Zero Touch Management: A Survey of Network Automation Solutions for 5G and 6G Networks

Estefanía Coronado*‖, Rasoul Behravesh†, Tejas Subramanya‡, Adriana Fernández-Fernández*, Shuaib Siddiqui*, Xavier Costa-Pérez*§, and Roberto Riggio¶**
*i2CAT Foundation, Barcelona, Spain;
Email: {estefania.coronado, adriana.fernandez, shuaib.siddiqui, xavier.costa}@i2cat.net
†Fondazione Bruno Kessler, Trento, Italy; Email: rbehravesh@fbk.eu
‡Nokia Standards, Munich, Germany; Email: tejas.subramanya@nokia.com
§ICREA, Barcelona, Spain
¶Polytechnic University of Marche, Ancona, Italy; Email: r.riggio@univpm.it
**RISE Research Institutes of Sweden AB, Stockholm, Sweden; Email: roberto.riggio@ri.se
‖Universidad de Castilla-La Mancha, Albacete, Spain; Email: estefania.coronado@uclm.es

***Abstract*—Mobile networks are facing an unprecedented demand for high-speed connectivity originating from novel mobile applications and services and, in general, from the adoption curve of mobile devices. However, coping with the service requirements imposed by current and future applications and services is very difficult since mobile networks are becoming progressively more heterogeneous and more complex. In this context, a promising approach is the adoption of novel network automation solutions and, in particular, of zero-touch management techniques. In this work, we refer to zero-touch management as a fully autonomous network management solution with human oversight. This survey sits at the crossroad between zero-touch management and mobile and wireless network research, effectively bridging a gap in terms of literature review between the two domains. In this paper, we first provide a taxonomy of network management solutions. We then discuss the relevant state-of-the-art on autonomous mobile networks. The concept of zero-touch management and the associated standardization efforts are then introduced. The survey continues with a review of the most important technological enablers for zero-touch management. The network automation solutions from the RAN to the core network, including end-to-end aspects such as security, are then surveyed. Finally, we close this article with the current challenges and research directions.**

***Index Terms*—Network Management, Autonomous Networks, Zero-touch Management, Mobile Networking, Wireless Networking, 5G, 6G.**

I. Introduction

The current Internet is a marvelous piece of engineering, connecting billions of people around the world. It is a gigantic feat and consists of many complex, interacting pieces of hardware and software, from the optical links, switches, routers, radio base stations, hosts, and the hundreds of other devices that form the network fabric to the thousands of different protocols and software that run it. Given the complexity of the resulting system, it is of capital importance to consistently and precisely ensure that all its components are operating within their expected parameters and to be able to detect and react in a timely fashion when this is not the case.

Starting from these considerations, we can provide a generic and encompassing definition of network management as all the tasks associated with *monitoring*, i.e., checking if a network subsystem, may it be hardware or software, is operating within parameters, and *repairing or reconfiguration*, i.e., adopting the necessary actions to ensure that the subsystem returns within the correct operating conditions. Opposed to network management is the term *network control*, which we use to refer to real-time operations happening at the network edges, e.g., Medium Access Control (MAC) scheduling at a cellular base station. A more formal definition of network management can be found in [1] which states: *network management includes the deployment, integration, and coordination of the hardware, software, and human elements to monitor, test, poll, configure, analyze, evaluate and control the network and element resources to meet the real-time, operational performance, and Quality of Service requirements at a reasonable cost*.

In this context, 5G and beyond 5G networks are a paradigm shift: their high performance in terms of latency, bit-rate, and reliability call for a technological and business convergence between cloud computing and the telecom worlds. Features such as slicing, edge computing, and better and more flexible and programmable radio connectivity can be used to enable different applications and services and to deliver a richer user experience, faster interactions, large-scale data processing, and better machine-to-machine communications. Nevertheless, the challenges to overcome the realization of this vision are significant. In particular, the growing diversity and complexity of the end-to-end mobile network is resulting in an overly complex system that operators and vendors are finding difficult to operate, manage, and evolve.

