host the UE end-points of the NS, RAN connectivity is of paramount importance, since it allows to connect the medical equipment to the NS regardless of the location as long as it is under the coverage of the provisioned E2E NS. A bad quality of signal will disturb the services running on top of the slice since it may cause a disruption of the communication between the paramedics, a degradation of the vertical quality of experience and a danger for the patient. As such, if
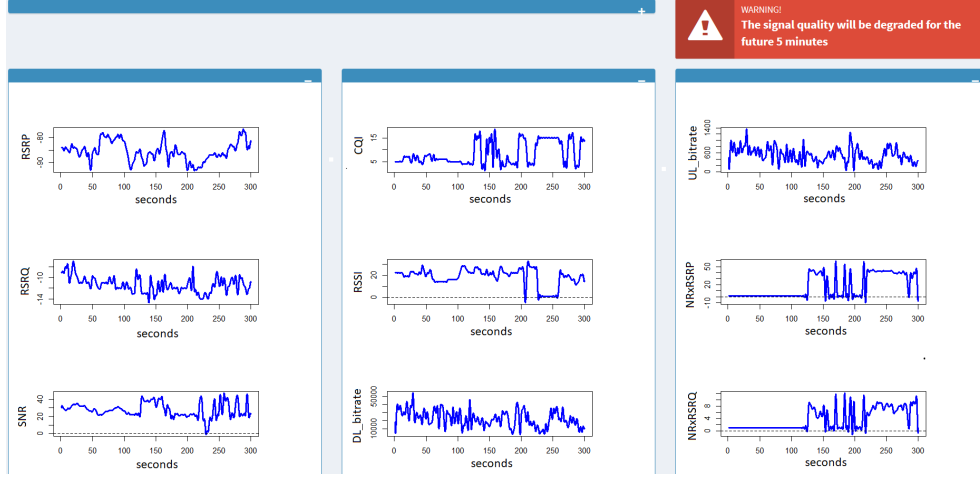
Figure 4: QoP prediction model objective

a bad signal quality is predicted, the ML model needs to alert the actuation framework that will trigger remedial actions.

Such a loop enables the real-time processing of the monitored UE data, its subsequent analysis and near-real-time actuation to overcome the predicted communication degradation. These requirements are essential in the e-health scenarios since these latter offer critical services that need a fast and up-to-date analysis and actuation to guarantee optimal quality levels.

Given the overall scenario and goal, the following sections proceeds on detailing the elaborated QoP prediction model and the actuation framework.

5. Quality of Perception prediction model

The data $\boldsymbol{X}$ under study are a set of n observations as illustrated in Figure 5. Each observation X_i corresponds to the curves of QoP metrics observed for the last 5 minutes. Hence, each observation is described by a set of p curves and a label: $\boldsymbol{X} = \{(\boldsymbol{X}_i(t), label_i)\}_{t \in [0,T], 1 \leq i \leq n}$ with $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))' \in \mathbb{R}^p$, $p \geq 1$ and $label_i$=1 if there is a signal degradation will be perceived by the vertical in the prediction horizon (the future 5 minutes) given $\boldsymbol{X}_i(t)$, 0 otherwise.

5.1. Imbalance management

Due to the imbalance in the class, the learning of the different patterns that are hidden in the data may be misled and it must be treated before the

| KPI 1	KPI 2	...	KPI p	Label
		...		1
		...		0
...	...	...	...	...
		...		1
...	...	...	...	...
		...		0
		...		1
...	...	...	...	...
		...		0

Figure 5: An illustration of the training set structure

training phase. For this reason, we use the SMOTE technique [26] that is a well-known approach for over-sampling the minority class in a data set. The over-sampling is achieved by generating "intelligent" copies of the minority class observations i.e. by artificially creating synthetic samples. The new examples are generated by using the nearest neighbors of these observations. It consists in perturbing one attribute at a time by a random amount within the difference to the neighboring instances. This approach effectively forces the decision region of the minority class to become more general.

5.2. Functional data extraction

In real life, the QoP metrics are extracted with discrete time points (every second). Therefore, a smoothing [27] is applied to every set of QoP metrics observed for the last 5 minutes in order to reflect their functional nature in a continuous time interval. Smoothing assumes that each observed curve x_{ij} ($1 \leq i \leq n$, $1 \leq j \leq p$) can be expressed as a linear combination of

basis functions $\{\phi_l\}_{\ell=1,\ldots,M}$: $x_{ij}(t) = \sum_{\ell=1}^{M} a_{ij\ell}\phi_l(t)$, $t \in [0,T]$, where $\{a_{ij\ell}\}_{\ell=1,\ldots,M}$ are the basis expansion coefficients. As a result, every KPI curve is a functional object that is represented by M values. These latter are the coefficients of the corresponding QoP curve after its projection in the new functional space.

