Application of Machine Learning in RAN Evolution for new Generation of mobile networks

Capstone Project

Presented by:
Seyedesmaeil Seyedalikhani

University of Alberta
Master of Science in Internetworking
Edmonton, Canada

Supervisor:
Sandeep Kaur

March 2023

Acknowledgement

I want to thank my mentor, Ms Sandeep Kaur, for her valuable support and all of the helpful advice, comments, and recommendations she gave me during the project.
I also want to express my gratitude to Pr. Mike MacGregor and Mr Shahnawaz Mir for allowing me to work on this project.
Finally, I want to thank God for providing me with strength and my family for their support during this project, which helped me stay motivated and accomplish it.

Table of Contents

Contents

ABSTRACT

The telecommunication industry has experienced considerable improvement and changes during the past years. Many standards and protocols have been introduced and implemented. This revolution in Radio Access Networks (RAN) is known as GSM, UMTS, LTE, 5G, and now B5G networks. Satisfying the user demands and keeping the level of QoS and QoE within the acceptable range have always been challenges for internet service providers and telecommunication operators. The researchers and studies are ongoing to address these massive requests and users' tendency to achieve reliable, low latency, and high throughput services.

Software Define Networking (SDN), Network Virtualization Function (NVF), Self-Organizing Networks (SON), and increasing capacity solutions (mmWave communication, Massive MIMO, Network Slicing, Beamforming, and RAN Evolutions) are the main proposed and implemented solutions during the past decade in 5G networks. However, whenever we talk about the data, we will see the brilliant role of machine learning.

In this study, we have researched and implemented machine-learning algorithms in new evolutions of RAN. We can mention RAN evolution as Distributed-RAN, Cloud-RAN, Virtual-RAN, and now Open-RAN. Open RAN is a novel method of setting up and running wireless networks Using standardized, interoperable hardware and software components. Instead of being dependent on the proprietary technology of a single vendor, an open RAN architecture separates the radio access network into interchangeable, functional components.

The proposed scenario uses supervised-learning algorithms to make predation (classification) of services and slices in Open-RAN 5G networks. This AI/ML scenario is implemented in the RIC (Radio Intelligent Controller) block of O-RAN, and we have evaluated and compared the performance of five different supervised-learning algorithms. A novel method based on hyperparameters tuning and K-fold cross-validating is proposed for Random Forest Algorithm. This technique will improve the classified results compared to the introduced baselines. The algorithms' training phase utilizes the KPI and KQI data of a 5G network. Moreover, simulation results prove that considering both KPI and KQI will improve the results compared to only KPI scenarios.

Key Words

RAN, 5G, Cloud RAN, Virtual RAN, Open RAN, Network Slicing, Machine Learning

List of Tables

List of Figures

1 Chapter 1) Telecommunication background and History

1-1) From GSM to LTE

Figure 1 depicts a brief chronological history of cellular radio systems from the 1970s, when they were originally developed (1G, the first generation), until the 2020s (i.e., 5G, the fifth generation).

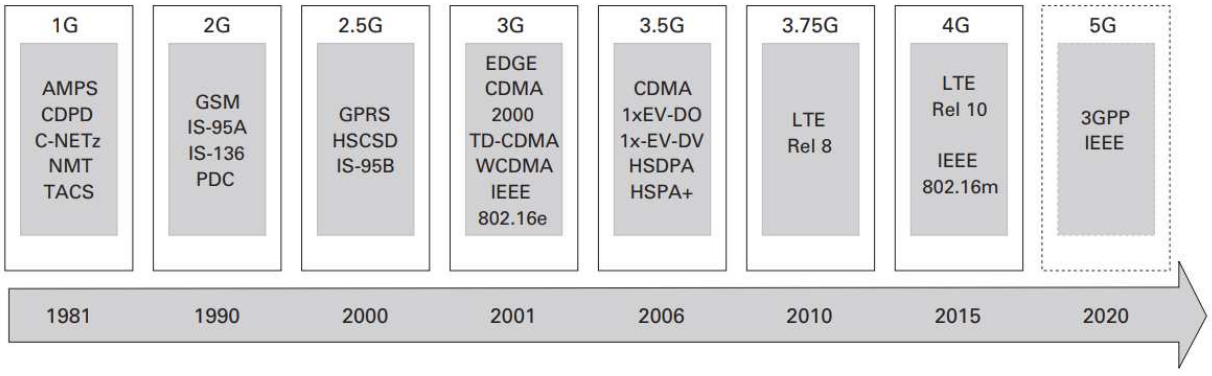

Figure 1: Evolution of cellular networks [1]

It shows that cellular mobile systems have undergone several major evolutionary phases. The first commercial analogue mobile communication systems were introduced with a modest penetration in the 1950s and 1960s. 1981 saw the commercial deployment of 1G mobile cellular standards, such as C-Netz in Germany and South Africa, Nordic Mobile Telephone (NMT), and Total Access Communications System in the United Kingdom (TACS). The Advanced Mobile Phone System was (AMPS) launched in the Americas. Analogue standards are also known as the 1G standards. They use analogue technology and typically

combine frequency-modulated radio waves with a digital signalling stream. In 1982, the European Conference of Postal and Telecommunications Administrations decided to create a Pan-European 2G mobile communications system. This was the point from which the Global System for Mobile Communications (GSM) was born, the international standard for 2G. Digital transmission and switching technology were critical to the introduction of 2G. Digital communication enabled significant improvements in voice quality and network capacity. It provided growth through supplementary services and advanced apps such as the Short Message Service (SMS) to store and forward textual information [1].

GSM's primary purpose created to allow international roaming through Europe. GSM uses a hybrid TDMA/FDMA instead of 1G systems based on FDMA. Other digital 2G systems were also developed parallelly to GSM and compete with one another. Other 2G standards include:

1. TIA/EIA-136, known to be the North American TDMA standard (NA-TDMA),
2. TIA/EIA IS-95A, known as CDMAOne
3. Personal Digital Cellular, used only in Japan.

The 2.5G evolution of 2G introduced packet-switched services to complement voice and circuit-switched information. General Packet Radio Service and TIA/EIA-951 were extensions of GSM, TIA/EIA IS–95A and other 2.5G standards. The General Packet Radio Service (GPRS) was improved by Enhanced Data rates for Global Evolution (EDGE), which emerged quickly from GSM (EGPRS). This was mainly due to the addition of higher-order modulation and coding. GSM/EDGE continues to evolve. The latest 3GPP standard supports greater bandwidths and carrier aggregation [1].

Industrial players began discussing and preparing the next generation of wireless standards shortly after 2G was operational. Parallel to this, the International Telecommunications Union Radio Communications (ITU-R) developed requirements for systems that would be eligible for the International Mobile Telecommunications 2000 classification (IMT-2000). The European Telecommunications Standards Institute adopted CDMA in two versions - Wideband Code Division Multiple Access (WCDMA) and Time Division CDMA(TD-CDMA), as a Universal Mobile Telecommunication System. UMTS was the most important 3G mobile communication system [1].

It was also one of the first cellular systems to be qualified for IMT 2000. Six radio interfaces have been qualified to meet IMT-2000 requirements, including three technologies based on CDMA, a version of GSM/EDGE known as UWC-1362, and two technologies based on OFDMA. New specifications were created within the 3rd Generation Partnership Project (3GPP). They have illustrated in Figure 1 as 3.5G. Two Radio Access Networks (RAN) and an evolution to the Core Network use this evolution [1].

1-2) History of RAN

The evolution steps of CDMA 2000 within 3GPP2 were the basis for the first RAN approach. They consisted of 1xEV DV, and 1xEV DO. High-Speed Packet Access was the second RAN approach. High-Speed Downlink Packet Access (HSDPA) and High-Speed Uplink Packet Access (HSUPA), which were included in 3GPP Release 5, were combined to become HSPA.

Each initially increased the packet data rate to 14.6 Mbps downlink and 5.76 Mbps uplink. However, they quickly evolved to handle higher data speeds with the introduction of MIMO. HSPA is backwards-compatible with WCDMA because it was built on WCDMA. CDMA 1xEV -DO was first deployed in 2003. HSPA and CDMA 1xEV -DV was introduced in 2006. The 3GPP standards adhere to the principle of adding new features and maintaining backward compatibility. HSPA+ is an evolution of HSPA that supports carrier aggregation to achieve higher peak data rates while not affecting existing terminals [1].

The commercially acceptable 4G evolution is Long Term Evolution (LTE). It consists of a new air interface based upon Orthogonal Frequency Division Multiple Access (OFDMA) and a new architecture and Core Network, the System Architecture Evolution/Evolved Packet Core. LTE is not compatible with UMTS. It was created in anticipation of higher spectrum block allocations for UMTS at the World Radio Conference (WRC) in 2007. The standard can also be used with component frequency carriers, which are highly flexible in their arrangement and support carriers from 1.4 MHz to 20 MHz.

The LTE standard provided significant capacity improvements and was intended to move cellular networks away from circuit-switched functionality. LTE evolution resulted in significant cost savings over previous generations. 3GPP approved the first LTE specifications at the end of 2007. They are now known as LTE Release 8. LTE Release 8 has a peak data rate of 326 Mbps, higher spectral efficiency, and significantly shorter latency (down 20ms) compared to previous systems. The ITUR was simultaneously developing IMT-Advanced requirements, a successor to IMT-2000 and the nominal definition of the fourth generation [1].
LTE Release 8 was not compliant with IMT-Advanced requirements and was initially considered a precursor for 4G technology. This statement was later relaxed, and LTE is now accepted as 4G. 3GPP LTE release 10 and IEEE 802.26 m (deployed under WiMAX) were the first air interfaces to meet IMT-Advanced requirements. WiMAX is an approved 4G technology but has struggled to gain widespread acceptance. LTE will replace it. LTE Release 10 introduced several technical features such as higher-order MIMO, carrier aggregation and improved throughput. Carrier aggregation of up to 100 MHz increases peak data rates to a maximum of 3 Gbps downlink and 1.5 Gbps uplink. Performance improvement can also be achieved by higher-order MIMO configurations, up to 8x8 downlink and 4x4 uplink. 3GPP standardization for LTE (i.e. Release 11 to Release 13 and the subsequent release are continuing and will be continued [1].

Several features were added to enable offloading the backhaul and core network. In LTE Releases 12, 13 and 14, new solutions (known as LTE-M (NB-IoT), were introduced to support massive Machine Types Communication devices like sensors and actuators. These solutions resulted in improved coverage, battery life and lower cost. Release 13 targets excessive broadband data rates by using carrier aggregation of up to 32 carriers. Mid-2015 saw a cellular global mobile market of 7.49 billion subscribers (10). The dominant Radio Access Network (RAN), the GSM/EDGE, included EGPRS data connectivity and was in use. GSM, which has a global market share exceeding 57% (correspondingly to 4.26 billion subscribers), is well above peak use and in decline. However, 3G subscribers, including HSPA, have increased to 1.94 billion since 2010, 26% of the total market share. According to the

Ericsson Mobility Report, WCDMA/HSPA subscriptions will reach their peak in 2020 and then start to fall after that.

LTE, the dominant 4G standard, has attracted approximately 910 million subscribers (or 12 per cent of the total market) as of 2015. It is predicted to have a staggering 4.1 billion subscriptions by 2022 [11]. This progress makes it the most popular mobile technology. Figure 2 shows the main features of 3GPP standards currently on the market. It highlights the trend toward widespread use of spectrum and higher bandwidths [1].

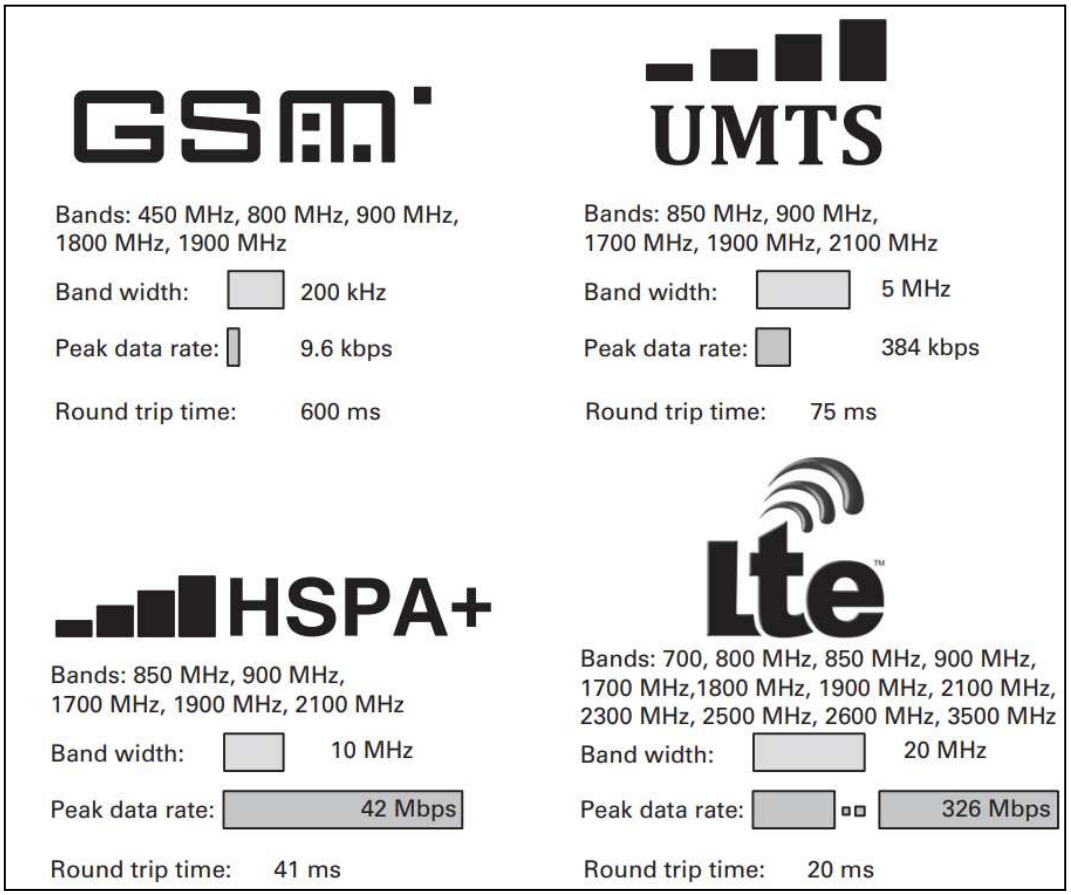

Figure 2: Main characteristics of 3GPP/ETSI standards [1]

2 Chapter 2) Global System Mobile, GSM (2G) and GPRS

2-1) Concept of GSM development

Ericsson began the basic technology development for GSM in the Nordic countries during the 1980s. Afterwards, it was transferred to a working group of standardization body Group Special Mobile (GSM), within the Conference Europeans des Posts et Telecommunications standard committees. The GSM system has undergone extensive modifications since the standardization process of GSM900. This was to meet the increasing demands of mobile operators participating in standardization bodies. The European Telecommunications Standards Institute's Special Mobile Group (SMG) was responsible for most further standardization. Since then, the 3rd Generation Partnership Project (3GPP) has taken over [2].

The GSM standard had one primary objective: to create a digital system that could be mass-produced at a low cost. GSM-Global System Mobile was required to provide equal or better speech quality and spectrum efficiency than existing analogue mobile systems. The system will be called GSM-Global System Mobile and should offer ISDN services. The fixed side, called GSM-specific services, includes:

- Global roaming (initially Pan-European).
- Authentication (fraud control)
- Ciphering (speech and data, signalling)
- Privacy of user (ciphered subscriber numbers on-air-interface

GSM was accepted worldwide by the end of 1990. GSM is still a significant source of revenue for mobile operators, even though its share in modern mobile networks is declining.

2-2) GSM Architecture

Figure 3 shows the components of the GSM network. It can be divided into three subnetworks or subsystems, the radio access network (RAN), the core network and the management network. In terms of subsystems, these are also known as the Base-Station Subsystem, the Network Switching Subsystem and the Operation Support Subsystem.

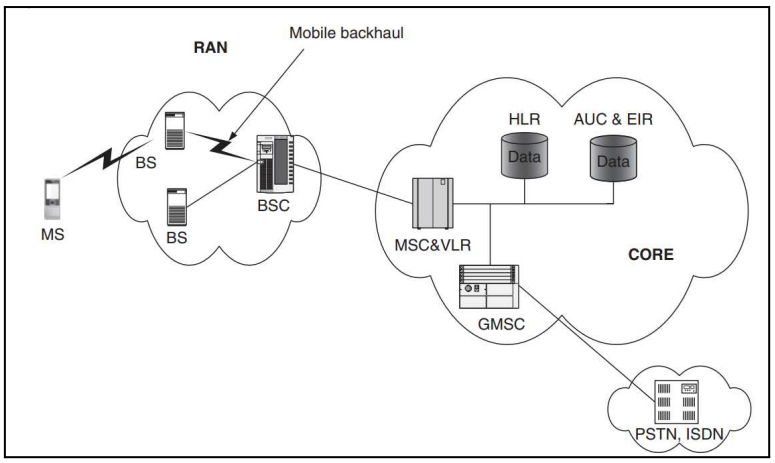

Figure 3. GSM architecture [2]

These are the abbreviations of the components shown in Figure 3.

MS: Mobile Station (Cell phone, User)
BS: Base Station (Site in GSM, NodeB in 3G, and eNodeB LTE)
BSC: Base Station Controller
MSC: Mobile Switching Center
GMSC: Gateway MSC
HLR: Home Location Register
VLR: Visited Place Register
AuC: Authentication Centre
EIR: Equipment Identification Register
PSTN: Public Switching Telephone Network

Mobile Station: User Equipment is a terminal that uses GSM to communicate over the air with a base station. This is called the Base Transceiver Station in GSM. The BS transceiver can be installed at an outdoor or indoor location, along with additional infrastructures such as antennas, power supply, and transmission equipment to connect to a Base Station Controller (BSC) [2].

- A radio cell is a logical object that is related to BS. It is a set of traffic and control channels. One BSC can control multiple BSs.
- The BSC manages radio resources on in-base stations. It is responsible for channel allocation and call setup and manages handovers.
- BSC and base stations are connected via fixed lines or point-to-point radio links. This part of the infrastructure is called Mobile Backhaul.
- The radio access network (RAN) comprises the BSs, BSCs, and mobile backhaul. The BS and BSC do different tasks to support communications over the air interface. Table 1 shows the task distribution among nodes.

Table 1: Functionalities of the base station and controller [2]

Main Functions	BS	BSC
Management of radio channels		✔

Mapping of upper layers to radio channels		✔
Channel coding and rate adaptation	✔	
Authentication		✔
Encryption	✔	✔
Frequency hopping	✔	
Uplink signal measurement	✔	
Traffic measurement		✔
Paging	✔	✔
Handover management		✔
Location update		✔

The RAN is linked to the Core network, which consists of a Mobile Switching Center (MSC), a Home Location Register (HLR), and other logical network nodes, such as Gateway MSC (GMSC), Equipment Identity Register (EIR), and Authentication Center (AuC). The MSC carries out all switching tasks, including path search, data forwarding, and service feature processing. The MSC and an ISDN switch vary primarily in that the MSC also needs to consider user mobility. The MSC must offer extra features for user location registration and connection handover management when a user switches between cells. There may be multiple MSCs in a cellular network, and each is in charge of a specific Location Area (LA).

A dedicated Gateway MSC (GMSC) manages calls that come from or end in the fixed network. The Interworking Function (IWF) performs the interworking of a mobile network and a fixed network (such as PSTN or ISDN). The cellular network's protocols must be translated into those of the corresponding fixed network. The MSC can implement GMSC or IWF as a standalone node or software feature with a few hardware interfaces.

The current position of a mobile user is kept in the Home Location Register (HLR) and Visited Location Register (VLR). The VLR is often a logical node that is implemented in MSC. User profiles are kept in HLR and VLR databases and are necessary for charging, billing, and other administrative tasks. The provisioning of new subscribers is done in the HLR database, which serves as a root database. Given how crucial the HLR database is to operator revenue, it frequently has a backup standby node that is geographically dispersed. Two additional databases carry out security-related tasks: Keys for authentication and encryption are stored in the Authentication Centre (AuC), while equipment data is registered in the Equipment Identity Register (EIR).

The network management is located in OMC[1]. The management of subscribers, terminals, charging information, network configuration, operation, performance monitoring, and network maintenance are all OMC activities. The hierarchical relationship between the network components, MSC, BSC, and BS are summarized in Figure 4. MSC is associated with a Location Area (LA), which is comprised of several BSCs and linked radio cell base stations. A BSC is assigned to each cell group, and both are connected by mobile backhaul [2].

[1] Operation and Maintenance Centre

There is at least one BSC for each LA; however distinct LAs may have different cells in the same BSC. The network operator determines how precisely the network region is divided among LAs, BSCs, and MSCs. The base station periodically broadcasts the Location Area Identity, or LAI, of each location area over a control channel. The mobile station records the most recent LAI while keeping an eye on the broadcast. The broadcasted LAI changes as the MS moves to a different LA.

The Public Switched Telephone Network (PSTN) is divided into islands by the Public Land Mobile Networks (PLMN) operated by various providers (PSTN). The call request is sent to the interface between the PSTN and the PLMN when the PSTN places a call to a mobile terminal that is part of a PLMN. The operator's Gateway Mobile Switching Center is the interface (GMSC). The Home Location Register (HLR) database contains information on each subscriber, part of the PLMN [2].

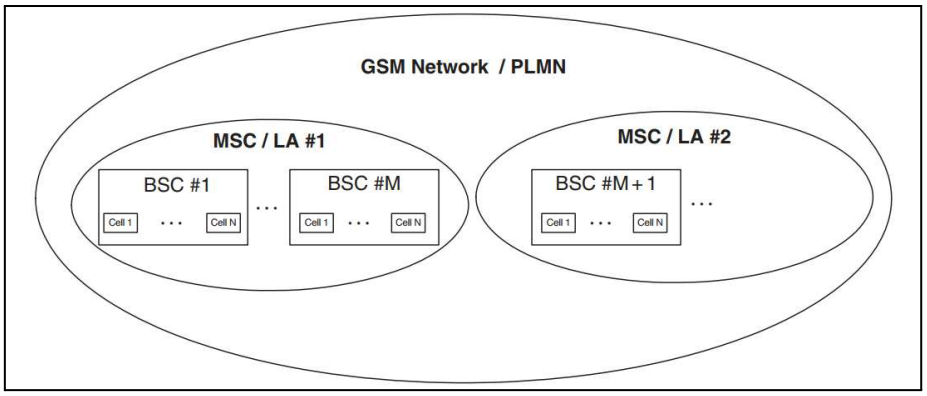

Figure 4: GSM system hierarchy [2]

2-3) GPRS/EDGE

The GSM Packet Radio Service, also known as GPRS, introduces the idea of allowing many users to utilize the cell's pool of accessible channels. The core subject is as follows:

- "Bundling" of timeslots is the practice of assigning multiple PTCH timeslots on a single carrier frequency to a single user. Timeslots may be bundled on both the uplink and the downlink; timeslots may be bundled.
- In contrast to circuit switch traffic, such as voice, a timeslot is not kept for one user; instead, a timeslot may be shared by many users according to their priority or on a round-robin basis.

The GPRS system was implemented as a GSM overlay with two new network nodes, SGSN and GGSN, as well as additional interfaces and functionalities in the base-station controller, BSC, Figure 5, to support the new idea of packet service. The enhanced Data rate for GSM Evolution is what EDGE stands for. With no adverse effects on other system components or nodes, it is a further development of GPRS that offers the option of a higher system data rate employing extended modulation methods at the air interface. EGPRS is the common acronym for GRS and EDGE working together [2].

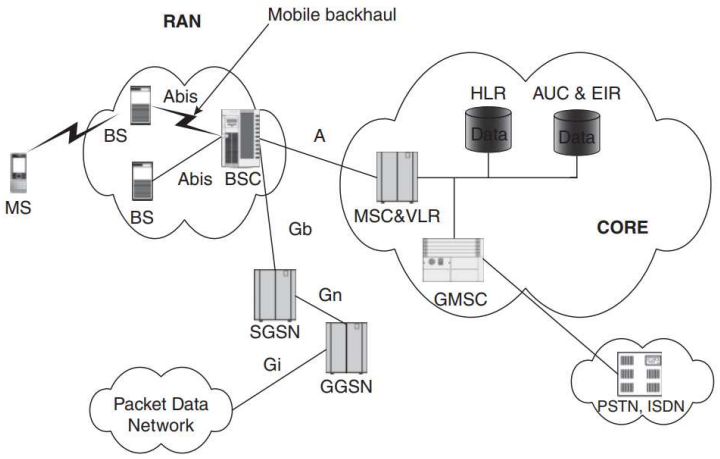

Figure 5: GPRS/EDGE Topology [2]

3 Chapter 3) Third Generation Network (3G), UMTS

3-1) Concept of 3G development

According to the European approach to 3G standardization, the third generation (3G) of mobile network technology (after GSM/EDGE) initially appeared in 1999 under the name of the UMTS[2]. UMTS and GSM are made backwards compatible by the 3GPP specification. Additionally, the GSM and UMTS networks can communicate with one another. Figure 6 depicts the general network architecture of the UMTS system, which is comparable to that of GSM [2].

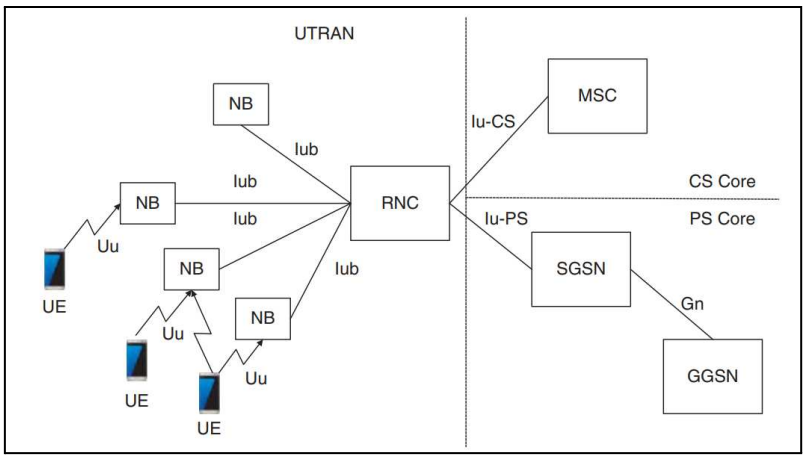

Figure 6: UMTS Architecture [2]

We can name the terminology used in 3G networks as follows:

- UE (User Equipment): It refers to the mobile terminal for the 3G system. Most terminals are dual-mode, multiband devices that can communicate over 2G and 3G networks.

[2] Universal Mobile Telecommunication System

- RAN (Radio Access Network): It stands for the 2G and 3G radio access components of the network.
- UTRAN (Universal Terrestrial Radio Access Network): It is the name of the radio access for WCDMA UMTS.
- RNC (Radio Network Controller): It is the base-station controller in UMTS.

GSM and UMTS switching systems may be interchangeable. However, UMTS includes several new protocols in the lower layers of the UTRAN and the Core Network (CN) that need specific GSM hardware, software, and interfaces for communication. The critical distinction between UMTS and GSM is that UMTS separates the radio network from the transport network, the access network from the core network, and the user plane from the control plane. Because the radio and network subsystems are separated, various RTAs (Radio Access Technologies) can be employed with the network subsystem. The GSM Core Network (CN) structure is comprised of two user-dependent domains that depend on traffic:

- circuit-switched traffic in the CS domain
- packet-switched traffic in the PS domain

The Home Location Register (HLR) and the Authentication Centre (AuC) or the Equipment Identity Register (EIR) are used by both traffic-dependent domains for subscriber administration, mobile station roaming and identification, and managing various services. As a result, the HLR contains subscriber data for GSM, GPRS, and UMTS. Both the GSM and the UMTS access networks are handled by two domains simultaneously, handling their respective traffic types.

