Chapter

Artificial Intelligence and Machine Learning in 5G and beyond: A Survey and Perspectives

Abdelfatteh Haidine, Fatima Zahra Salmam, Abdelhak Aqqal and Aziz Dahbi

Abstract

The deployment of 4G/LTE (Long Term Evolution) mobile network has solved the major challenge of high capacities, to build real broadband mobile Internet. This was possible mainly through very strong physical layer and flexible network architecture. However, the bandwidth hungry services have been developed in unprecedented way, such as virtual reality (VR), augmented reality (AR), etc. Furthermore, mobile networks are facing other new services with extremely demand of higher reliability and almost zero-latency performance, like vehicle communications or Internet-of-Vehicles (IoV). Using new radio interface based on massive MIMO, 5G has overcame some of these challenges. In addition, the adoption of software defend networks (SDN) and network function virtualization (NFV) has added a higher degree of flexibility allowing the operators to support very demanding services from different vertical markets. However, network operators are forced to consider a higher level of intelligence in their networks, in order to deeply and accurately learn the operating environment and users behaviors and needs. It is also important to forecast their evolution to build a pro-actively and efficiently (self-) updatable network. In this chapter, we describe the role of artificial intelligence and machine learning in 5G and beyond, to build cost-effective and adaptable performing next generation mobile network. Some practical use cases of AI/ML in network life cycle are discussed.

Keywords: Next Generation mobile Networks, 5G, Artificial Intelligence, Machine Learning, Deep Learning, Physical Layer, Big Data, Network Control

1. Introduction

The massive deployment of LTE (Long Term Evolution) or 4G mobile network has solved one of the major challenges of wireless communications, which is high capacities, to build real broadband mobile Internet. This was possible mainly through very strong physical layer, based on orthogonal frequency division multiplexing (OFDM) and multiple input multiple output (MIMO) among others, and flexible network architecture. However, new bandwidth-hungry services have been developed in unprecedented way, reaching capacities up to 1 Gbps, such as virtual reality (VR), augmented reality (AR), etc. Furthermore, mobile networks are facing other new services with extremely demand of higher reliability and almost zero-latency performance, like vehicle communications or Internet-of-Vehicles (IoV).

The 5G systems solved the major problems related to the capacity through use of new radio interface, massive MIMO, beamforming, high modulation orders, etc. Furthermore, 5G is planned to include a high level of flexibility to optimize the network utilization by integrating software defined networking (SDN) and network function virtualization (NFV) technologies. This should allow the network operators to support current and new more demanding future services. The main challenge is to be ready to support services for customers in completely different vertical markets/industries, like e-health, Internet-of-Vehicles (IoV), Industry 4.0, smart grids, etc. Furthermore, the network operators have to establish more partnerships on multiple layers for sharing of the 5G infrastructure through network sharing relationship among different mobile operators, delivery of Infrastructure as a Service, Platform as a Service or Network as a Service by assets providers. Facilitating such partnerships may act as catalyst for the deployment of 5G networks, considering the large investments for mobile network operators (MNOs) in capital expenditure CAPEX and operational expenditure OPEX are still not being followed by significant revenue increase, [1]. Network operators are also forced to consider a higher level of intelligence in their networks, in order to deeply and accurately learn the operating environment and users behaviors and needs. The adoption of Artificial Intelligence (AI), and Machine Learning (ML) approaches as core part of AI, is crucial to forecast the evolution of the environment and users/services behavior/demand to build a pro-actively and efficiently (self-) optimizing and (self-) updating networks. This is true for each layer of the system and each level of the network. For example, AI/ML are crucial for massive MIMO to identify dynamic change and forecast the user distribution by analyzing historical data, dynamically optimize the weights of antenna elements using the historical data or to improve the coverage in a multi-cell scenario considering the inter-site interference between multiple 5G massive MIMO cell sites, etc.

In this chapter, we describe the role and the integration method of AI and different ML approaches as core part of AI in the next generation mobile networks. The rest of the chapter is built by giving a short overview on AI and ML definitions, historic and (sub-) classes in second section. The third section shows the role of big data as prerequisite for a full exploitation of AI/ML advantages. The components of 5G are described in fourth section and how to make AI/ML a main component of next generation mobile networks. Practical use cases are illustrated and discussed in fifth section.

2. AI and ML in mobile communications networks

2.1 Overview on AI and ML

2.1.1 What is AI?

AI is the scientific field that deals with programming machines to mimic human behavior in solving tasks that humans are good at (natural language, speech, image recognition, etc.). AI involves the intersection of many fields of computer science and applied mathematics. The position of artificial intelligence is rather to consider that we, as human beings, have an intuitive understanding of what intelligence is and therefore we can judge whether a machine is intelligent or not. This operational definition of AI was promoted by Alan Turing in 1950, who introduced his famous "Turing test". The Turing test is an operational test; according to which a machine is considered intelligent if it can converse in such a way that (human) interrogators cannot distinguish it from a human being [2].

Initial efforts at AI involved modeling the biological neurons in the brain. In 1943 McCulloch and Pitts [3] modeled for the first time the artificial neural as a binary variable that is switched to either on or off. Later in 1949, Donald Hebb developed an algorithm for learning neural networks. In 1951, Marvin Minsky and Dean Edmonds built the Stochastic Neural Analog Reinforcement Calculator (SNARC), the first neural network computer. Following this accomplishment, a small group of scientists interested in the study of intelligence met in a 2-month workshop at Dartmouth University in 1956. According to common belief, the term AI was first introduced and defined by John McCarthy at this workshop, as:" AI involves machines that can perform tasks that are characteristic of human intelligence".

Over the last few decades, AI has gained increasing interest among researchers and industry. This is due to the wide variety of applications in which AI has been used, such for example, natural language processing (e.g. broadcast news transcription, speech-to-speech translation), healthcare (e.g. assisting in surgeries, computer aided diagnosis), smart cars and drones (e.g. self-driving cars, obstacle detection) and also mobile networks (e.g. performance optimization, traffic prediction).

2.1.2 The connection between machine learning, deep learning and AI

Today AI is a collection of different technologies working together to enable machines to sense, comprehend, act, and learn with human-like levels of intelligence. Rule-based techniques as well as expert system are the first approaches to AI. Technologies such as ML, Deep Learning, and Big Data are all part of the AI landscape, as illustrated in **Figure 1**. As the most recent advances in AI have been in the field of machine learning, people often mistakenly conflate AI with ML. Eventually, the field of MML was created from the desire to design AI with the ability to learn and acquire knowledge.

ML is a subset of AI, which aims to give the ability for a computer to perform tasks without being given explicit instructions on how to solve it. It is a paradigm that aims to build a computer that can learn, just like humans do. The learning

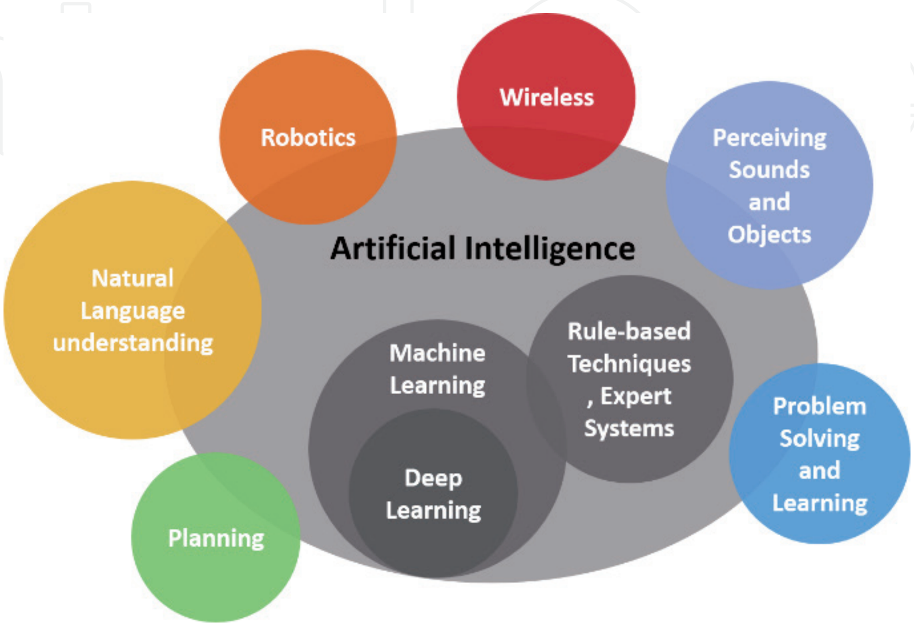

Figure 1.
Connection and overlap between machine learning, deep learning, and artificial intelligence.

process consist of providing a ML algorithm with examples of the task we want to solve (data), and letting the computer finding patterns and making inferences that optimizes the decision making according to a user-defined objective. In general, ML could be used to accomplish different type of tasks including classification, clustering and making predictions about data.