5.3. Dimension reduction

By considering that the numbers of the observed QoPs and the observation duration might be huge, a dimensionality reduction seems inevitable. Therefore, a Functional Principal Components Analysis for multivariate data (FPCA, [27]) is applied. It consists in computing the principal components C^h and principal factors f^h of the Karhunen-Loeve expansion: $X(t) = \mu(t) + \sum_{h\geq 1} C^h f^h(t)$, $t \in [0,T]$. It allows to have an optimal representation of curves from the set of functional data into a functional space of reduced dimension. The number of principal components $m \leq M$ is chosen so that at least 90% of the information is covered. As a result, each curve is defined by a vector of its principal components of size m. For more details, we refer the reader to [27] and [28].

5.4. Supervised learning

At this stage, the QoP measurements are transformed in terms of principal components, a training data with numeric values and a binary class is ready for a training phase. The training is assured by a Random Forest [29] which is a classification technique that does not require much memory. Random forest run-times are quite fast and it is able to deal with imbalanced and missing data. It starts by splitting the data set into a number of samples. For each sample, a decision tree is trained. It is a "divide-and-conquer" approach that aims to form a "strong learner" (a forest) from a group of "weak learners" (decision trees). From the set of the decision tree classifiers that have been trained on various sub-samples of the data set, the random forest uses averaging to improve the predictive accuracy and to control over-fitting. To classify a new observation, it puts this latter down each of the trees in the forest. Each tree gives a classification i.e. it "votes" for a class and the forest chooses the classification having the most votes.

In case of imbalanced data sets, the prediction error between classes is highly imbalanced. One class have a low prediction error compared to the other. Then random forests, by trying to minimize overall error rate, will keep the error rate low on the large class while letting the smaller class to

have a larger error rate. As a result, the accuracy of the model will not be misled by the imbalance. Figure 6 illustrates the training phase of the model.
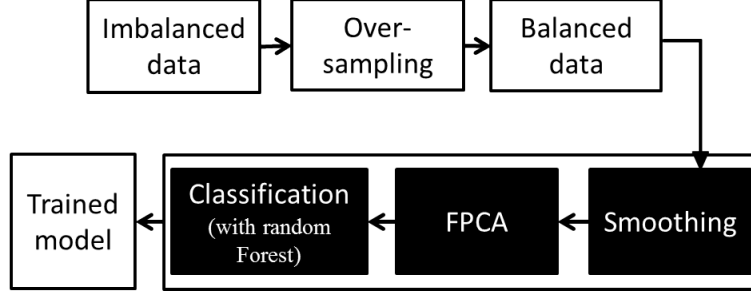

Figure 6: Training phase

Given the smoothing basis, the FPCA basis and the learned random forest, prediction of future anomalies for new observations is possible as described in Figure 7.
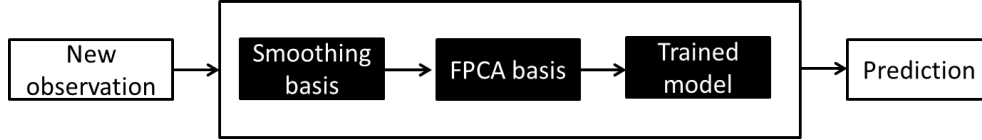

Figure 7: Prediction phase

For each new observation, a smoothing is first applied for every QoP metric. The observation in terms of functional features is then projected on the same FPCA basis used in the training phase. Given the obtained principal components, the learned random forest will then predict whether a degradation of the signal quality will be perceived in the future 5 minutes or not.

6. Actuation framework for QoP optimization

Aside from the presented model, which opens the door for QoP-aware management of e-Health NS, it is also important how the model engages with other elements in order to close the cognition loop as to maintain top notch quality levels. With this in mind, an actuation framework has been developed within the presented architecture inside the cognition sub-plane.

Following the explanations of previous sections, and to complete the cognitive loop for QoE-aware management of e-Health NSs, this section elaborates about the developed actuation framework to be triggered as response to the outputs from the presented QoP prediction ML model. As illustrated in Figure 3, the actuation framework consists in a Policy Framework and a QoE Optimizer, which serves to execute remedial actions once degradation on the slices quality is observed.

The QoE Optimizer is the element responsible to trigger remedial action in cases which the quality of deployed E2E slices is compromised or is not within satisfactory levels. To do so, it listens to events (outputs) of the ML models within the cognition sub-plane (for the case at hand, the developed QoP prediction model) and, in case the events mark that quality levels of the slice are not as expected or some anomaly can affect them, it triggers the necessary (re-)configurations of the E2E NS and composing NSSIs to regain optimal quality levels.