The CS domain handles all circuit-switched traffic for the GSM and UMTS access networks, and the PS domain handles all packet-switched traffic for both access networks. The UMTS has a modular design that divides the protocol stack from the relevant network nodes to facilitate information flow, connectivity, and mobility functionalities. As shown in Figure 7, these modules define UMTS in a different domain structure that consists of an Access Stratum (UE and UTRAN) and a Non-Access Stratum that includes the USIM (Universal Subscriber Identity Module) [2].

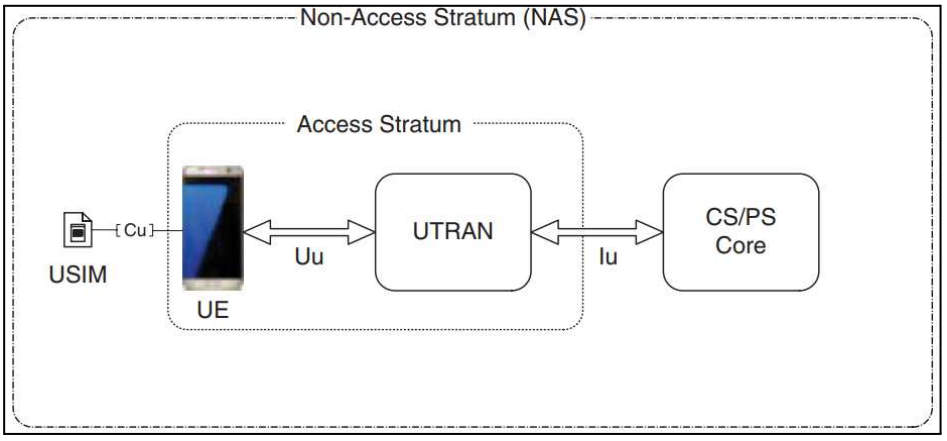

Figure 7: Modular functionality split in the UMTS [2]

According to this definition, the Access Stratum (AS) is a functional layer supporting the protocol stack between the mobile and radio access network. In contrast, the Non-Access

Stratum (NAS) is a functional layer in the UMTS protocol stack between the core network and user equipment at the application layer. NAS allows for transparent radio network communication between mobile and core network nodes (such as MSC and SGSN). A few of the features that NAS supports are as follows:

- Identity management
- Mobility Management
- Establishing, continuing, and ending communication sessions
- Call Control

The 3GPP TS 24.301 defines the NAS protocol stack. The 2G Mobile Station is replaced by the User Equipment (UE) in the UMTS (MS). The UE features a modular design made up of numerous components, as seen in Figure 8:

- The radio interface in the UE is terminated by the mobile termination (MT) module.
- A terminal adapter module terminates application-specific protocols.
- USIM is a user subscription module that provides access to the subscribed network.

The USIM differs from a GSM SIM in that it may be downloaded, accessed through an air interface, and updated by the network. A GSM SIM has much less capacity than a USIM, a Universal Integrated Circuit Card (UICC). It can store Java applications and profiles with user management and user rights [2].

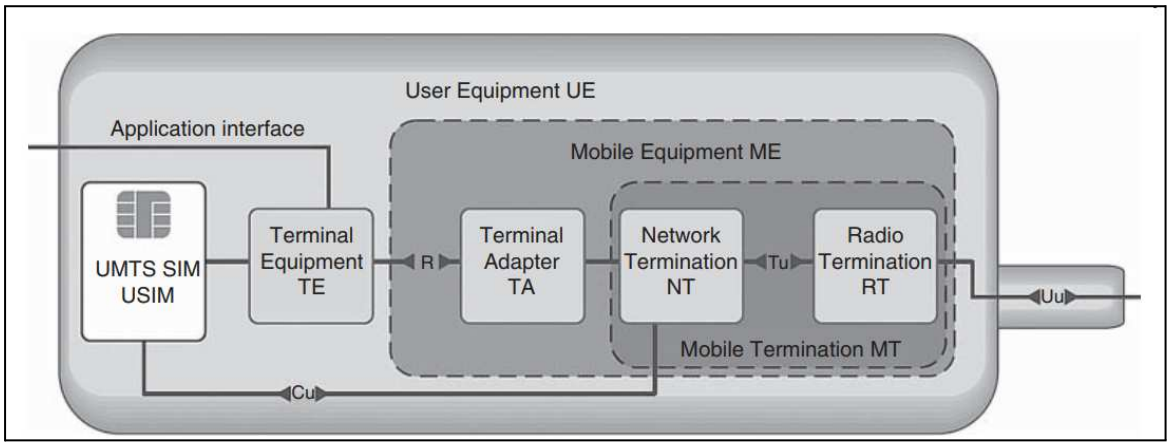

Figure 8: Modular architecture of UE [2]

3-2) WCDMA Concept

Wideband Code Division Multiple Access (WCDMA), with types FDD (Frequency Division Duplex) and Time-Division Duplex (TDD), were chosen by the European Telecommunications Institute (ETSI) in 1998, is the primary radio technology used in the UMTS. The spread spectrum is the primary technology for WCDMA, just like it is for 2G CDMA (IS-95). Spread spectrum techniques are used differently in 3G WCDMA than in IS-95, with various control and signalling channels that enhance call control and link performance management. The following fundamental ideas are applied in the WCDMA system [2]:

- Channelization (Spreading) and Scrambling
- Channel Coding
- Power Control
- Handover

A spread spectrum technique called channelization is used in WCDMA to transmit a radio signal over a frequency range far more extensive than the message bandwidth. To spread the signal spectrum in WCDMA, each information symbol is filled with pseudo-noise, such as a spreading sequence of "0" and "1" (chips), at a rate that is substantially higher than the symbol rate. Despite the variable symbol rate, the constant chip rate of 3.84 Mcps results in a variable amount per symbol. The information is transmitted in time slots mapped to radio frames. The WCDMA air-interface time arrangement is shown in Figure 9. The spreading coder generates 3.84 Mega chips per second (Mcps). This chip stream is divided into one-hundred 10ms radio frames, and each radio frame contains 15 slots leaving 2560 chips per time slot and 38400 chips per radio frame (or 3.84 Mega Chips per second: 3.84 Mcps).

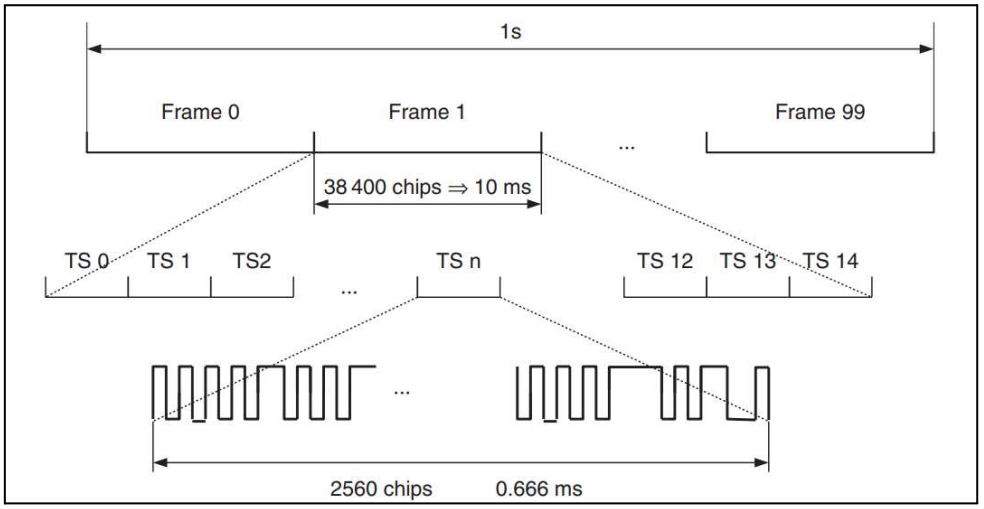

Figure 9: WCDMA timing [2]

3-2) UMTS Network Architecture and Interfaces

In the Radio Access Network, WCDMA adds two new nodes called Radio Network Controller (RNC) and NodeB (NB), as shown in Figure 10. These two nodes carry out functions comparable to the BSC and BTS in the GSM. Several new interfaces are defined between new network nodes in 3G WCDMA networks [2]:

- Iub Interface: In the radio access network, a NodeB and RNC are logically connected by the Iub interfaces. One or more NodeBs are connected to the RNC. It also enables the RNC and NodeB to negotiate radio resources and transmit uplink and downlink frames.
- Iur interface: It is designed to permit soft handover (HO) between RNCs within the same UTRAN.

A logical interface linking the radio access network and the core network is the Iu interface. The radio access and core network are separated from the system by the open interface, Iu. The core network handles switching, routing, and service control.

Implementation of the interface satisfies the 3GPP interface requirements. It has two significant examples of categories that differ:

- Iu-CS: This link is between the core network's circuit-switched domain and RAN. The interface is where an RNC, MSC Server, and MGW communicate. Additionally, it carries the RNC-transparent direct connection between the MSC and user equipment.
- Iu-PS: This link is between the packet-switched domain in the core network and RAN. An RNC and an SGSN communicate with one another through the interface. Additionally, it carries the RNC-transparent direct communication between the user equipment and the SGSN [2].

The main functions of the Iu interface are:

- Establishment, keeping, and releasing RAB[3]s.
- Handling RNC relocations as well as intra- and inter-system handovers.
- Reporting issues that do not pertain to a specific user's equipment
- Dividing up each UE at the protocol level to manage user-specific signalling.

The other interfaces in the operator network are deployed in practical networks for network administration and value-added services.

3-2-1) The NodeB in 3G RAN Networks

The primary function of the NodeB is to process the air-interface physical layer (channel coding and interleaving, rate adaptation, spreading, etc.). It is also responsible for Radio Resource Management (RRM) operations such as closed-loop power control [2].

3-2-2) Roles of RNC in 3G RAN network

The RNC is responsible for the following:

• CAC[4]: The current traffic load for each cell must be determined by the RNC. CAC feature uses this data to determine whether the interference level is tolerable and, if necessary, to reject the call.
• RRM[5]: The RNC manages all associated cells' radio resources. Calculating interference and usage levels and priority control are part of this.
• Radio bearer Setup and release: The establishment of a logical data connection is what the radio bearer setup is all about; it makes no distinction between PS and CS data transmission across the radio bearer.
• Code Allocation: The RNC can assign different portions of the code tree to different mobile stations.
• Power Control: The target control values are defined in the RNC; however, each NodeB performs the quick power control.
• Packet scheduling: The same resource is shared by several mobile stations. The RNC periodically assigns transmission capacity to each MS while considering negotiated QoS.

[3] Radio Access Bearers
[4] Call Admission Control
[5] Radio Recourse Management

• Hanover management. The RNC selects a handover, performs signalling with the new cell, and notifies the MS about the new channel based on the measurements provided by NodeB and MS.
• Encryption of CS services. The mobile terminated call is encrypted in RNC before transmitting over Iub (air) interface.

Three logical RNC types are established in the 3G network regarding the connection between NodeB and RNC to provide soft handover. The first stop is the Iub interface toward NodeB with a Controlling RNC (CRNC). All load and congestion control of connected cells are handled by the CRNC, which also handles admission control and code distribution for new radio connections. The Serving and Drift RNC idea is shown in Figure 10 [2].

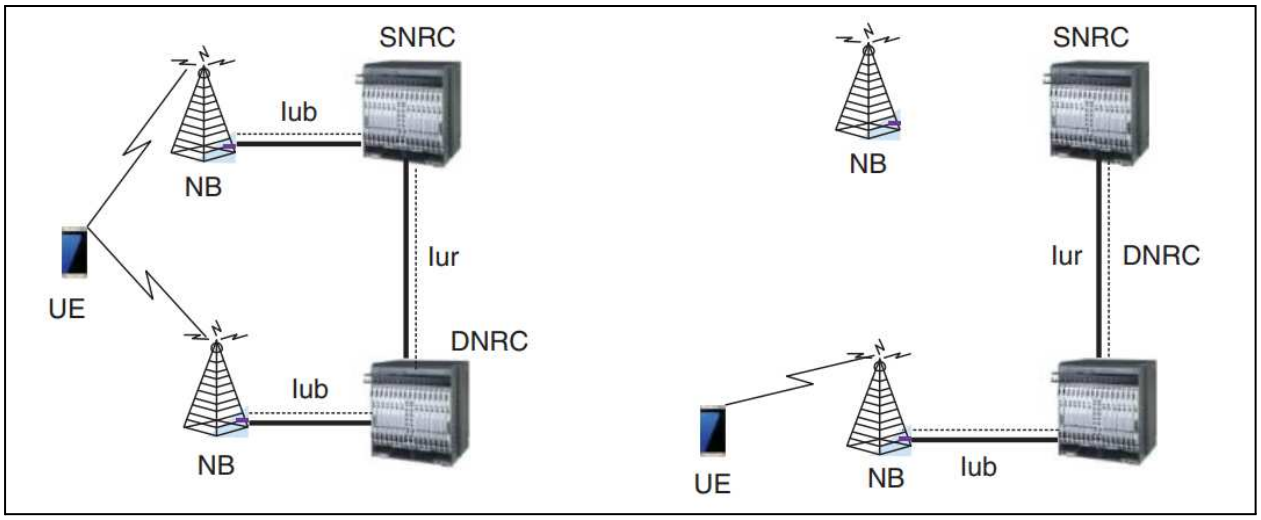

Figure 10: Logical role of the RNC for one UE UTRAN connection [2]

4 Chapter 4) 4G-Long Term Evolution (LTE) Systems

4-1) Introduction to LTE

Widespread acceptance of the 3G/HSPA system led to tremendous growth in the usage of mobile data. That was also stimulated by the availability of affordable mobile devices and flat data pricing by the operators. Mobile internet access extended from laptop usage to smartphones, thus facilitating development in mobile network performance towards very high instant peak data rates and very low latencies. The considerable growth in mobile users and the traffic to be carried by mobile networks demands a significant increase in system capacity that, in turn, instigates a new technological solution to network design. When high capacity and high performance at flat pricing are offered to the end customer, then cost per bit becomes a critical issue for the service provider [2].

These three key drivers, capacity, user experience and lower cost per bit, have led to the specification of a Long-Term Evolution (LTE) of UTRAN. In contrast, mobile system core specification is defined as a System Architecture Evolution (SAE), also called Enhanced Packet Core (EPC). LTE, together with EPC forms the Evolved Packet System (EPS). To satisfy capacity demands, a different portion of the radio spectrum was released for LTE in the 2.6 GHz to 700 MHz range. New radio technology deployed in LTE delivers high spectrum efficiency and capacity per site, reducing CAPEX and OPEX for service providers. A significant reduction in cost per bit is ensured with flat IP-based LTE network architecture, cost-efficient high bandwidth backhaul and transport network. The 3GPP has set performance targets for an LTE of peak data rates >100 Mbps in DL and >50 Mbps in UL with a latency of less than 5 ms on the air interface per link. The spectral efficiency of LTE can exceed the one of UMTS Release 6 by a factor of 3–4 in DL and a factor of 2–3 in UL. The access scheme in LTE is OFDMA in the downlink and SC-FDMA in the uplink. OFDM allows for improved interference control, advanced scheduling techniques and ease of implementation of MIMO to improve spectrum efficiency [2].

Further, OFDM enables scaling of user bandwidth dynamically from meagre bit rates required; for example, for control up to very high instantaneous peak data rates above 100

Mbps in the downlink and 50 Mbps in the uplink. With scalable RF bandwidth, OFDM allows for scaling the operator bandwidth from 1.4 or 3 MHz in re-farming scenarios up to 20 MHz for very high capacities. OFDM technology can be used in both FDD and TDD multiple-access schemes so that both LTE-FDD and LTE-TDD systems are standardized, thus allowing flexibility in implementation [2].

4-2) Architecture of Evolved Packet System

EPS (Evolved Packet System) is known as Evolved UTRAN (E-UTRAN), Evolved Packet Core (EPC), and connectivity to older 3GPP access and non-3GPP access systems. The EPS for the 3GPP access system is depicted in Figure 11 and was created as an extension of the 2G and 3G architecture. In comparison to GPRS/UMTS, the EPS architecture features fewer network elements on the data channel, supports RAN capability in a single node, and separates the control and user-plane network elements (MME and Serving Gateway). These are the new network components [2]:

.

- Mobility Management Entity (MME):

It is the control plane (C-plane) functional element in EPC and as a terminating point for Non-Access Stratum (NAS) signalling, MME controls and saves UE context, creates temporary identities, and assigns them to UEs authenticate the user, regulates mobility and bearers, and manage UE context.

- Serving Gateway (S-GW):

It is the user plane (U-plane) gateway to the E-UTRAN. S-GW acts as an anchor point for both intra-3GPP mobility and inter-eNB handover (i.e. inter-3GPP access mobility between LTE and 2G or 3G). For UEs in the ECM-IDLE state, it is also in charge of packet forwarding, routing, and buffering downlink data.

- Packet Data Network Gateway (P-GW):

The operator's IMS[6] or any other PDN is accessed through this U-plane gateway, which is known as the Packet Data Network. P-GW is in charge of allocating the user's IP address, supporting charging and enforcing rules.

- E-UTRAN:

It is the radio access part of the LTE Network. The legacy network elements interfacing with LTE are as follows:

- Gateway GPRS Support Node (GGSN):

It is in charge of terminating the Gi interface toward the PDN for legacy 2G/3G access networks. LTE interfaces this node only as a part of P-GW functionality and from the perspective of inter-system mobility management.

- Serving GPRS Support Node (SGSN):

[6] IP Multimedia Subsystem

It is in charge of transmitting packets between the Core Network and the 2G/3G RAN. LTE interfaces the SGSN only in case of inter-system mobility management.

- Home Subscriber Server (HSS):

It is the IMS Core Network entity in charge of managing user profiles and performing the authentication and authorization of users. The user profiles managed by HSS; consist of subscription and security information as well as details on the user's physical location. While IMS is not a mandatory network element, the HSS is a necessary node for the operation of the LTE system.

- Policy Charging and Rules Function (PCRF):

It is responsible for brokering QoS Policy and Charging Policy on a per-flow basis.

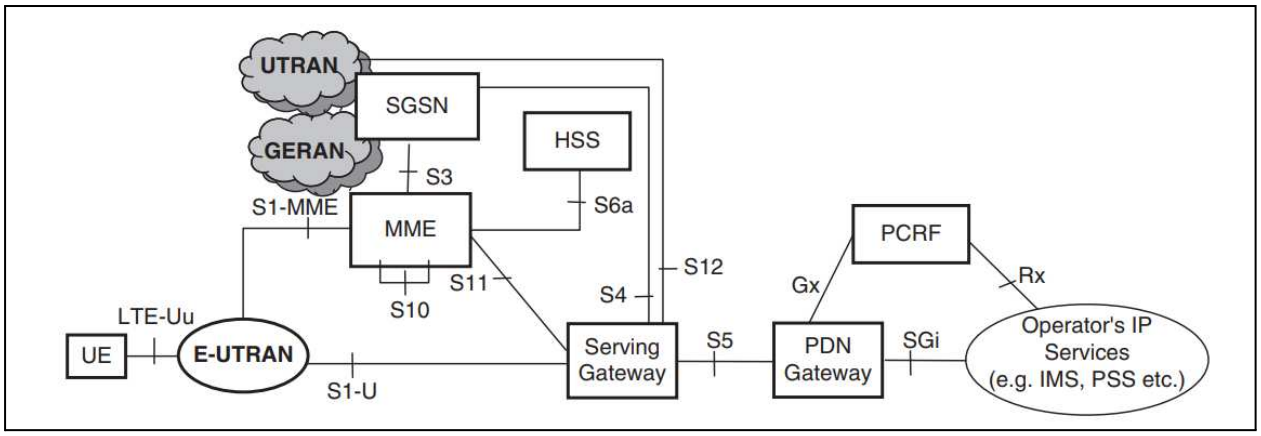

Figure 11: EPS architecture for 3GPP accesses [2]

4-3) EUTRAN Interfaces

eNodeB (eNB) is the only node type in E-UTRAN, which gives UE the air interface. The X2 interface allows eNBs to communicate with one another, while the S1 interface allows them to communicate with MMEs and S-GWs. Multiple MMEs and S-GWs can connect to a single eNB. This feature, known as S1-flex, offers flexibility and reliability. Figure 12 shows the eNB connection. The eNB handles radio transmission to and reception from UE. The RNC node does not exist in the LTE network, as shown in Figure 12. Instead, the eNB handles RNC functionalities. This contains scheduling user data, radio bearer control, management of the radio resources (including admission control), and control of the signalling through the air interface [3].

However, because e-UTRAN lacks an anchor point and capability, the X2 interface will only link eNodeBs with nearby cells. The X2 supports relocation capabilities with packet forwarding rather than drift-like RNC functionality. The S1 interface is used to link the eNB to the core network. The Iu-PS interface in the 3G system and the S1 interface are relatively comparable. The IP-based S1 and Iu-PS user planes are transport tunnels independent of the content of the sent packets. The EPC or the eNB puts the end user's IP packets into the S1 IP tunnel.

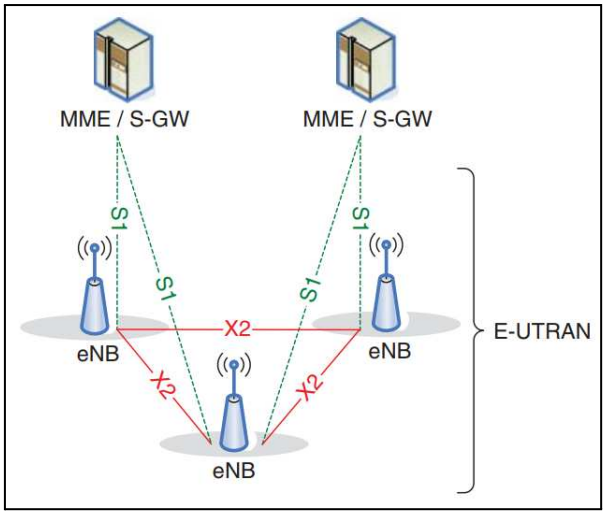

Figure 12: E-UTRAN and EPS with S1-flex interface [3]

4-4) LTE-Advanced

International Mobile Telecommunications-Advanced refers to this idea as it relates to the overall network system (IMT-Advanced). With the first release of 3GPP Release 10, the organization began to create the LTE-Advanced (LTE-A) radio network standard. The criteria for radio networks from IMT-Advanced include [3]:

- support for up to 100 MHz bandwidth
- peak data speeds up to 1 Gbps for nomadic (low mobility scenario) and 100 Mbps for the high mobility case
- increased spectral efficiency in various environments

Another 3GPP requirement is backward compatibility with 3GPP Release 8 LTE. It includes support for inter-RAT mobility between LTE-A and LTE, GSM/EDGE, HSPA, and cdma2000. Additionally, the LTE-A must allow flexible spectrum allocation. LTE-A should allow asymmetric bandwidth allocation for FDD downlink/uplink and non-contiguous spectrum allocation due to variations in the spectrum available in different use cases and scenarios. The following are the system performance targets for 3GPP Release 10 (LTE-A) [2]:

- cost-efficient data rates of 3 Gbps for downlink and 1.5 Gbps for uplink;
- spectral efficiency up to 30 bps/Hz; (Spectral efficacy is defined as the rate divided into the system bandwidth)
- higher capacity with an increased number of simultaneously active clients;
- improved performance at some weaker covered locations, same as cell edges.

The main new functionalities introduced in LTE-Advanced are:

- Carrier Aggregation (CA),
- Enhanced MIMO,
- Support for Relay Nodes (RN)
- Coordinated multipoint transmission and reception (CoMP).

5 Chapter 5) 5G Development

5-1) Introduction to 5G

A few years after 2020, 5G, or the Fifth Generation of mobile communication technology, is expected to hit the market. An increasingly networked human society's demands are one of the main reasons for the advent of the new generation of technologies. The Internet of Things (IoT) sector, Machine-to-Machine (M2M) communications, cloud computing, and many other technologies are predicted to grow exponentially in connection and traffic density. So, 5G technology will be required to take network performance to the next level. Additionally, 5G will need to solve the present LTE and LTE-Advanced (LTE-A) limit in terms of latency, capacity, and reliability. Some of the specifications for 5G networks that are frequently mentioned in the most recent research are [4]:

- Address the growth required in coverage and capacity;
- Address the growth in traffic;
- Provide better Quality of Service (QoS) and Quality of Experience (QoE);
- Support the coexistence of different Radio Access Networks (RAN) technologies;
- Support a wide range of applications;
- Provide peak data rates higher than 10 Gbps and a cell edge data rate higher than 100 Mbps;
- Support radio latency lower than 1ms;
- Support ultra-high reliability;
- Provide improved security and privacy;
- Provide more flexibility and intelligence in the network;
- Reduction of Capital and Operational Expenditures (CAPEX and OPEX);
- Provide higher network energy efficiency.

The 5G research and studies performed by the Next Generation Mobile Network Alliance (NGMN) describe a multi-faceted 5G system which can support multiple combinations of reliability, latency, throughput, positioning and availability. This is achievable by introducing new technologies in access and the core parts of the network. According to ITU-R studies, the major usage scenarios for the new 5G system can be largely classified into five categories [2]:

1- Enhanced Mobile Broadband (EMBB):
Mobile Broadband addresses the human-centric use cases for access to multimedia content, services and data. In these use cases, we need to provide users with more throughput, and a huge amount of payload is downloaded an uploaded in this category. This usage scenario covers a range of cases, including wide-area coverage and hotspot, which have different requirements.

2- Ultra-reliable and low-latency critical communications (URLLC):
In this category; throughput, latency, and availability requirements for this use in the case are very strict. This could include communications from and to interactive games, sports, drones, robotics, and emergency communications. Examples of machine forms of communication include wireless control of industrial manufacturing or production processes, remote medical surgery, distribution automation in a smart grid, transportation safety, and others.

3- Machine Type Communications (MTC):
Massive and critical MTC can be distinguished from one another. Massive MTC is defined as the transmission of non-delay-sensitive data by a very large number of connected devices, usually at a very low rate. Devices must be reasonably priced and have a long battery life. Applications like traffic safety/control, critical infrastructure control, and wireless communication for industrial processes are all examples of critical MTC. These applications need wireless communication with extremely high levels of availability and reliability as well as very low latency. Wide instantaneous bandwidth is essential to meet capacity and latency requirements, even though the average amount of data transferred to and from devices may not be very much. A later study by 3GPP added two other 5G use cases:

4- Network Slicing:
Network operation that is enhanced with network slicing, routing, migration and interworking and energy saving.

5- Enhancement of Vehicle-to-Everything:
for example, autonomous driving, safety and non-safety aspects associated with vehicles.

5-2) 5G Architecture

5G technology has been designed for high throughput and low latency applications. Massive connections are a few examples of burst data traffic patterns and models. 5G is intended to

offer various services (e.g., IP data traffic, non-IP data traffic, short data bursts and high throughput data transmissions). Different PDU session types, such as IPv4, IPv6, IPv4v6, Ethernet, and Unstructured, are supported in 5G networks [4].

The primary features of 5G are the introduction of a new radio interface and New Radio (NR), which provides the adaptability required to handle diverse types of services. The 5G Access Network's ability to connect to both LTE and 5G Core Network is another important aspect of 5G. This is referred to as the NSA architecture, whereas the SA architecture refers to a 5G AN (Access Network) connected to a 5G CN (Core Network). The 5G System delivers a wide range of features on the Core Network side, including enhanced Network Slicing, mobile edge computing, and network capability exposure. These ideas are all presented below:

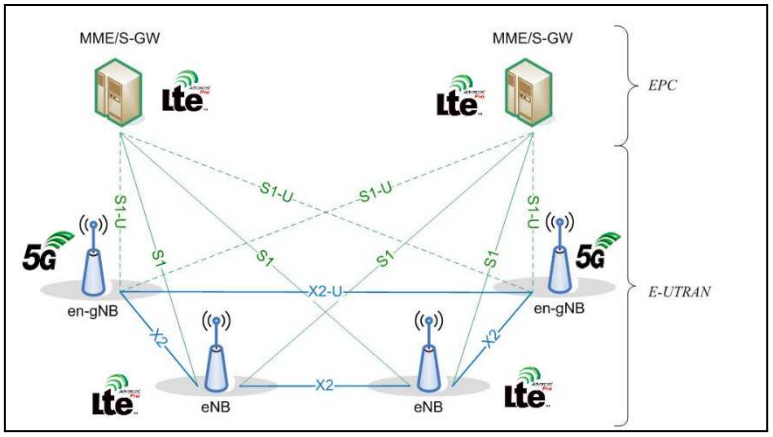

Figure 13: The NSA Architecture [4]

The NSA architecture, where the 5G Access Network couples to the 4G Core Network, is considered an initial solution toward "full 5G" implementation. The LTE eNB and the 5G NR base station (en-gNB) connect via the X2 interface in the NSA design. As mentioned earlier, the X2 interface was developed to link two eNBs. In Release 15, linking an eNB and en- gNB is also supported to offer NSA. Figure 14 below is an illustration of the SA architecture.