Artificial Neural Networks (ANN), referred also as Neural Networks (NN), are a popular machine learning models inspired by the biological processes of the brain. The first NN algorithm is the Perceptron developed by Rosenblatt in 1958; [4]. This finding was inspired by McCulloch mathematical models of neurons in the human brain [3]. In the following decades, different types and architectures of neural networks have been proposed as well as algorithms to train them effectively. Following these accomplishments, the term Deep Learning was introduced to the ML community. Around the year 2000, Deep Learning is a subcategory of ML focused on parameterizing multilayer (deep) neural networks that can learn representations of the data.

In recent years, deep learning based methods has gained increasing interest. This is due to their high performances in image classification [5], speech recognition [6] and natural language processing tasks [7]. In fact, deep learning techniques have largely outperformed stat-of-the-arts results in these tasks. However, deep learning, and indeed most of ML techniques, have several limitations. The first limitation is the amount of data required during training in order to achieve human like performances. Another limitation is the computational capacities required to train deep learning based models on large datasets.

2.1.3 The categories of machine learning

As pointed out in the previous paragraphs, ML is a complex landscape. Based on the training strategy, ML can be divided into three classical categories; which different learning approaches are illustrated in **Figure 2**:

Supervised Learning: Learning with a labeled training set. This category includes Classification and Regression tasks.

Unsupervised Learning: The process of training a model using training data that is unlabeled. The model had to discover patterns in unlabeled data. The widely used task in unsupervised learning is Clustering.

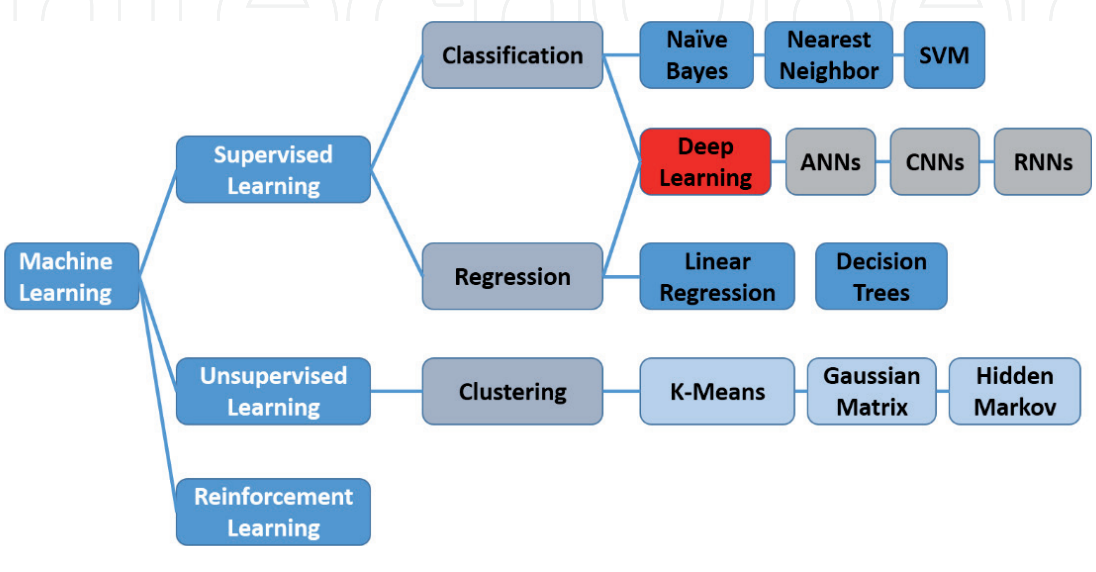

Figure 2.
Classification of different ML approaches.

Reinforcement Learning: The process of training a model on a series of actions that lead to a particular outcome, where the system receives rewards for performing well and punishments for performing poorly directly from its environment. Reinforcement Learning is used in robotics and games.

2.2 Introducing AI and ML in mobile communications

2.2.1 Needs for intelligence in mobile networks prior to 5G

As one of the worldwide leading mobile systems manufacturers, Ericsson has led a study to analyze the state-of-the-art and the expectation of adopting AI by the mobile network operators and global service providers. The study found that operators have mainly adopted AI as means to enable them to switch to 5G and to guarantee optimized investment; [8]. Furthermore, already with 4G/4.5G there has been an increased complexity in the management of a vast number of devices and huge amount of data. Operators are hoping that AI and ML will help to reduce this complexity. Other main findings of the study are as follows; [8]:

- AI is already being incorporated into networks, with a primary focus on reducing capital expenditure, optimizing network performance and building new revenue streams. Operators from all over the world are already reaping the benefits of integrating AI into their networks. More than half of service providers (53 percent) expect to have fully integrated some aspects of AI into their networks by the end of 2020.

- AI will be vital for improving customer service and enhancing customer experience, generally referred as "Quality of Experience (QoE)" to. AI is expected to help providers further improve customer experience in many ways, including improving network quality and providing personalized services.

- AI will help recoup the investments communications service providers (CSPs) are making in their networks to switch to 5G. Lowering operational costs and ensuring returns on network investments are key priorities that service providers are looking to achieve using Artificial Intelligence. **Figure 3** shows the prioritized domains for the integration of AI to optimize the costs as well as the management of the always-increasing network complexity. Network intelligence and automation are crucial to the evolution of 5G, IoT and industrial digitalization. As 5G-enabled technologies develop; operators will need to increase their network capacity. However, the added capacity brings additional complexity.

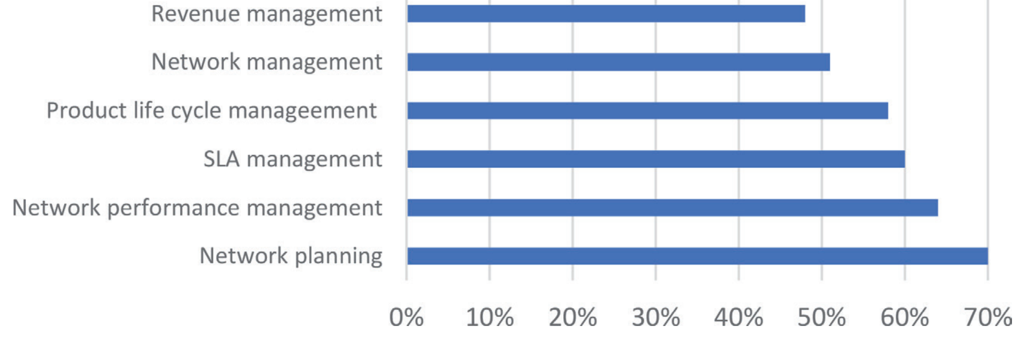

Figure 3.
Core areas with the highest potential for returns on AI investments [8].

- Adopting AI is creating new data challenges, even as it solves network complexities. Network providers agree that they need to develop effective mechanisms for collecting, structuring and analyzing the huge volumes of data that AI is capable of amassing.