To do so, the QoE Optimizer interacts with the service orchestrator at the DSP level (Service), which, in turn interact with the multiple single-domain slice orchestrators present at the multiple NSPs. The slice orchestrators may engage directly with control plane functions if configuration details of the NSs need to be modified (e.g. change the assigned bandwidth, apply a UE handover) or may engage with the resource orchestrator at the same NSP (NMR-O) in case that characteristics of the virtual resources need to be modified (e.g. CPU/memory assigned to a virtual function). Thanks to that, the full actuation is orchestrated.

This presented mechanism is governed through policies in place at the QoE Optimizer. These policies are disseminated from a Policy Framework at the DSP level, following an Event-Condition-Action (ECA) model [30] which specifies for what event (an output from the ML model) and condition (e.g. an anomaly is predicted) what action should be applied (e.g. increase the bandwidth at the RAN segment). Thanks to that, it is possible to follow a rather declarative approach in which the Actuation Framework specifies the "what" and "when" of the remedial action while the "how" is implemented thanks to the logic of the elements responsible for orchestration or configuration of slices/infrastructure elements (i.e. orchestrator, control plane).

Aside from that, policies also carry information about the end-points that need to be listened for monitoring the information related to the events as well as a small logic that dictates simple extraction/aggregation operations

for this information, named as Sensors. In this way, all the elements that belong to the monitoring pipeline will be properly configured following the specifications of the Sensors present in the policies for a specific E2E NS.

7. Experimental proof of concept and testing

In the previous sections, we presented the architecture of the QoP-aware management for e-Health 5G NS, the ML model in charge of QoP predictions and the Actuation Framework developed to execute remedial actions. In this section, we proceed on detailing the performed tests to showcase the capabilities and the performance of the presented framework. With this in mind, this section details the experimental set-up employed to emulate an e-Health scenario as well as the tests performed on top of the testbed.

7.1. E-Health testbed

The experimental testbed is based on an infrastructure deployed within Dell premises in Ovens, Ireland, which replicates the capabilities of the presented e-Health scenario. Figure 8 depicts a schematic of the testbed, highlighting the most important components.

On top of the infrastructure, an open source security software, pfSense, is running to provide different security services. Below the firewall and security services, the infrastructure is spanning across 3 racks. A first rack is dedicated to running management services such as OpenSource MANO (OSM) and the software artifacts presented in the proposed architecture; a second rack is running Core network components and cloud operating system, running VIO VMware. A third rack is dedicated to the edge running OpenStack. This edge Virtual Infrastructure Manager (VIM) is hosting the e-Health services. Finally, two stand-alone evolved Node B (eNB) are running the OpenAirInterface (OAI) software to provide the RAN network segment. To have access to this eNBs, open-cells Subscriber Identity Modules (SIMs) are reprogrammed with uicc/sim programming software provided by open-cells. These SIMs are the ones that emulate the UEs associated to e-Health service, such as wearables inside a roaming ambulance. Given this testbed, the following sub-sections details the specific experiments done to test both the ML model and the Actuation Framework.

7.2. QoP prediction model evaluation

The QoP prediction model aims to train from data that are collected from the vertical service perception. Several real traces are captured having
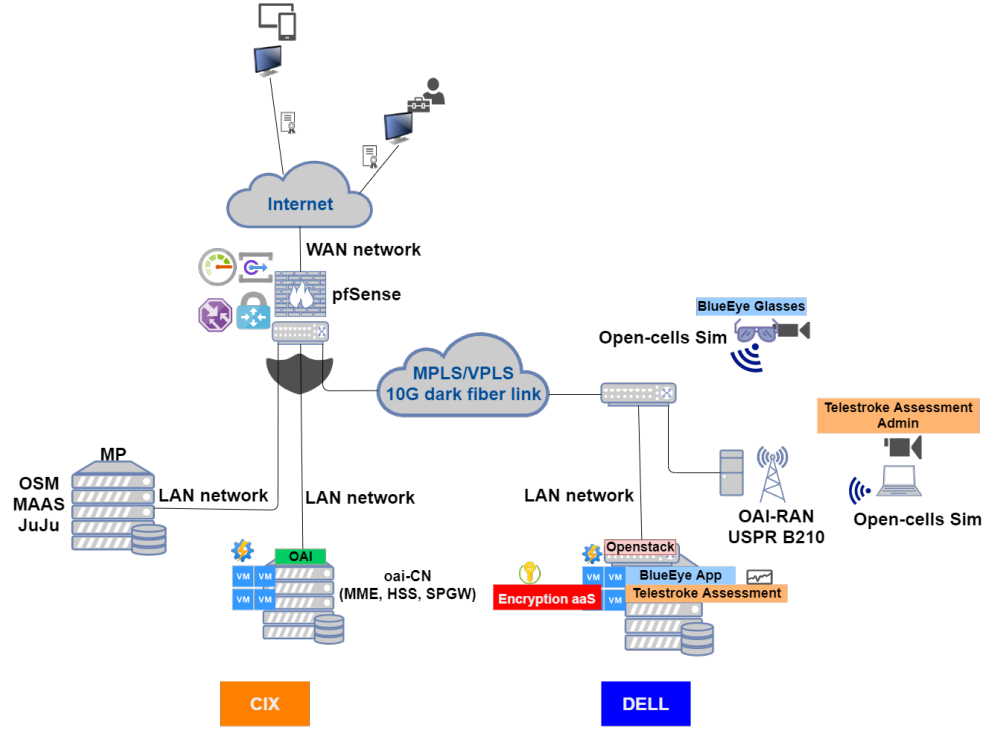
Figure 8: E-Health infrastructure at Dell premises, Ovens, Ireland