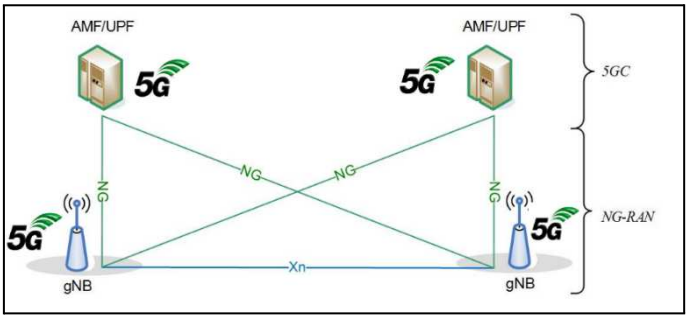

Figure 14: The SA Architecture [4]

The SA architecture can be seen as the "full 5G deployment", not needing any part of a 4G network to operate. The NR base station (gNB) connects via the Xn interface, and the Access

Network (called the "NG-RAN for SA architecture") connects to the 5GC[7] network using the NG[8] interface.

5-2) Operational Requirements for 5G Network

In addition to many new services, the 5G system will support most of the existing EPS[9]. The existing EPS may be accessed using the new 5G access technologies even where the EPS specifications might indicate E-UTRA(N) only. The following exceptions will apply [4]:

- CS voice service continuity and/or fallback to GERAN[10] or UTRAN,
- Handover between NG-RAN[11] and GERAN,
- Handover between NG-RAN and UTRAN,
- Access to a 5G core network via GERAN or UTRAN.

The 5G system will support mobility procedures between a 5G core network and EPC with minimum impact on the user experience (QoS, QoE).

5-3) 5G Device Requirements

The network will be able to configure the access technology and transport protocol of 5G terminals to a high degree over the air. This ability will eliminate terminal-type dependence while enabling effective logical division (slicing) for various services. Depending on QoS requirements, radio parameters and conditions, 5G UE can dynamically select particular profiles. The 5G UE must handle several frequency bands and various transmission techniques (TDD, FDD, and mixed). The 5G device will be able to aggregate data flows from several technologies and carriers while maintaining concurrent Multi-RAT (multiband) connectivity. These standards significantly improve the 5G devices' resources, signal processing, signalling, and power efficiency [2].

5-4) 5G Capabilities

The following items are regarded as essential 5G characteristics. The values listed below are IMT-2020 research and study goals and may be expanded upon or changed in future ITU recommendations.

- The peak data rate for 5G Enhanced Mobile Broadband is predicted to exceed 10 Gbps in both indoor and dense outdoor locations.
- The following data rates for user experience are anticipated:
  At least 10 Mbps for everywhere
  100 Mbps for wide-area coverage
  1 Gbps for indoor coverage

[7] 5G Core
[8] New Generation
[9] Evolve Packet System
[10] GSM EDGE Radio Access Network
[11] New Generation-Radio Access Network

- The spectrum efficiency is anticipated to be three times higher than LTE-A For improved eMBB applications. For an indoor scenario and small cell, 5G is anticipated to support a 10 Mbps/m2 area traffic capacity.
- The energy consumption for the 5G RAN should not be greater than LTE networks deployed today while delivering enhanced capabilities.
- Over-the-air latency is to be reduced to 1 ms to support services with very low-latency requirements.
- 5G is also expected to enable high mobility up to 500 km/h with acceptable QoS.
- 5G is expected to support a connection density of up to $100km^2$ massive MTC scenarios.

5-5) 5G Spectrum Allocation

The NB-IoT will likely be deployed in frequency bands below 2 GHz, providing high capacity and in-depth coverage for many connected devices. Different 5G scenarios, such as increased mobile broadband, ultra-reliable and low-latency communications, and massive machine-type communications, would require different spectrum bandwidths. Considering a wideband contiguous spectrum allocation over 6 GHz would be necessary for those applications requiring bandwidth from several hundred MHz up to 1 GHz. Various frequency bands and bandwidths are essential for 5G services; for instance, MTC services may utilize relatively narrow bandwidth, while eMBB requires very wide bandwidths for high-capacity usage cases.

An overall solution for 5G will include a spectrum below 6GHz as well as the spectrum at the higher frequencies in the range of 6–100 GHz. By around 2020, most mobile communications will be equipped with LTE. Therefore, an evolution of LTE to 5G has to support dual connectivity between the LTE operating below 6 GHz and New Radio (NR) technology operating in the range above 6 GHz. The schematic spectrum allocation for further evolution of LTE and 5G New radio is illustrated in Figure 15. In [26], we can find the findings of a study on the radio propagation environment of IMT in bands between 6 GHz and 100 GHz. The research discusses solutions based on MIMO and beamforming with several antenna elements to address the frequency-dependent propagation loss and offers performance simulation results for a variety of deployment scenarios [2].

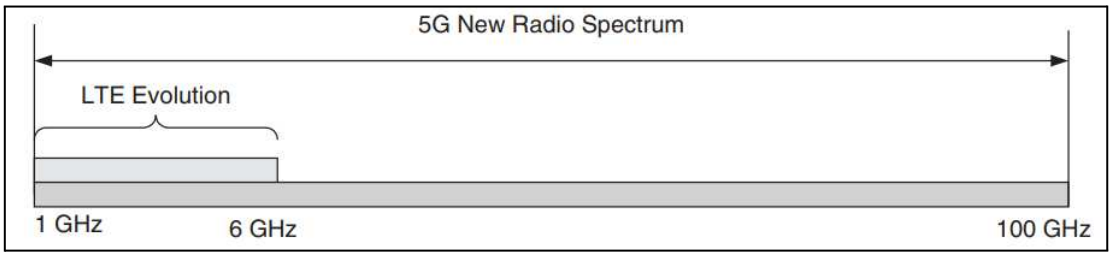

Figure 15: Planned frequency spectrum allocation for 5G [2]

5-6) 5G Technology Components

Here, I have mentioned the headlines of 5G technology components:

- **Technologies to enhance the radio interface**
  Advanced Modulation and Coding
  Non-Orthogonal Multiple Access (NOMA)
  Active Antenna System (AAS)
  3D beamforming and Multi-user MIMO (MU-MIMO)
  Massive MIMO
  Full-Duplex mode
  Self-Backhauling
- **Technologies to enhance Network Architectures**
  Self-Organized Network (SON)
  Software-Defined Networking (SDN)
  Network Slicing
  Cloud RAN and Virtual RAN
  Open RAN
- **Technologies to support a wide range of emerging services**
- Technologies to enhance user experience
- Technologies to improve network energy efficiency
- Terminal technologies
- Technologies to enhance privacy and security.

The whole spectrum of 5G technology enhancements is out of this project's scope. We briefly reviewed a few technological developments related to the radio part. However, we will dive deep into the RAN part (Cloud RAN and Open RAN) and the use of machine learning in this 5G development. However, before starting that, it is essential to look into 5G open issues and challenges before addressing the topic of 5G-RAN.

5-7) Challenges and Open Issues in 5G

Here, we have presented 5G technical issues, including mmWave communication, D2D communication, backhaul transmission, technology maturity, and security concerns [5].

5-7-1) mmWave Communication

Application:
An important part of the 5G mobile network is to provide eMBB services such as VR, AR and ultra-high definition video (UHDV). It can support the requirements of high growth of mobile

traffic demand and reduce the bottleneck effects of wireless bandwidth; it is a key problem of 5G networks.

Drawback:
Blockage of electromagnetic signals and designing integrated circuits are the challenges of mmWave communications. These waves in the 60 GHz band are sensible to blockage by barriers (e.g., humans and furniture), E.g. penalizing 20-30 dB due to blocking by a person).

5-7-2) D2D communications

Application:
In the evolution to 5G, traditional performance indicators, such as network capacity and spectral efficiency, must be continually improved. A wider variety of communication modes and applications must be provided with enhanced user experience. Device-to-device (D2D) technology has drawn widespread attention in the industry for its potential to improve system performance, enhance user experience, and expand cellular applications. With cellular-based D2D communication, also called proximity service (ProSe), user data can be directly transmitted between terminals without routing via eNBs and core networks. D2D communication structure differs from a traditional cellular network (Figure 16).

D2D communication helps increase spectral efficiency, enhance user experience, and expand communication applications. Applications of 5G D2D include local service, emergency communication, and IoT enhancement. A typical application of D2D-based IoT enhancement is vehicle-to-vehicle (V2V) communication in the Internet of Vehicles (IoV). When running at high speeds, a vehicle can warn nearby vehicles in D2D mode before it changes lanes or slows down. D2D can also be applied in other potential scenarios, such as multiuser MIMO enhancement, cooperative relaying, and virtual MIMO. D2D can also help to solve problems in new wireless communication scenarios and support indoor positioning in 5G networks at a low cost [5].

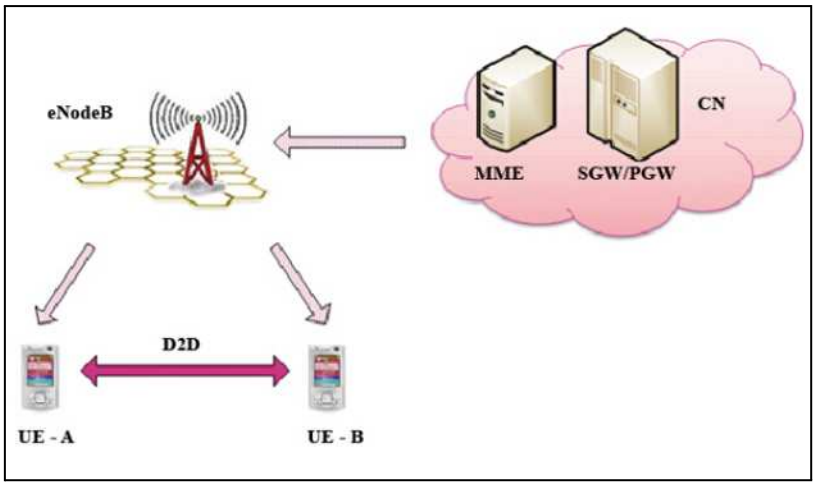

Figure 16: D2D Communication [6]

Drawback:

There are two main issues for D2D communications in the 5G era. The first one is controlling and limiting interference among D2D devices and microcell users because there is no operator control as a central body for direct communications, interference management, and resource allocation. Another issue is security and privacy in D2D communication because of routing users' data through other users' devices [18].

5-7-3) Backhaul

Application:
To meet the anticipated capacity of the 5G network, vendors and players need to develop new technologies in telecommunications [19-20]. The backhaul network is responsible for transmitting this volume of traffic. Backhaul (backhaul network or backbone or transport), in cellular networks, is defined as a network that connects the access network (e.g. eNB) to the core network and is composed of fibre, copper, microwaves and sometimes satellite [21].

Drawback:
Utilizing backhaul networks for small cells to support high data rates and low latency is a significant challenge for operators due to the need for adequate fibre networks in many different areas. These backhaul networks are necessary for transferring the heavy traffic of the high-dense cells with capacity constraints such as delay and delay. There is no one unique solution to address 5g backhaul requirements. Future 5G backhaul can be designed by utilizing existing transmission networks such as xPON and new technologies such as mmWave. In this regard, authors in [22] have suggested that technology adoption, such as SDNs can help in the evolution of 5G backhaul to facilitate backhaul management in a heterogeneous environment.

5-7-4) Technology maturity

Currently, operators have started 5G service with eMBB cases, and other service types, e.g. URLLC is only available for a while due to a lack of technology maturity. Despite the presenting architecture and some implementation, however, a maturity level to propose different services is
required for used technologies in the 5G era. Because the growth of 5G requires the development of enablers such as SDN, orchestration and NFV and RAN technologies. Maturity in technology requires concentration on a specific one and avoiding fragmentation in technology [24]. For example, each vendor works individually instead of focusing on a specific one, such as NFV. This could delay the maturity of the NFV implementation and therefore limit us in providing. Today, Cloud RAN and Open RAN are promising solutions for technology maturity in 5G to make no dependence on particular hardware and vendor equipment [5].

5-7-5) Security challenges

The 5G network leverages cutting-edge technologies, including virtualization, software-defined networking (SDN), and network function virtualization (NFV) to provide services and use cases. On the other side, service security can only be offered if the network architecture is secure. Traditional networks include components that are isolated from one another, whereas 5G networks have virtualized services and shared infrastructure resources.

In this setting, many virtual network slices are formed, requiring various levels of protection. In addition, security heterogeneity in 5G networks is a novel issue that needs to be considered. ITU service framework states that 5G supports numerous services with varied requirements, such as mMTC, URLLC, and eMBB. Different levels of security are required for each of them. IoT services, for instance, require minimal security, but URLLC services, such as industrial services, demand more effective protection. A multi-level architecture security framework is required in 5G networks to support policies, threats detection, and threat mitigation on a dynamic basis [5].

5-8) Overview of 5G-Radio Access Network (NG-RAN)

Figure 17 below, has been extracted from TS 38.40 and shows the overall architecture of the Access Network in the 5G scenario:

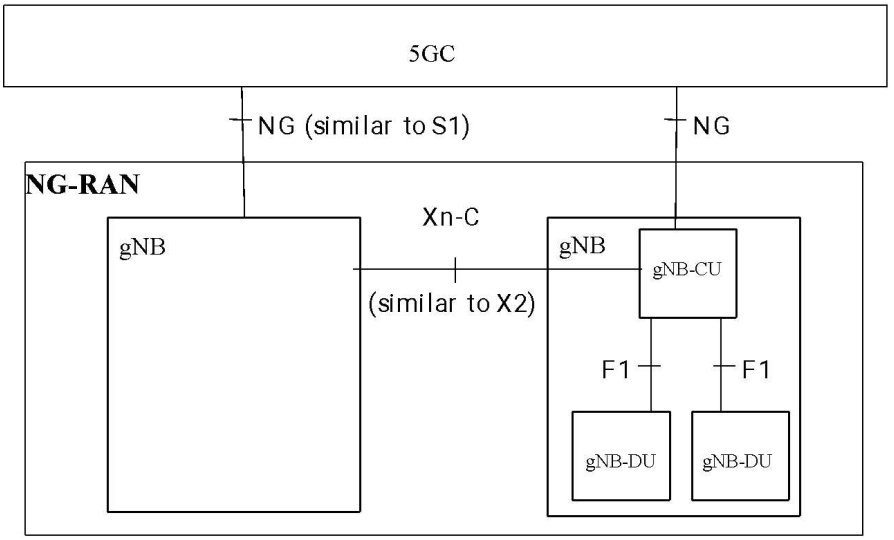

Figure 17: Overall NG-RAN architecture [7]

The NG-RAN consists of gNBs connected to the 5GC through the NG interface. This connection is similar to the LTE's S1 interface. As briefly mentioned in section 5-1, the gNB (5G Node B) can be connected to another gNB through the Xn interface, based on (and very similar to) the LTE's X2 interface. The gNB may be further split into a gNB-Central Unit (gNB-CU) and one or more gNB- Distributed Unit(s) (gNB-DU), linked by the F1 interface.
The gNB performs the following tasks [7]:

- Functions for Radio Resource Management: Radio Bearer Control, Radio Admission Control, Connection Mobility Control, and Dynamic allocation of resources to UEs.
- IP header compression, encryption and integrity protection of data.
- Selection of an AMF at UE attachment when no routing to an AMF can be determined from the information provided by the UE.
- Routing of User Plane data towards UPF(s).
- Routing of Control Plane information towards AMF.
- Connection setup and release.
- Scheduling and transmission of paging messages.
- Scheduling and transmission of system broadcast information
- Measurement and measurement reporting configuration for mobility and scheduling.
- Transport level packet marking in the uplink.
- Session Management.
- Support of Network Slicing.
- QoS Flow management and mapping to data radio bearers.
- Support of UEs in RRC_INACTIVE state.
- The distribution function for NAS messages.
- Radio access network sharing.
- Dual Connectivity.
- Tight interworking between NR and E-UTRA.

5-9) NG-RAN Interfaces

Understanding the control and user plane concepts is best before introducing the 5G interfaces. Since the functions of interfaces depend on this classification, for instance, NG-C is in charge of managing control plane operations between G-NG and 5GC. In contrast, Xn-U manages user plane operations between two G-NBs [2].

5-9-1) The NG and S1 Interfaces

Before introducing 5G-RAN interfaces, one must know about two 5GC components connected to 5G-RAN.

- Access and Mobility Management Function (AMF): It is one of the control plane network functions of the 5G core network (5GC) with the following main functions:
  Registration Management
  Reachability Management
  Connection Management
  Mobility Management

- User Plane Function (UPF): It is the function that does all of the work to connect the actual data coming over the Radio Area Network (RAN) to the Internet.

The NG interface connects the 5GC to the NG-RAN. Figure 18 shows the logical segmentation of the NG interface. From the NG perspective, the 5GC access point is either the control plane AMF logical node or the user plane UPF logical node, whereas the NG-RAN access point, is an NG-RAN node, either an ng-eNB or a gNB. Accordingly, depending on the 5GC access point the NG-RAN node is linked to, NG interfaces are defined at the boundary: NG-C towards an AMF and NG-U towards a UPF [2].

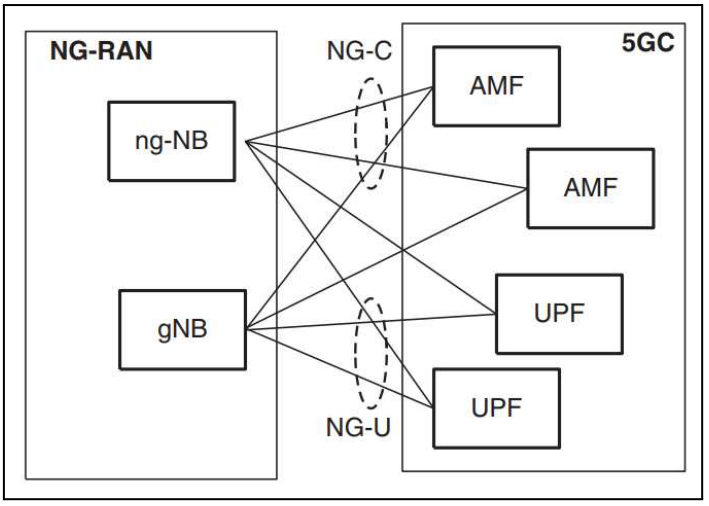

Figure 18: NG interface architecture [2]

The NG-RAN may thus have several NG access points towards the 5GC. The NG interface has the following features [2]:

- It is open;
- It is a point-to-point logical interface between an NG-RAN node and a 5GC node, and it is feasible in the absence of a direct physical connection between the NG-RAN and 5GC;
- It supports control plane and user-plane separation;
- It separates Radio Network Layer and Transport Network Layer;
- It is decoupled with the possible NG-RAN deployment variants.

There may be multiple NG-C logical interfaces towards the 5GC from any one NG-RAN node. The selection of the NG-C interface is then determined by the NAS Node Selection. There may be multiple NG-U logical interfaces towards the 5GC from any one NG-RAN node. The selection of the NG-U interface is done within the 5GC and signalled to the NG-RAN Node by the AMF. The NG interface support:

- Protocols for setting up, managing, and releasing the NG-RAN portion of PDU sessions;
- The methods for performing intra-RAT and inter-RAT handovers;
- The protocol-level isolation of each UE for user-specific signalling management;
- The exchange of NAS signalling messages between the UE and the AMF;
- The mechanisms for packet data stream resource reservations.

The gNB/eNB and the User-Plane Function define the NG user-plane interface (NG-U) (UPF). The user-plane PDUs are carried between the gNB/eNB and the UPF using GTP-U, which is built on top of UDP/IP at the transport network layer. The gNB/eNB and the AMF are described as the NG control plane interface (NG-C). Figure 18 demonstrates the control plane protocol stack of the NG interface. IP transport is the foundation of the transport network layer. To reliably carry signalling messages, SCTP is introduced to IP. The NG-AP stands for the application layer signalling protocol (NG Application Protocol). Application layer messages are guaranteed to arrive thanks to the SCTP layer. IP layer point-to-point communication is used in transit to signalling PDUs [2].

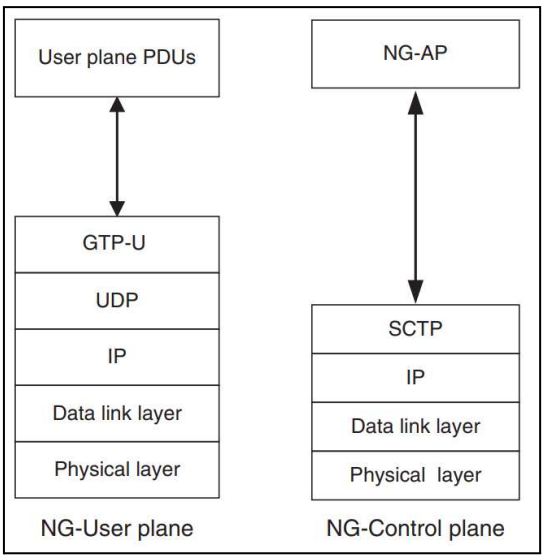

Figure 19: NG interface protocol stack [4]

5-9-2) The Xn and X2 Interfaces

The Xn User-plane (Xn-U) interface is specified between two gNBs connected to 5GC, between a gNB and an ng-eNB connecting to 5GC, and between two ng-eNBs connecting to 5GC. Figure 20a depicts the user-plane protocol stack on the Xn interface. The user-plane PDUs are transmitted by GTP-U on top of UDP/IP at the transport network layer, which is based on IP transport. Data forwarding and flow control are performed through connectionless, non-guaranteed delivery of user-plane PDUs delivered by Xn-U [2].

Figure 20b depicts the Xn interface's control plane protocol stack. On top of IP, SCTP is used to construct the transport network layer. Xn-AP is the name of the application layer signalling protocol (Xn Application Protocol). Application layer messages are guaranteed to arrive thanks to the SCTP layer. The signalling PDUs are delivered using point-to-point transmission at the transport IP layer. The UE mobility for connected modes between nodes in the NG-RAN is managed through the Xn-C interface [2].

The Xn-C may perform the following tasks:

- Interface management and error handling (e.g. setup, reset, removal, configuration update).
- Connected mode mobility management (handover procedures, sequence number status transfer, UE context retrieval).
- Support of RAN paging.
- Dual connectivity functions (secondary node addition, reconfiguration, modification, and release).

The user-plane Xn-U supports:

- Context transfer from old serving NG-RAN node to new serving NG-RAN node.
- Control of user-plane tunnels between the old serving NG-RAN node and the new serving NG-RAN node.

When gNB connects to the eNB that is still connected to the EPC, the legacy X2 interface will be used for signalling related to UEs connected to the EPC while the new Xn interface will be used for UEs connected to the 5GC.

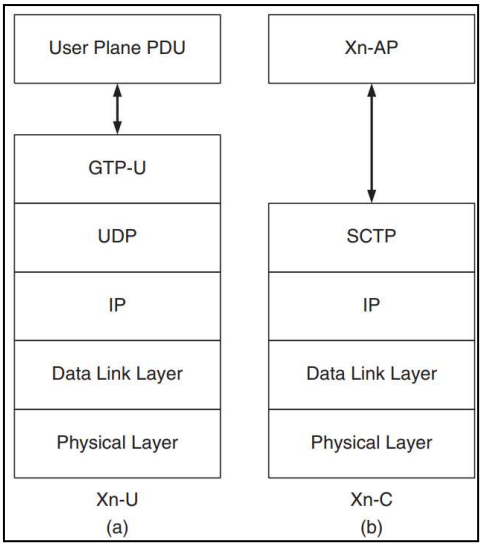

Figure 20: Xn protocol stack: (a) user plane and (b) control plane [4]

5-9-3) The F1 Interface

The F1 interface is specified for the case where the en-gNB is further subdivided into gNB-CU and gNB-DU logical nodes. In this case, the gNB-CU hosts the RRC and PDCP protocols, while the gNB-DU hosts the RLC, MAC and PHY functions. The F1AP protocol provides the following functions [2]:

- System information management function
- RRC message transfer function
- Paging function

6 Chapter 6) Cloud RAN

6-1) Cloud RAN Introduction

Data traffic and mobile subscriptions have increased dramatically over the past few years. According to one study, there are now 6.2 billion potential mobile customers worldwide by the end of 2023, versus 4.4 billion in 2013 and 5.4 billion in 2017. Other contributing factors are [8]:

- The introduction of smart devices with improved capabilities
- A variety of user-friendly applications, and
- An acceleration in deploying 4G and 5G cellular systems globally.

Mobile devices will become even more common thanks to the projected introduction of 5G systems and more data-intensive applications like virtual reality and augmented reality technology. As depicted in Figure 21, the prediction indicates that the total monthly worldwide mobile data traffic will increase from the current level of 13.8 ExaBytes[12] (EB) in 2017 to 110 EB by the end of 2023 at a compound annual growth rate (CAGR) of 42% [8].

[12] 1 EB = 1024 PB = 1048576 TB

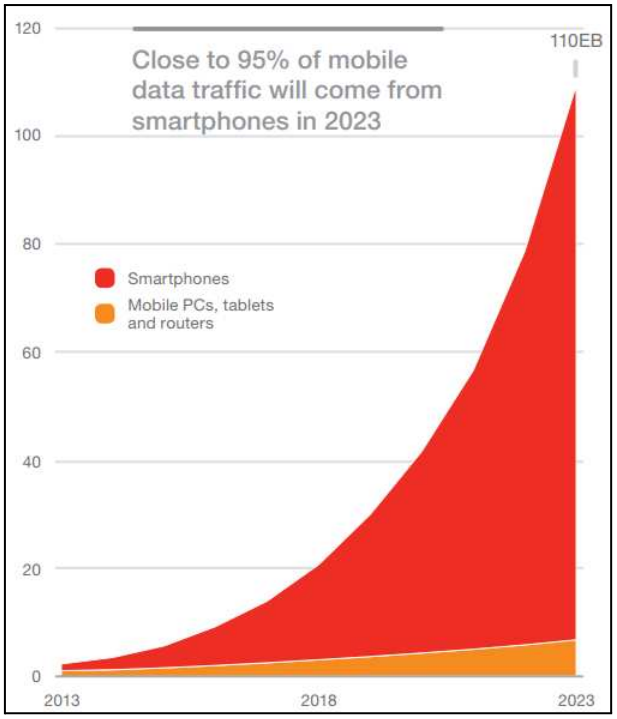

Figure 21. A forecast of global mobile data traffic in EB per month up to 2023 [9]

It will not be straightforward for network operators to support such a large amount of mobile data traffic in the coming years. In short, mobile networks must add more capacity to keep up with demand. There are some practices to improve the capacity as follows [8]:

- Adding more cells to the network.
- Implementing heterogeneous networks (HeNets) introducing small cells.
- Beamforming (BF)
- Deployment of relays and repeaters
- Adopting distributed antenna systems (DAS)
- Cognitive Radio (CR)
- Multi-user MIMO and Massive MIMO

The main problems with these increasing capacity technologies are inter-cell interference, complex network operation and maintenance, greater energy consumption, and decreased operator profit margin due to increased CAPEX and OPEX. As I t was mentioned in the 5G chapter, 5G networks will need to support an enormous number of connected devices, high data rates, improved energy efficiency (EE), robust reliability, nearly 'anytime anywhere' connectivity, low latency, and increased capacity [10]. Current technologies cannot satisfy such fundamental needs. Therefore, sectors and researchers must make certain basic adjustments in future networks of both technologies.