2.2.2 Advantages of using AI in mobile networks

The history of mobile communication is evolving from one generation to a next one over the last three decades. Nowadays, before to jump to a next generation, i.e. 5G, the mobile network operator need to understand deeply the current situations and future scenarios. The MNO needs to learn more about the real behavior of the subscribers, their profiles, the traffic patterns, wished and possibly adequate services for each profile and how all these will evolve in the future. In the communications market, whether it is network operators, equipment manufacturers or solution providers, etc., the market players hope to take advantage of AI to assist in areas that become very challenging, such as in designing, operating, maintaining and managing communication networks and services. Mainly AI will offer the MNO the ability of learning more about their network and the need of customers, ability of understanding and reasoning to make the ideal decision/actions for different scenarios and environment conditions and finally the ability to collaborate between high heterogeneous widely expanded and densified network infrastructures.

Operators need intelligent decisions to manage complex resources and dynamic traffic. However, so far no one single model has the ability to model accurately the network traffic characteristics. Fortunately, AI has entered into the cognitive age, and deep learning can be used. Through deep learning, the machine system can use the existing training data to process large amounts of data through data mining. AI can also learn the characteristics of data traffic, management, controls and other characteristics automatically and master expert experience of operating, managing and maintaining networks. By these efforts, the accuracy of analysis can be enhanced, and the intelligent management and services of communication networks can be realized. Detailed description can be found in [9].

Due to the high dynamics of the network system, the state information of a resource may have changed when it is transmitted to the network management system. Therefore, the network management can only know the local state information without the knowledge about the system internal state. ML has the strength to deal with this kind of fuzzy logic and uncertainty reasoning. In order to make the classification or prediction of the state easier, deep learning constructs a multi-hidden layer model and uses the hierarchical network structure to transform the feature representation of the sample into a new feature space layer by layer, as detailed in the first sections of the chapter. The major advantage is the fact that AI does not need to describe the mathematical model of the system accurately, and therefore has the ability to deal with uncertainty or even 'unknowability.

Due to the expansion and high densification of the network infrastructure, both in scale and size, the structure complexity of communication networks, and especially for the next generation, are increasing quickly. Concepts such as distribution and hierarchy are often talked about in the network management. Management tasks and controls are distributed to the entire network, in order to avoid centralization of the management functions that requires more data overhead and reliability issues. As a result, network operators have to deal with issues such as tasks' distribution, communication and collaboration between management nodes. The introduction of the multi-agent collaboration of distributed AI into the network management will support the ability to collaborate between network managers distributed in every layer. However, such collaboration requires a high

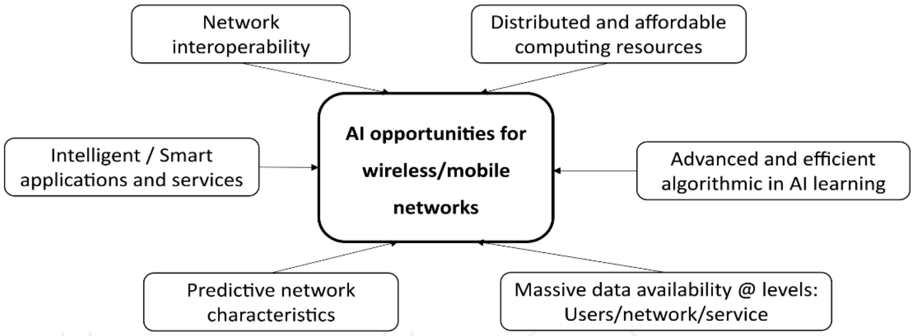

Figure 4.
Factors contributing to a full and efficient AI integration into mobile networks.

level of interoperability between the heterogeneous networks building the entire communications infrastructure. Furthermore, this network interoperability is just of the many other factors that must be fulfilled by the environment, in order to take benefits of the above-cited advantages of AI. The other factors allowing taking full potential from AI are depicted in **Figure 4**.

One of the major enablers of a profitable AI integration is the availability of high performing and efficient computing resources. Indeed, 5G systems seek to provide high throughput and ultra-low latency communication services, to improve users QoE. Implementing deep learning to build intelligence into 5G systems, to meet these objectives is expensive. This is because powerful hardware and software is required to support training and inference in complex settings. Several tools are emerging, which make deep learning in mobile networks tangible as discussed in [10]. Authors discussed a hierarchy of advanced computing tools as deep learning integration enablers; namely: (i) advanced parallel computing, (ii) distributed ML systems, (iii) dedicated deep learning libraries, (iv) fast optimization algorithms, and (v) fog computing.

3. Big data as prerequisite for integrating AI in mobile networks

The integration of intelligent algorithms and learning approaches requires the availability of big data sets, which represent the starting point. However, because the AI and ML can be integrated at different levels of the next generation mobile work, or at least in 5G, different big data sources can be explored and must be selected carefully to be able to extract as much as possible of knowledge/learning. Different ways are proposed in the literature to classify the needed types of big data sets, which should build the basis for the AI and ML, such the ones proposed in [10–12]. For example, a classification of different big data sources is illustrated in **Figure 5**, which contains three pools of big data sets: general wireless data, social network-aware data, social data and cloud data [11].

Wireless data: This class represents the big data generated by wireless users, and which contains useful information about the activity patterns in time, frequency and space (locations). For example, this can help to infer from the data traffic/demand variation over time, the interference power at different frequencies, and the congestion level distribution at different locations, etc. The exploitation of these spectral patterns will allow an efficient management of the wireless resources, at radio resource management/allocation (RRM/RRA) functionalities, for improving the systems spectral efficiency and for enhancing the delivered quality of service experienced by the end-user. One of the possible intelligent applications can be the load balancing relying on proactive RRA. In such context, the network operator

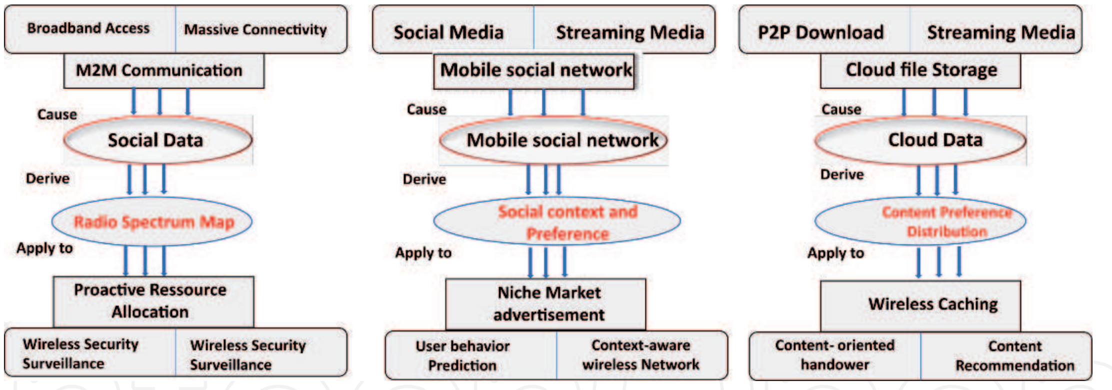

Figure 5.
Classes of big data in mobile network and their applications [11].

can adjust the transmit power, frequency or direction, through the beamforming mechanism adopted for 5G or simply through sectorized antennas, of the different base station transmitters relaying on the mobile users distributions. In addition, the operator can dispatch mobile base stations in advance when anticipating a regional surge of data traffic, which could be caused for example by sport events, music concerts, etc. In [10], authors subdivided the mobile big data into classes: Network level-data, which is similar to "wireless data", and App-level data. It is worth to notice that network level-data are further subdivided into sub-classes, namely:

- Infrastructure data: Infrastructure locations, capability, equipment holders, etc.
- Key performance indicator (KPI) data: Data traffic, data throughput, end-to-end delay, QoE, jitter, bit/packet error rates, etc.
- Call details records (CDR) data: Session start and end times (i.e. inter-arrival time, holding time), type, sender and receiver, etc.
- Radio information data: Signal power, frequency, spectrum, modulation, serving BS, etc.