between 2 to 4 hours of length. The total number of instances is 11820. For each instance, seven QoP metrics are considered for determining the quality of the network signal perceived by the ambulance as described in Table 1. Among the real traces, 75% have been used for the training phase and the rest is kept for the testing phase.

As detailed in Section 5, the first step aims to apply a sliding window that will transform the data into a 3D matrix. Each window will observe the KPIs for the last five minutes and it will label the data by observing the future five minutes. The window will advance each 5 seconds. The labeling aims to define the perceived signal strength label for each instance by according thresholds to RSRP, RSRQ and SNR metrics. The result of the previous step is a 3D matrix composed of 11820 instances. Each instance is defined by 7 KPIs where each KPI is defined by 300 values that correspond to 5 minutes of observation. The resulted training set is imbalanced. After applying SMOTE technique, 25 % of all instances correspond to problems in signal strength.

In order to evaluate the model, two techniques are used: (1) Evaluation using cross-validation technique over the training set; and (2) evaluation over the test set. Table 2 presents the results over the training data using a cross validation with 5 folds where the classification is achieved using a random forest. It also illustrates the results over the test data. Table 2 illustrates the numerical results of our experiment where four performance metrics [31] are used: (1) "Accuracy": is the ratio of correctly predicted data over total data; (2) "Recall": is the fraction of the elements of a class that are successfully predicted; (3) "Precision": is the fraction of correct predictions among all the predictions of a class; (4) "F1-score": is a harmonic mean of precision and recall.

Table 2: QoP forecasting model results

Metric	Results with cross-validation	Results over the test set
Accuracy	96%	99%
Precision	91%	99%
Recall	93%	99%
F1-score	93%	99%

As shown in Table 2, the results are promising since it indicates that 93% of the signal strength degradation in the data is detected and 91% of the detected problems are not false alarms. Once the model predicts low perceived signal strength, the Actuation Framework will trigger the event in order to apply a remedial action and to solve the problem before it occurs. This offers the opportunity to the vertical to supervise the performance of its slice and to express its feedback to which SliceNet framework will react accordingly.

7.3. Actuation mechanisms experimentation

The presented QoP prediction model is devoted to predict future perceived degradation in signal quality at the RAN segment in which E2E NS are being supported. Communications over the RAN are of paramount importance, since they enable the necessary mobility of the UE over a geographical area which is a key requirement of e-Health services. As such, proper actuation mechanisms must be proposed and tested in order to react to the predicted degradation.

In the framework of the presented work, in order to overcome potential anomalies in the RAN segment of the NS, two specific actuations have been

developed. A first one consists on the handover of the UEs present in the ambulance to another RAN slice that is available in the physical infrastructure. This stems from the fact that signal degradations may be caused due to the mobility of the end-points connected to the e-Health NS, leading to attenuations, delays and other impairments due to the geographical surroundings of the area in which the ambulance is traveling. This being the case, a remedial action entails that the UEs connected to the RAN slice supported over a particular eNB should be re-associated to another RAN slice of another eNB, closer to the current location of the ambulance (i.e. handover).