Cellular networks based on C-RAN (Cloud-RAN; it can also stand for Centralized-RAN), a leading candidate for future 5G cellular systems, have the potential to satisfy the objectives above. It is a paradigm-shifting evolutionary concept that IBM introduced in [8], and it proposes completely different cellular network architecture and operations from the ones used in traditional cellular networks. Baseband units (BBUs) from every base station (BS) in a

RAN are pooled together and virtualized in C-RAN before being distributed across the BSs [8].

6-2) C-RAN Architecture

The architecture of the C-RAN is provided in this part, along with detailed explanations of its essential elements and functional split options. It should be noted that the PHY layer and several upper layer features are all located in the BS inside traditional RANs, resulting in a high cost for network deployment, updating, and maintenance. Contrary to typical RANs, the baseband unit of the C-RAN simplifies the performance of BS by shifting a substantial portion of its operations to the cloud server (BBU). As a result, it is practical and economical to deploy more remote radio heads (RRHs), which are APs. Future wireless communication systems are anticipated to be designed and deployed on a new network architecture paradigm [11].

6-3) C-RAN Components

Figure 22 shows the architecture of C-RAN. The RRHs and BBUs are separated geographically, and the BBUs are located on a cloud server. In addition, a front-haul link connects BBU and RRH, while a back-haul link connects BBU and the core network. Next, we will see that when BBUs are centralized in a BBU pool, we refer to it as Option 1 in C-RAN architecture [11].

BBU: The BBU is enabled by cloud computing, which achieves flexible spectrum management and advanced network coordination. A considerable portion of the baseband signal processing for the entire network can also be handled by BBU, which manages the signalling to RRHs. The joint signal processing across a larger coverage area can be carried out at the BBU side in a centralized manner, as opposed to a standard BS, which has the potential to reduce interference and enhance performance. BBU carries out several tasks in the cloud, including modulation, coding, Fast Fourier Transform, and choosing an appropriate frequency or channel.

RRH: RRH mainly handles the radio frequency function and some basic signal processing. Even in hot spot zones, the implementation of RRH can offer seamless connection. Furthermore, RRHs placed close together can provide better coverage and higher data rates. Antenna-equipped RRHs forward baseband signals from users to the cloud for processing and send radio signals from the BBU cloud to users in the downlink. RF amplification, up/down conversion, filtering, digital processing, A/D and D/A conversion, and interface adaptability are among the primary responsibilities of RRHs. Since most signal processing tasks are now carried out in the cloud, RRHs can be kept relatively simple in large-scale applications while still being cost-effective.

Fronthaul link: The communication channels between BBUs and RRHs are provided via fronthaul. This fronthaul connectivity can be implemented using various technologies,

including optical fibre communication, wireless communication, or even millimetre wave (mmWave) communication. [9]. High transmission capacity can be supported via optical fibre communication fronthaul, but at a high cost and with limited deployment flexibility. While wireless fronthaul using 5 GHz to 40 GHz carrier frequencies is more affordable and flexible. However, it comes with decreased capacity and other limitations. Depending on the application scenarios, the fronthaul links may use wired or wireless media and are often capacity-limited. For practical system designs inside the two-hop C-RAN architecture, the fronthaul restrictions should be properly considered.

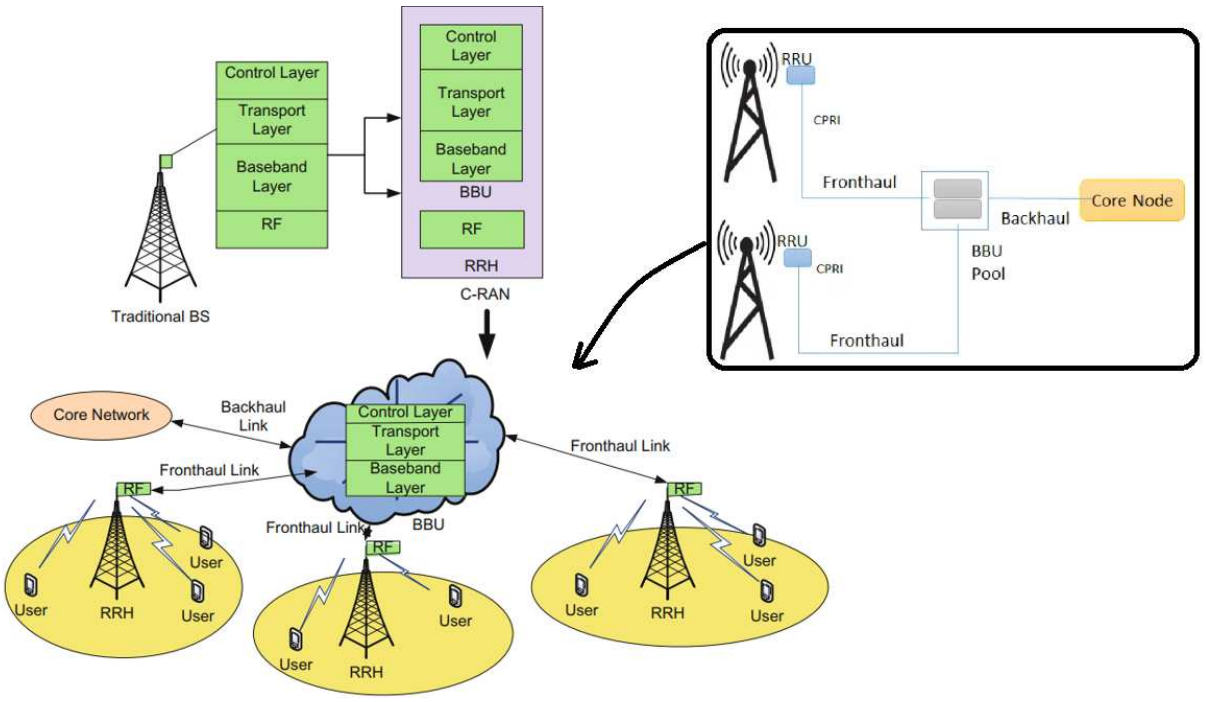

Figure 22: C-RAN Architecture (Option 1) [11]

In addition, the BBUs are further divided into DU and CU in a second variant of the centralized RAN architecture. As CU moves closer to the core network in this instance, a new interface known as mid-haul is created (Figure 23). It is important to note that the C-RAN architecture has many noteworthy benefits. For instance, C-RAN can offer more bandwidth by combining heterogeneous spectrum resources, expanding the user base served. The BBU's centralized management also helps lower its operating and capital costs. Cooperative processing and advanced MIMO techniques can potentially increase the gains of many RRHs.

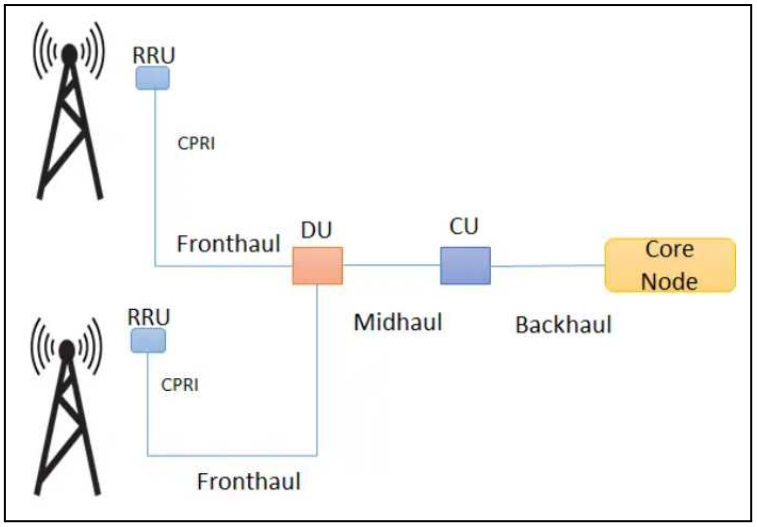

Figure 23: C-RAN with BBU Split [12]

6-4) C-RAN Functional Split

Option 1: The concentration of baseband signal processing in the C-RAN design results in high cooperative processing gains, which increases the flexibility of network coordination. However, there are trade-offs among the split option, fronthaul capacity and signal processing complexities. Three ideas for a functional split between BBU and RRH are shown, along with a short discussion of the trade-offs. Figure 24(a) illustrates the first option, which is suggested for the initial C-RAN architecture. The majority of PHY layer functions have been shifted to BBU. Almost all PHY layer functionalities are moved to BBU [11].

Meanwhile, RRH acts as a simple relay with RF, ADC/DAC and digital front-end. In this option, BBU and RRH are connected by the standard public radio interface (CPRI). This centralized PHY architecture may achieve the highest cooperative processing gain. However, forwarding I/Q samples via fronthaul links requires very high transmission bandwidth.

Option 2: The second approach is depicted in Figure 24(b), where some PHY layer processing is kept reserved at RRH, and some baseband processing is partially centralized. This functional split results in considerable cooperative processing improvements while reducing the fronthaul's transmission bandwidth. However, implementing cooperative processing becomes complicated due to the distributed deployment of PHY functionalities. This architecture aims to balance fronthaul capacity and signal processing complexity.

Option 3: Figure 24(c) demonstrates the third option, where all PHY layer functions are transferred to the RRH. This architecture requires the minimum fronthaul bandwidth by limiting the fronthaul link's transmission capacity to the maximum medium access control (MAC) layer throughput requirement compared to the previous two options. The price paid for this reduction is the increased scheduling delay in the fronthaul link, which may degrade system performance and network throughput. At the same time, the benefit of this

architecture is the saving of power consumption at BBU and higher flexibility to support radio resource allocation towards users.

In conclusion, separating baseband signal processing into BBU and RRH enables the C-RAN several choices for deploying with various fronthaul capacity restrictions. However, it is easier to implement the cost-effective deployment of C-RANs with adequate PHY layer signal processing [11].

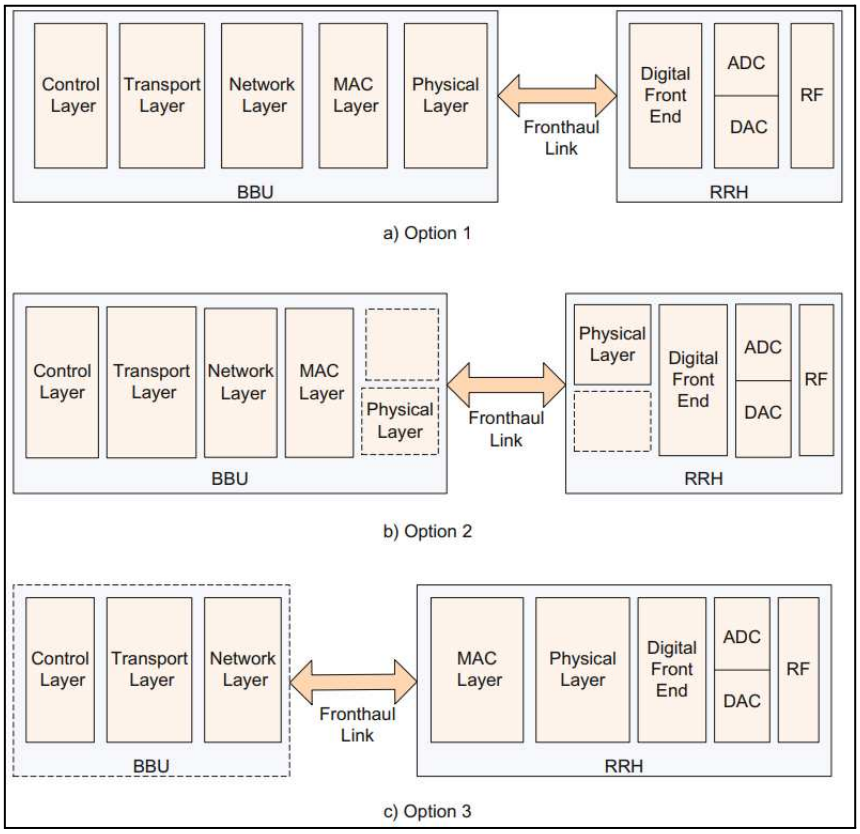

Figure 24: Functional split between BBU and RRH in C-RAN [11]

6-5) C-RAN Advantages and Drawbacks

Traditional BSs are split up into two components in C-RAN: distributed RRHs and pooled BBUs. The pool is placed at a single, cloud-based location with several BBUs. To meet the dynamic user demand having spatial variation, it is possible to share the radio resources of many BBUs. The amount of BBUs can be adjusted over time thanks to the cloud's ability for reconfiguration, which also controls the RRHs. The cloud performs baseband processing as a virtual base station using general-purpose CPUs. In the cloud, signal processing resources are dynamically assigned based on demand. Several operations are carried out, including modulation, coding, Fast Fourier Transform, and channel or frequency selection. On the other hand, RRHs with antennas forward baseband signals from users to the cloud for processing while transmitting radio signals from the BBU cloud to users in the downlink [8].

The main functions of RRH are Radio frequency (RF) amplification, up/down conversion, filtering, digital processing, analogue-to-digital conversion, digital-to-analogue conversion, and interface adaptability. Since most signal processing tasks are now being performed in the cloud, RRHs can be kept relatively simple in large-scale applications while still being cost-effective. Fronthaul, the third element, offers communication channels between BBUs and RRHs. These fronthaul links can be implemented using various technologies, including optical fibre communication, conventional wireless technology, and even millimetre wave (mmWave) communication.

Optical fibre communication can be utilized for high transmission capacity via fronthaul, but at a high cost and with limited deployment flexibility. While wireless fronthaul using 5 GHz to 40 GHz carrier frequencies is more affordable and flexible, it comes at the expense of decreased capacity and other limitations [8].

6-5-1) C-RAN Advantages

C-RAN has some significant advantages over the existing counterpart, as summarized below [8]:

• **Reduced CAPEX and OPEX**: Macrocell BS (MBS) deployment and commissioning are costly and time-consuming. In contrast, deploying and running RRHs for C-RAN requires less money, time, and space. Additionally, C-RAN can enable more efficient equipment sharing, resulting in lower CAPEX. According to a quantitative estimate published, C-RANs might lower CAPEX by up to 15% per kilometre. Additionally, C-RAN9 aggregates processing resources into a small number of massive clouds, leaving more specific functions in RRHs, which can reduce OPEX and administrative costs significantly.

• **Improved EE**: A C-RAN requires fewer BBUs than a traditional RAN, which results in lower power usage. Additionally, as RRHs are naturally cooled by air hanging on building walls or masts, air conditioning for radio modules in C-RANs can be reduced by about 90%. Moreover, C-RANs enable UEs and MBSs to offload their energy-intensive calculations that require much data to a nearby cloud, saving energy. According to a study done by ZTE, C-RAN can save up to 80% more energy than regular RAN.

• **Improved spectral efficiency (SE):** C-RAN can help cellular networks' SE. Implementing coordinated and cooperative transmission/reception techniques is much simpler and more effective, resulting in higher SE [12].

• **Reduced latency:** C-RAN can lower latency when carrying out various tasks. For instance, since handovers can be completed inside a cloud rather than between BSs, the time required to do a handover will be less. It is also possible to reduce the handover failure rate. Additionally, C-RAN allows for a reduction in the overall volume of signalling data delivered to the core network, which in turn reduces latency.

• **Facilitate the switching of BBUs:** By dynamically distributing processing capacity and moving tasks in the BBU pool, power consumption and load congestion can be decreased because all processing operations are implemented in a remote cloud. As a result, many BBUs can be placed in low-power sleep mode or even turned off to save energy.

• **Improved interference management**: C C-RAN can facilitate sharing channel status information (CSI), traffic information, and control information for mobile services among participating BSs. As a result, multi-point cooperation will be more efficient, and more streams can be multiplexed on the same channel with significantly less mutual interference. Link quality and throughput will both considerably increase as interference declines.

• **Ease of maintenance and expansion:** The centralized architecture of C-RAN makes it naturally scalable, which makes updating and maintaining cellular networks easier. As an illustration, a network operator needs to split the cell for increased capacity or add new RRHs connecting to the cloud to cover more service areas. Additionally, C-RAN makes it easier to install virtual resources in the cloud as needed.

• **Adaptability to non-uniform traffic:** There is a significant amount of tempo-spatial variability in traffic in contemporary cellular networks. However, because BSs are designed for peak usage, processing power is lost while the system is not in use. Based on the instantaneous traffic demand, resources to the BSs can be distributed optimally, improving the overall resource utilization rate. In C-RAN, baseband processing of numerous BSs is performed in the centralized BBU pool.

•**Wireless technology coexistence***:* C-RAN and its centralized BBU design can support multiple wireless standards, effectively managed and utilized based on the user demands, leveraging a fully heterogeneous wireless system.

• **Spectral efficiency improvement and reduction in inter-channel interferences:** Multiple cells can collaboratively and dynamically exchange resources through the centralized BBU (i.e. RRUs). As a result, the resources can be effectively used to meet service demand. The coordinated scheduling and processing will help to reduce inter-channel interference.

• **Throughput improvement:** C-RAN can facilitate dense RRU deployment schemes in areas that require high throughput services.

• **Business Model Transformation:** The C-RAN concept will generate more business models, such as the BBU pool resource rental system, cellular system as a service, as well as more freemium services.

6-5-2) C-RAN Drawbacks

Here we can find some of the main drawbacks in the deployment of cloud RAN [8]:

• **Need for high fronthaul capacity:** C-RAN architecture brings a significant overhead on the optical fronthaul links between RRHs and the cloud, which can be as high as 50 times compared to the backhaul requirements. Fronthaul link between BBUs and RRUs is required to have high bandwidth capability with low delay and cost requirements. The fully centralized approach is the most adopted structure in C-RAN, which require a considerable communication overhead on fronthaul link. As a result, the high bandwidth requirement is required at the fronthaul, which cannot be met by wireless communication. Optical fibre communication systems can give the high bandwidth required to solve such a problem.

However, optical fibres usually come with the problem of very high cost, which most cellular providers might not afford. Hence, a compromised solution between delay, bandwidth and cost must be considered in such systems before they come to reality.

• **BBU Cooperation:** BBUs in the same pool must cooperate to support sharing users' data, scheduling and channel feedback collection. Such cooperation is not defined and introduces a challenge in dealing with user privacy, high bandwidth and low latency communication between such BBUs.

• **Cell clustering:** Optimal clustering of the cells and BBU pool assignability with minimal overhead and maximum gain is still challenging. One BBU pool should achieve the maximum number of send and receive channels while minimizing the fronthaul delay and overhead. In addition, one BBU should support multiple distributed geographical locations, such as offices in different states, to consolidate them into one BBU. Therefore, such clustering and BBU assignment are still challenging to resolve in C-RAN systems.

•**Virtualization Technique:** Another issue with C-RAN is the distributed processing and resource sharing that virtualization approaches encourage between several BBUs. It must be real-time and dynamic for processing to support changing cell loads. Additionally, the specifications for the clouds on which BBUs will be deployed will differ from those for IT clouds already in use. As a result, cloud infrastructure needs to be modified to satisfy these needs. Thus, virtualization is yet another significant issue impacting the actual adoption of C-RAN at this time.

• **Security:** Protecting user privacy and trusted third parties from intrusion is a significant concern for C-RAN. Due to its transmission and self-deploying characteristics, C-RAN will face more severe security risks and trust issues than conventional wireless networks, such as the primary user emulation attack (PUEA) and spectrum sensing data falsification (SSDF) assault. Furthermore, resource sharing amongst BBUs makes it possible to violate user privacy and access data presumed to be private, especially in such a distributed design. Additionally, parties in C-RANs, including BBUs and RRUs, are taken for granted. Such presumptions might not be true, particularly given the vast user base of these platforms. Such a sizable, virtualized system is vulnerable to misbehaviour and threats from hacked users. Consequently, classic cellular systems also have vulnerabilities. C-RAN would pose a new security challenge that had not been considered or less complicated. As a result of the BBUs of numerous BSs being bundled together in the cloud, C-RAN also carries a significant single-point failure risk; in other words, if the cloud fails, the entire network will be inoperable [8].

• **Latency and Jitter between cloud and RRHs:** There is a potential to increase latency in some cases due to centralized signal processing.

• **Complex BS operations:** There is the risk of complexity in the BS operation due to centralizing a more significant number of BBUs in a unique pool in C-RAN architecture.

6-6) C-RAN Challenges and Open Issues

In this section, we mention some of the open issues and challenges of C-RAN deployment [13]:

Edge Cache in C-RANs: Several technical challenges must be overcome when applying edge cache to C-RANs. The edge cache approach is the first. In particular, care should be taken in selecting the frequency of data updates. Although it uses more fronthaul resources, the high-frequency update improves the QoE for UEs, whereas the low-frequency update does the opposite. The edge cache technique is easier to implement when there is sufficient information on data popularity, which is the primary aspect that should be considered for the edge cache.

Data fetching strategy is the second problem. The data requested by UEs can now be found in nearby edge devices thanks to the involvement of edge cache in C-RANs. An effective data fetching strategy should be created to help determine where to retrieve the data and the appropriate route if the data requested by a UE still needs to be cached in these devices. The RRH association strategy is the final issue. For small cell networks, base station association techniques have been developed that take edge cache into account. However, these methods are created for the scenario in which a single UE can only be connected to a single base station. While in C-RANs, a single UE is frequently supplied concurrently by several RRHs. Therefore, an advanced RRH association technique that considers edge cache should be used.

Big Data Mining in C-RANs: Big data in the context of mobile networks encompasses subscriber level, cell level, core network level, and other level data. It can help the network become more proactive. It is possible to use big data technology to extract intriguing patterns or knowledge to improve the self-organizing capabilities in C-RANs due to the rapid development of big data mining techniques and the powerful computing capability of the BBU pool. For instance, by analyzing past content request data, it is possible to predict user preferences for watching movies, enabling edge devices to store videos. However, there are still a few technical difficulties with C-RAN's massive data mining. For instance, the fronthaul will be heavily taxed for transmitting the substantial amount of data acquired by edge devices. Additionally, computing sparse, ambiguous, and incomplete data is a significant challenge that calls for advanced data mining methods.

Social-Aware D2D in C-RANs: D2D communication, which underlies cellular systems, has been the subject of extensive investigation due to its superiority in enhancing SE[13] and EE[14]. To fully take advantage of D2D, it is essential to handle a variety of difficulties, including peer identification, mode selection, resource allocation, and interference control. Recently, some works have used social network features like social community and social connection to address these issues. As a result, the social-aware D2D is proposed.

[13] Spectral Efficiency
[14] Energy Efficiency

A significant amount of data capacity can be offloaded from fronthaul links for the C-RAN using social-aware D2D, easing fronthaul restrictions and reducing transmission latency. Social-aware D2D can typically be implemented with and without BS support in a cellular network. Since RRHs are typically installed to provide high capacity in specific zones, the socially aware D2D with a BS requires the BS to send control signalling, which is challenging to implement in C-RANs. Therefore, in C-RANs, the socially aware D2D without a BS is favoured. Additionally, choosing between C-RAN mode and D2D mode is crucial to achieving excellent QoS. Additionally, when D2D communication uses the licenced band, there may be significant interference between D2D users and RRH users and possible mutual interference between D2D users. Therefore, it is essential to consider how to stop those interferences.

CR[15] in C-RANs: The spectrum resource is getting scarce because of the constantly rising capacity needs. On the other hand, a significant portion of the given spectrum is only partially utilized because of sporadic usage. Utilizing cognitive radio (CR) technology, which enables secondary users to share the spectrum with authorized users in an under or overlay fashion, is a viable way to address the issue of spectrum shortage.

While a centralized BBU pool can increase SE for C-RANs, CR is suggested to raise the spectrum utilization rate further. RRHs can interact with the radio environment and locate temporally unused spectrum assigned to C-RANs thanks to their cognitive capability. The operating bandwidth of RRHs can be significantly increased when combined with the CA method, increasing data rates. While the difficulties primarily stem from the complexity and expense of implementation. For instance, the complexity of the transceiver is relatively large to achieve the spectrum sensing function.

Additionally, secondary users typically need to detect the radio spectrum to find the vacant spectrum continuously. As a result, their transceivers must always be in the active state, which results in them using almost the same amount of power as the transmit state. Future research into the structure of CR usage in C-RANs is necessary to simplify implementation and boost EE.

SDN with C-RANs: SDN decouples the control and data planes to make a centralized controller and network programmability. Many advantages can be attained by implementing SDN for C-RANs. The software-defined fronthaul suggested can provide a flexible mapping between BBUs and RRHs for optimizing the network to traffic volume and user mobility. In particular, to implement DAS[16], reduce handoffs, and provide diversification benefits, many RRHs are logically mapped to a BBU for a specific region with high mobility users. Two controllers are defined: infrastructure manager and service manager, comparable to the SDN in wired networks.

- Infrastructure manager: It is responsible for pooling hardware resources and offering them as service slices to the service manager
- Service manager: It consumes service slices according to the requirement of a BBU virtual instance.

[15] Cognitive Radio
[16] Distributed Antenna Systems

By doing this, the BBU pool's hardware resources may be used more effectively, lowering the price of additional hardware. For the implementation of SDN to C-RANs, there are still specific fresh issues that need to be handled [13]:

1. If a single central controller controls the entire C-RAN, this controller's failure could cause the entire C-RAN to crash.
2. As the placement significantly affects processing latency and other network performance metrics, the controller placement should be improved, particularly in situations with several controllers and the large-scale C-RAN.
3. Operators must deal with scalability concerns in C-RANs due to the SDN controller's service capability limitations.

Physical Layer Security in C-RANs: Wireless channels are broadcasting by nature, making the network open to numerous emerging assaults like an attack from eavesdroppers. Physical layer security, which takes advantage of the physical properties of wireless channels to achieve perfect secrecy against eavesdropping in an information-theoretic sense, has become an attractive topic recently in this context. This is because encryption frequently requires significant computational resources and communication overheads. It has been suggested that opportunistic relaying can lower the cognitive communications system's secrecy outage floor. Relay-security-reliability selection's trade-off significantly outperforms the standard straight transmission. The C-RAN, one of the anticipated future wireless networks, ought to guarantee communication security. RRHs can be utilized to boost security performance to address this issue. A UE is often supplied by a cluster of RRHs in the downlink C-RANs, and some RRHs can act as jammers to create artificial noise to obstruct eavesdroppers. The choice of jammers from the available RRHs is one of the difficulties that could be faced [13].

7 Chapter 7) Moving from Cloud RAN to Virtual and Open RAN

7-1) V-RAN Concept

V-RAN separates the software from the hardware by virtualizing Network Functions and utilizing Software Defined Everything (SDx). Virtualization technologies like NFV or containers are used to deploy CU and DU over an x86 server. This is comparable to running software functions. Therefore, the only distinction between V-RAN and C-RAN is that V-RAN employs Network Functions on the server platform, while C-RAN generally uses proprietary hardware. In reality, V-RAN is a variety of C-RAN [14].

Many RRHs can be connected to a single BBU using the V-RAN technique (usually fibre optics) using high-speed fronthaul. In contrast to a conventional base station (BS), where the RRH and BBU are paired together, it isolates the two, enhancing network resource sharing. Additionally, V-RAN can improve the wireless system's flexibility and scalability while addressing several issues with legacy wireless systems, including interference and power consumption. Networks can also become programmable, flexible, cost-effective, and centrally managed thanks to V-RAN. Additionally, it makes it simpler to transition to newer wireless technology generations while still maintaining existing technologies through software and component separation

Because of V-RAN HW/SW decoupling flexibility, we can achieve scalability. As a result, hardware costs and application agility may reduce. In addition, applications may be upgraded or replaced entirely (which is not more manageable with traditional hardware). Figure 25 presents the general design of a V-RAN network for option 1 (BBU pool) and option 2 (BBU split into CU and DU units).

