Leveraging the Power of AI/ML in 5G & Beyond 5G (B5G) Networks

By:

Om Yadav (Solutions Architect)

GLOBALLOGIC

January 2023

Contents

Executive Summary

This paper introduces the main relevant mechanisms in Artificial Intelligence (AI) and Machine Learning (ML), currently investigated and exploited for 5G and B5G networks. The study explains about the various applications of AI/ML in the telecom industry. A family of neural networks is presented which are, generally speaking, non-linear statistical data modeling and decision making tools. They are typically used to model complex relationships between input and output parameters of a system or to find patterns in data. Feed-forward neural networks, deep neural networks, recurrent neural networks, and convolutional neural networks belong to this family. Reinforcement learning is concerned about how intelligent agents must take actions in order to maximize a collective reward, e.g., to improve a property of the system. Deep reinforcement learning combines deep neural networks and has the benefit that it can operate on non-structured data. Hybrid solutions are presented such as combined analytical and machine learning modeling as well as expert knowledge aided machine learning. Finally, other specific methods are presented, such as generative adversarial networks (GANs) and unsupervised learning and clustering.

In the sequel this paper elaborates on use cases and optimisation problems that are being tackled with the help of AI/ML, partitioned in three major areas namely, Network Planning, Network Diagnostics/Insights and Network Optimisation and Control.

In Network Planning, attention is given to AI/ML assisted approaches to guide planning solutions. As B5G networks become increasingly complex and multidimensional, parallel layers of connectivity are considered a trend towards disaggregated deployments in which a base station is distributed over a set of separate physical network elements which ends up in the growing number of services and network slices that need to be operated. This climbing complexity renders traditional approaches in network planning obsolete and calls for their replacement with automated methods that can use AI/ML to guide planning decisions.

In Network Diagnostics, attention is given to the tools that can autonomously inspect the network state and trigger alarms when necessary. The contributions are divided into network characteristics forecasts solutions, precise user localizations methods, and security incident identification and forecast. The application of AI/ML methods in high-resolution synthesizing and efficient forecasting of mobile traffic; QoE inference and QoS improvement by forecasting techniques; service level agreement (SLA) prediction in multi-tenant environments; and complex event recognition and forecasting are among network characteristics forecasts methods discussed.

In regard to the Network Optimisation and Control, attention is given to the different network segments, including radio access, transport/fronthaul (FH)/backhaul (BH), virtualisation infrastructure, end-to-end 5G/B5G AI/ML for Networks, E2E network slicing, security, and application functions. Among applications of AI/ML in radio access, the slicing in multi-tenant networks, radio resource provisioning and traffic steering, user association, demand-driven

power allocation, joint MAC scheduling (across several gNBs), and propagation channel estimation and modeling are discussed. Moreover, these solutions are categorized (based on the application time-scale) into real-time, near-real-time, and non-real-time groups. On transport and FH/BH networks, AI/ML algorithms on triggering path computations, traffic management (using programmable switches), dynamic load balancing, efficient per-flow scheduling, and optimal FH/BH functional splitting are introduced. Moreover, federated learning across MEC and NFV orchestrators resource allocation for service function chaining, and dynamic resource allocation in NFV infrastructure are among introduced AI/ML applications for virtualisation infrastructure. In the context of E2E slicing, several applications such as automated E2E service assurance, resource reservation (proactively in E2E slice) and resource allocation (jointly with slice-based demand prediction), slice isolation, and slice optimisation are presented.