Social data: Nowadays, the main cause of the soaring data volume in the Internet is the online social networks. The penetration of the mobile Internet into our daily lives and behaviors makes convenient multimedia communications easily accessible for everyone, independently of the age or social level. The volume of social data, which represents the data exchanged on social media, has reached an unprecedented astonishing magnitude and it is expected to continue an explosive increase in the next years. On one hand, its strong ties to public events in the physical world feature social network data. So that, an important football game or political event may inspire heated online discussions that may last for day in case of games and some weeks or months in case of an elections year. On the other hand, social network data contains rich information about the context/preferences of individuals or social groups. The mobile network operator can exploit such data to build a social-network-aware wireless concept. For example, operators may offer more bandwidth in touristic sites so that users/tourists can share their tourist-experience on the social media with highly satisfying quality of service even if they have only paid pre-paid service. Example of connecting the social media big data source to a AI learning framework is illustrated in a simple way in **Figure 6**, showing the most promising expected AI applications, like mobile caching, drone-based mobile BS

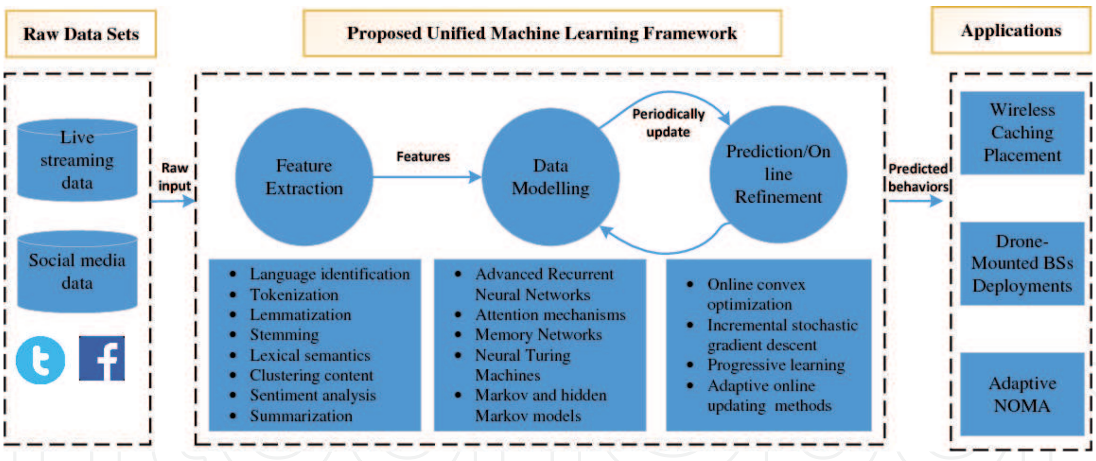

Figure 6.
Example of connecting big data source to learning framework before feeding the predicted behavior models to the AI applications layer [11].

placement or radio resource allocation in case of adaptive non-orthogonal multiple access (NOMA).

Cloud data: one of the major cause for the apparition of the data tsunami in mobile network is the transmission of the multimedia contents stored in the cloud servers. Furthermore, it was found that most contents transferred over Internet are based on its popularity. It is possible therefore to reduce the tele-traffic of the system by exploiting the users' preference for special cloud contents. For example, the most popular videos, for a specific geographic region, can pre-cached at the edge servers so that no real-time backhaul data downloading is needed for frequent request. In addition, individual user's preferences for specific multimedia contents can be used to predict the user's future content demand, based on which the network operator can perform a pre-feeding or recommendation actions.

4. AI and ML as main components for 5G

4.1 Main goals and components of 5G

While the evolution toward 4G/LTE was driven by more mobile data speed, the 5G system is confronted by more stricter and diverse requirements; as summarized in **Table 1**. The deployment of 5G systems seeks to provide high throughput and ultra-low communications latencies, to improve users' quality of experience (QoE). To meet these requirements, 5G targets three evolution axes to cope with the new applications fields; such autonomous cars/driving, industrial automation/smart manufacturing (Industrie 4.0), virtual reality, e-health, etc. These axes are:

Enhanced mobile broadband (eMBB): Allows for new bandwidth-hungry applications with extreme high data rate demands over a uniform coverage area. Examples include ultra-high-definition video streaming and virtual/augmented reality (VR/AR).

Massive machine-type communications (mMTC): A key characteristic of 5G communication services is the scalable connectivity demand for expanding the number of wireless devices with efficient transmission of small amounts of data over extended coverage areas. Applications like body-area networks, smart homes, IoT, and drone delivery will generate this type of traffic. mMTC must be able to support massive new uses and others uses that would appear in the future.

Ultra-reliable low-latency communications (URLLC): Connected healthcare, remote surgery, mission-critical applications, autonomous driving, unmanned aerial vehicles (UAV), vehicle-to-vehicle (V2V) communications, high-speed train

Requirement	Desired value	Application example
Data rate	1 to10Gbps	Virtual reality office
Data volume	9GB/h (busy period) 500GB/month/user	Stadium, Dense urban, information society
Latency	Less than 5 ms	Traffic efficiency and safety
Battery Life	One decade	Massive deployment of sensors and actuators
Connected devices	300,000 devices/AP	Massive deployment of sensors (massive IoT)
Reliability	99.999%	Teleprotection in smart grid network, Traffic efficiency and safety

Table 1.
Summary of major requirements for 5G mobile system [13].

connectivity, and smart industry applications. These applications prioritize high reliability, low latency, and mobility over data rates.

To achieve the above listed objectives, 5G systems use a very large number different advanced technologies or enablers; such as new PHY layer, new MAC layer, SDN, NFV, etc. To have a structured view of these different enablers, **Figure 7** classifies them into two main categories; the system-level and network-level technologies. More details about of the listed technologies and paradigms can be found in the survey [14]. If we consider the breakthrough of the unprecedented achieved capacities, three enabling technologies can be cited, namely: massive MIMO that allows a new level of bit rates, use of millimeter-wave that finally give a hope to overcome the spectrum scarcity and network densification.

The introduction of MIMO in 4G/LTE was decisive in achieving real broadband speeds for mobile Internet. With the same philosophy, the massive MIMO is bringing the next mobile network to the next level of throughputs up to Gbps. Therefore, massive MIMO is considered as a leading candidate technology for 5G, where the high number of antennas at the base station requires large number of power amplifiers. According to [15], the primary problem with power amplifiers is known as the trade-off between linearity and efficiency: amplifiers can be designed to attain good linearity at the cost of efficiency. As the highly linear power amplifiers are