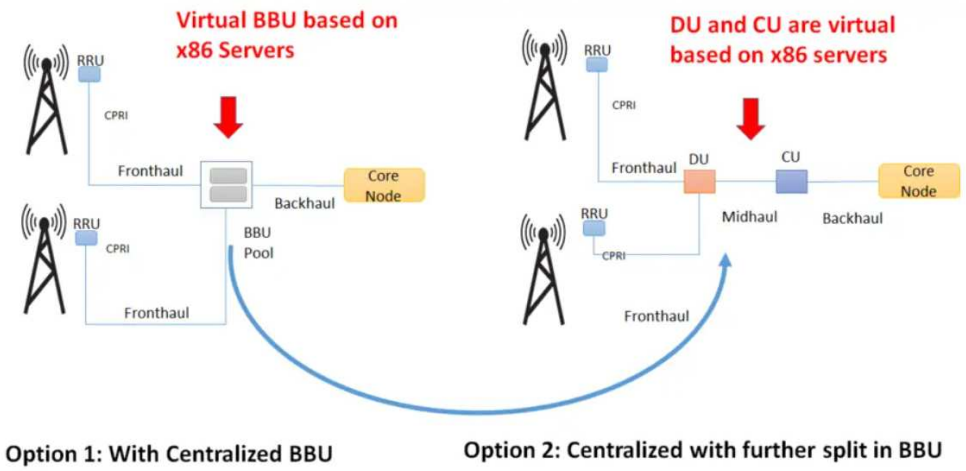

Figure 25: V-RAN architecture [14]

7-2) V-RAN: Evolution of C-RAN

7-2-1) V-RAN Benefits

Virtualization techniques facilitate the construction of conceptually separated instances over abstracted physical hardware, which may be shared dynamically, efficiently, and flexibly. The fields of cloud computing and data storage have constantly been using virtualization. For the actual deployments of the C-RAN idea, network virtualization represents a novel evolution of virtualization. In contrast to C-RAN, the V-RAN design supports flexible control, low cost, efficient resource consumption, and varied applications. It can also address many fundamental problems associated with RAN deployment in the cloud [14].

The Virtual RAN, or V-RAN, strategy encourages the separation of hardware and software operations. Massive machine access, mission criticality, tactile internet, and other future services can all be made possible by the V-HW/SW RAN's decoupling, flexibility, scalability, and inherent centralized nature. Based on the underlying traffic conditions, V-RAN uses the wireless radio and BBU resources to be shared by several RRHs. This encourages increased energy efficiency and lowers running the wireless system's operational and investment expenses. Additionally, it promotes innovations and helps new companies enter the market more affordably [14].

7-2-2) Virtualization Technologies

Virtualization and cloud infrastructure have already been thoroughly researched and developed for IT applications. The V-RAN concept, however, places entirely new demands on

the cloud infrastructure. For instance, two essential virtualization technologies that can be used in V-RAN are container-based virtualization and hypervisor-based virtualization. Both technologies use various virtualization, orchestration, and resource scaling techniques. Numerous virtualization frameworks, including VMware, OpenStack, Kubernetes, Docker, Hyper-V, etc., serve as examples of these principles. Most current ICT installations use the OpenStack (virtualization based on hypervisors) or Docker (virtualization based on container-based technology) frameworks. Several initiatives have recently aimed to develop and integrate multi-functional orchestration engines, such as OMF, OSM, etc., into the V-RAN environment. However, their current applicability could be more extensive and far from commercial deployments due to instability, resource consumption or limited scope of features [14].

7-3) Virtual RAN Towards Open RAN

The V-RAN is growing from the concept of C-RAN towards the concept of Open-RAN (O-RAN), focusing on two essential pillars: openness and intelligence, to address the most pressing issues. Open interfaces are essential to help smaller suppliers and operators quickly roll out new services and allow operators to customize the network based on their own needs. Additionally, openness makes multivendor V-RAN deployments possible, creating a more vibrant and competitive market. Furthermore, while maintaining backward compatibility with legacy systems, open-source software and hardware designs can speed up and improve innovation and commercial implementation [15].
Future wireless systems such as 5G and beyond 5G will also become significantly more complex due to network densification and more successful and demanding applications. As a result, mobile network manufacturers and operators should be self-organizing. They ought to be able to take advantage of cutting-edge innovations like machine learning (ML) and artificial intelligence (AI) to automate network operations and cut expenses. The telco sector has acknowledged the initiatives to provide an open virtualized RAN as the first significant evolutionary step towards the 5G standard. The relationship and development of the C-RAN concept and its offshoots, V-RAN and O-RAN, are shown in Figure 26.

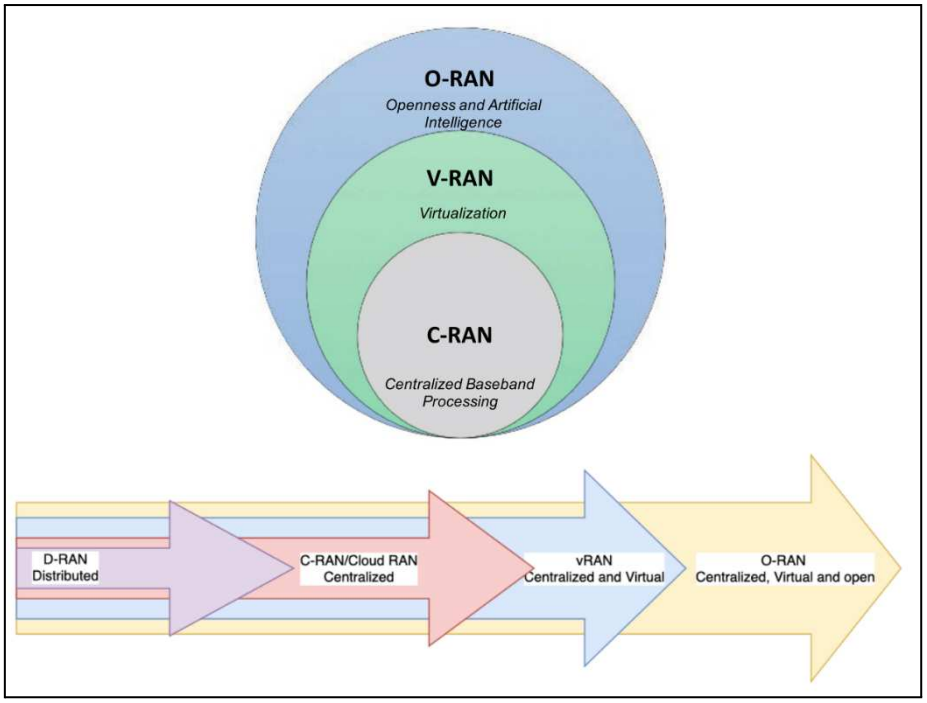

Figure 26: RAN evolution from D-RAN to O-RAN [14, 15]

The number of research and development activities that focus on the O- RAN concept is constantly increasing under the umbrella of the O-RAN alliance that provides the necessary framework for O-RAN commercial development and deployment. Three significant concepts form the foundation of the O-RAN initiative:

- Set the industry standard for open, interoperable interfaces, RAN virtualization, and RAN intelligence powered by big data.
- Specify APIs and interfaces, pushing standards to accept them where necessary and, when necessary, investigating open source.
- Increase the use of commercial silicon and commonly available hardware while reducing the use of proprietary hardware.

RAN Virtualization, Open Interfaces, White Box Hardware, and Open Source Software are a few cutting-edge technologies that the O-RAN effort heavily utilizes. Three technologies are upon these fundamental ideas. Their main traits are as follows [14]:

• **Software Defined, AI-Enabled RAN Intelligent Controller:** The O-RAN architecture seeks to further the SDN idea of separating the control plane (CP) from the user plane (UP) in RANs by promoting embedded intelligence. This strategy extends the CP/UP divide and further improves the standard RRM functions with embedded intelligence by adding a RAN Intelligent Controller (RIC). Decoupling offers the potential for increased UP standardization, which is its principal advantage. As a result, the UP may use simple scaling and affordable solutions. The second benefit provides increased efficiency and improved radio resource management, which is the ability to use more advanced control functionality.

Then, advanced ML/AI technologies will be used with analytics and data-driven methodologies to implement these control functionalities. O-RAN aspires to be the industry

leader in creating RICs with AI capabilities. The O-RAN project creates specifications and software reference designs, promotes operator proofs-of-concept and aids in operator field tests. The O-RAN AI-enabled RIC support even the most complicated networks. These networks are expected to have the inherent capability to provide a practical, optimized device and radio resource management through closed-loop control [15].

• **RAN Virtualization***:* RAN virtualization is one of the fundamental tenets of the O-RAN architecture. The primary goal of O-RAN is to provide Network Function Virtualization Infrastructure/Virtualized Infrastructure Manager (NFVI/VIM) specifications to improve virtualization platforms in support of various splits throughout the protocol stack (i.e. network slicing). For instance, a low layer split within the PHY, a high layer split between PDCP and RLC, etc. To develop essential solutions like programmable hardware accelerators, real-time processing, and lightweight virtualization technologies, the O-RAN initiative also focuses on evaluating the advantages and effectiveness of pertinent open-source virtualization communities (such as OPNFV, ONAP, Akraino, M-Cord, OpenStack, etc.).

• **Open Interfaces**: The O-RAN architecture is built on critical interfaces between multiple decoupled RAN components. Improved 3GPP interfaces for effective multi-vendor interoperability are among them. O-RAN also proposes an open interface between the RIC and V-RAN and an open fronthaul interface between the BBUs and RRHs.

• **White Box Hardware:** O-RAN's reference designs call for high-performance, spectral, and energy-efficient white-box base station hardware to fully benefit from the economies of scale of an open computing platform strategy. The O-RAN reference systems contain specific hardware and software architecture plans, support a decoupled approach, and make it easy to create and deploy BBUs and RRHs.

• **Open Source Software:** Most of the O-RAN architecture's parts, including the RIC, protocol stack, virtualization platform, and others, are already available as open-source solutions and will continue to be so through organizations like the Linux Foundation, OMF, and others. In addition, to implement the 3GPP interface standards, the O-RAN open-source software framework anticipates providing the reference design for the upcoming RIC-based RRM.

8 Chapter 8) Open RAN

8-1) O-RAN Concept

According to the preceding sections, next-generation wireless systems based on a variety of heterogeneous technologies and frequency bands are increasing the complexity of cellular networks. Massive MIMO, communications using millimetre waves and sub-terahertz communications, network-based sensing, network slicing, and machine learning (ML)-based digital signal processing are a few recent developments. The network's operators will encounter increasing capital and operational costs. This happened due to continuously improving and maintaining the infrastructure to keep up with new technological advancements and client demands [16].

Solutions enabling the Radio Access Network are needed to manage and optimize these new network infrastructures (RAN). This exposes data and analytics, as well as data-driven automation, closed-loop control, and optimization [17–19]. However, the current methods for cellular networking are far from flexible. These days, RAN components are all-encompassing monolithic systems that implement every layer of the cellular protocol stack. A small group of vendors offers them, and the operators classify them as "black boxes." Using black-box solutions exclusively has led to the following:

- The RAN's low compatibility, with hardware whose operations cannot be adjusted to serve a variety of installations and traffic profiles
- Limited network node coordination prevents cooperative optimization and management of RAN components.
- Vendor lock-in, with limited options for operators to deploy and interface RAN equipment from multiple vendors.

Real-time adaptation for optimal spectrum uses and optimized radio resource management is challenging under these conditions [20]. To get beyond these restrictions, the Open RAN has been recognized as the new paradigm for future RAN in many recent research and standardization initiatives. To connect RAN nodes, the O-RAN Alliance is standardizing a virtualization platform for the RAN and expanding the definition of 3GPP and eCPRI interfaces [21]. The design, implementation, and functionality of the upcoming generations of cellular networks will be fundamentally altered by the Open RAN paradigm and, more specifically, O-RAN networks. They will enable, among other things, transformative applications of ML for optimization and control of the RAN [20].

8-2) Open RAN Key Architectural Principles

The Open RAN vision is based on years of research on open and programmable networks. In the past 15 years, these concepts have driven the Software-defined Networking (SDN) revolution in wired networks, and more recently, they have begun to spread to wireless networks. A standardized fronthaul interface, for instance, has been proposed by the operator-led xRAN Forum, and open, standardized interfaces for integrating external controllers in the RAN have also been proposed [20].

In parallel, as previously mentioned, the Cloud RAN (C-RAN) architecture has emerged as a method for centralizing the majority of the RAN's baseband processing in virtualized cloud data centres that are linked to remote radio units through high-speed fronthaul interfaces. By utilizing centralized data and control routes, C-RAN enabled more advanced signal

processing and load balancing techniques while lowering costs by multiplexing computer resources. With the overarching objective of establishing an architecture and a set of interfaces to actualize an Open RAN, these two initiatives came together in 2018 to form the ORAN Alliance [22]. Overall, four guiding principles for the Open RAN architecture can be identified. These include virtualization, open interfaces, intelligent, data-driven control via RICs, and disaggregation.

8-2-1) Disaggregation

The O-RAN reference architecture's primary characteristics and functional modules include the following (Figure 27):

- RIC non-Real Time (non-RT) layer
- RIC non-Real Time (near-RT) layer
- Multi-RAT Centralized Unit (CU) protocol stack and platform,
- A Central Unit (CU), a Distributed Unit (DU), and a Radio Unit (RU) make the gNB (called O-CU, O-DU, and O-RU in O-RAN

The Control Plane (CP) and the User Plane (UP) are the following two logical divisions of the CU (UP). This logical separation enables the deployment of various capabilities among hardware platforms and network locations [16].

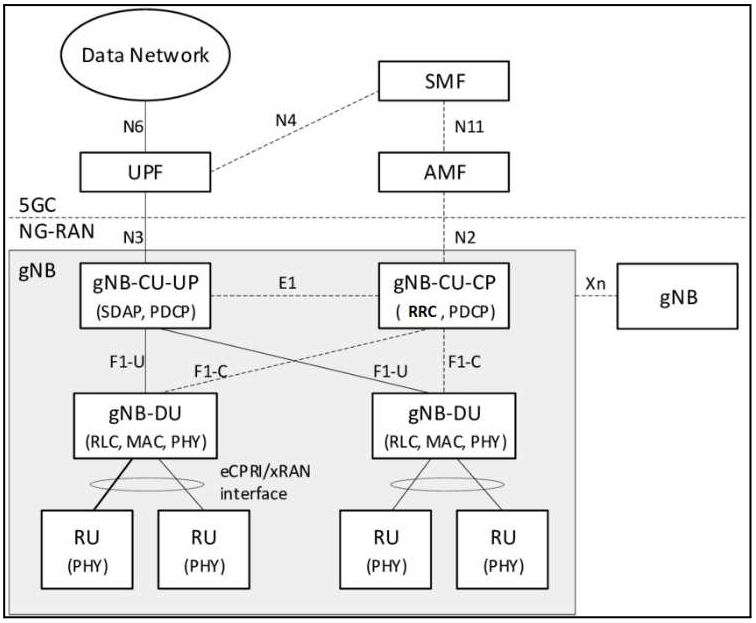

Figure 27: NG-RAN architecture with a CU-DU split deployment [16]

The functional disaggregation paradigm provided by 3GPP for the NR is effectively adopted and extended by RAN disaggregation, as shown in Figure 28. It divides base stations into several functional pieces (gNBs).

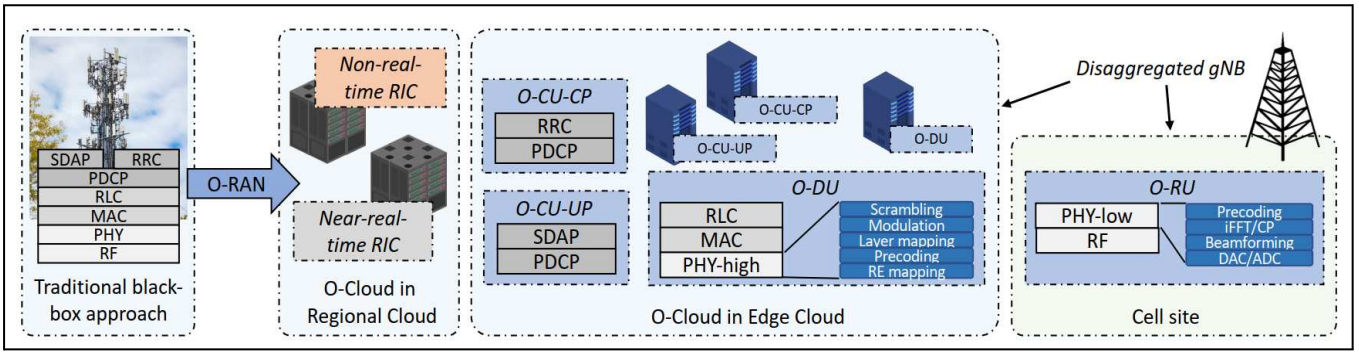

Figure 28: Evolution of the traditional black-box base station architecture (left) toward a virtualized gNB with a functional split [16]

CUs and DUs can be virtualized on white box servers at the edge (with hardware acceleration for some physical layer functionalities). At the same time, the RUs are generally implemented on Field Programmable Gate Arrays (FPGAs) and Application-specific Integrated Circuits (ASICs) boards and deployed close to RF antennas.

The O-RAN Alliance has evaluated the different RU/DU split options proposed by the 3GPP, with a specific interest in alternatives for physical layer split across the RU and the DU. The 7.2x split creates a balance between the RU's simplicity and the data speeds and latency needed at the RU and DU interface. The RU is cheap and straightforward to deploy in split 7.2x[17] because it executes FFT and cyclic prefix addition and removal operations. The remaining functions of the physical layer and those of the Medium Access Control (MAC) and Radio Link Control (RLC) levels are then handled by the DU.

The operations of these three layers are generally tightly synchronized, as the MAC layer generates Transport Blocks (TBs) for the physical layer using data buffered at the RLC layer. The higher layers of the 3GPP stack are implemented by the CU units (CP and UP). It encompasses the Radio Resource Control (RRC) layer for controlling the connection's life cycle; the Service Data Adaptation Protocol (SDAP) for controlling the QoS of traffic flows; and the Packet Data Convergence Protocol (PDCP) for managing packet duplication, reordering, and encryption for the air interface [16].

8-2-2) RAN Intelligent Controllers and Closed-Loop Control

The RICs, which introduce programmable elements that can conduct optimization procedures with closed-loop control and coordinate the RAN, represent the second breakthrough. The O-RAN vision includes two logical controllers with a centralized, abstract

[17] Split 7.2x Minimize impact on transport bandwidth while maximizing virtualization in gNB CU and gNB DU. Enable simple, low-cost RRU designs for wide adoption. Eliminate performance loss compared to integrated solutions with ideal fronthaul.

network view. The two RICs use AI and ML algorithms to process this data and develop control rules and actions to be taken on the RAN [16].

This presents data-driven, closed-loop control that may automatically optimize various processes, such as network and RAN slicing, load balancing, handovers, and scheduling strategies. The O-RAN Alliance has developed specifications for a near-real-time RIC that functions on a time scale between 10ms and 1s, drives control loops with RAN nodes, and a non-real-time RIC that integrates with the network orchestrator and operates on a time scale longer than 1s. (for real-time RIC, this time scale is less than 10ms). The disaggregated O-RAN infrastructure's closed-loop control and the evaluated real-time enhancements for future work are shown in an overview in Figure 29.

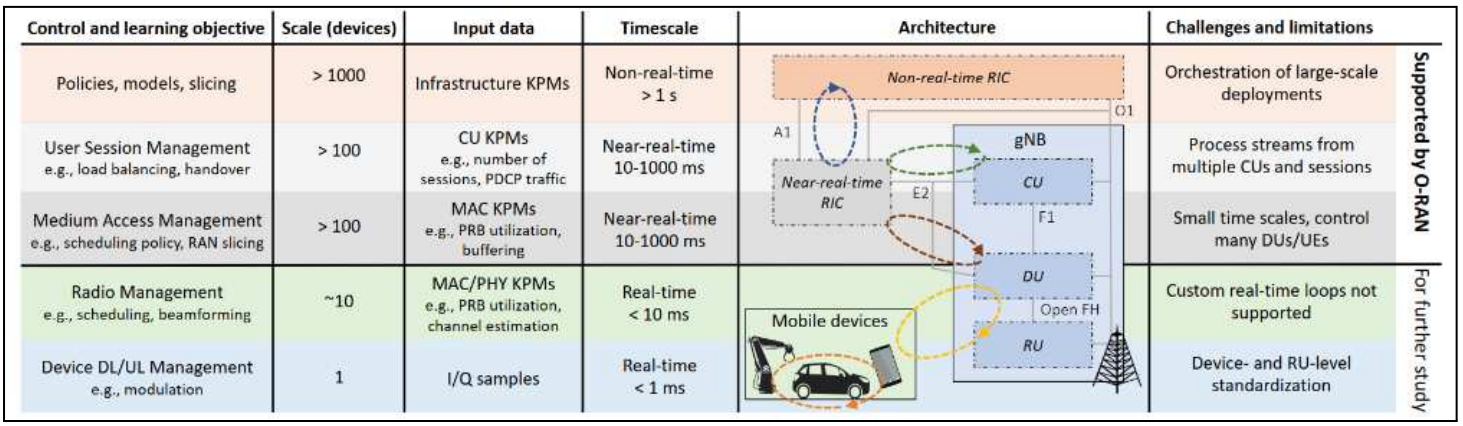

Figure 29: Closed-loop control enabled by the O-RAN architecture, and possible extensions, adapted from. The control loops are represented by the dashed arrows over the architectural diagram [16].

Non-real-time RIC and Control Loop. The non-real-time RIC (non-RT RIC) is a component of the SMO[18], as illustrated in Figure 30, and complements the near-RT RIC for intelligent RAN operation on a time scale more prominent than 1 second. It provides value-added services to support and facilitate RAN optimization and operations, including policy guidance, enrichment information, configuration management, and data analytics, through the execution of third-party applications or rApps.

The non-RT RIC manages ML models for the near-RT RIC and provides guidance and enrichment information using the non-real-time control loop. Additionally, the non-RT RIC has the potential to affect SMO operations, giving it indirect control over all O-RAN architectural components connected to the SMO. As a result, it can apply policies and make decisions impacting thousands of devices. As depicted in Figure 29, this introduces scalability difficulties that must be resolved by effective process and software design

Near-real-time RIC and Control Loop. The near-real-time RIC (Near-RT RIC) is deployed at the network's edge and operates control loops with a periodicity between 10ms and 1s. Figures 29 and 30 show that the near-RT RIC communicates with older O-RAN-compliant LTE evolving Node Bases and DUs and CUs in the RAN (eNBs). Since the near-RT RIC

[18] Service Management and Orchestration

connects to several RAN nodes, the near-RT closed-loop control can impact the quality of service for hundreds or thousands of pieces of user equipment (UE).

The near-RT RIC contains many apps that support custom logic, or "xApps," and the services necessary to support their execution. A microservice known as a "xApp" can be used to manage radio resources through defined interfaces and service models. It receives information from the RAN (such as user, cell, or slice KPMs, as depicted in Figure 29), calculates any necessary control actions, and then sends the results back. The near-RT RIC has the following to support xApps [16]:

- A database that serves as a shared data layer for xApps and contains data on the RAN (such as a list of linked RAN nodes, users, etc.);
- Messaging infrastructure enabling the subscription of RAN parts to xApps across all platform components;
- Terminations for open interfaces and Application Programming Interfaces (APIs),
- Conflict resolution mechanisms to orchestrate control of the same RAN function by multiple xApps.

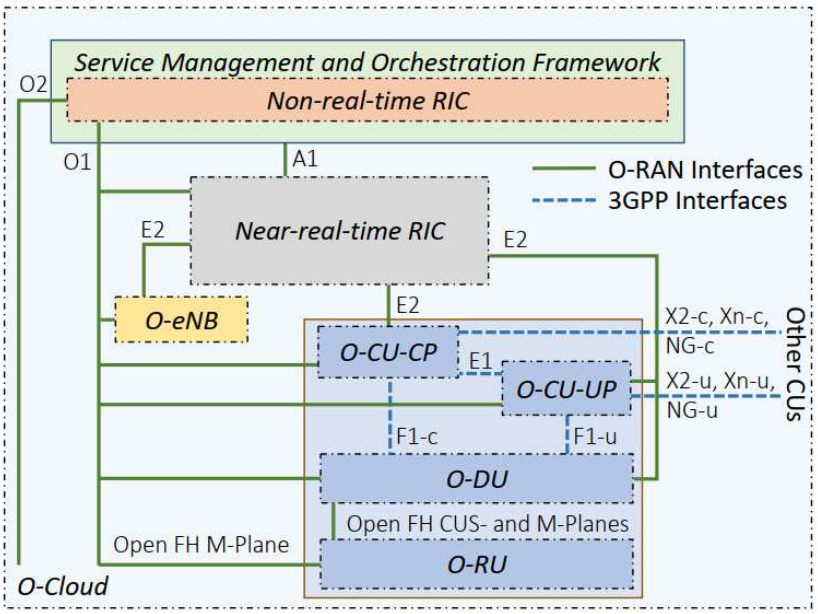

Figure 30: O-RAN architecture, with components and interfaces from O-RAN and 3GPP. O-RAN interfaces are drawn as solid lines, 3GPP ones as dashed lines [16]

Real-Time Control Loops. Figure 29 also shows loops that manage radio resources at the level of RAN nodes in the real-time domain, i.e., below 10ms, or even below 1ms, for device management and optimization. Real-time control commonly uses scheduling, beam management, and feedback-less physical layer parameter detection (e.g., modulation and coding scheme, interference recognition). Although these loops are separate from the present O-RAN design and only have a small scale in terms of the number of devices they may optimize, we cannot observe a variety of machine-learning methods in these control loops.

8-2-3) Virtualization

The third principle of the O-RAN architecture is the introduction of additional components for the management and optimization of the network infrastructure and operations, spanning from edge systems to virtualization platforms. In [21], it is stated that the O-RAN architecture depicted in Figure 30 may be implemented on the hybrid cloud computing platform, O-Cloud. The O-Cloud is specifically a collection of computing assets and virtualization technology that are consolidated in a single or several physical data centres. This platform specializes in the virtualization paradigm for O-RAN by combining physical nodes, software components (such as the operating system, virtual machine hypervisors, etc.), management and orchestration features, and software components. It permits:

- Decoupling between hardware and software components;
- Standardization of the hardware capabilities for the O-RAN infrastructure;
- Sharing of the hardware among different tenants,
- Automated deployment and instantiation of RAN functionalities.

The O-RAN Alliance Working Group (WG) 6 is also creating common APIs for dedicated hardware-based logical processors and the O-RAN software infrastructure, such as channel coding, decoding, and Forward Error Correction, under the name Acceleration Abstraction Layers (AALs) (FEC). These efforts also translate into faster, virtualized RAN implementations on commercial hardware that can serve 3GPP NR use cases, such as flows for Ultra-Reliable and Low Latency Communications (URLLC) (e.g., the NVIDIA Aerial platform, NEC Nuberu [23], and [24] from Intel). In addition, WG 7 is defining the requirements for white box hardware to implement an item of equipment that complies with O-RAN, such as indoor picocells, outdoor microcells, and macrocells (all operating at sub-6 GHz and mmWaves), integrated access and backhaul nodes, and fronthaul gateways.

These cover architectural components from Figure 28, such as the RAN nodes (CU, DU, and RU) and fronthaul interface enablers. The specifications define the hardware properties of the nodes (such as accelerators, computation, and connection) and the functional parameters relevant to the scenarios of interest (such as frequency bands, bandwidth, inter-site distance, and MIMO configurations). It is anticipated that the virtualization of the O-RAN computer parts and the RAN components will result in power consumption reductions and optimization for the RAN. With virtualization, it is simple and dynamic to scale up or down the computing resources needed to satisfy user requirements, restricting the power consumption to the necessary network functions. In this way, the base stations and RF components—which often account for most of the power consumption in cellular networks—can also benefit from more precise and dynamic sleep cycles thanks to the closed-loop control capabilities discussed above and the virtualization in the RAN.