The second considered actuation consists on enforcing changes on the provisioned resources of the RAN slice, more in concrete, the allocated bandwidth. Bandwidth in the RAN is provisioned by reserving a set of contiguous Radio Resource Blocks (RRBs), which provide for the enough network capacity to support the communications over the configured NS. In this regard, it may happen that due to the dynamic nature of the data streams (audio, video) produced by the UEs at the ambulance, in certain points of time the provisioned RRBs are not enough to withstand these data streams, leading to poor video/audio quality and thus affecting the QoE of users. This also translates to some of the QoP metrics being affected (the ones related to the bit-rate), that the QoP prediction model is able to detect. To overcome bandwidth limits, the Actuation Framework engages with the rest of the control/management elements in order to increase the provisioned RRBs at the RAN slices so as to maintain optimal QoE/QoP levels. In both cases of actuation, a policy is placed that ties the output of the ML model to the desired action, handover or bandwidth increase. The setting of the proper policy can be made by a system administrator or by some automated method.

For the developed work in this paper, the FlexRAN SDN controller, which is part of the Mosaic 5G project, has been employed [32]. FlexRAN exposes a REST API that allows external entities to access to the control and configuration operations of an underlying RAN segment. For the presented work, FlexRAN exposes two operations of interest that allow for the materialization of the defined actuations. The first operation of interest allows to change the association of a UE attached to a particular RAN slice in a specific eNB to another RAN slice of the same eNB or other eNB. By exploiting this operation, it is possible to effectively produce a handover of the UEs in the ambulance across RAN slices. The other operation of interest entails the posting of a new configuration for an existing RAN slice, covering aspects such as priority, sharing, scheduling and, more important, the number and position of

the RRBs for both up and down links of the slice. Through this operation it is possible to change the RRBs of the slice, thus effectively increasing the bandwidth that is allocated for the communications that originate at the UEs.

In order to provide a PoC of the developed actuation system, we focus on the case of the bandwidth increase actuation. For that, an experimental scenario has been prepared on top of the previously described e-Health testbed. Figure 9 depicts the employed set-up.
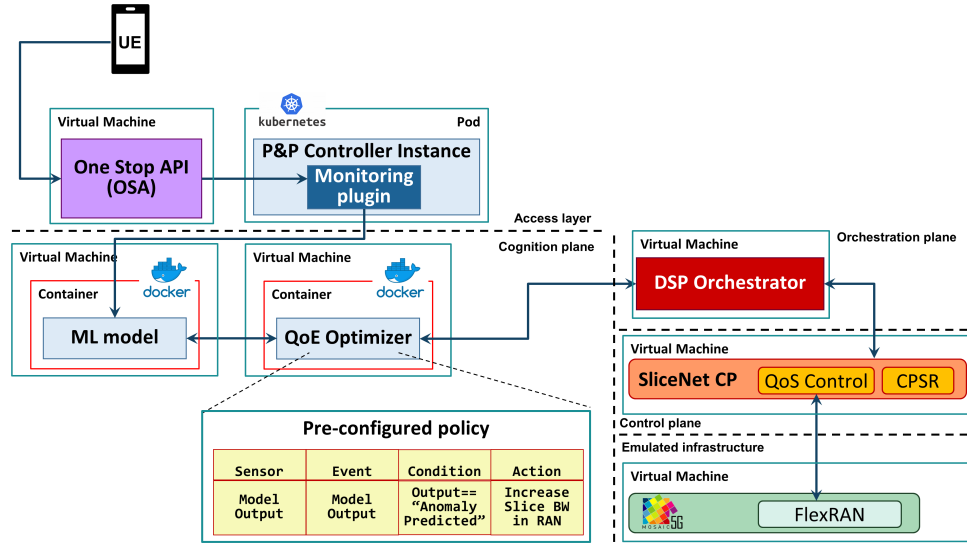

Figure 9: Experimental set-up for actuation PoC.

The set-up consists on an OSA instance and a P&P control instance that allow collecting the UE-related QoP metrics described previously. A specific Monitoring plugin deployed within the P&P control instance exposes the monitoring operations towards the UE so as to correctly fetch the desired metrics. Then, these metrics are fed to an instance of the Anomaly Detection ML model which, given the inputs, produces an output that directly is consumed by the QoE Optimizer. A pre-configured policy is in place in the QoE Optimizer which dictates that, in the case that the output states that an anomaly is going to happen, it is necessary to increase the bandwidth in the RAN segment of the E2E NS. For the enforcement of the actuation, the QoE Optimizer contacts an instance of the DSP orchestrator. For the particular PoC, the orchestration system has been simplified, with only having one level

of orchestration that directly contacts the control functions responsible for
the (re-)configuration of slice parameters. The function responsible for that
is the QoS Control function within the SliceNet control plane, which allows
to enforce any change of QoS parameters (such as bandwidth) of deployed
slices by properly engaging with the underlying controls of the physical in-
frastructure. Lastly, the RAN infrastructure has been emulated employing
the aforementioned FlexRAN SDN controller, which, aside of emulating the
RAN segment, exposes the control operations necessary to modify the char-
acteristics of RAN slices.