Several attempts have been made in the past to embed intelligence and reasoning into the mobile network for automation and optimization purposes. This includes Active Networks (ANs) [2], Self Organizing Networking (SON) [3], and Autonomic Network Management (ANM) [4], just to name a few. More recently, Zero-touch Network Management (ZTM) has emerged as a fully autonomous network management solution with human oversight. According to this paradigm, the network can reason about its current state, interpret it,

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2022.3212586

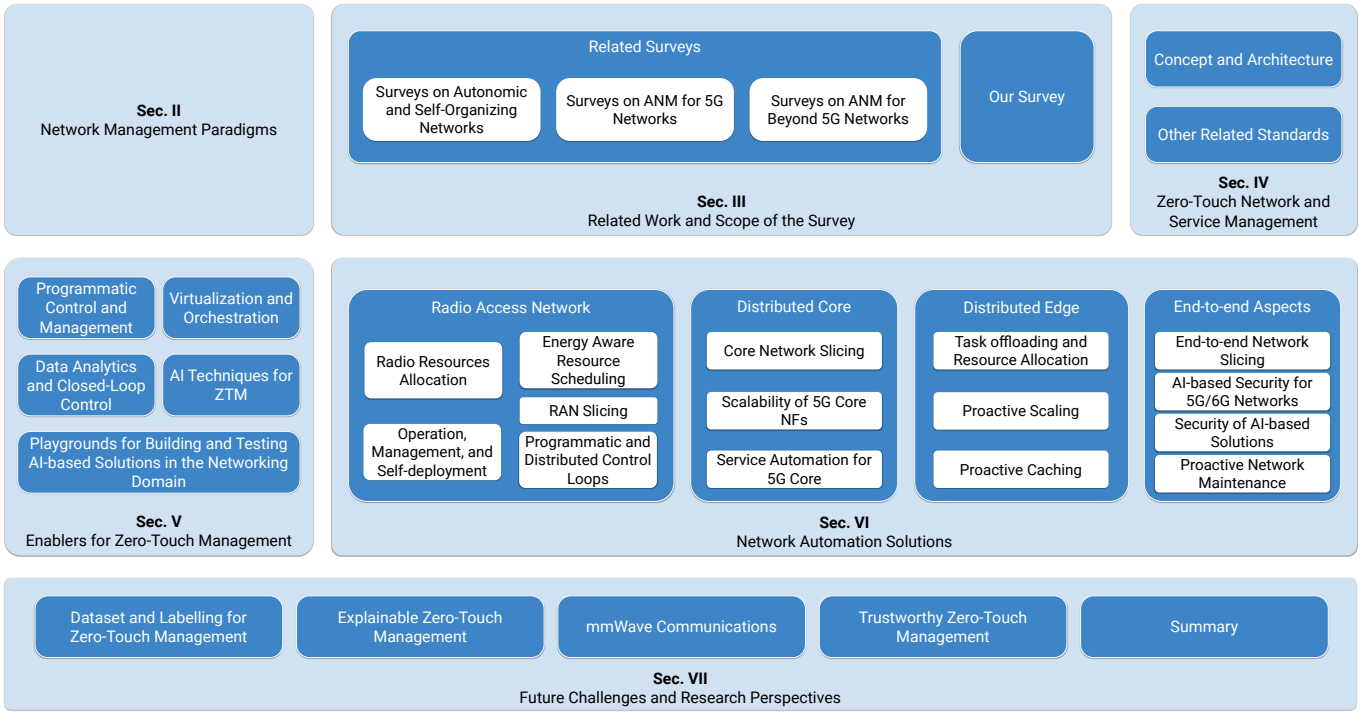


Fig. 1. Overview of the organization of the survey.

and provide recommendations about possible reconfigurations, always leaving a human operator the task of validating and accepting the suggested changes.