8-2-4) Open RAN Interfaces

Lastly, the O-RAN Alliance has unveiled technical standards that outline open interfaces connecting various O-RAN architecture components. Figure 30 lists the 3GPP specifications'

intra-RAN interfaces and the new, open interfaces that O-RAN defines. However, the O-RAN Open Fronthaul between the DU and the RU performs the gNB disaggregated architecture, which is only partially enabled by the latter. The O-RAN interfaces assist in overcoming the old RAN black box approach by exposing data analytics and telemetry to the RICs and enabling various control and automation operations. Without O-RAN, Radio resource management and virtual/physical network function optimization would be closed and rigid without O-RAN. It means that operators would have a different level of access to the hardware in their RAN, or it would be carried out using a specialized, fragmented method [16].

It is crucial to standardize these interfaces to break the vendor lock-in in the RAN. For example, this will allow a near-RT RIC from one vendor to communicate with base stations from a different vendor or enable the interoperability of CUs, DUs, and RUs from various manufacturers. Additionally, this encourages innovation, market competition, and quick updates and makes it simpler to build and implement new software components in the RAN ecosystem [20]. The E2 interface, one of the O-RAN-specific interfaces, links the RAN nodes and the near-RT RIC. Through the streaming of telemetry from the RAN and the feedback with control from the near-RT RIC, E2 makes it possible for the near-real-time loops to function, as depicted in Figure 29.
The near-RT RIC is linked to the non-RT RIC Through the A1 interface, enabling the deployment of policy, guidance, and intelligent models in the near-RT RIC and a nonreal-time control loop. The O1 interface, which connects to every other RAN component for management and orchestration of network activities, is also terminated by the non-RT RIC. Finally, through the O2 interface, the non-RT RIC and SMO connect to the O-RAN O-Cloud through the O-RAN fronthaul interface, which links DUs and RUs. The O-RAN Alliance has also defined a set of standardized tests to promote interoperability across different interface implementations, with an initial focus on the fronthaul interface and E2. The O-RAN architecture described in Figure 30 can be deployed by selecting different network locations (cloud, edge, cell sites) for different equipment with multiple configurations [16].

8-3) AI/ML Workflows in Open RAN

The AI/ML workflow is being standardized by O-RAN WG2 (Working Group 2), with its specifications described in [22]. However, only some procedures, features and functionalities have been finalized, with some left for further studies. This workflow is composed of six main steps [16]:

- Data Collection and Processing
- Data Training
- Data Validation and Publishing
- Deployment
- AI/ML Execution and Inference
- Continuous Operations

In the following, we will outline the steps involved in the AI/ML lifecycle within O-RAN systems for explanation and clarity. We use the scenario where a network operator wants to

create, train, and deploy a xApp that manages RAN slicing policies by modifying them almost instantly in response to traffic demand and network strain. In the following chapter, we will approach and implement this service classification issue in the 5G network, using various machine learning methods to arrive at a good prediction. Each base station in this chapter's example hosts three slices: An Enhanced Mobile Broadband (eMBB) slice for high-throughput traffic (like file transfers and video streaming) and a Massive Machine-Type Communications (mMTC) slice for traffic from things like tiny sensors and Internet of Things (IoT) devices. The URLLC slice is for ultralow latency services. The eMBB slice is for high-throughput traffic (like file transfers and video streaming). By allocating the available PRBs to each slice to satisfy each slice's various performance needs, the xApp controls RAN slicing policies.

8-3-1) Data Collection and Processing

Data is first gathered across the O1, A1, and E2 interfaces and then stored in sizable datasets (such as data lakes or centralized repositories) from which it may be requested. The O-RAN specifications consider a preliminary data pre-processing (or preparation) step because different AI/ML solutions may use different KPI types. These KPIs were gathered over different periods and with different granularity. For example, we can mention throughput, latency, Modulation and Coding Scheme (MCS), Channel Quality Information (CQI), delay and jitter. In this step, data is shaped and formatted to match the input size of the particular AI/ML model under consideration for training and online inference.

Data and performance metrics related to the xApp regulating RAN slicing policies are collected over the O1 interface in this step (i.e., training phase). This step is done to create a training dataset used in the following step. For instance, the data collected must show how many PRBs are required to transmit the data requested by each user of the three slices, as well as throughput (eMBB), number of transmitted packets (mMTC), and latency (URLLC) measurements. This is necessary so that the xApp can adjust RAN slicing policies for the various slices according to the current data demand and required minimum performance levels. Data processing can use well-known ML techniques like normalization, scaling, and autoencoders [16].

8-3-2) Data Training

The O-RAN specifications do not allow the deployment of any untrained data-driven solution. All the AI/ML models are required to be trained offline to ensure the reliability of the intelligence and avoid inaccurate predictions, classifications and actions that might result in outages or inefficiencies in the network. Online training is still supported by O-RAN as long as it is only used to enhance and update a previously trained model offline. Therefore, this does not rule it out. For instance, the operator can train several Deep Reinforcement Learning (DRL) agents and decision trees, experiment with various combinations of input formats (such as the precise subset of KPIs and their quantity), and examine various designs (e.g., depth and width of a DRL agent, number of neurons, among others). This process aims

to train many ML algorithms and determine which ones are best suited to complete a given task.

8-3-3) Data Validation and Publishing

Models go through a validation step after they are trained to ensure that they are reliable, robust, and capable of handling classification, prediction, or control tasks. (The situation covered in the following chapter involves a classification problem.) The models are published and kept in an AI/ML catalogue on the SMO/non-RT RIC if the validation is successful and they are determined to be ready for deployment. Otherwise, until the validation tests are successful, they must go through extra re-design and retraining steps.

After training is complete, the various AI algorithms are evaluated against various validation datasets, including data that has never been seen before, to determine which models are the most successful at controlling RAN slicing strategies. For instance, a typical validation test looks at the performance of several AI solutions under various traffic patterns and demands, user numbers and distribution, bandwidth availability, and operational frequencies. The operator can leverage the AI solution that is most appropriate for a particular deployment. Using this procedure to identify AI solutions, we can provide side information on the ideal network conditions (such as network load, mobility pattern, and deployment size) under which the specific AI solution delivers the best performance [16].

8-3-4) Deployment

Models kept in the AI/ML catalogue can be downloaded, deployed, and executed using either the image-based or the file-based deployment method. The model is deployed over the O1 interface in both scenarios, and the node that runs the model is referred to as the inference host. In the image-based deployment, the AI/ML model runs as a containerized image inside an O-RAN application (such as xApps and rApps), which is installed at the O-RAN nodes and used to make an online inference. These nodes are currently restricted to RICs, and further research will be done on AI execution at CUs and DUs [16].

The file-based deployment considers the scenario in which the AI/ML model is downloaded as a standalone file that runs in an inference environment—outside the ORAN application domain and transmits the model's inference output to one or more O-RAN applications. As an illustration, the operator will choose the pre-trained AI-based RAN slicing models in our scenario from the AI/ML catalogue and deploy them as xApps that will be run in the near-RT RIC [16].

8-3-5) AI/ML Execution and Inference

After being placed on the inference host, models are provided with data to carry out various online inference tasks. These involve managing and controlling operations, determining policies at both RICs (sent across the A1 and E2 interfaces), and performing classification and prediction tasks (over the O1 and E2 interfaces, respectively). The xApp is fed with KPIs (such as requested PRBs, latency, and throughput measurements) collected over the E2 interface

and computes control actions that are used to pilot the DU and assign the available PRBs to the different slices in near-RT. This is done by executing the operations after the xApp has been deployed on the near-RT-RIC [16].

8-3-6) Continuous Operations

Monitoring and analyzing the intelligence deployed across the network is a crucial part of the AI/ML workflow. It is essential to confirm that the inference outputs of AI/ML models are reliable and correct and do not adversely impact the network's performance. Models that perform poorly in the real world can be improved and retrained through continuous operations. Example: In our scenario, the operator can continuously check the RAN slicing xApp's performance and, if any anomalies or inefficiencies are found, can choose to retrain the AI/ML model built within the xApp using new data gathered through the O1 and E2 interfaces [16].

9 Chapter 9) Machine Learning Algorithms

9-1) Introduction

The science (art) of programming computers to learn from data is known as machine learning. Our spam filter, for instance, is a machine learning programme that can be taught to identify spam based on examples of spam emails (such as those reported by users) and examples of regular (sometimes known as "ham") emails. The training set refers to the examples the system utilizes to learn. The term "training instance" refers to each training example (or sample). The job T in this situation is to mark new emails as spam. *Experience* E is the training data, and the *performance* measure P needs to be determined; one option is to utilize the percentage of successfully classified emails. *Accuracy* is a specific performance metric that is frequently employed in classification jobs. Figure 31 shows a machine-learning approach [25].

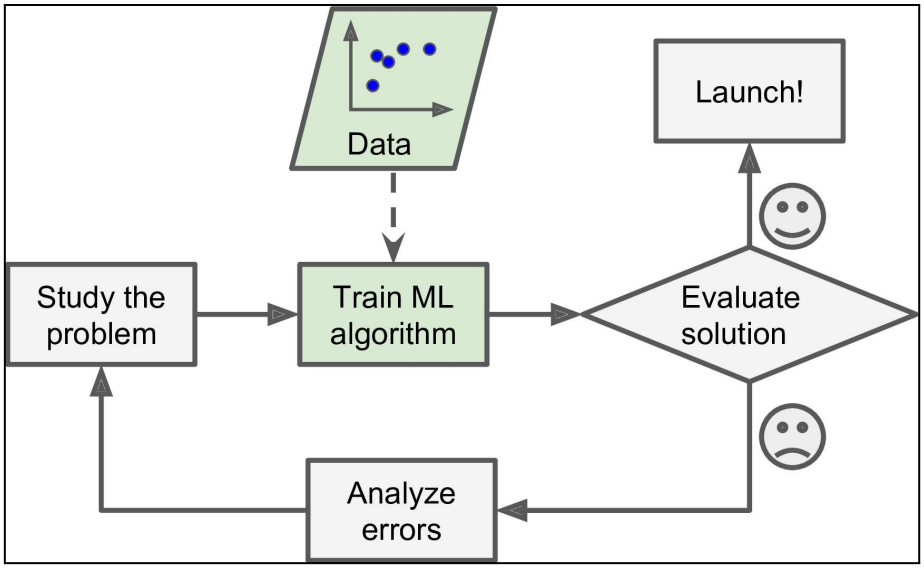

Figure 31: Machine Learning Approach [25]

Machine Learning is great for [25]:

- Simplifying the code and performing better for issues where current solutions require a lot of manual tweaking or lengthy lists of rules.
- Complex issues for which traditional methods offer no viable solutions can be solved using the best machine-learning techniques.
- Adapting to new data in fluctuating environments.
- Gaining knowledge about complicated issues and vast amounts of data.

9-2) Types of Machine Learning Algorithms

Machine learning algorithms are classified based on the amount and type of supervision that algorithm gets during training. There are four categories based on this criterion [25]:

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

Because in this project, we will use supervised learning algorithms in 5G classification, we focus on this algorithm in the rest of this chapter.

9-2-1) Supervised Learning Algorithms

Machine Learning systems can be classified according to the amount and type of supervision they get during training. There are four major categories: supervised learning, unsupervised learning, semi-supervised learning, and Reinforcement Learning [25].

In supervised learning, the training data we feed the algorithm includes the desired solutions, called labels (Figure 32).

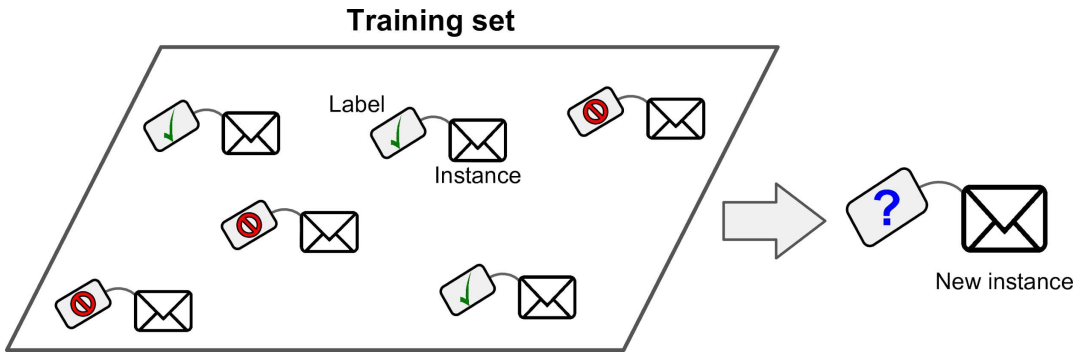

Figure 32: A labelled training set for supervised learning (e.g., spam classification) [25]

We have two typical kinds of supervised learning:

- Classification algorithms
- Regression Algorithms

  - **Classification Algorithm:** An excellent example is the spam filter, which must learn how to categorize new emails after being taught several instances of emails along with their classification (spam or ham). (Label: Ham or Spam)
  - **Regression Algorithm:** It predicts a target numeric value, such as the price of a car and the benefit of selling a particular product. The algorithm performs this prediction given a set of features (mileage, age, brand, etc.) called predictors. This sort of task is called regression. To train the system, we need to give it many examples of cars, including their predictors (mileage, age, brand) and labels (prices).

So, we can introduce marching learning terminology as follows:

Attribute: Data type (Mileage)
Feature (Predictor): An attribute with its value (For example, Mileage: 15000)
Label: Output (Car price or Spam/Ham)

Here it has mentioned some well-known types of supervised learning algorithms:

- K-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVMs)
- Decision Trees and Random Forests
- Neural Networks

9-3) Main Challenges of Machine Learning

In short, since our main task is to select a learning algorithm and train it on some data, the two things that can go wrong are "bad algorithm" and "bad data.". Therefore, this section will review some of the main challenges in data and machine learning [25].

9-3-1) Bad Data

We can mention bad data for the training step in machine learning into four categories [25]:

- Insufficient Quantity of Training Data
- Non-representative Training Data
- Poor-Quality Data
- Irrelevant Features

- **Insufficient Quantity of Training Data:**

Most machine learning algorithms require a large amount of data to operate correctly. We usually need thousands of instances, even for elementary issues. Millions of examples may be required for complicated problems like voice or image recognition (unless you can reuse parts of an existing model).

- **Non-representative Training Data**

Our training data must accurately reflect the new cases we wish to generalize to achieve good generalization. Whether we employ model-based learning or instance-based learning, this is true.

- **Poor-Quality Data**

Our system is less likely to function effectively if our training data contains many errors, outliers, and noise (for instance, due to low-quality measurements). Outliers are those data points significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution and arise due to inconsistent data entry or erroneous observations (Figure 33)

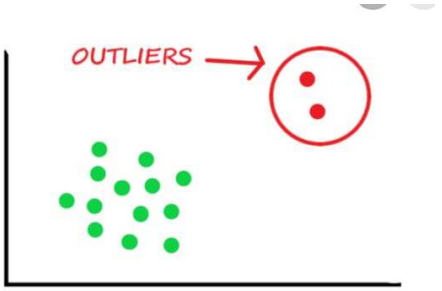

Figure 33: Outliers

It will be more difficult for the system to identify the underlying patterns. Spending time cleaning the training data is frequently well worth the effort. Most data scientists devote a sizable portion of their time to doing just that. For instance:

• If some instances are apparent outliers, it can be helpful to ignore them or try to correct the mistakes personally.
• If a few instances lack one or more features (for example, 5% of our clients failed to provide their age), we must determine whether to discard the attribute entirely, account for the missing values, train one model with the feature and another without it, etc.

- **Irrelevant Features**

Our system can only learn if the training data has an appropriate balance of valuable and irrelevant features. Finding a solid set of features to train on is essential to the success of a machine learning project. This procedure, known as feature engineering and includes the following:

• Feature selection: choosing from among already-existing features which ones will be most valuable for training.
• Feature extraction: combining the already-existing features to create a more beneficial one.

9-3-2) Bad Algorithm

We can mention bad algorithms for training step in machine learning into two categories [25]:

- Over-fitting the Training Data
- Under-fitting the Training Data

- **Over-fitting the Training Data**

Over-fitting occurs when the model could be more complex in comparison to the volume and granularity of the training data. There are a variety of potential remedies, including the following:

• Choose a model with fewer parameters to reduce the number of attributes in the training data

• Increase the size of training data.
• Decrease the noise of training data noise (e.g., fix data errors and remove outliers)
• Regularization: Regularization constrains a model to make it more straightforward and less prone to overfitting.

A hyperparameter can regulate how much regularisation is used during learning. This parameter must be established before training and remains constant during the training. The learning algorithm itself does not impact it. We will not face overfitting for large values of hyperparameters, but it will be less likely to find a good solution. In the proposed machine learning algorithm in the next chapter, we will deal with setting hyperparameters before starting the data training.

- **Under-fitting the Training Data**

The reverse of overfitting is under-fitting, which happens when the model is too straightforward to understand the underlying structure of the data. For instance, a linear life satisfaction model is prone to under-fitting since reality is inevitably more complex than the model, leading to unreliable forecasts even on training data.

- Under-fitting solutions:
  Choosing a more robust model with more variables.
  Enhancing the learning algorithm's features (feature engineering);
  Reducing the constraints on the model (e.g., reducing the regularization hyperparameter)

9-4) Testing and Validating of Data

Like any machine learning algorithm, we must divide our data into two sets (training and testing). In the next chapter, we will use hyperparameter tuning and cross-validation for data mining. These concepts will be described in more detail in this part before the implementation phase [25].

A model can only be tested on new cases to determine how effectively it generalizes to new situations. Putting a model into production and keeping an eye on how it does is one method to achieve that. This is effective, but there are better ideas than this if the model is good. The training and testing sets should be separated into two sets as a preferable alternative. These titles suggest that the training set is used to train the model while the test set is used to evaluate it. By assessing the model on the test set, we may estimate the generalization error (also known as out-of-sample error). This value informs us how well the model functions in hypothetical situations. It indicates that the model is overfitting the training set if the generalization error is significant while the training error is low (i.e., the model makes a few mistakes on the training set). Typically, 80% of the data are used for training, and the remaining 20% are kept for testing.

9-5) Hyperparameter Tuning and Model Selection

Using a test set, we can quickly evaluate a model. However, how would we choose between linear and polynomial models? One choice is to teach both and assess their generalization errors using the test set. The linear model generalizes more effectively, but we still want to

use regularisation to prevent overfitting. How do we determine the regularisation hyperparameter's value precisely? One possibility is to use 100 distinct values for this hyperparameter to train 100 different models. Assuming we discover the optimal hyperparameter value that results in a model with the slightest generalization error, let us say it is 5%. We put this model into operation, but it could perform better and generate 15% more mistakes than anticipated. What happened just now? The issue is that we modified the model and hyperparameters to build the best model for that specific. Despite many assessments of the generalization error on the test set, this modification was done. This indicates that the model's performance with new data is unlikely to be as good [25].

Holdout validation is a popular approach to solving this issue. We withhold a portion of the training data to compare various candidate models and choose the best one. The validation set is called the development set or the dev set. The reduced training set (i.e., the whole training set minus the validation set) is used to train multiple models with different hyperparameters, and the model that performs the best on the validation set is chosen. The best model is trained on the entire training set (including the validation set) following this holdout validation approach, which yields the final model. We assess this final model on the test set to determine the generalization error. In most cases, this solution is very effective. The model evaluations, however, will be imprecise if the validation set is larger; we can choose a suboptimal model.

The remaining training set will be considerably less than the entire training set if the validation set, on the other hand, is extensive. Why does this matter? Comparing candidate models learned on a significantly smaller training set is not optimal because the final model will be trained on the entire training set. It would be comparable to picking the fastest sprinter to run a marathon. Repeated cross-validation utilizing multiple small validation sets can help overcome this issue. After being trained on the remaining data, each model is evaluated once for each validation set. We obtain a far more precise assessment of a model's performance by averaging all of its evaluations. However, there is a drawback: the training time is multiplied by the number of validations sets [12].

9-6) Cross-Validation

Cross-validation is a crucial tool for data scientists. It helps create machine learning models that are more accurate and evaluate how well they perform on a different test dataset. Cross-validation is a common technique for evaluating the predictive abilities (or skills) of various models and selecting the best one since it is simple to comprehend and put into practice. It helps when limited data is available and is an excellent approach to seeing how a predictive model performs in real-world situations [26].

In cross-validation (CV), a specific sample from a dataset on which the model hasn't been trained is set aside. Later, this sample is used to test the model and assess it. This technique prevents overfitting in models, mainly when data is lacking. It is frequently employed when the model's aim is prediction and is also known as rotation estimation or out-of-sample testing. Another application of cross-validation is to adjust a machine-learning model's

hyperparameters. This process is known as randomized grid search cross-validation. We will use a method in the following chapter on the suggested machine learning algorithms.

Types of cross-validation

Cross-validation methods can be broadly classified into exhaustive and non-exhaustive methods. Exhaustive cross-validation methods aim to examine every possible option to split the initial data sample into a training and a testing set, as the name suggests. Non-exhaustive approaches, on the other hand, only compute some possible ways to divide the original data into training and assessment sets. The five most popular cross-validation methods are listed below. Since we will use the K-fold cross-validation approach in the next chapter, it is described here in more detail [26].

1. Holdout method
2. K-fold cross-validation
3. Stratified k-fold cross-validation
4. Leave-p-out cross-validation
5. Leave-one-out cross-validation

K-fold cross-validation

The holdout approach has been improved with the k-fold cross-validation method. Because the model's score is independent of how the training and testing datasets are chosen, it is more stable. The holdout method is applied k times to each dataset split in this non-exhaustive cross-validation strategy. For example, if K is two, there will be two equal-sized subgroups. In the initial iteration, the model is trained on one subsample and validated on the other. The model is evaluated on the additional subset and trained on the subset that served as its initial validation. This technique is referred to as 2-fold cross-validation.

Similarly, the method is known as the K-fold cross-validation method and requires K subsets and K iterations. Moreover, K's value is chosen at random. Typically, K is set to be equal to 10. The same is advised if we need help deciding on a value. The initial step in the K-fold cross-validation technique is randomly dividing the original dataset into K folds or subsets. The model is trained on each iteration's K-1 subsets of the total dataset. The model is then tested on the Kth subset to see how well it performs [12].

This process is repeated until the k-folds have served as the evaluation set. The cross-validation accuracy is the result of averaging the findings from each iteration. As a performance statistic, cross-validation accuracy is used to evaluate the effectiveness of various models. Since every data point from the original dataset will present in both the training and testing sets, the K-fold cross-validation technique typically results in less biased models. This approach is ideal when we only have a small amount of data. Nevertheless, since the algorithm must be done K times from scratch, this procedure could take some time. Additionally, it takes K-1 times as much computing as the holdout technique [12].

10 Chapter 10) Implementation of Proposed ML Algorithm in O-RAN

10-1) 5G network slicing and service classification

There are situations in 5G with such divergent requirements where it is necessary to use the same network functions (NFs) but in various placements to meet requirements like those for latency, for example. There are also instances where certain NFs are required for one use case but not another. This might be accomplished through network slicing using a single physical infrastructure, allowing the functionality to be placed where and when needed based on the use case and application. As a result, this idea enables the customization of a dedicated virtual network to meet a particular demand [28].

10-2) 5G network slicing in Cloud RAN

A 5G idea called "network slicing" involves leveraging a single physical infrastructure to support numerous logical networks through virtualization. These networks can satisfy various demands or those that are specified for various tenants, such as MNOs who wish to offer their services. A network slice is a logical network made up of a collection of network functions (NFs) implemented on a shared physical infrastructure, enabling communication services for a specific use case. These use cases could refer to a variety of applications with various specifications, such as V2X (Vehicular-to-Anything), IoT (Internet of Things), or MBB (Mobile Broadband). The exact use case for the same operator but for various objectives, such as private networks or separate services of the same user type but for different operators, could also be meant by this.

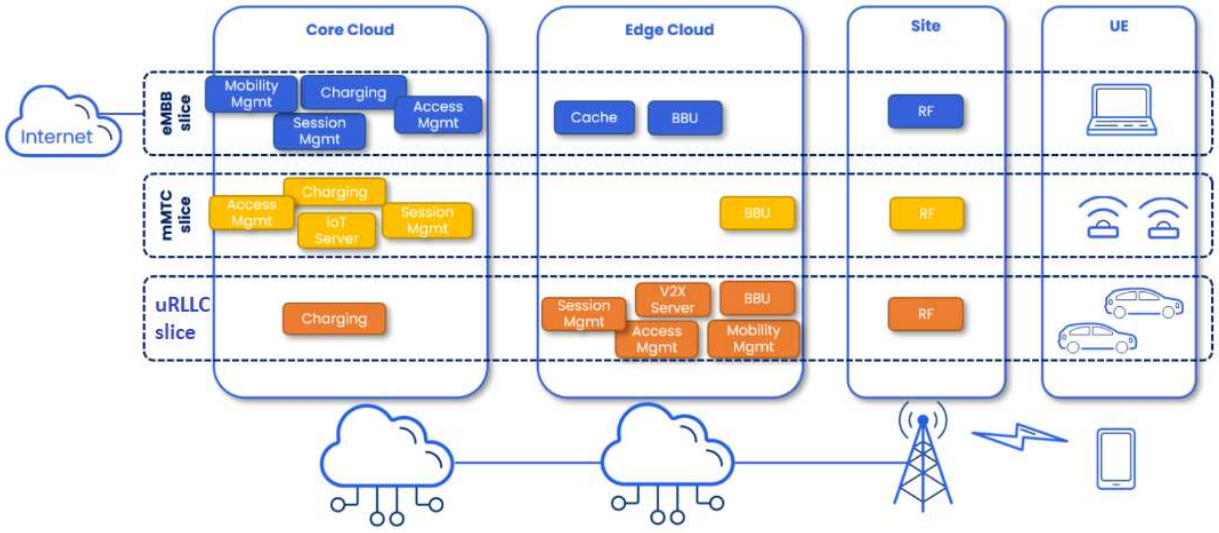

Figure 34: Network slicing concept in Cloud-RAN [28]

Figure 34 shows an example of a single infrastructure with cloud computing platforms at various locations along with the cell site. A core cloud is a hub that houses a lot of computing power for an operator, generally a local data centre. Edge clouds are situated close to the network edge to guarantee cloud computing resources at a remote location and reduce data processing delay. As virtual network instances are made up of various functional components to fulfil various requirements, three different slices are offered.

10-3) 5G Network Slicing in Open RAN

O-RAN could be used to realize many parts of the network, notably RAN slicing, because of its native virtualization and embedded intelligence. However, traditionally implementing network slicing is significantly more challenging and significantly restricted. As a result, Network Slicing is one of the main applications for O-RAN. O-RAN alliance discussions cover resource management inside a slice and resource optimization between slices. One of the main issues is keeping resources used by one slice separate from those used by others. Another subject is the proper scaling of resources to ensure SLA inside a specific slice. In this study, we present a novel model that outperforms network slicing using machine learning methods in the RIC block of the Open RAN [29].

A sample RAN slicing deployment of O-RAN network functions based on the selected initial deployment option is shown in Figure 35, with some of the network functions shared between RAN slice subnets (such as O-CU-CP, O-DU, O-RU) and some network functions dedicated to a particular RAN slice subnet (such as O-CU-UP).