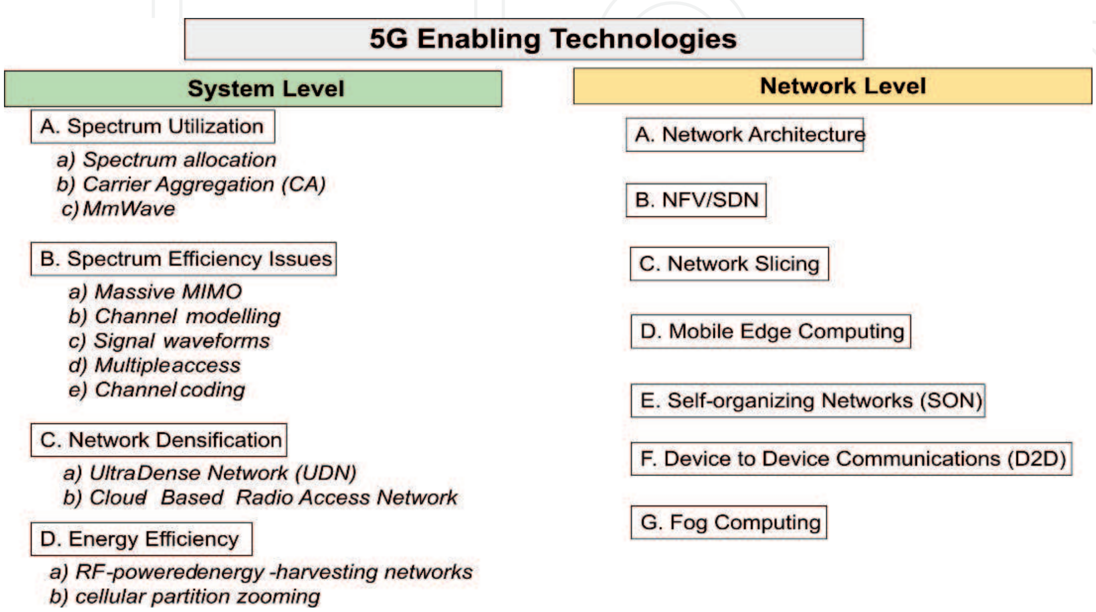

Figure 7.
Classification of major enabling technologies of 5G systems and networks.

expensive and power inefficient, the excessive use number of antennas at the base station makes the use of inexpensive elements strongly targeted to keep the overall network costs, capital expenditure and operational expenditure, manageable. However, emerging energy and spectrum-efficient wideband wireless communications systems are vulnerable to non-linear distortions that are attributed to the radio frequency front-ends. For example, those types of high-power amplifier affect the performance of the intended receiver and thus the entire network.

Furthermore, because the 5G systems should operate in more dynamic environments and in different bands (extending from cm-wave to mm-waves), the dynamic range requirements will likely become more demanding. Therefore, the power amplifiers need to meet stricter linearity specifications while at the same time maintaining an acceptable efficiency for the overall systems. The aimed high efficiency of power amplifiers is achievable when constantly feeding the amplifier at the limit of its high-power linear zone. However, this is not realistic for 5G base station, because this is not feasible solution for the peak-to-average power ratio.

The previously cited 5G enablers have a direct contribution in the network performance; however, an operative and efficient 5G network cannot be complete without AI. For example, 5G enables simultaneous connections to multiple IoT devices, generating massive amounts of data that must be processed using ML and AI. When ML and AI are integrated within, wireless providers can, for example [16]:

- Identify dynamic change and forecast the user distribution by analyzing historical data,
- Forecast the peak traffic, resource utilization and application types; and optimize and fine tune network parameters for capacity expansion,
- Eliminate coverage holes by measuring the interference and using the inter-site distance information,
- High level of automation from the distributed ML and AI architecture at the network edge,
- Application-based traffic steering and aggregation across heterogeneous access networks,
- Dynamic network slicing to address varied use cases with different QoS requirements,
- ML/AI-as-a-service offering for end users, etc.

4.2 Machine learning/deep learning in 5G

With the increasing advances and advantages of ML in the wireless communications, each research community has tried to evaluate the impact of ML on 5G in its discipline. As result, we have several publication for ML impact on physical layer, security aspects, radio resource managements, etc. This makes it very difficult to give short overview on the utilization and the impact of AI/ML in 5G. Therefore, we can summarize the works applying ML to 5G according to two principles: a "general ML categorization", where we consider all possible ML approaches from the literature, and a "Deep Learning-based Categorization", which focuses only the deep learning, because several leading publications consider deep learning as the most promising approach of ML for the high complexity of 5G.

Learning classes	Learning models	Example of applications in 5G
Supervised learning	ML and statistical logistic regression techniques.	Dynamic frequency and bandwidth allocation in self-organized LTE dense small cell deployments
	Support Vector Machines (SVM)	Path loss prediction model for urban environments
	Neural-Network-based approximation	Channel Learning to infer unobservable channel state information (CSI) from an observable channel
	Supervised ML Frameworks	Adjustment of the TDD Uplink-Downlink configuration in XG-PON-LTE Systems to maximize the network performance based on the ongoing traffic conditions in the hybrid optical-wireless network
	Artificial Neural Networks (ANN), and Multi-Layer Perceptrons (MLPs).	Modeling and approximations of objective functions for link budget and propagation loss for next-generation wireless networks
Unsupervised Learning	K-means clustering, Gaussian Mixture Model (GMM), and Expectation Maximization (EM).	Cooperative spectrum sensing and Relay node selection in vehicular networks.
	Hierarchical Clustering.	Anomaly/Fault/Intrusion detection in mobile wireless networks
	Unsupervised Soft-Clustering ML Framework.	Latency reduction by clustering fog nodes to automatically decide which low power node (LPN) is upgraded to a high power node (HPN) in heterogeneous cellular networks.
	Affinity Propagation Clustering.	Data-Driven Resource Management for Ultra-Dense Small Cells
Reinforcement Learning	Reinforcement Learning algorithm based on long short-term memory (RL-LSTM) cells.	Proactive resource allocation in LTE-U Networks, formulated as a non-cooperative game, which enables SBSs to learn which unlicensed channel, given the long-term WLAN activity in the channels and LTE-U traffic loads.
	Gradient follower (GF), the modified Roth-Erev (MRE), and the modified Bush and Mosteller (MBM).	Enable Femto-Cells (FCs) to autonomously and opportunistically sense the radio environment and tune their parameters in HetNets, to reduce intra/inter-tier interference.
	Reinforcement Learning with Network assisted feedback.	Heterogeneous Radio Access Technologies (RATs) selection.

Table 2.
Learning approaches and their 5G applications for the three ML classes.

A general ML categorization in case of 5G follows the general structure of ML as seen in the first sections, which uses three classes of ML: supervised learning, unsupervised learning and reinforcement learning. **Table 2** shows an example of such classification giving the used learning approaches from each class and a concrete example of application in 5G, [17]. In the next section, some 5G use cases will be described and solution for AI/ML integration in mobile network operators will be proposed.

Some research works focus only on the deep learning, because it is considered as most powerful learning approach of AI/ML. This gives a very large spectrum of applications in 5G and its different aspects, as detailed in the most recent surveys from examples [10, 18]. **Figure 8** shows just a small part of possible applications in

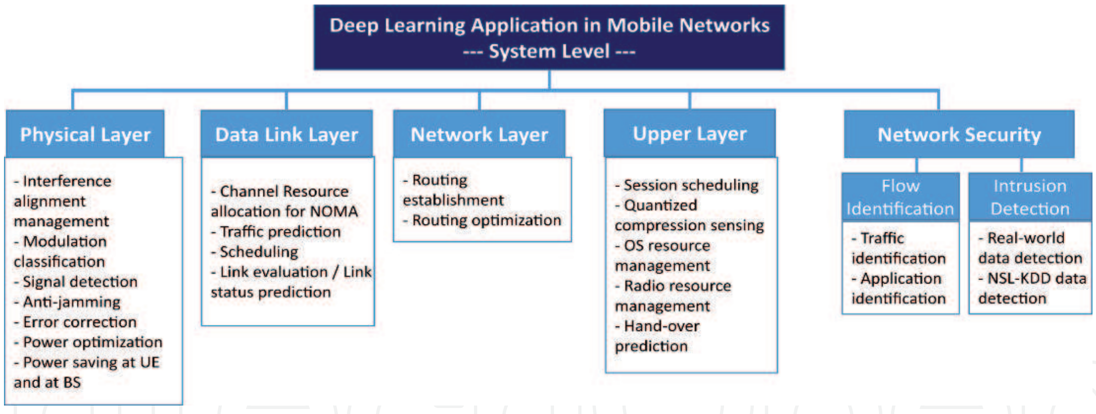

Figure 8.
Applications of deep learning in different layers of 5G systems (extended version of [18]).