Table 3: Configured initial RAN slice.

```
sliceConfig: {
   dl: [
      {
         id: 0,
         label: xMBB,
         percentage: 10,        Percentage of total available bandwidth
         isolation: false,
         priority: 10,
         positionLow: 0,        Position of RRBs
         positionHigh: 25,       Position of RRBs
         maxmcs: 28,        Position of RRBs
         ...
      }
   ],
   ul: [
      {
         id: 0,
         label: xMBB,
         percentage: 10,        Percentage of total available bandwidth
         isolation: false,
         priority: 10,
         firstRb 0,        Position of RRBs
         maxmcs: 28,        Position of RRBs
         ...
      }
   ],
   ...
```

Following the depicted set-up, the PoC consisted on simulating a signal
quality degradation to the RAN slice in which the UEs of the ambulance
are connected. To this end, as a first step, a RAN slice has been deployed

employing the operations exposed through FlexRAN. The slice has been assigned with a 10% of the RRBs of the total RAN spectrum. Table 3 depicts in a Java Script Object Notation (JSON) format a sample of the details of both the down- and up-link of the slice as exposed through the monitoring API of FlexRAN. The highlighted fields in boldface are both the percentage of bandwidth occupied by the slice as well as the positioning of the RRBs.

Then, the UE is connected to the deployed slice. Given the configuration of resources and the monitored QoP metrics, the Anomaly Detection ML model detects that a potential anomaly may happen due to the insufficient bit-rate that the UE is experiencing, as a result of the low number of RRBs that have been provisioned. Once this is detected, the full actuation workflow is triggered, which results in a request of increasing the amount of assigned bandwidth of the deployed RAN slice. This is ultimately enforced by the QoS Control module, which contacts the FlexRAN SDN controller specifying the new configuration following the format of the exposed operation. Table 4 depicts the JSON of the modification operation, stating that the slice needs to be provisioned with 25% of the available RAN spectrum instead of the initial 10%.

Table 4: New RAN slice configuration as a result of the actuation.

```
{
    dl: [
        {
            id: 0,
            percentage: 25,        New percentage of provisioned bandwidth
        }
    ],
    ul: [
        {
            id: 0,
            percentage: 25,        New percentage of provisioned bandwidth
        }
    ],
    ...
```

Once the FlexRAN controller is contacted, it then takes care of applying the necessary re-configurations at the RAN segment so as the indicated slice is provisioned with the new specified bandwidth. Table 5 depicts the characteristics of the slice after the re-configuration process. Indeed, it can

be appreciated how the desired increase of bandwidth is reflected in its new characteristics by an increased percentage of the provisioned RRBs. After this process, the communications that are supported over the RAN slice find themselves with enough bit-rate to be supported, resulting in the anomaly being adverted.

Table 5: Characteristics of the re-configured RAN slice.

```
sliceConfig: {
   dl: [
      {
        id: 0,
        label: xMBB,
        percentage: 25,        Percentage of total available bandwidth
        isolation: false,
        priority: 10,
        positionLow: 0,        Position of RRBs
        positionHigh: 25,       Position of RRBs
        maxmcs: 28,        Position of RRBs
        ...
      }
   ],
   ul: [
      {
        id: 0,
        label: xMBB,
        percentage: 25,        Percentage of total available bandwidth
        isolation: false,
        priority: 10,
        firstRb 0,        Position of RRBs
        maxmcs: 28,        Position of RRBs
        ...
      }
   ],
   ...
```

Figure 10 depicts the time evolution for one of the radio parameters corresponding to the input of the anomaly detection ML model, specifically, the bit-rate for the uplink (UL) of the deployed slice. The reasoning behind the graph is to showcase how the actuation is triggered due to the ML model predicting that the QoP would be affected due to changes on the monitored radio parameters. For the initial period of 5 minutes, the model forecasts that no anomalies will be experimented, thus, the assigned RRBs remain