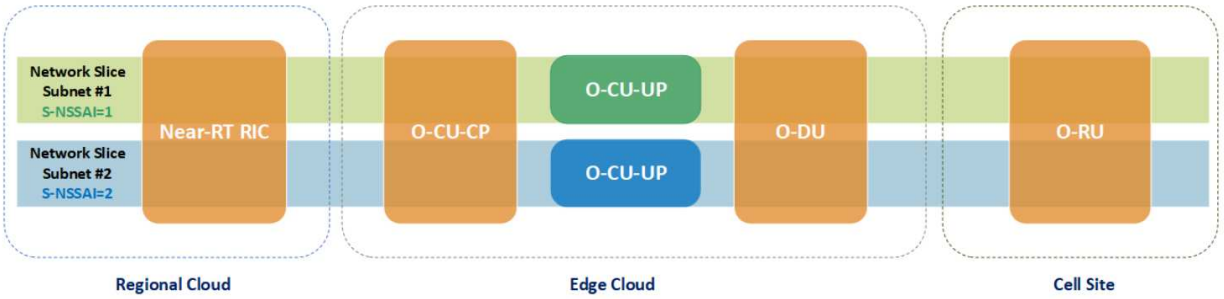

Figure 35: Network slicing concept in Open-RAN [29]

10-4) O-RAN Slicing Use Cases

There are three essential use cases for Network Slicing in the new generation of RAN networks, and in this section, we just mentioned the title of three important cases. AI/ML is used in RAN Slice SLA Assurance and Resource Allocation Optimization cases [29].

- RAN Slice SLA[19] Assurance
- Multi-vendor Slices
- NSSI[20] Resource Allocation Optimization

10-5) Proposed 5G Network Slicer and Service Classifier

This section will address two issues (SST[21]) in 5G RAN networks that are crucial for satisfying user QoS and QoE and intelligent resource allocation. This classification uses cases in RAN Slice SLA Assurance and NSSI Resource Allocation Optimization.

- **Service Classification**
  Ultra-High Definition (UHD) Video Streaming (UHD)
  Immerse Experience (IE)
  Smart Grid (SG)
  Intelligent Transport Systems (ITS)
  Voice over 5G (VO)
  E-Health (EH)
  Connected Vehicles (CV)
  Industry Automation (IA)
  Video Surveillance (VS)

- **Network Slicing**

[19] Service Level Agreement
[20] Network Slice Subnet Instance
[21] Slice/Service Type

eMBB
uRLLC
mMTC

We can provide a mapping with services and slices in 5G networks according to table 2. It is necessary to point out a brief description of Vo5G. Since 5G networks are designed to be standalone, i.e. work without relying on legacy networks (e.g. 4G LTE), they need to continue offering real-time services like voice calls and messaging that legacy networks already deliver. VoNR or Vo5G is the technology in 5G that allows them to continue offering voice, messaging and potentially other real-time services without any dependency on legacy networks. The delivery of voice calls and text messages is one of many use cases for 5G networks. However, due to persisting low amounts of latency for call establishment, we put it in the slice of uRLLC.

Table 2: 5G Service and Slice mapping

5G Service	5G Slice
Ultra-High Definition (UHD) Video Streaming (UHD)	eMBB
Immerse Experience (IE)	eMBB
Smart Grid (SG)	mMTC
Intelligent Transport Systems (ITS)	uRLLC
Voice over 5G (VO)	uRLLC
E-Health (EH)	uRLLC
Connected Vehicles (CV)	mMTC
Industry Automation (IA)	uRLLC
Video Surveillance (VS)	uRLLC

A new approach to categorizing 5G services using ML has been developed by the Network Machine Learning Research Group (NMLRG). [30]. Results from using their models to categorize 5G services are presented in some publications released by the NMLRG. When doing system classification, they concentrated on network traffic-related KPIs as the primary considerations. The authors compared it using SL methodologies and concluded that Decision Trees and Random Forests are the best solutions for this issue. KQIs represent a change from conventional network-based performance parameters (KPIs) to an arbitrary quality-based metric known as QoE that the end-user sees. KQIs were not promoted or used for a while after they were defined [6]. Most studies on service classification have focused on KPI requirements, and none considered the KQI parameters as elements for the classification of 5G services. When KQIs are included, the ML algorithm becomes even more complicated. However, KQI offers a framework that can objectively reflect service performance and quality from an E2E perspective. These indicators can be obtained through direct testing and statistical analysis of the network [31].

Here we have simulated different supervised learning algorithms to address these two classifying issues. The algorithms are trained using a database containing 5G RAN KPIs plus KQIs. All learning and evaluation are done in the block of RIC of Open RAN, and the results

can be sent to different telecommunication operators through APIs. AI/ML training models can be deployed in Non-RT and Near-RT RICs blocks (see Figure 36). Because non-real-time RIC, integrates with the network orchestrator and operates on a time scale longer than 1s, and a near-real-time RIC, drives control loops with RAN nodes with a time scale between 10ms and 1s. This timing scale is enough to have time for training and fit the machine learning model with the dataset. Future research and proposals may investigate the use of AI/ML models in RT RIC which we have fewer timing scales and closed-loop control times between 1ms and 10ms [31].

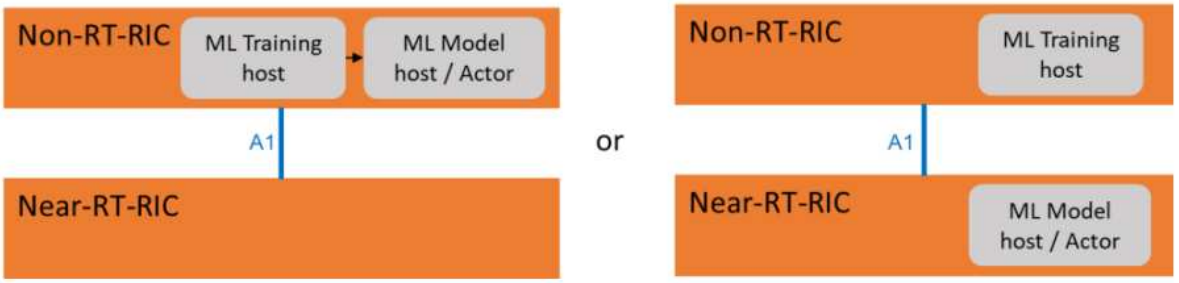

Figure 36: Supervised learning model training and actor locations [32]

We suggest enhancing the classification (identification) of the services while also considering KQI characteristics and KPI. We predicted that adding KQIs to the classification of 5G/B5G services will enhance QoS and QoE while also increasing the accuracy of SLA determination and compliance. Service Level Objectives are the specifications for establishing SLAs in 5G based on the services offered and the provider's infrastructure (SLOs). It is essential to consider KPIs and KQIs because the QoS, which serves as the foundation for determining the SLA, depends on the network, the applications, and other factors like user experience (QoE) [7]. There are two bold improvements that we are presenting at the end. The first shows that adding KQI to the KPI will improve accuracy and other evaluation factors, which we shall discuss later. Second, we will provide a brand-new random and grid search technique for hyperparameter tuning that will significantly enhance the Random Forest method for forecasting the slices and services for the 5G network.

10-6) Proposed block diagram for intelligent 5G service and slice classification

Figure 37 shows our block-level diagram of the proposed system for classifying services in 5G networks. The planned scheme first operates offline until the predictive model has been validated, and then it will learn to classify services effectively with few values of the error. In the next phase, the system is implemented online by the network operators, and the predictive model then classifies new services requested by the UEs. The proposed algorithms can be deployed for SLA assurance and Resource Allocation in the new generation of RAN networks (Figures 38 and 39). For a more straightforward graphical presentation, we name the block diagram of Figure 37 ISSC (Intelligent Service and Slice Classifier).

The system's output corresponds to the requested service classification and is feedback to the ML algorithm, making the predictive model more efficient. The proposed system can classify services in next-generation networks. However, it is essential to clarify that this forms only one part of a system that a network operator can use to offer services. Our system needs to

interact to connect with the rest of the operator's system, and we must therefore consider two options:

1. Interpreting the suggested system in the operator's language. This has the drawback of necessitating reprogramming, which is not very practical, of the systems utilized by each operator, including the cloud (due to future maintenance or update issues).

2. Incorporating a suitable Application Programming Interface (API) into the proposed system to allow connection with the operator's system (accessible from a public or private server). The required security must be offered to guarantee that this is only used in an authorized manner.

The second option is preferable since 5G systems provide appropriate APIs to allow a trusted third party to create, modify, delete and monitor instances of the network segments used by the third party and to manage a set of devices or capabilities, including QoS functions.

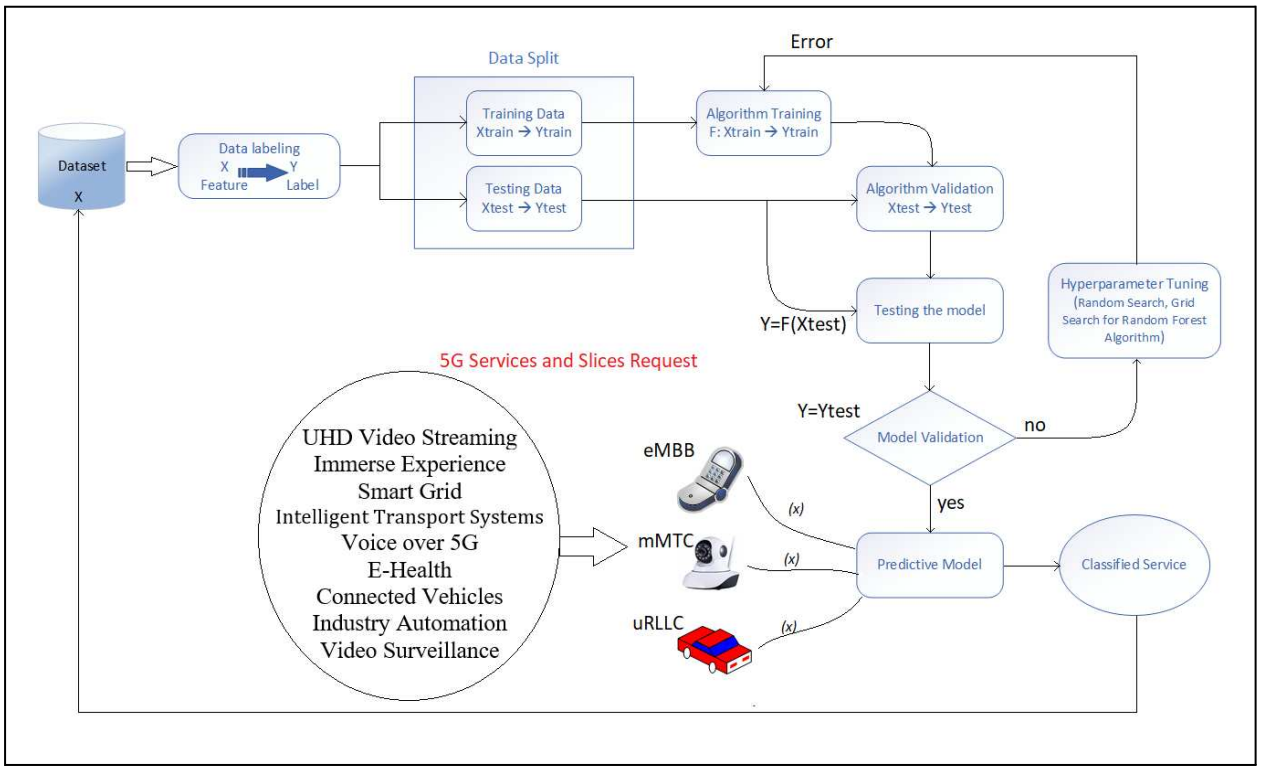

Figure 37: Proposed block-level diagram for intelligent 5G service and slice classification (ISSC)

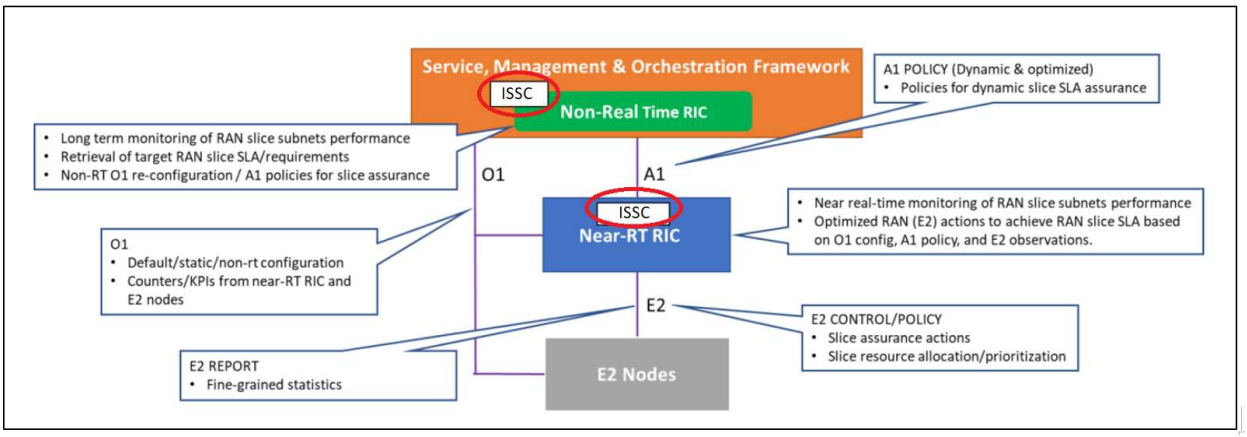

Figure 38: Use of AI/ML in SLA assurance for Open RAN systems [29]

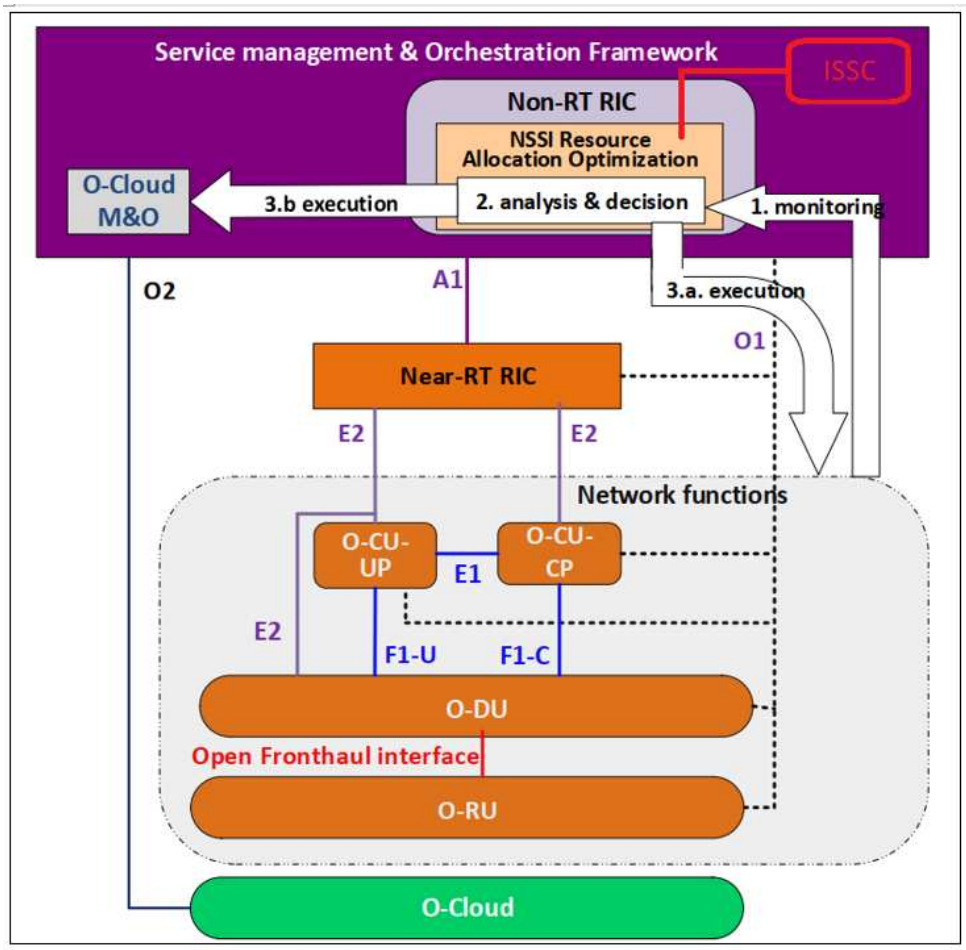

Figure 39. Use of AI/ML in Optimized resource allocation for Open RAN systems [29]

10-7) 5G KPI and KQI datasets

The main limitation in this project and many other ML-based projects was the achievement of a real dataset with 5G operating systems parameters. It was searched extensively for different references and online valid repositories on the internet. The drawback is that we can rarely

find a useful and labelled 5G dataset from telecommunication vendors dataset freely on the internet. The reason might be the intense competition between different vendors and operators. However, better access to the 5G and B5G databases might be found through the Open RAN alliance. I decided to use one of the valuable datasets recently used for 5G service classification [33]. The main idea behind our proposed algorithms comes from the comprehensive research done in [33].

This dataset was created manually and by examining KPI and KQI parameters that were taken from ITU standards publications and other European studies and analytical documents created by telecom firms. The dataset was created by randomly fluctuating the threshold values from the bibliography, resulting in different values for each KPI/KQI and each service.

The first block of the scheme shown in Figure 37 corresponds to the database we used to train the ML algorithm to validate and verify the predictive model. The database was manually generated using parameter values in Comma Separated Values (CSV) format that matched the KPIs and KQIs of the chosen services. In this phase, documents from the International Telecommunications Union (ITU), Huawei, the 5G Public-Private Partnership (5G-PPP), NGMN, Speed, 5G America, and other suppliers were studied. These documents dealt with standards and various project reports on 5G networks. Standard threshold values were the chosen parameter values, which were changed at random until we got results sufficiently close to their limiting limits. Table 3 shows the thresholds for extracted KPI/KQI parameters.

Table 3: Thresholds for the extracted KPI/KQI parameters [33]

Servil	E2E Latency (ms)	Jitter (ms)	Bit Rate (Mbps)	Packet Loss Rate (%)	Peak Data Rate DL (Gbps)	Peak Data Rate UL (Gbps)	Mobility (km/h)	Service Reliability (%)
UHD Video Streaming	Min: 4 Max: 20	5.84	10	Max: 1	20	10	Min: 0 Max: 500	Min: 95
Immersive Experience	Min: 7 Max: 15	20	50	Max: 5	20	10	Min: 0 Max: 30	Min: 95
Smart Grid	Min: 5 Max: 50	1	1	Max: 0.0001	20	10	Min: 0 Max: 0	Min: 99.9
E-Health	Min: 1 Max: 10	10	16	Max: 0.00000001	0.3	0.3	Min: 0 Max: 120	Min: 99.9999
ITS	Min: 10 Max: 100	20	0.5	Max: 0.1	20	10	Min: 50 Max: 500	Min: 99.999
Vo5G	Min: 20 Max: 150	30	10	Max: 1	20	10	Min: 0 Max: 500	Min: 99.9
Connected Vehicles	Min: 3 Max: 100	0.44	10	Max: 0.001	1	0.025	Min: 50 Max: 250	Min: 99.999
Industry Automation	Min: 1 Max: 50	0.1	1	Max: 0.0000001	20	10	Min: 0 Max: 30	Min: 99.999
Video Surveillance	Min: 10 Max: 50	5	10	Max: 0.001	0.05	0.12	Min: 0 Max: 320	Min: 99

The used database contains 165 rows and 14 columns. The first 13 columns include the KPI and KQI values, while the last column matches the labels of the 5G services. The rows represented the parameter values of the 5G services that needed to be classified. Table 4 is a portion of the database where we can see specific KPI, KQI, and 5G service statistics. In this project, we tackle a classification issue using various labels (5G services). We must assign a label to separate the elements that need to be classified.

We are talking about the supervised learning scheme for the classification of 5G services when applying ML to a set of labelled data, which uses both the characteristics of the services and their labels to solve a classification problem. In the following section, we will give a different block diagram and dataset that were utilized for slice classification. The 5G services detected in the database must be labelled (represented by the variable y); the labels Y correspond to the parameters or characteristics for the incoming services, which the variable X. represents (see Figure 37). Before constructing the algorithm's predictive model, the data is labelled, making it possible to determine which label (Y) corresponds to the parameters (X) of each 5G service in the database (see Table 4).

Table 4: Fragment of ten entries of the database [33]

Index	Latency (ms)	Jitter (ms)	Bit Rate (Mbps)	Packet Loss Rate (%)	Peak Data Rate DL (Gbps)	Peak Data Rate UL (Gbps)	Mobility (km/h)	Reliability (%)	Service Availability (%)	Survival Time (ms)	Experienced Data Rate DL (Mbps)	Experienced Data Rate DL (Gbps)	Experienced Data Rate UL (Mbps)	Experienced Data Rate UL (Gbps)	Interruption Time (ms)	Service
1	15	5	11	0.1	18	7	260	95	99	8	1000	1	500	0.5	1000	UHD_Video_Streaming
2	5	5.5	10	1	20	10	20	95	99.2	9	990	0.99	440	0.44	2000	UHD_Video_Streaming
3	8	10	50	3.8	15	7	15	97	99.9	10	1000	1	50	0.05	0.2	Immerse_Experience
4	40	1	0.5	1.00E-05	18	9	0	99.92	99.999	10	5	0.005	8	0.008	0	Smart_Grid
5	90	18	0.2	0.08	13	2	480	99.9995	99.9999	100	10	0.01	10	0.01	1000	ITS
6	130	5	8	0.9	14	6	400	99.94	95	100	50	0.05	25	0.025	0	Vo5G
7	10	19	32	4.7	13	5	26	95.6	99.92	8.9	900	0.9	40	0.04	0.1	Immerse_Experience
8	2	3	15	8.00E-09	0.2	0.2	100	99.99996	99	1	10	0.01	100	0.1	0	e_Health
9	5	0.5	10	0.00075	0.8	0.024	80	99.9992	99	1	50	0.05	25	0.025	0	Connected_Vehicles
10	1	0.05	0.6	1.00E-07	15	6	28	99.999	99.9999	0	1	0.001	10	0.01	100	Industry_Automation

10-8) Proposed Machine Learning Algorithm and Predictive Model

The basis for our classification of services is that each service is represented by a set of parameters (x) that are determined by the KPIs and KQIs that describe it. These parameters must assign each label (y). Based on table 2, we will forecast the services, and the slice will

then be mapped. As was already said, only one database contains all of the descriptive information about the services that need to be classified; as a result, it must be split into two datasets, one for training the algorithm and the other for predicting or validating its results. Four additional variables are created when the database is split: Xtrain, Xtest, Ytrain, and Ytest. The input values for the 5G services that were chosen to train the ML and Ytrain algorithms with their respective output labels correspond with the training variable Xtrain (Features). The input and output variables for the testing and validation stage of the predictive model are represented by the other two variables, Xtest and Ytest. The training phase involves passing training data to the ML algorithm to allow it to learn. The ML algorithm develops a function based on the training data (Xtrain) that provides the correct answer (Ytrain). Using *Xtrain* and *Ytrain*, the algorithm learns, and a function *f*(*Xtrain*) = *Ytrain* is generated that: identifies patterns in the training data, allows the attributes of the input data to be assigned to the target data (representing the answer to be predicted) and generates a model that captures these patterns.
The next step is to use a machine learning method to create a function y = f(x) that can predict the value associated with any input item x. (in the proposed system for classifying services, these represent the KPIs and KQIs of 5G services). A set of parameters or characteristics from the various services must be classified to train the machine learning algorithm. Since the predictive model may now produce results for additional data after the initial training, this training enables the assignment of new known input values (x) and new unknown labels (y). This means that after the machine learning algorithm is trained, the predictive model can predict or classify the required services (see Figure 37).

The result of using the Xtrain data to train the ML algorithm is a predictive model that can classify 5G services. The Xtest data were not included in the training of the ML algorithm. So, the predictive model cannot correctly classify services. Therefore, it is essential to validate the ML algorithm to ensure the predictive model is successful. If the predictive model displays overfitting, it will not be helpful because a model that repeats the labels of the samples it has just seen would score 100 per cent, but it is unable to predict unknown labels. To avoid this problem for the validation block of the ML algorithm, we will apply a method based on the cross-validation technique.

If the validation results are comparable to those from the evaluation and training, we can say that the trained model is accurate, and there is no sign of overfitting. This validation indirectly impacts the predictive model's final evaluation. The metrics' values must match those from the validation stage when the model is evaluated with the new Xtest data, as this proves that the chosen algorithm is efficient. Testing the prediction model to see if it can forecast new and future data comes after the algorithm has been trained and validated. This is dealt with by the test block of the model, which is shown in Figure 37 as Y = f(Xtest), by performing a prediction test with Xtest. The output Y from this block corresponds to the test results of the predictive model with the variable Xtest and is a vector of the various 5G services generated by the predictive model.

It is essential to label the 5G services located in the database (represented by variable y); labels y relate to features or qualities for the arriving services, which are represented by variable x. (see Figure 37). Before developing the algorithm's predictive model, the data is labelled, making it possible to identify which label (y) corresponds to the parameters (x) of each 5G service in the database. To evaluate the accuracy of the model's predictions, a prediction Y can be compared with previously recorded data as the target response (Ytest). This test can then be used to establish the predictive accuracy of future data. Consequently,

an analysis of the Y and Ytest vectors can describe the verification or validation ability of the model.

Five different supervised learning algorithms have been used. In the end, we conclude that the Random Forest algorithm using our proposed random search and grid search for hyperparameter tuning performs better than the other algorithms. Here is the name of the supervised learning algorithms.

1- Decision Tree Algorithm (DT)
2- Random Forest Algorithm (RF)
3- Support Vector Machines Algorithm (SVM)
4- K-Nearest Neighbor Algorithm (KNN)
5- Multi-Layer Perceptron Algorithm[22] (a Neural Network algorithm) (MLP)

10-9) Validation of the Predictive Model

It is necessary to determine whether the values obtained for *Y* are the expected ones. This is required to validate the predictive model. Using metrics to measure performance can allow us to confirm the model's effectiveness. The relationship between *Ytest* and Y is used to generate the performance measures and to construct the confusion matrix shown in Table 5.

Table 5: Confusion matrix for binary classification

Confusion Matrix		**Prediction (Y)**	
		Positive	**Negative**
Label (Ytest)	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

A confusion matrix is so named because it visualizes the predictive model's performance and observes confusion in two labels. The columns of the matrix represent the number of predictions for each label (*Y*) made by the predictive model, while each row represents the current label for the test values (*Ytest*) as follows:

- **True Positives**: The number of current values classified as belonging to a particular class for which the model´s prediction is correct.
- **False Positive**: These are the current values classified as belonging to an incorrect class. The model considers them to be positive, but the prediction is wrong.
- **False Negative**: These values belong to a particular class but are classified differently (incorrect prediction).
- **True Negative**: These are observations that do not belong to a given class and are classified correctly.

An example of these four states is the famous instance of diabetic cases; Suppose that the output is: diabetic (+ve) and healthy (-ve)

· True Positives: The prediction is +ve, and the patient also has diabetes (We want it)

[22] A fully connected class of feedforward artificial neural network

· False Positive: The prediction is +ve, but patient is healthy (False alarm)
· False Negative: The prediction is -ve, but the patient has diabetes (The worst case)
· True Negative: The prediction is -ve, and the patient is healthy (We want it)

A series of metrics can be derived from the results in the confusion matrix of Table 5 and used to evaluate the performance of the predictive model as follows:

Accuracy: This is the relationship between the number of correct predictions (TP and TN results) made by the model and the total number of predictions. In other words, this reflects how often the predictive model's classification is correct. It is the most direct measure of the quality of the classification. However, it is less appropriate when the labels of the output variables are not balanced (unbalanced data), i.e., labels are not of similar quantities.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Precision: This measures the precision with which the predictive model ranks services by their performance due to optimistic predictions. It is the relationship between the number of correct predictions and the total number of correctly predicted predictions.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall: This is the relationship between the number of correct predictions to the total number of positive predictions. In other words, it represents the predictive model's sensitivity in detecting positive instances.

$$Precision = \frac{TP}{TP+FN} \quad (3)$$

F1 score: This is a weighted average of recall and precision. A higher score represents a better model. Thus, it provides a good indicator of the overall accuracy of the predictive model, while the accuracy and recall provide information on explicit areas.

$$Precision = \frac{2}{\frac{1}{Accuracy}+\frac{1}{Precision}} = \frac{2\times Accuracy\times Precision}{Accuracy+Precission} \quad (4)$$

Matthews correlation coefficient (MCC): As an alternative measure unaffected by the unbalanced datasets issue, MCC is the only binary classification rate that generates a high score only if the binary predictor correctly predicted the majority of positive and the majority of negative data instances. It ranges in the interval [-1, +1], with extreme values -1 and +1 reached in case of perfect misclassification and classification, respectively. At the same time, MCC = 0 is the expected value for the coin-tossing classifier.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (5)$$

Suppose the values of the metrics for the predictive model are satisfactory. In that case, the offline work phase is terminated, and the model is ready to be used online by a network operator to classify new services requested by the UEs. The entire cycle must be repeated, beginning with the training of the ML algorithm until an acceptable success rate is seen so that the model will generate fewer errors in the future. This is necessary if the results obtained in terms of the metrics are different from what was anticipated. In the latter scenario, any of the subsequent steps may be taken:

• Increasing the volume of training data and testing the predictive model.
• Choosing another ML algorithm.
• Making the ML algorithm used in the simulation more straightforward or complex to achieve better precision.