Introduction

The fast adaptation of 5G is promising a staggering number of new devices. According to the Cisco Annual Internet Report 2018-2023 forecasts that machine to machine (M2M) connection will increase up to 2.4 fold from 6.1 billion to 14.7 billion by the end of 2023. There will be 1.8 M2M connections for each member of the global population by 2023”. The exponential growth in connected devices along with the introduction of 5G technology is expected to cause a challenge for the efficient and reliable network resource allocation. Moreover, the massive deployment of Internet of Things and connected devices to the Internet may cause a serious risk to the network security if they are not handled properly. During the 5G era, network operators will have a chance to dynamically create and deploy different use cases or services such as massive Internet of Things (mIoT), massive Machine Type Communication (mMTC), Ultra-Reliable Low Latency Communication (URLLC), and enhanced Mobile Broadband (eMBB). This will be achieved via the concurrent support of several different logical networks (i.e., “network slices”) that will operate on top of the same physical infrastructure and will be fine-tuned to serve different requirements for different vertical sectors. To tackle this level of flexibility and network complexity, service providers should come up with solutions to ensure the security, reliability and allocation of the necessary resources to their customers in a dynamic, robust and trustworthy way. The use of Artificial Intelligence (AI) and Machine Learning (ML) as a key enabler for future networks has been recognized at European and global level.

The identified challenges and corresponding opportunities given by AI/ML will affect different network aspects, layers, and functions and even create new requirements for the architecture of future mobile networks. Despite the hype of the previous few years, the adoption of AI/ML methods in cellular networks is still at its early stages. A lot of work is still needed to identify the most suitable solutions for the dynamic network management and control via AI/ML mechanisms. Ongoing research activities need to take into consideration diverse aspects, such as the availability and usability of data sets needed for specification and testing of AI/ML solutions, regulatory aspects and practical implementation issues. The aim of this paper is to discuss at a high-level the potential applications of AI and ML mechanisms in 5G as well as Beyond 5G (B5G) and 6G networks.

AI/ML for Network Optimization

Overview

The recent paradigm shift that characterized mobile and fixed networks architectures allowed to evolve traditionally centralized and dedicated architectures to evolve into a common pool of resources, which can be dynamically orchestrated and tailored to service-specific requirements, e.g. in terms of communication latency and bandwidth. In this context, Artificial Intelligence (AI) is quickly becoming a key-feature in both network management and operational aspects of mobile networks. The wide availability of monitoring and operational data coming from heterogeneous networking domains allows gathering substantive insights on real-time networking processes. Decisions that previously took slow human interactions, based on traditional network characterization and optimization methods, can now be autonomously performed by Machine Learning (ML) algorithms with a holistic view of the network, enabling software components to directly contribute into decision-making activities related with the mobile network resource management. This not only improves the overall operational efficiency of the infrastructure, but also has a significant impact on the reduction of management and energy related costs.

Despite the general applicability of ML-based solutions, their practical application often relies on the possibility to access real-time data to perform analytics and diagnostics. Most of the solutions available nowadays derive from the combination of a few well-known frameworks. Therefore in the following, we will provide an overview of the existing and emerging ML frameworks as enablers for the adoption of Machine Learning solutions into the network management operations1 , as follows:

- Neural Networks
  - Feed-forward neural networks
  - Deep neural networks
  - Recurrent neural networks
  - Convolutional neural networks
- Reinforcement Learning
  - Basics/overview
  - Deep Reinforcement Learning
- Hybrid Solutions
  - Combined analytical and Machine Learning modeling
  - Expert knowledge aided Machine Learning
- Further Specific Methods

- Generative adversarial networks
- Kalman type filtering – and it relation to AI
- Unsupervised learning and clustering

The latest ML developments assume different neural network topologies distributed over multiple (hidden) layers. As topologies become more complex, deep-learning model benefits for training purposes from the adoption of Graphics Processing Units (GPUs) or programmable integrated circuits (FPGA) over common Central Processing Units (CPUs), mainly thanks to parallel computing platforms optimized for such intensive applications, e.g. the Nvidia CUDA package.

General machine learning tasks can be easily performed exploiting python programming language and its Scikit-learn library, which provides a wide selection of machine learning algorithms for classification, regression, clustering, dimensionality reduction, but lacks methods to develop deep or reinforcement learning tasks. However, the majority of current applications require more advanced multi-layered MLbased models combining the best characteristics of each singular framework.