equal to the initially provisioned during the slice deployment, which was a 10% of the total radio spectrum, as explained before. For the next 5 minutes, it can be appreciated an increase of the UL bit-rate, due to the utilization imposed by the eHealth application being executed at the slice. Since the initial assigned RRBs are not enough to withstand the increased bit-rate (the horizontal dashed line marks the initial maximum achievable bit-rate), the actuation framework is notified of this situation. As a result, the RAN slice is provisioned with an increased number of RRBs (25%, as noted before), overcoming the undesired situation before it takes place. These results highlight how the combination of advanced ML-based sensors with an actuation framework dedicated to the reconfiguration of deployed slices can adapt the runtime resource assignment to the dynamic requirements of the services executed on top, eHealth video services for the particular targeted scenario in this work.
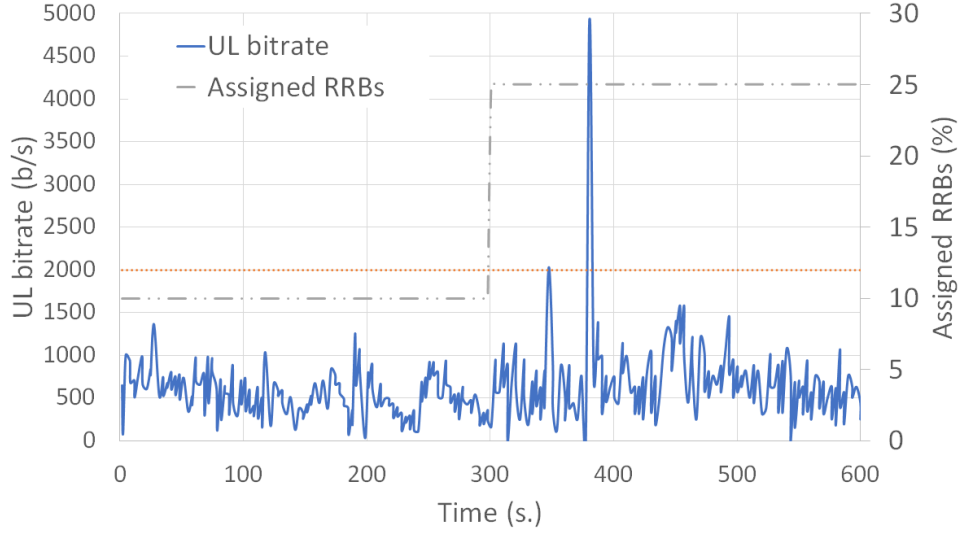

Figure 10: Time evolution of the monitored UL bit-rate and assigned RRBs

8. Conclusion and Future works

Preserving a top communication quality in 5G network slices between the paramedic and the medical teams is a necessity for eHealth verticals. In this regard, a framework has been proposed in this paper allowing the vertical to

supervise its slice and to express its feedback regarding the perceived quality of services. Thanks to the developed architecture, it is possible to apply remedial actions to overcome situations in which the quality of the slice, and thus, the supported services, is compromised.

We have shown how the developed ML for QoP estimation is able to digest data coming from UE equipment an transform it into QoP estimations that indicate the quality of the communication over the provisioned E2E slice. From the results, it can be seen that the performance of the model is promising, with high levels of accuracy, showcasing its value as an advanced monitoring function to allow for the quality-aware and cognition-enabled management of e-Health network slices. In addition, we showcased how the developed framework is able to react to outputs coming from the ML model and trigger the necessary remedial actions to overcome the forecast anomalies. We showed how the process is automatically triggered and is able to engage from top to bottom to all involved layers to make effective the necessary changes at the infrastructure level. All in all, the presented experiments show that the proposed architecture is a valuable framework that enables 5G systems with the cognitive management network slices, catering to the special requirements and challenges posed by e-Health vertical services.

As future lines of research, additional prediction models to detect anomalies not only at the RAN segment but also at other segments involved at the E2E slice deployment, such as core, may be investigated. This would allow to have a more precise way of tackling down the network segment responsible for bad quality levels at the service. In addition to that, a method to automate the choice of the corrective action and of the choice of the corresponding policy, not based on static policies disseminated from a policy framework, but rather elaborated based on gathered knowledge and monitoring data, can be investigated. This would allow to adapt the remedial actions to the most updated state of the underlying physical infrastructure and the deployed slices. One way to achieve this automation is by using re-enforcement learning or optimisation algorithms, which could then feed the policy framework for the dynamic creation of policies.

9. Acknowledgment

References

[1] Ericsson Mobility Report, Tech. rep. (june 2019).
URL https://www.ericsson.com/49d1d9/assets/local/mobility-report/documents/2019/ericsson-mobility-report-june-2019.pdf

[2] 3GPP, TR 28.801: Telecommunication management;Study on management and orchestration of network slicing for next generation network, Tech. rep. (2017).

[3] ITU-T, P.10/G.100: Vocabulary and effects of transmission parameters on customer opinion of transmission quality, Tech. rep. (2017).

[4] ITU-T, E.800: Definitions of terms related to quality of service, Tech. rep. (2009).

[5] NGMN, 5G White Paper, Tech. rep. (2015).