We currently have a predictive algorithm that can classify 5G services according to their KPIs and KQIs. When this is done online, users (UEs) submit requests for new services (shown in the lower portion of Figure 37), and the model receives a vector of the KPIs and KQIs of those services as input. Our system uses the service's KPIs and KQIs as well as an output tag to classify the services and feeds the information into the system database. This strategy aims to benefit from each service requested by adding it to the database and repeatedly retraining the ML algorithm until a new, more reliable predictive model is formed that offers a better classification.

10-10) Implementation of ML algorithms

We perform two simulations to determine whether the inclusion of KQIs improves the predictive service classification model. The first considers only the KPIs, while the second also incorporates the KQIs. We first explain and define the scenario and conditions used in the simulations. The necessary elements are the SML algorithms, a programming language, a development platform, the 5G services to be classified, and the parameters of their KPIs and KQIs.

For the validation scenario and to simulate the proposed system, we used SL algorithms: Decision Tree, Random Forest (with five trees), Support Vector Machine (SVM) with a linear kernel, K-Nearest Neighbors (KNN, K = 3) and Multi-Layer Perceptron Classifier (MLPC), using the Python language, and Anaconda Navigator platform with Jupyter Notebook as IDE. We considered nine essential 5G services to be classified (Table 2): Ultra High Definition (UHD) video streaming, immersive experience, connected vehicles, e-health, industry automation, video surveillance, smart grid, Intelligent Transport Systems (ITS) and Voice over 5G (Vo5G). Based on Table 4, the selected KPI parameters were E2E latency, jitter, bit rate, packet loss rate, peak data rate Downlink (DL), Uplink (UL), mobility and service reliability. The KQI parameters were service availability, user experience data rate DL/UL, survival time and interruption time.

In the first simulation, we worked with the KPIs. The dataset had dimensions of 165 × 9, where the first eight columns represented the KPIs, and the last contained the labels of the services. We divided the database into two parts, where 80% (132) of the data (*Xtrain*) were used to train the algorithms created and, once trained, generated the predictive model. The remaining 20% (*Xtest*) was used to test the model.

10-11) Simulation Results (KPI as training set)

The models may be prone to underfitting or overfitting, which means that while they may perform perfectly with known training data (Xtrain), their accuracy may be poorer with new services (Xtest). There are two ways to prevent overfitting: increasing the database's size or reserving extra data by separating the dataset into three sections (training, validation and testing). Because there is a need for known data from the 5G service, increasing the amount of data is challenging. As a result, extra data were reserved, and the K-Folds cross-validation technique was used here with K = 10, yielding the results shown in Table 6 for each algorithm. Notably, all the variables were included in the initial dataset and stayed the same.

Table 6: Results of the accuracy in the cross-validation stage for the first simulation (KPIs)

SL Algorithms	K-Folds (K = 10) Cross-Validation Results
Decision Tree	99.23
Random Forest	99.23
SVM	92.42
KNN	59.83
MLPC	86.26

The confusion matrices for each model in the first simulation, in which we took the KPIs into account, are shown in Figure 40. The primary diagonal displays how many accurate predictions the predictive model made. Values outside the main diagonal show the model's incorrect predictions. The acronym used for each row of this matrix was mentioned in sections 12-4 and repeated here for convenience.

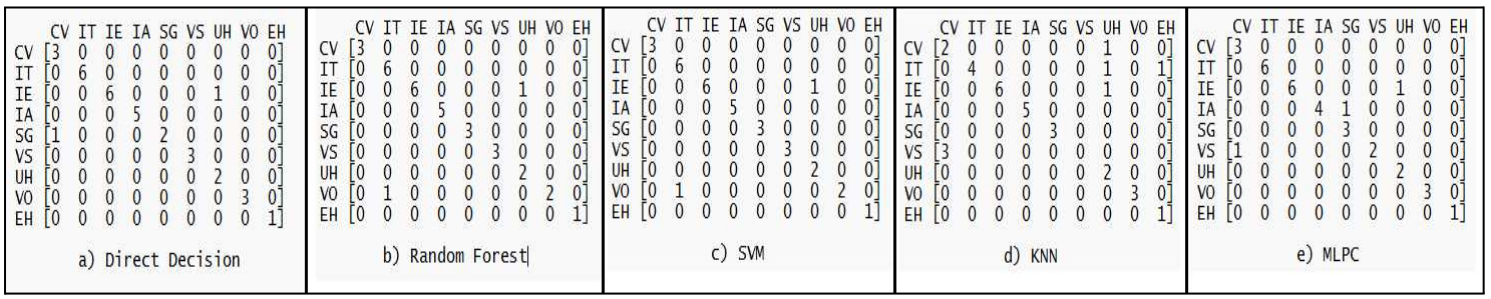

Figure 40: Confusion matrices for the first simulation (KPIs)

CV: Connected Vehicles
IT: Intelligent Transport Systems

IE: Immerse Experience
IA: Industry Automation
SG: Smart Grid
VS: Video Surveillance
UH: Ultra-High Definition (UHD) Video Streaming (UHD)
VO: Voice over 5G
EH: E-Health

We applied Equations (1) to (5) to the metrics obtained from the confusion matrix to evaluate the performance of the predictive models. The results are as follows in Table 7.

Table 7: Model metric results for the first simulation (KPIs)

SL Algorithms	Accuracy %	Precision Macro %	Recall %	F1-Score %	MCC %
Decision Tree	93.9	93.5	94.7	93.1	93.19
Random Forest	93.9	94.7	94.7	93.8	93.17
SVM	97	96.3	98.4	96.9	96.6
KNN	78.8	70	79.9	71.8	76.89
MLPC	90.9	90.7	92.5	90.3	89.8

10-11-1) Proposed Random Forest Algorithm

We have used hyperparameter tuning using two random seraph methods and gird search methods in the Random Forest algorithm to improve the metric result obtained for this algorithm shown in Table 7. Gathering more data and feature engineering usually has the most significant payoff in terms of time invested versus improved performance. However, it is time to move on to model hyperparameter tuning. This optimization for the random forest model is done in Python using Scikit-Learn[23] tools.

10-11-2) Hyperparameter Tuning

Back to our description for Hyperparameter Tuning (section 11-5), the best way to think about hyperparameters is like the settings of an algorithm that can be adjusted to optimize performance. While model parameters are learned during training, hyperparameters must

[23] Scikit-Learn is a free machine-learning library for Python. It supports supervised and unsupervised machine learning and provides various algorithms for classification, regression, clustering, and dimensionality reduction.

be defined before training the algorithm. Here, the number of decision trees in a random forest algorithm and the number of features of each tree are considered by each tree when splitting a node. (The variables and thresholds used to split each node identified during training are known as a random forest's parameters.) Scikit-Learn offers a set of reasonable default hyperparameters for all models (Table 8 shows these default values for the random forest technique). However, these defaults are not always the best for a given situation. The easiest way to find the ideal settings for hyperparameter tuning is to experiment with various combinations and assess each model's performance because experimental data rather than theory is used to tune these parameters.

Table 8: Random Forest Default Parameters

'bootstrap': True,
'ccp_alpha': 0.0,
'class_weight': None,
'criterion': 'entropy',
'max_depth': None,
'max_features': 'auto',
'max_leaf_nodes': None,
'max_samples': None,
'min_impurity_decrease': 0.0,
'min_impurity_split': None,
'min_samples_leaf': 1,
'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0,
'n_estimators': 5,
'n_jobs': None,
'oob_score': False,
'random_state': 0,
'verbose': 0,

'warm_start': False

10-11-3) Deployed Cross Validation

This strategy was discussed entirely in 9-6. The best way to teach cross-validation (CV) is with an example utilizing the most popular approach, K-Fold CV. Before starting a machine learning challenge, we divided our data into training and testing sets. We further divide our training set into K number of folds in K-Fold CV. We then fit the model iteratively K times, training the data on fold K-1 and evaluating fold K. (called the validation data). Here we used K=3 for cross-validation (Figure 41).

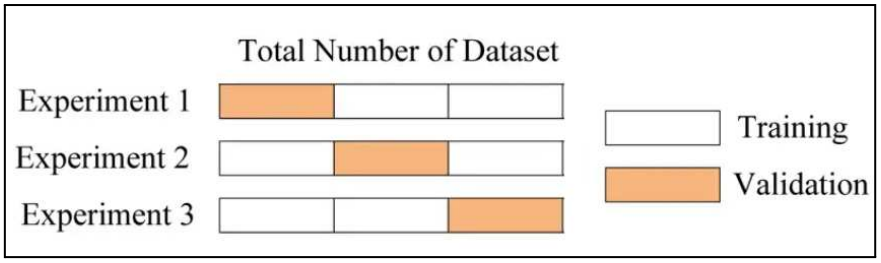

Figure 41: 3-Fold Cross Validation

We perform many iterations of the entire K-Fold CV process for hyperparameter tuning, each time using different model settings. We then compare all the models, select the best one, train it on the whole training set, and evaluate it on the testing set. Each time we want to assess a different set of hyperparameters, we have to split our training data into K fold and train and evaluate K times. Here we modify six parameters of the random forest algorithm, which is shown in Table 9. So, considering 3-Fold CV, we will have (6*3 = 18) training loops.

Table 9: Modified parameters for random search in Random Forest algorithm (KPIs)

'bootstrap': True,
'max_depth': None,
'max_features': 'auto',
'min_samples_leaf': 1,
'min_samples_split': 2,
'n_estimators': 5,

10-11-4) Random Search Cross-Validation in Scikit-Learn

Using Scikit-Learn's RandomizedSearchCV method, we can define a grid of hyperparameter ranges, and randomly sample from the grid, performing K-Fold CV with each combination of values. We will try adjusting the following set of hyperparameters because, based on the

mathematical theory of the Random Forest algorithm, these six parameters influence the model's accuracy and final prediction at the end.

n_estimators = number of trees in the forest
max_features = max number of features considered for splitting a node
max_depth = max number of levels in each decision tree
min_samples_split = min number of data points placed in a node before the node is split
min_samples_leaf = min number of data points allowed in a leaf node
bootstrap = method for sampling data points (with or without replacement).

Table 10 shows the achieved parameters after performing the random search for this algorithm when considering KPIs as the features of the random forest algorithm.

Table10: Achieved parameters for random search in Random Forest algorithm (KPIs)

'bootstrap': True,
'max_depth': 30,
'max_features': 'sqrt',
'min_samples_leaf': 1,
'min_samples_split': 5,
'n_estimators': 400,

10-11-5) Grid Search Cross-Validation in Scikit-Learn

Random searches allowed us to narrow down the range for each hyperparameter. Now that we know where to concentrate our search, we can explicitly specify every combination of settings to try. We do this with GridSearchCV, a method that evaluates all combinations we define instead of sampling randomly from a distribution. To use Grid Search, we make another grid based on the best values provided by random search. Table 11 shows the obtained parameters after the grid search.

Table 11. Achieved parameters for grid search in Random Forest algorithm (KPIs)

'bootstrap': True,
'max_depth': 20,
'max_features': 2,
'min_samples_leaf': 1,
'min_samples_split': 3,
'n_estimators': 100,

The results of the simulation provers that random search and grid search gain the same result, and it shows that the optimum parameters have been tuned for this algorithm. The

proposed algorithm gained a considerable improvement compared to the first baseline algorithm, which is shown in Table 12 (3.1% increase in accuracy and 3.43% increase in MCC)

Table 12. Model metric results for the first simulation (KPIs) (Comparison between the proposed algorithm and the other SL algorithms)

SL Algorithms	Accuracy %	Precision Macro %	Recall %	F1-Score %	MCC %
Decision Tree	93.9	93.5	94.7	93.1	93.19
Random Forest	93.9	94.7	94.7	93.8	93.17
SVM	97	96.3	98.4	96.9	96.6
KNN	78.8	70	79.9	71.8	76.89
MLPC	90.9	90.7	92.5	90.3	89.8
Proposed Random Forest (Random Search)	97	NA	NA	NA	96.6
Proposed Random Forest (Grid Search)	97	NA	NA	NA	96.6

10-12) Simulation Results (KPI plus KQI as training sets)

In the second simulation, we incorporated the user quality parameters (KQIs) and repeated the procedure (with a few differences from the previous simulation). The KQI parameters considered were service availability, user experience data rate DL/UL, survival time, and interruption time. A database containing 165 rows was kept, with five additional columns corresponding to the KQI parameters. We used the same functions to create and train the ML algorithms, resulting in the same SL algorithms. Again, we used the K-Folds cross-validation technique with K = 10 to validate the ML algorithm and obtained the results in Table 13. Figure 2 demonstrates the confusion matrix obtained for each model in this simulation.

Table 13: Results of the accuracy in the cross-validation stage for the first simulation (KPIs+KQIs)

SL Algorithms	K-Folds (K = 10) Cross-Validation Results
Decision Tree	97.69
Random Forest	98.52
SVM	91.59
KNN	83.35
MLPC	87.86

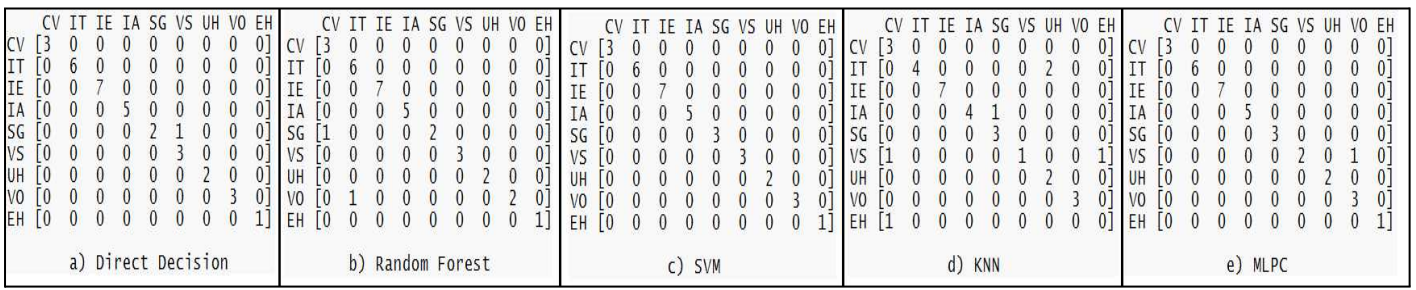

Figure 42. Confusion matrices for the first simulation (KPIs+KQIs)

We obtained the performance metrics for the predictive models based on the newly generated confusion matrices. The results are as follows in Table 14.

Table 14: Model metric results for the first simulation (KPIs+KQIs)

SL Algorithms	Accuracy %	Precision Macro %	Recall %	F1-Score %	MCC %
Decision Tree	97	97.2	96.3	96.2	96.6
Random Forest	97	97.2	96.3	96.2	96.6
SVM	100	100	100	100	100
KNN	81.8	76.1	75.6	71.8	79.8
MLPC	97	97.2	96.3	96.2	96.6

From Table 14, it is possible to know that the KNN model does not apply to our problem because it had inadequate accuracy. Furthermore, we can see that the other models increased their metrics in this second simulation, and the best metrics obtained are Decision Tree, Random Forest, MLPC, and SVM. To verify if the predictive model was satisfactory, we created a function to compare the accuracy obtained in the cross-validation stage versus the accuracy of the testing stage. We considered the model acceptable if the difference does not exceed 5%. The result obtained for the SVM had a difference of 8.41% (100-91.59), so this model may be overfitting. Also, this difference for MLPC is 9.14% (97-87.86). The result differed from 0.79% (97.69-96.9) and 1.62% (98.52-96.9) in the Decision Tree and Random Forest. This result balances the decision tree and the Random Forest algorithms' predictive model. If the predictive model is overfitting, we choose the third option mentioned above, for example, making a Random Forest with maximum depth. We can use both Decision Tree and Random Forest to solve the service classification problem presented. Although in this project, we proved random forest gains with proposed search algorithms.

10-12-1) Random Search Cross-Validation in Scikit-Learn

Table 15 shows the achieved parameters after performing the random search for this algorithm when considering KPIs+KQIs as the features of the random forest algorithm.

Table 15: Model metric results for the first simulation (KPIs+KQIs)
(Comparison between the proposed algorithm and the other SL algorithms)

SL Algorithms	Accuracy %	Precision Macro %	Recall %	F1-Score %	MCC %
Decision Tree	97	97.2	96.3	96.2	96.6
Random Forest	97	97.2	96.3	96.2	96.6
SVM	100	100	100	100	100
KNN	81.8	76.1	75.6	71.8	79.8

MLPC	97	97.2	96.3	96.2	96.6
Proposed Random Forest (Random Search)	100	NA	NA	NA	100

Table 16 shows the achieved parameters after performing the random search for this algorithm when considering KPIs+KQIs as the features of the random forest algorithm.

Table 16. Achieved parameters for random search in Random Forest algorithm (KPIs+KQIs)

'bootstrap': True,
'max_depth': 30,
'max_features': 'sqrt',
'min_samples_leaf': 1,
'min_samples_split': 5,
'n_estimators': 400,

As it is evident in Table 16, after using random search in the random forest algorithm and considering KPIs+KQIs as the features to train the algorithm, we could achieve 100% results in both accuracy and MCC evaluation criteria.

10-13) Necessary time for training

The other important parameter we have considered for simulation and comparison between algorithms is the necessary time for training each algorithm. This time is essential to know which algorithm is suitable for use in the three control loop mechanisms of Open RAN (Non-RT, Near-RT, and RT). This selection will be made based on the timing description in section 10.1. For calculating the time, we have used a cut-down timer in python, and each algorithm's fit () function has been considered for essential time for training. Table 17 demonstrates the average result for N=100 times of training for each algorithm. As mentioned before in section 10-1, Non-Real-Time RIC (Non-RT RIC) integrates with the network orchestrator and operates on a time scale longer than 1s, and a Near Real-Time RIC (Near-RT RIC), drives control loops with RAN nodes with a time scale between 10ms and 1s (RT RIC is less than 10ms). This timing scale is enough to have time for training and fit the machine learning model with the dataset. The below result shows that the falsest algorithms are DT and KNN, and the lowest speed is MLPC.

Table 17. Necessary time for training each algorithm

Average time for N=100 times of training RT (t<10ms) Near-RT (10ms<t<1s) Non-RT (t>1s)			
Training time with KPI (ms)	Usable ORAN Blcok for RIC	Training time with KPI+KQI (ms)	Usable ORAN Blcok for RIC
3.39	RT	4.58	RT

8.87	RT	10.62	Near-RT
19.37	Near-RT	5.8	RT
2.07	RT	1.32	RT
1150	Non-RT	625	Near-RT

Conclusion and Future Scope

5G service requirements include low latency, high reliability, and high throughput. Appropriate services and slices are essential for 5G operators to address client demands and IoT applications besides reducing the CAPEX and OPEX. During the past years, the telecommunication industry has migrated from distributed solutions in RAN to Cloud RAN, and Open RAN. In parallel with this evolution, machine-learning algorithms have been a valuable tool for data scientists and researchers to improve network performance, reduce costs, and enhance customer experience. Network optimization, predictive maintenance, fraud detection, customer experience management, network security, and QoS optimization are well-known use cases of machine learning in 5G networks.

In this project, we combined two new concepts of AI/ML and Open-RAN to utilize machine learning in Open-RAN networks. The Open RAN alliance is gaining traction, with numerous telecommunication operators, vendors, and industry organizations backing the project. O-RAN is seen as a crucial enabler for 5G and B5G networks, which need more sophisticated and adaptable network architectures to utilize 5G technology fully. The future classification, regression and optimization problems are not solvable using the traditional methods. To acquire the precise prediction, we must inevitably deal with a massive amount of KPI and KQI data in the new networks. Here, the role of machine learning in dealing with this amount of data will be bold and is presented as a novel solution to fix new network problems efficiently.

The proposed solution of network slicing in this project is a novel one which uses a supervised-learning algorithm for 5G service and slice classification. We evaluated and compared the result of five different algorithms based on accuracy, precision, recall, F1-Mscoe and MCC. We realized that the Random Forest algorithm is the best solution among the proposed algorithms. Moreover, the demonstrated training time for each algorithm can make it a good idea to find the fastest conversion algorithm. This project will be expandable in the future by using more complicated machine learning algorithms like unsupervised and reinforced learning. Furthermore, we implemented this scenario in the Non-real time and Near-real time RIC block of Open-RAN. Implementing more advanced machine learning algorithms in Real-time RIC is an open issue and might be the subject of future research and case studies.

References

[1] Rajatheva, N. “5G Mobile and Wireless Communications Technology”; Osseiran, A., Monserrat, J.F., Marsch, P., Eds.; Cambridge University Press: Cambridge, UK, 2016; ISBN 978-1-107-13009-8.

[2] Alexander Kukushkin; “Introduction to Mobile Network Engineering: SGM, 3G-WCDMA, LTE and the Road to 5G”; Print ISBN:9781119484172 |Online ISBN:9781119484196 DOI:10.1002/9781119484196; © 2018 John Wiley & Sons Ltd

[3] IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond, ITU-R M.2083–0

[4] NR; NR and NG-RAN Overall Description; Stage 2 (Release 16); 3GPP TS 38.300 V15.6.0 (2019–06).

[5] M. Taheribakhsh, A. Jafari, M. M. Peiro, and N. Kazemifard, ‘‘5G implementation: Major issues and challenges,’’ DOI:10.1109/CSICC49403.2020.9050110; 2020 25th International Computer Conference, Computer Society of Iran (CSICC)

[6] https://www.nairaland.com/3437130/d2d-device-device-communications-4

[7] NG-RAN; Architecture description (Release 17); 3GPP TS 38.401 V0.2.0 (2022–12).

[8] Md. Farhad Hossain, Ayman Uddin Mahin, Topojit Debnath, Farjana Binte Mosharrof, Khondoker Ziaul Islam, “Recent research in cloud radio access network (C-RAN) for 5G cellular systems - A survey” Journal of Network and Computer Applications, Volume 139, 2019, Pages 31-48, ISSN 1084-8045

[9] P. Jonsson, S. Carson, J. S. Sethi, M. Arvedson, R. Svenningsson, P. Lindberg, K. Ohman, and P. Hedlund, “Ericsson Mobility Report,” Ericsson, Nov 2017.

[10] I. B. Sofi and A. Gupta, “A survey on energy-efficient 5G green network with a planned multi-tier architecture,” Journal of Network and Computer Applications, vol. 118, pp. 1–28, 2018

[11] Makhanbet, Meruyert & Zhang, Xuewei & Gao, Hui & Suraweera, Himal. (2017). “An Overview of Cloud RAN: Architecture, Issues and Future Directions”. 44-60. 10.1007/978-3-319-52171-8_3.

[12] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, “Cloud Radio Access Network: Virtualizing Wireless Access for Dense Heterogeneous Systems,” Journal of Communications and Networks, vol. 18, no. 2, pp. 135–149, Apr. 2016.

[13] M. Peng, Y. Sun, X. Li, Z. Mao and C. Wang, "Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues," in IEEE Communications Surveys & Tutorials, vol. 18, no. 3, pp. 2282-2308, third quarter 2016, doi: 10.1109/COMST.2016.2548658.

[14] Gavrilovska, Liljana & Rakovic, Valentin & Denkovski, Daniel. (2020). “From Cloud RAN to Open RAN.” Wireless Personal Communications. 113. 10.1007/s11277-020-07231-3.

[15] https://telcocloudbridge.com/blog/c-ran-vs-cloud-ran-vs-vran-vs-o-ran/

[16] Michele Polese, Leonardo Bonati, Salvatore D'Oro, Stefano Basagni, Tommaso Melodia, “Understanding O-RAN Architecture Interfaces Algorithms Security and Research Challenges”, submitted for IEEE publication, Aug 2022. https://doi.org/10.48550/arXiv.2202.01032

[17] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, “A survey of machine learning techniques applied to self-organizing cellular networks,” IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2392–2431, Fourth quarter 2017.

[18] U. Challita, H. Ryden, and H. Tullberg, “When machine learning meets wireless cellular networks: Deployment, challenges, and applications,” IEEE Communications Magazine, vol. 58, no. 6, pp. 12–18, June 2020.

[19] M. Polese, R. Jana, V. Kounev, K. Zhang, S. Deb, and M. Zorzi, “Machine Learning at the Edge: A Data-Driven Architecture With Applications to 5G Cellular Networks,” IEEE Transactions on Mobile Computing, vol. 20, no. 12, pp. 3367–3382, Dec 2021.

[20] L. Bonati, S. D'Oro, M. Polese, S. Basagni, and T. Melodia, "Intelligence and Learning in O-RAN for Data-driven NextG Cellular Networks," IEEE Communications Magazine, vol. 59, no. 10, pp. 21– 27, October 2021.

[21] O-RAN Working Group 1, "O-RAN Architecture Description 5.00," ORAN.WG1.O-RAN-Architecture-Description-v05.00 Technical Specification, July 2021.

[22] xRAN Forum, "xRAN Forum Merges With C-RAN Alliance to Form ORAN Alliance," 2018. [Online]. Available https://www.businesswire.com/news/home/20180227005673/en/

[23] G. Garcia-Aviles, A. Garcia-Saavedra, M. Gramaglia, X. Costa-Perez, P. Serrano, and A. Banchs, "Nuberu: Reliable RAN Virtualization in Shared Platforms," in Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, ser. MobiCom '21. New Orleans, Louisiana: Association for Computing Machinery, 2021, p. 749–761.

[24] J. S. Panchal, S. Subramanian, and R. Cavatur, "Enabling and Scaling of URLLC Verticals on 5G vRAN Running on COTS Hardware," IEEE Communications Magazine, vol. 59, no. 9, pp. 105–111, Sep. 2021.

[25] Aurélien Géron. (2019). "Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems (2nd ed.)". O'Reilly publication.

[26] https://learn.g2.com/cross-validation

[27] O-RAN Working Group 1, "O-RAN.WG1.Slicing-Architecture-v08.00" Technical Specification, October 2022

[28] https://rimedolabs.com/blog/network-slicing-in-o-ran/

[29] O-RAN Working Group 1, "O-RAN.WG1.Slicing-Architecture-v08.00" Technical Specification, October 2022

[30] Demestichas, P.; Tsagkaris, A.G.K.; Vassaki, K.S. Service Classification in 5G Networks. November. Seoul, Korea. 2016, p. 13. Available online: https://datatracker.ietf.org/meeting/97/materials/slides-97-nmlrg-service-classification-in-5g-networks-00 (accessed on 26 May 2021).

[31] Chen, W.; Zhao, Q.; Duan, H. Research on the Key Concepts and Problems of Service Quality. In 2nd International Conference on Mechatronics Engineering and Information Technology; Atlantis Press: Paris, France, 2017; pp. 651–654.

[32] O-RAN Working Group 2, "O-RAN AI/ML Workflow Description and Requirements 1.03," O-RAN.WG2.AIML-v01.03 Technical Specification, July 2021.

[33] Preciado, Jorge & Gonzalez-Franco, Joan & Anías Calderón, Caridad & Nieto, Juan & Rivera-Rodríguez, Raúl. (2021). 5G/B5G Service Classification Using Supervised Learning. Applied Sciences. 11. 4942. 10.3390/app11114942.