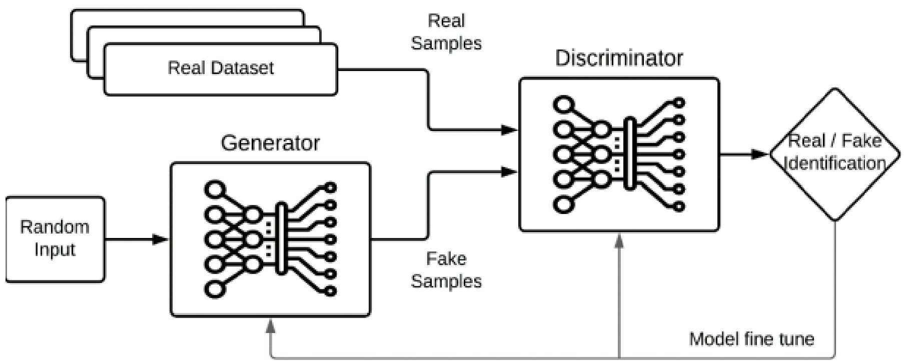
ML Techniques for Networking Problems

Generative Adversarial Networks (GANs)

Generative adversarial networks (GANs) are deep learning models whose main objective is to generate realistic new data samples with the same properties that the training data. These models use two neural networks in a combined way, one called generator and another called discriminator.

In a summarized way, the generator network is in charge of creating new data, also known as "fake samples", whose features have a distribution similar to the training data, while the discriminator network is in charge of differentiating the data generated by the generator from the real samples. When training begins, the generator produces obviously fake data, and the discriminator quickly learns to tell that it's fake, but the data generated then is perfect to fool the discriminator. In this sense, both networks "indirectly" train each other, so when the generator improves the quality of the output samples, the discriminator improves its capacity to differentiate real and fake samples, and vice versa.

The main advantage of these models lies in their ability to generate data with a quality, at a realistic level, much higher than other generative models such as Variational Autoencoders. In addition, the discriminator network can be used directly as a classifier of false and real data. Besides, these networks can handle high dimensional spaces in a much more efficient way than other methods such as Boltzmann machines. However, as drawbacks, training this type of model is often complex because of instability in the combined network training process, and it requires a high number of computational resources. In addition, false patterns can be generated when working when discrete data, such as text.



These models have been applied for network optimization from several perspectives, being its applicability highly relevant in 5G self-organizing networks. These modern networks require exhaustive labeled data to train the models in charge of automatic network management. However, low data comes already labeled from direct network monitoring, and the labeling process can be highly expensive and slow. In this context, realistic synthetic data generated using GANs is very relevant, as it can help to solve the previous issue, increasing the amount of data available at a low cost.

In this sense, GANs allow inferring fine-grained mobile traffic maps from coarse-grained, which are collected by traffic monitoring probes. As the traffic monitoring probes cannot be placed in every network point, the data that are gathered for traffic maps are not enough for the purpose of training a complex neural network. The presence of fine-grained mobile traffic maps allows the prediction of anomalous events in mobile traffic, such as network congestion, burst prediction or traffic pattern recognition. This allows the fast response of MNOs in order to enable early countermeasures, such as the reallocation of UEs in different cells, improving resource usage and planning.

Besides, QoS management has also been improved by GANs enhancing the prediction of metrics such as delay, packet loss, jitter, etc. GANs have been applied not only in traffic-related scenarios but in other key aspects in 5G networks, such as call frequency and duration forecasting. In the physical layer, Massive MIMO antenna management has also been improved using these models for channel state information generation. Similarly, cell coverage planning and performance optimization has been also enhanced using GAN models.