5G systems and classify them according to each system layer where AI/ML is integrated. Furthermore, some works have focused on of the application of some very promising sub-variants of deep learning, like the application of Deep Q-learning for caching/offloading, network security and connectivity preservation, traffic engineering/resource scheduling, [19]. Deep learning presents several strengths to cope with the challenges in wireless communications, and especially in case of 5G, which are explained in detail in [10] and can be summarized as follows:

- Feature extraction: Deep neural networks can automatically extract high-level features through layers of different depths. This allows reducing the expensive hand-crafted feature engineering in processing heterogeneous and noisy mobile big data.

- Big data exploitation: Unlike traditional ML tools, the performance of deep learning usually grow significantly with the size of training data. Therefore, it can efficiently utilize huge amounts of mobile data generated at high rates.

- Unsupervised learning: Deep learning is effective in processing un−/semi-labeled data, enabling unsupervised learning. This is very important in handling large amounts of unlabeled data, which are common in mobile system.

- Multi-task learning: Features learned by neural networks through hidden layers can be applied to different tasks by transfer learning. This reduces computational and memory requirements when performing multi-task learning in mobile systems.

- Geometric mobile data learning: Dedicated deep learning architectures exist to model geometric mobile data, which revolutionize geometric mobile data analysis.

5. Use cases: network planning, optimisation and management

5.1 AI in network life cycle: from planning to control and management

The life cycle of a communication network starts with its planning, dimensioning and deployment in a first life phase. Here, the network operator aims an investment optimization, minimal capital expenditure (CAPEX), while respecting the design

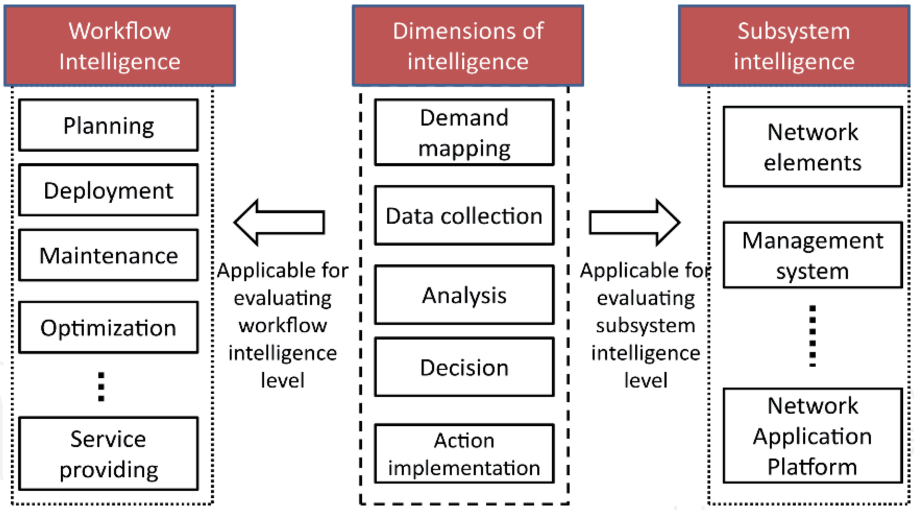

Figure 9.
Three pillars for integrating AI in network life cycle, by ITU-T [20].

requirements, especially the quality of service delivered or experimented by the end-users. In the second phase of the network life, a continuous control and management should guarantee a continuity of the service in a certain quality and a reliability of the services. In addition, network optimization must assist the management in order to keep the quality of service when necessary through upgrade the network hardware/software components to cope with the changes in the operating environment. Such changes can be in form of increase of the subscribers' number along the years or the apparition of new services with high demand of capacity, etc. In order to integrate a certain level of artificial intelligence in the above cited workflow processes (planning, dimensioning, etc.), **Figure 9** elaborated by the ITU-T illustrates the different intelligence pillars needed in the network intelligence landscape 19]. Therefore, the intelligence in the workflow requires different big data sets, like the demand mapping (and/or its forecasting for the coming years), a continuous collection of large data volume for the optimization tasks during the entire life cycle of the network. Moreover, the execution of such AI decisions and outputs requires intelligent sub-systems, which are able to interpret and to learn from the collected and analyzed big data streams.

The ideal case is to reach ZSM (Zero-touch network & service management), which is based on self-optimisation and self-healing network at different level of complexity either in the optimisation or in control and management. ITU-T gives three most desired cases, namely:

- Radio resource management for network slicing: Providing performance guarantee with high reliability, while ensuring efficient utilization of radio resources. For this purpose, the network should support the continuous collection of data, analysis of network slice behaviour and resource utilization patterns.

- End-to-end network service design automation: Automatically translating service requirements of application services to network parameters/requirements. Here also the big data sets are necessary. Therefore, network has to support data models to specify service requirements, integrate automated network configuration methods.

- End-to-end fault detection and recovery: Predictive detection and root cause analysis, and automated recovery decision making. This requires the collection of performance data on real-time basis, as well as generation of training data using testing environments.

Network Intelligence Level		Dimension of Intelligence				
		Action Implementation	Data Collection	Analysis	Decision	Demand Mapping
L0	Manual Operation	Human	Human	Human	Human	Human
L1	Assisted Operation	Human & System	Human & System	Human	Human	Human
L2	Preliminary Intelligence	System	Human & System	Human & System	Human	Human
L3	Intermediate Intelligence	System	System	Human & System	Human & System	Human
L4	Advanced Intelligence	System	System	System	System	Human & System
L5	Full Intelligence	System	System	System	System	System

Table 3.
Five possible degrees of intelligence in next generation networks [20].

However, it is clear that the migration toward a full intelligent network will not be an easy and one-dimensional task. Therefore, ITU-U has defined different degrees or levels of network intelligence according to five dimension, as listed in **Table 3** [20]. It is up to the network operator to define its own roadmap by prioritizing its objective concerning the investment, introduction of new services, etc. Nevertheless, in order to take a full benefit of the AI the operator should reach the fifth level “L5: Full Intelligence” in all the five dimensions. In addition, from telco's perspective the migration has to be in coherence with the business model as well as the return on investments.

5.2 AI/ML revolutionizing the planning and optimization process

One of the most critical processes to determine the final performance of a mobile network, and about its success in technical as well as financial aspects, is the initial planning process. Because the initial planning will determine also the way of functioning, operating, control and management processes, a bad-dimensioned network will always require more interventions from the control and management teams to try to bring the network performance to an acceptable level. In this process, decisions must be made about infrastructure (node deployment), spectrum, parameters and configuration setting procedures, energy consumption, network capacities to serve the worst-cases (peak or busy-hours traffic), evolution of the bandwidth demand over the years, etc. Furthermore, the planning and deployment of the next generation mobile network is not a greenfield task, i.e. starting from scratch. In fact, this planning task should take into consideration the already existing legacy systems and assets, such as point-of-presence, already existing base station, optical fiber for connecting the core network elements, the data center, etc.