[6] S. M. Ahsan Kazmi, L. U. Khan, N. H. Tran, C. S. Hong, Network Slicing for 5G and Beyond Networks, Springer, 2019.

[7] P. L. Vo, M. N. H. Nguyen, T. A. Le, N. H. Tran, Slicing the edge: Resource allocation for ran network slicing, IEEE Wireless Communications Letters 7 (6) (2018) 970–973.

[8] S. Husain, A. Kunz, A. Prasad, K. Samdanis, J. Song, Mobile edge computing with network resource slicing for internet-of-things, in: 2018 IEEE 4th World Forum on Internet of Things (WF-IoT), 2018, pp. 1–6.

[9] T. Taleb, I. Afolabi, K. Samdanis, F. Z. Yousaf, On multi-domain network slicing orchestration architecture and federated resource control, IEEE Network 33 (5) (2019) 242–252.

[10] 3GPP, TS 28.533: Technical Specification Group Services and System Aspects; Management and orchestration; Architecture framework (2021).

[11] 3GPP, TS 23.288: Technical Specification Group Services and System Aspects; Architecture enhancements for 5G System (5GS) to support network data analytics services (2021).

[12] 3GPP, TS 29.520: Technical Specification Group Core Network and Terminals; 5G System; Network Data Analytics Services (2021).

[13] SESAME: Small cEllS coordinAtion for Multi-tenancy and Edge services.
URL https://www.sesame-h2020-5g-ppp.eu/

[14] SelfNet: A Framework for Self-organized Network Management In Virtualized and Software Defined Networks.
URL https://selfnet-5g.eu/

[15] CogNet – Building an Intelligent System of Insights and Action for 5G Network Management.
URL https://5g-ppp.eu/cognet/

[16] ENI, Experiential Networked Intelligence (ENI); System Architecture, GS ENI 005 v1.1.1, Tech. rep. (2019).

[17] NGMN, 5G End-to-End Architecture Framework v3.0.8, Tech. rep. (2019).

[18] A. Martin, J. Egaña, J. Flórez, J. Montalbán, I. G. Olaizola, M. Quartulli, R. Viola, M. Zorrilla, Network resource allocation system for qoe-aware delivery of media services in 5g networks, IEEE Transactions on Broadcasting 64 (2) (2018) 561–574.

[19] W. Kellerer, P. Kalmbach, A. Blenk, A. Basta, M. Reisslein, S. Schmid, Adaptable and data-driven softwarized networks: Review, opportunities, and challenges, Proceedings of the IEEE 107 (4) (2019) 711–731.

[20] V. Sciancalepore, F. Z. Yousaf, X. Costa-Perez, z-torch: An automated nfv orchestration and monitoring solution, IEEE Transactions on Network and Service Management 15 (4) (2018) 1292–1306.

[21] Y. Syu, C. Wang, Y. Fanjiang, Modeling and forecasting of time-aware dynamic qos attributes for cloud services, IEEE Transactions on Network and Service Management 16 (1) (2019) 56–71.

[22] V. Chandrasekhar, Y. Heng, J. Cho, J. Lee, J. Zhang, J. G. Andrews, Experience-centric mobile video scheduling through machine learning, IEEE Access 7 (2019) 113017–113030.

[23] M. Torres Vega, C. Perra, F. De Turck, A. Liotta, A review of predictive quality of experience management in video streaming services, IEEE Transactions on Broadcasting 64 (2) (2018) 432–445.

[24] J. J. A. Esteves, A. Boubendir, F. Guillemin, P. Sens, Optimized network slicing proof-of-concept with interactive gaming use case, 23rd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN) (2020).

[25] SliceNet: End-to-End Cognitive Network Slicing and Slice Management Framework in Virtualised Multi-Domain, Multi-Tenant 5G Networks. URL https://slicenet.eu/

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, J. Artif. Int. Res. 16 (1) (2002) 321–357.

[27] J. O. Ramsay, B. W. Silverman, Functional data analysis, 2nd Edition, Springer Series in Statistics, Springer, New York, 2005.

[28] Y. B. Slimen, S. Allio, J. Jacques, Anomaly prevision in radio access networks using functional data analysis, in: 2017 IEEE Global Communications Conference, GLOBECOM 2017, Singapore, December 4-8, 2017, IEEE, 2017, pp. 1–6.

[29] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

[30] IETF, SUPA Policy-based Management Framework, Tech. rep. (2017).

[31] N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011.

[32] FlexRAN: First Open-source Implementation of a Flexible and Programmable Platform for Software-Defined Radio Access Networks. URL http://mosaic-5g.io/flexran/