AI Enabled Network Tomography

Network tomography (NT) has been proposed as a methodology for the efficient inference of network performance characteristics based on measurements realized at a subset of accessible network elements. It can be broadly classified into the following three categories

1. Link-level NT that regards the estimation of per link QoS parameters (e.g., loss rates, delays, jitter) based on end-to-end path measurements
2. Path-level NT that concerns the estimation of the origin-destination traffic intensity matrix based on link-level measurements
3. Topology inference level NT for reconstructing an unknown network topology.

Compared to conventional monitoring techniques involving direct measurement of all objects of interest, NT alleviates the need for special-purpose cooperation of all devices and reduces the measurement traffic overhead. An example of employing a deep learning-based NT method for inferring a traffic matrix from the available link counts and routing information is present.

More recent works have attempted to extend the input of the employed neural networks (BPNN or convolutional neural network) by including routing information, either implicitly in the form of the Moore-Penrose pseudoinverse of the routing matrix, or explicitly using graph embedding.

Federated Learning

Federated learning is a recent addition to the distributed ML approaches, which aims at training a machine learning or deep learning algorithm across multiple local datasets, contained in decentralized edge devices or servers holding local data samples, without exchanging their data — thus addressing critical issues such as data privacy, data security, and data access rights to heterogeneous data. The Federated learning approach is in contrast to traditional centralized learning techniques where all data samples are forwarded to a centralized server and to classical distributed machine learning techniques, which assume that the local data samples are identically distributed and have the same size.

Training ML model at the network edge ensures network scalability by distributing the processing from centralized architectures of the Mobile Core/Cloud to the edge located closer to the user. This allows faster response to user requests since computations, data aggregation, and analytics are handled within user proximity. Moreover, it provides latency improvements for real-time applications as ML models are executed near the user. Many 5G applications are characterized by latency stringency and demand; therefore, the latency induced by communicating and executing ML models in the Mobile Core/Cloud may violate these requirements; hence, the edge is preferable for Mobile Network Operators (MNOs).

Unsupervised Learning Clustering Methods

Unsupervised learning and specifically clustering algorithms are considered as one of the key solutions in order to improve the performance of 5G and beyond networks. In particular, clustering algorithms discover previously undetected patterns in raw data with no pre-existing knowledge and with a minimum of human supervision and divide them in different groups that share common characteristics. More precisely, the clustering algorithms select the relevant attributes for the data analysis to identify the different points of interest and understand the groups of observations. Since the data variables usually vary in range, they need to be standardized prior to clustering. Some clustering algorithms such as K-means and Hierarchical Agglomerative Clustering use Euclidean distance to group the similar data.

One example of the use of unsupervised learning in a 5G network is to minimize the latency in communications by employing a novel fuzzy clustering algorithm. The network model is a fog model which consists of a data plane and a control plane. Within the data plane, the fog computing achieves key objectives by using novel methods such as dense geographical distribution, local resource pooling, latency reduction and backbone bandwidth savings to improve the Quality of Service (QoS). While in the control plane, interference mitigation between multiple devices is coordinated within the fog network. The fog networks are comprised of high power node (HPN) and low power node (LPN)

AI/ML Use Cases

AI/ML solutions can be applied in several network domains. In this section we provide a compilation of solutions that are designed, implemented, and tested in the context of 5G PPP projects. More specifically, the following sections describe AI/ML applications to:

- **Network Planning:** Encompassing mechanisms that can guide the planning and dimensioning decisions taken prior to network deployment.
- **Network Diagnostics and Insights:** Including mechanisms used to obtain insights that help operators run the network in a better way (e.g., data traffic forecasts, prediction of failure events etc).
- **Network Optimization and Control:** Characterized by mechanisms that use AI/ML techniques to dynamically reconfigure the network at different time scales. The solutions included in this category are further classified according to their network operational domain, i.e., RAN, transport or compute.

Network Planning

5G, beyond 5G (B5G), and 6G networks will become increasingly complex due to their multi-RAT nature, where parallel layers of connectivity are considered a trend towards disaggregated deployments in which a base station is distributed over a set of separate physical network elements, and the growing number of services and network slices that need to be operated. This growing complexity renders traditional approaches in network planning obsolete, and calls for new automated methods that can use AI/ML to guide planning decisions.