In this very complex optimization problem, i.e. network planning, where several input parameters are uncertain random variables or distribution the AI can play a very important role in mastering the high complexity as well as delivering high efficient solutions. **Figure 10** shows how the AI can be integrated in the planning process [1]. The AI integrating module contains three parts:

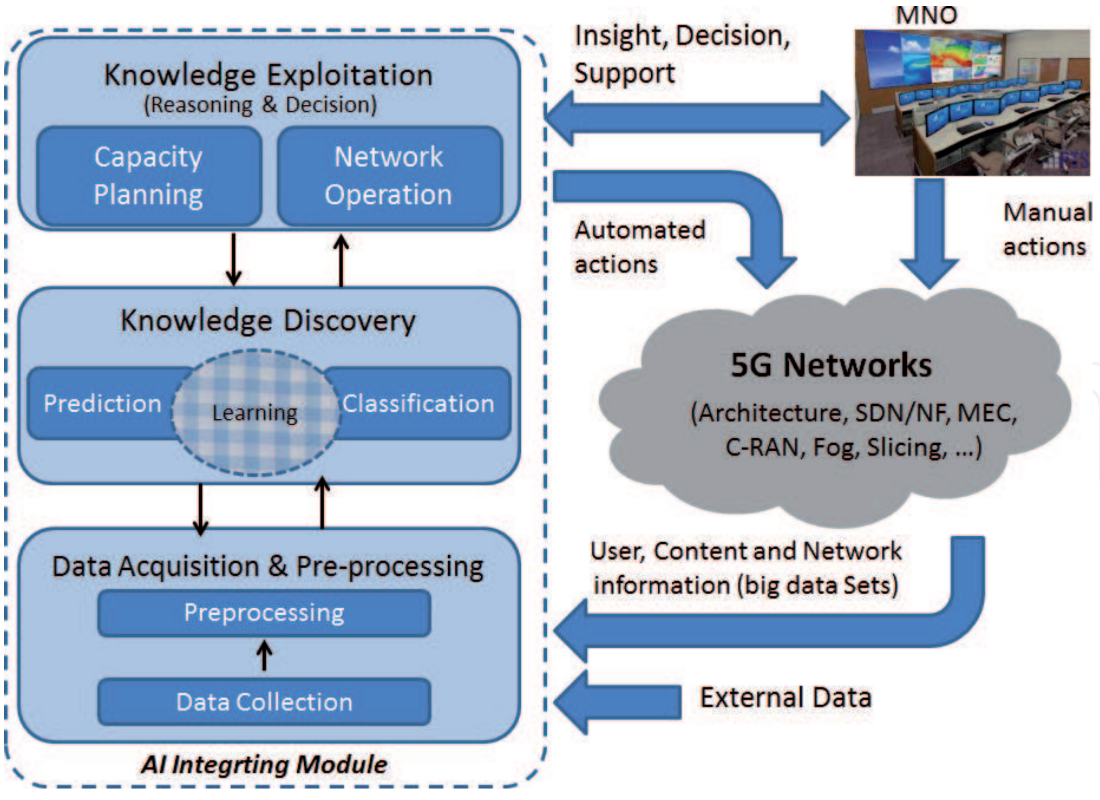

Figure 10.
Integrating AI/ML in the planning process of mobile networks (adapted from [1]).

- Data Acquisition and Pre-processing: Mobile Network Operators (MNOs) operate generally with complex, disparate sets of data, with useful information residing in multiple systems such as Operation and Management Systems (OMS), billing systems, inventories, network elements, Customer Relationship Management (CRM) systems, etc. However, to gain the challenge for achieving high performing future mobile network MNOs are forced to adopt efficient big data tools to bring together all necessary and profitable data sets. An AI-based planning system should be able to smartly analyze and correlate all these different data sources. A smart and efficient data management has to include all the necessary functions of collecting data, cleaning data, filtering data, correlating data from multiple sources and finding the relevant data.

- Knowledge discovery: this is the learning stage, where we try to learn and understand the traffic pattern and congestion, user behavior, resource usage, QoS, future location, faulty/problem elements/areas and their impacts on network efficiency. This stage should help to understand network performance and QoE, identify network anomalies, perform optimization, predict performance, disruption and requirements automate the control and operation, for example, self-configuration, self-optimization, self-healing.

- Knowledge exploitation: this stage will make use of the extracted knowledge (e.g. the prediction of the traffic patterns analysis in time/frequency/spectrum, the identified users' behaviors, etc.) to make decisions about the actions to be applied to the network elements/configurations. Such actions may either correct some system errors or improve the performance to be adapted to current/upcoming

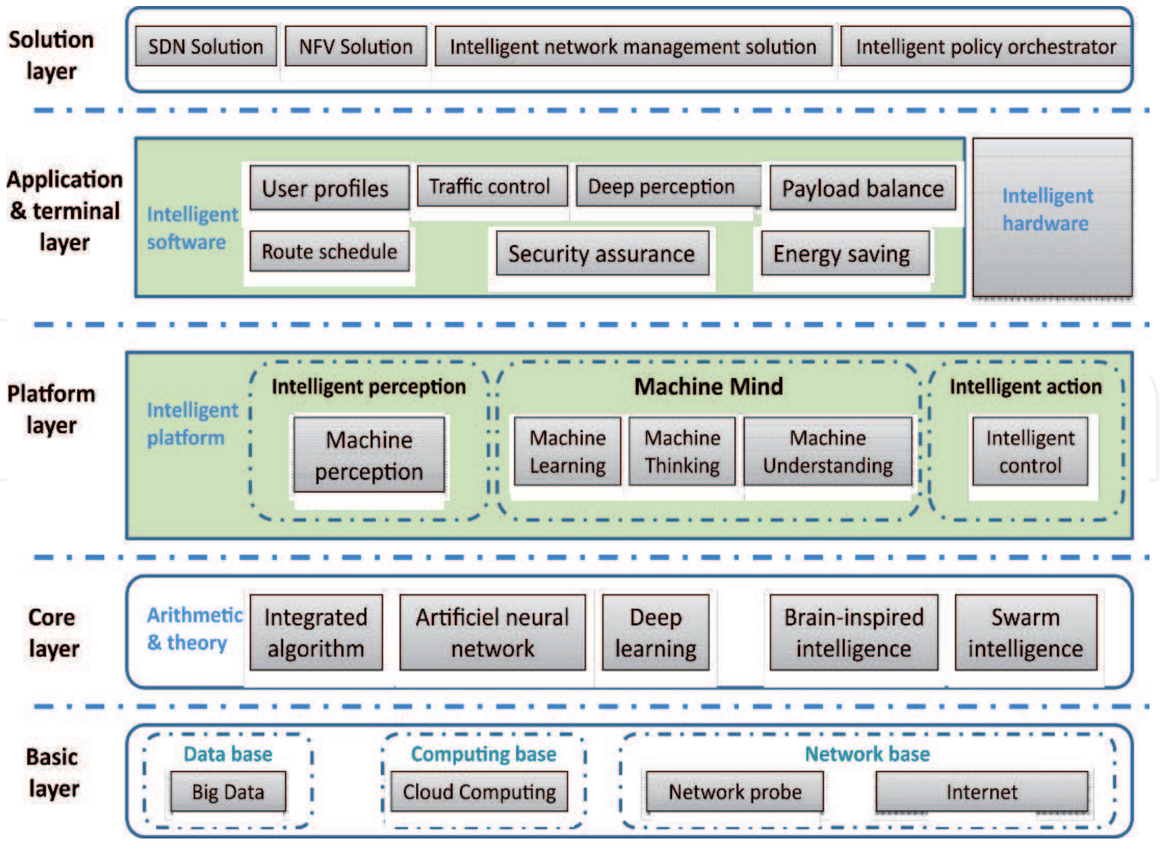

Figure 11.
Intelligence plane for the integration of AI in SDN, NFV and network control in the platform "FINE" [9].

situations and scenarios in the operating environment. In other words, this part gives out options and/or planning for slicing, virtualization, edge computing and impact of each decision option and/or planning, decisions about network expansion plan or resource utilization plan, suggestions on corrective actions.

5.3 AI building efficient collaboration NFV/SDN and network management

Already with 4G, the mobile network operators were facing an increasing network densification as response to the increasing demand for capacity and coverage, while with 4.5G operators were facing an exponential increasing number of end-devices, essentially in case of M2M and NB-IoT LTE. Therefore, research works have been dealing with the integration of AI in different levels of mobile architecture; independently of the access technology, either 4G or 5G. For example, authors in [9] proposed a functional architecture of the integration of AI to exploit and serve SDN, NFV and network control/monitoring. The authors proposed a framework of an intelligent communication network, called future intelligent network (FINE). The framework architecture is constituted of three planes: intelligence plane, agent plane and business lane.