Network Element Placement Problem

Future networks will get increasingly complicated due to densification and employment of heterogeneous radio technologies. This leads to large numbers of network elements that make the network deployment very difficult. Regardless of computer aided cellular network design tools, such as 3D-map-based ray tracing tools and field-measurement-based coverage maps, one of the well-recognized problems for radio network design is the network element placement problem. These are network design aspects where ML and AI are seen as potential aids for providing the best possible solutions towards maximum coverage with minimum hardware.

ML and AI play an important role in the future network design. However, there are aspects that still require human intervention. It often happens that the desired base station location is not available due to various reasons, such as lack of electricity or property/land owner refusing to give space for the equipment, among other possible reasons. Thus, the optimal network may not be possible, but ML techniques can be utilized to take into account limitations or revised network designs and adapt a suboptimal solution that maximizes QoS criteria for any propagation environment.

Dimensioning considerations for C-RAN clusters

In C-RAN environments, it is very important to identify optimal allocation of BBU functions to the appropriate servers hosted by the CU, as it is expected to give significant efficiency gains (such as power consumption). Currently, this is performed without taking into consideration the details and specificities of the individual processing functions that BBUs entail. Given that the operation of future C-RAN networks will be supported by virtualized BBU that will operate in a combination of general and specific purpose servers, it is necessary to analyze the specificities and characteristics of the individual processing functions forming the BBU service chain.

Towards this direction, purposely developed NN models can be used to estimate the BBU processing requirement of individual LTE PHY under various wireless access requirements and traffic load scenarios. Typical examples of NN models that can predict (a) the appropriate PHY layer parameters (i.e., MCS, PRBs, CQI, etc.) and (b) the associate processing requirements of each individual BBU function (i.e. FDMA demodulation, sub-carrier demapper, equalizer and transform decoder, etc.), include the LSTM and MLP NN models.

Network Diagnostics & Insights

Traditionally, operators have relied on expert knowledge to identify problems on a running mobile network. However, the growing complexity of 5G and B5G mobile networks calls for new tools that can autonomously inspect the network state and trigger alarms when necessary. Here we will discuss about forecasting and diagnosing techniques in three specific domains:

- AI/ML techniques to forecast network characteristics and events, such as predicting traffic demands or inferring SLA violations.
- AI/ML for high precision user localization, where user location is a critical network insight in several vertical domains.
- AI/ML techniques that can be used to identify and forecast security incidents.

Forecasting Network Characteristics & Events

AI/ML techniques can be used to forecast network characteristics and events, including forecasting of traffic distributions in time and space, and forecasting of QoE levels or SLA violations. Below operations and events can be optimized using AI/ML

- Synthesizing high resolution mobile traffic
- Efficient mobile traffic forecasting
- Improving QoS & QoE Inference using forecasting techniques
- SLA prediction in multi-tenant environment
- Complex Event Recognition(CER) & forecasting

Integrating the QoE prediction model into the network management system is a prerequisite to realize QoE-driven network service optimization. The integrated management system can estimate the QoE based on network monitoring data and reconfigure networks to assure a QoE score predefined in SLA. Some work has been done to integrate QoE ML models with 5G networks, e.g., SDN, after evaluating several deep learning (DL) models, the ML QoE predictor for multimedia services is designed to consists of a DL classifier based on a combination of a convolutional neural network (CNN) and a RNN with a final Gaussian process (GP) classifier based on Laplace approximation. It not only produces a QoE score but also detects and isolates seven common anomalies that lead to the QoE score. The combination of DL and GP classifier generates the optimal performance as shown in the diagram below.

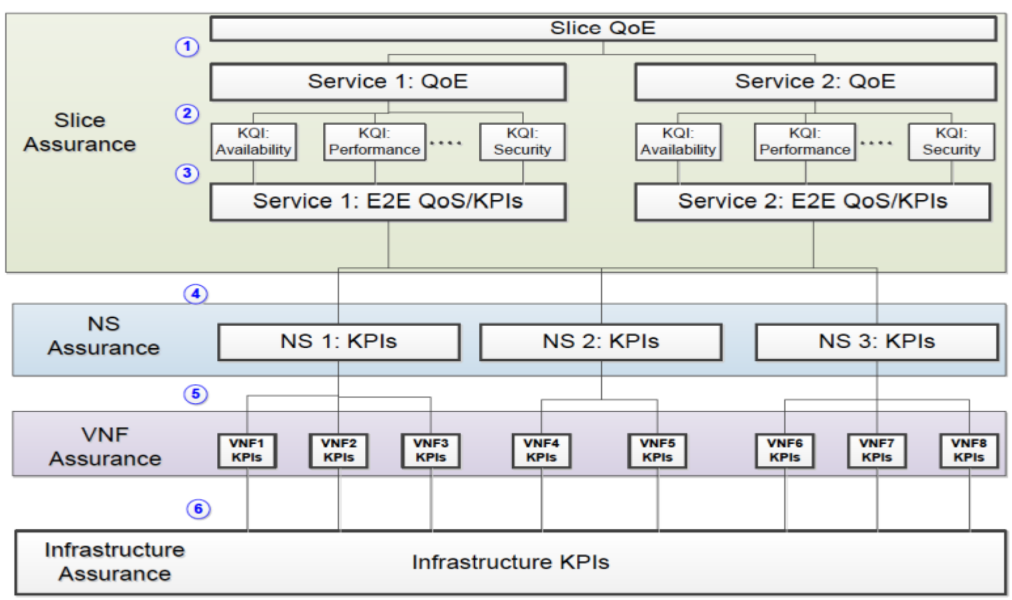


Figure2: Multi-layer QoS-QoE mapping in network slicing

The combination of DL and GP classifier generates the optimal performance. As shown in the figure below, a sequence of two CNN layers extracts new features from a 2-D time series of samples and adds these new features to create a new dimension. Then two LSTM layers are inserted to process the flattened time sequence information, creating a final embedding of the data into a 1-D vector, delivered to a fully connected final network for generating the expected prediction.

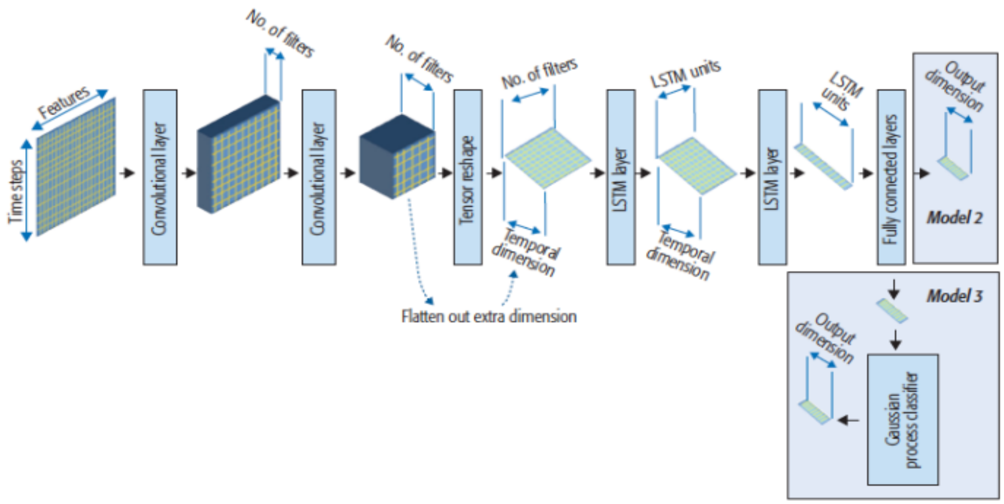


Figure3: Architecture for the best DL network to predict QoE and classify anomalies

Estimating User Locations

Accurate user positioning is of paramount importance for industry verticals such as Industry 4.0, where real-time monitoring of assets and robots is critical to the overall business efficiency. User positioning is already included in 5G standards, where it targets meter level accuracies, still far from the cm-level accuracies required in some domains. User positioning is a problem well-suited for the application of AI/ML techniques that can fuse positioning data from different technologies, or aid in determining the line-of-sight path in a multi-path propagation environment.