In this section, we focus on the integration of the AI in SDN/NFV and network management, which is achieved through the intelligence plane that acts as the brain of the entire framework, **Figure 11**. Therefore, FINE is an intelligent network with an AI core. The intelligence plane can be composed of the basic layer, the core layer, the platform layer, the application/terminal layer and the solution layer. The basic tasks of each layers are summarized in **Table 4**.

Layers	Main tasks
Basic Layer	Provides support in data, calculation and the network for the intelligent plane. The data here is big data, not only including static data such as expert knowledge data, network infrastructure data, user profile data and others, but also including dynamic original data collected by the network probes from the business layer, such as status data of various types of equipment, applications and services.
Core Layer	Provider of intelligent algorithms in the intelligent plane, such as integrated algorithms, an artificial neural network, depth learning, brain-inspired intelligence and swarm intelligence. It is the kernel of the FINE core.
Platform Layer	Provides intelligent planes for the realization of the intelligent logic of AI ability and behaviour, such as intelligent perception, machine mind, intelligent action etc. The intelligent perception function can make use of theories and algorithms of the core layer, and deal with the big data of the basic layer supported by the computing resources, so as to perceive the development trends of networks and services. The machine mind function includes machine learning, machine thinking, machine understanding, etc. The ML consists of machine learning abilities generated by algorithms such as deep learning, brain-inspired intelligence and swarm intelligence. The machine thinking function provides the ability of knowledge mapping and knowledge reasoning. The machine understanding function provides the abilities of understanding based on the existing knowledge and the phenomenon, solving the ambiguity problem in reasoning, etc.
Application & Terminal Layer	Provides abilities of modular realization of functions needed by the solution layer. The functions here may include the user portrait, the flow control, the load balancing, the depth perception, the routing, the security, the energy saving, etc. These realizations may be in software or hardware using the abilities of perception, thinking and action provided by the platform layer.
Solution Layer	In charge of designing flexible policies and related activities related to satisfy the requirements to operate or manage the network, the network element, the network management system, etc.

Table 4.
Main tasks and layers in the intelligence plane of the platform "FINE" [9].

6. Conclusion

After building 4G network, network operators offered real broadband mobile Internet with capacities up to 600 Mbps. However, services and subscribers requirements have evolved to a very demanding and unprecedented level for higher quality of service. Virtual reality and augmented reality are demanding extreme high capacities and Internet of vehicles are requiring ultra-reliable communication and extreme low latency. This pushed the mobile network operators to start the migration toward 5G. Indeed, evolved technologies allowed 5G to reach bit rates over 1Gbps, through new radio interface, massive MIMO, beamforming, etc. However, the operators has to increase also the intelligence in their network, to learn more concisely about their operating environment and forecasting its evolution to optimize the resources utilization, adapt and configure automatically the network to cope with the wide variety of services. In this chapter, we have shown that this became possible through the integration of artificial intelligence and different machine learning approaches. We have presented some interesting use cases, which allow the operator to build self-healing and self-upgrading networks.

Nomenclature

4G/5G	Fourth/fifth generation of mobile network
AI	Artificial Intelligence
CAPEX	Capital Expenditure
Gbps	Giga bit per second.
LTE	Long Term Evolution
M2M	Machine-to-Machine communication
MAC	Medium Access Control
MIMO	Multi-Input Multi-Output
ML	Machine Learning
mMTC	Massive Machine-Type Communications
MNO	Mobile Network Operator
NB-IoT	Narrowband Internet of Things
NFV	Network Function Virtualization
NOMA	Non-Orthogonal Multiple Access
OFDM	Orthogonal frequency Division Multiplexing
OPEX	Operational Expenditure
QoE	Quality of Experience
QoS	Quality of Service
RRM/RRA	Radio Resource Management/Allocation
SDN	Software-Defined network

Author details

Abdelfatteh Haidine[1]*, Fatima Zahra Salmam[2], Abdelhak Aqqal[1] and Aziz Dahbi[1]

1 National School of Applied Sciences, Chouaib Doukkali University, El Jadida, Morocco

2 Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco

*Address all correspondence to: haidine.a@ucd.ac.ma

References

[1] J. Pérez-Romero, O. Sallent, R. Ferrús and R. Agustí, "Artificial Intelligence-based 5G network capacity planning and operation," 2015 International Symposium on Wireless Communication Systems (ISWCS), Brussels, 2015, pp. 246-250, doi: 10.1109/ISWCS.2015.7454338.

[2] Turing, A., Computing Machinery and Intelligence,"Mind, Vol. 59, 1950

[3] McCulloch, W., and W. Pitts, A Logical Calculus of the Ideas Immanent in Nervous Activity, Bulletin of Mathematical Biophysics, Vol. 5, 1943.

[4] Rosenblatt, F., The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain," Psychological Review,Vol. 65, No. 6, 1958

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770-778, 2016.

[6] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6):82-97, 2012.

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In NIPS, pages 3104-3112, 2014.

[8] Ericsson AB: "Employing AI techniques to enhance returns on 5G network investments Ericsson AI & Automation Report, 2019. Online available https://www.ericsson.com/en/networks/offerings/network-services/ai-report (last retrieved 31.01.2021)

[9] G. Xu, Y. Mu and J. Liu: "Inclusion of Artificial Intelligence in Communication Networks and Services," ITU Journal: ICT Discoveries, Special Issue No. 1, 13 Oct. 2017

[10] C. Zhang, P. Patras and *H. Haddadi*, "Deep Learning in Mobile and Wireless Networking: A Survey," in IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2224-2287, third quarter 2019, doi: 10.1109/COMST.2019.2904897.

[11] Y. Liu, S. Bi, Z. Shi and L. Hanzo, "When Machine Learning Meets Big Data: A Wireless Communication Perspective," in IEEE Vehicular Technology Magazine, vol. 15, no. 1, pp. 63-72, March 2020, doi: 10.1109/MVT.2019.2953857.

[12] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu and F. Kojima, "Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-Generation Wireless Networks," in IEEE Access, vol. 6, pp. 32328-32338, 2018, doi: 10.1109/ACCESS.2018.2837692.

[13] FP7 ICT project: Mobile and Wireless Communications Enablers for the Twenty-Twenty Information Society 5G. [Online]. Available: https://www.metis2020.com/

[14] S. Hassani, A. Haidine and H. Jebbar: "Road to 5G: Key Enabling Technologies," Journal of Communications Vol. 14, No. 11, November 2019

[15] M. Yao, M. Souhoul. V. Marojevic and J. Reed: "Artificial Intelligence-Defined 5G Radio Access Networks", in IEEE Communications Magazine, vol. 57, no. 3, pp. 14-20, March 2019, doi: 10.1109/MCOM.2019.1800629.

[16] O. Dharmadhikari: "Leveraging Machine Learning and Artificial Intelligence for 5G", [Internet]. Avaiblae

from : https://www.cablelabs.com/leveraging-machine-learning-and-artificial-intelligence-for-5g (last accessed: 11/01/2021)

[17] M. E. M. Cayamcela and W. Lim, "Artificial Intelligence in 5G Technology: A Survey," 2018 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, 2018, pp. 860-865, doi: 10.1109/ICTC.2018.8539642.

[18] Q. Mao, F. Hu and Q. Hao, "Deep Learning for Intelligent Wireless Networks: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 2595-2621, Q4 2018, doi: 10.1109/COMST.2018.2846401.

[19] N. C. Luong et al., "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," in IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3133-3174, 4th Q 2019, doi: 10.1109/COMST.2019.2916583.

[20] V. P. Kafle: "Network control with AI/ML – Standardization progress in ITU –," Presentation of National Institute of Information and Communications Technology (NICT), for the Institute of Electronics, Information and Communication Engineers (IEICE), Japan, 2020 May 21. Online available https://www.ieice.org/~nv/NV2020-01-Kafle.pdf. (Last retrieved 05.01.2021)