Below are the applications of AI/ML to positioning use cases described in this section.

- AI assisted sensor fusion
- 5G Localization based on Soft Information
- 5G Localization based on Sequential Autoencoding
- ML assisted LoS/NLoS discrimination

Forecasting Security Incidents

Future networks are incorporating new technologies, such as SDN and NFV, which however give rise to new security threats, requiring new security solutions. In this sense, the use of ML and DL techniques is gaining more importance in the last years within cybersecurity research. Modern attacks being addressed in this direction are:

Device-centered attacks: These attacks vary depending on the purpose or objective of the attacker. We have identification attacks, whose objective is to discover device hardware and software characteristics to gain information from the network environment, or uniquely identify each one of the present devices, highly affecting network privacy. There are also Bidding down attacks, which degrade performance of a device by degrading it to older networks such as 2G or 3G. Besides, some attacks are centered on Battery draining, targeting resource constrained IoT devices with the objective of making them inoperative.

Base station attacks: In this category, attacks affect the network access points, preventing service to users or enabling more advanced attacks. Examples of this type of attacks include bandwidth spoofing attack, where fake APs use the same frequencies and identifiers that a legitimate one to perform Man-in-the-middle and Eavesdropping, or Denial of Service (DoS) performed using mobile botnets or jamming techniques.

Attacks on multi-tenant network slices: In contrast to previous generations, 5G networks include multi-tenant networks addressed through network slicing. These network slices present a new attack vector to perform network attacks. DDoS flooding attacks in this scenario can cause service disruption in the entire slice, affecting even to slice-shared physical link and core network components, impacting the proper performance of other slices. DDoS attacks are already common in current networks but attacks directly related to slice management have also emerged, such as Slice-initiated attacks, which focus on the modification of the VNF/slice configuration to exhaust hardware resources, or side channel attacks, which focus on data leakage by performing information gathering from other slices running in shared hardware.

Vulnerabilities in Firmware, Software and Protocols: The explosion in the number of services offered brings with it an exponential increase in the software protocols to be developed. Thus, the detection of vulnerabilities and their correction before they are exploited is a key aspect of the networks of the future. In this area, one of the key points is the maintenance of service security over time, not leaving vulnerable versions operational.

Traditional network attacks: Network attacks common in earlier networks are still present in modern networks, and are even enhanced by the increased number of devices and improved network performance. Then, the methods for detecting and mitigating common network attacks, such as massive horizontal and vertical port scanning, botnets, service DoS/DDoS or ransomware, need to be improved according to the evolution of the attacks themselves.

Network Optimization & Control

The application of AI/ML techniques to network optimization and control is the ultimate goal behind the introduction of AI/ML in networking, where AI/ML functions act on the network, rather than only assisting in network planning or in forecasting events. AI/ML based network optimization and control is the most challenging application of AI/ML in mobile networks and is therefore widely investigated within the different communities.

In this section, we introduce AI/ML use cases for network optimization and control classified according to the network domain where they apply, namely:

1. Radio Access Networks (RANs)
2. Transport
3. NFV infrastructures
4. E2E network slices
5. Security
6. Application Function (AF) domain.

Radio Access Networks

Radio resource provisioning in a multi-technology RAN

Radio resource provisioning in a multi-technology RAN ML algorithm can be used to solve the resource provisioning issue in this kind of private network. This network can be considered as a Standalone Non-Public Network (SNPN) managed by a unique private network operator that separates his offered services into several network slices. These network slices need to be provisioned with enough radio resources in order to preserve the service quality offered by the slices. However, radio resource provisioning becomes a complex task, especially in the considered scenario in which the radio access network is composed of non-3GPP access technologies (e.g. Wi-Fi) in addition to pure 3GPP ones (5G/LTE).

As depicted in the figure below, a multi-technology radio access network represents the environment. The agent state shall be defined with some performance indicators such as the slice throughput, the cell resource allocation, the slice resource quota and other indicators that are slice-specific (e.g. delay for URLLC slices). The action triggered by the agent will be the modification of the slice resource quota, which could be to increase, decrease or maintain the quota of allocated resources.

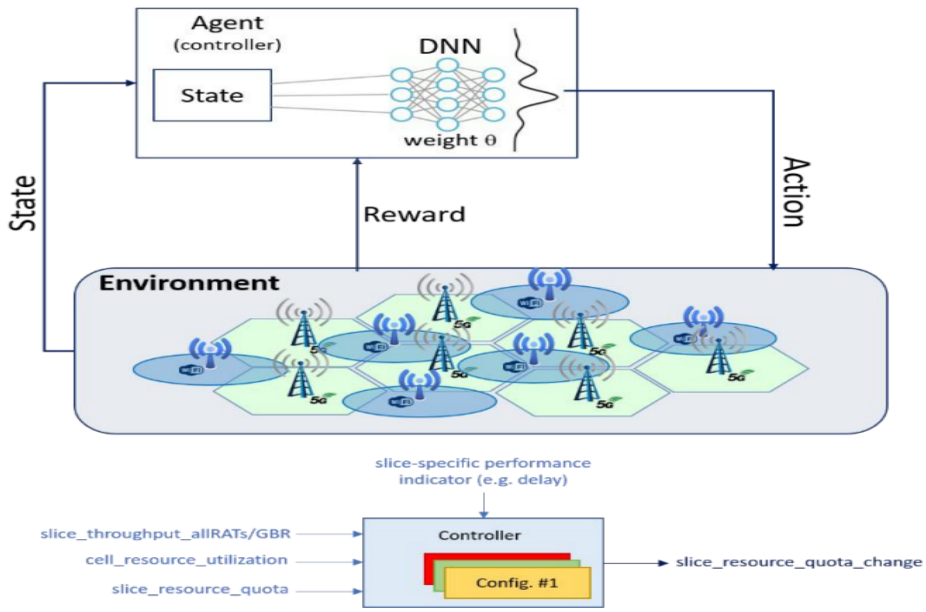


Figure4: Radio resource provisioning using Reinforcement Learning

Transport Networks Fronthaul/Backhaul

The transport network domain has been subject of intense research during the 5G development phase, where the functional split adopted in the 5G RAN architecture directly impacted the capacities that need to be provided by the transport network. This work gave rise to differentiated transport network services such as the backhaul, the fronthaul or the midhaul. Optimizing the transport network is critical in 5G and beyond 5G networks, and thus AI/ML also finds a natural application in this network segment. Below are 5 transport network related AI/ML use cases:

1. Triggering path computation based on AI/ML inputs
2. ML-based Traffic management using programmable switches
3. Dynamic Load Balancing
4. Efficient per flow scheduling in programmable transport networks
5. Determine optimal FH/BH functional split

Globallogic has inhouse AI/ML tools like **Intellisight** which can be used to train different models used in 5G or B5G networks and solve different use cases related to these networks. This will help in building different levels of capabilities in the telecommunication domain.

References

https://en.wikipedia.org/wiki/5G
https://ieeexplore.ieee.org/document/9626968
https://research.ece.ncsu.edu/ai5gchallenge/
https://www.cablelabs.com/blog/leveraging-machine-learning-and-artificial-intelligence-for-5g
https://5g-ppp.eu/
https://www.intechopen.com/chapters/77411