The introduction of ZTM concepts and technologies in the cloud-network convergence process will be crucial to help operators achieve a higher level of automation, increase network performance, and decrease the time-to-market of new features. This trend is reflected in ZTM-based solutions for problems ranging from energy-aware Radio Access Network (RAN) resource scheduling to service automation and from proactive caching to secure end-to-end slicing. The trend toward ZTM-based solutions has also been fostered by the European Telecommunications Standards Institute (ETSI), which founded in 2017 a new Industry Specification Group (ISG) named Zero-touch Service Management (ZSM) aiming to define the requirements and architecture of a network automation framework based on ZTM concepts.

Despite the growing interest in ZTM and ANM techniques in the mobile networking area, a comprehensive survey was still missing. This article aims precisely at filling this gap between ZTM/ANM and mobile networks, with a particular focus on 5G and beyond systems, by providing a survey of up-to-date literature that considers commonality across these two domains. Beyond reviewing the most relevant literature, we also discuss how ZTM and ANM can be applied to specific areas of the mobile networking landscape, including RAN, core, edge cloud, and end-to-end aspects.

Structure of the paper. The structure of this survey is sketched in Fig. 1, which is divided into seven sections. In Sec. II, firstly, we provide the reader with a tutorial on some basic network management concepts and nomenclature. Then, Sec. III presents a review of the related works, covering surveys on autonomic management and SON, on ANM for 5G networks, and on ANM for beyond 5G networks. We conclude the section with a clear definition of the scope and contribution of this paper with respect to existing works.

In Sec. IV, we provide some background on ZTM, highlighting its major advantages and architectural choices as well as the most relevant standards. There exist many enabling technologies required to implement ZTM and ANM; they include programmable networks, network virtualization, data analytic and closed-loop control, Artificial Intelligence (AI), etc. Such technologies are surveyed in Sec. V. In Sec. VI, we review recent ZTM and ANM literature in the mobile networking domain, grouping it by network segment, i.e., radio access, distributed core, cloud and edge, and end-to-end. Finally, we conclude this survey with a discussion of open challenges and future research directions in Sec. VII. The abbreviations used in this survey are listed in Table I.

TABLE I
LIST OF ACRONYMS USED IN THIS SURVEY.

Acronym	Definition
3GPP	3rd Generation Partnership Project
4G	4th Generation of Broadband Cellular Networks
5G	Fifth Generation
ACK	Acknowledgment
AI	Artificial Intelligence
AIOps	Artificial Intelligence for IT Operations
AMF	Access and Mobility Management Function
AN	Active Network
ANM	Autonomic Network Management
ANN	Artificial Neural Network

Acronym	Definition
ANR	Automatic Neighbour Relation
AP	Access Point
API	Application Programming Interface
ARIMA	Auto-Regressive Integrated Moving Average
ATM	Asynchronous Transfer Mode
AuSF	Authentication Server Function
BBU	Base-Band Unit
BS	Base Station
BSS	Business Support System
C-RAN	Cloud-Radio Access Network
CDN	Content Delivery Network
CMIP	Common Management Information Protocol
CNM	Cognitive Network Management
CN	Core Network
CNN	Convolutional Neural Network
COE	Container Orchestration Engine
COBANETS	COgnition-BAsed NETworkS
COSLA	Closed Loop Communication Service Assurance
CPU	Central Processing Unit
CSI	Channel State Information
CSMF	Communication Service Management Function
CU	Central Unit
D2D	Device To Device
DCAE	Data Collection, Analytics and Events
DDL	Distributed Deep Learning
DDoS	Distributed Denial of Service
DL	Deep Learning
DLT	Distributed Ledger Technologies
DN	Data Network
DNN	Deep Neural Network
DPE	Distributed Processing Environment
DQN	Deep Q Learning
DRL	Deep Reinforcement Learning
DU	Distributed Unit
E2E	End-to-End
eMBB	enhanced Mobile Broad Band
EMS	Entity Managment System
eNB	Evolved Node B
ENI	Experiential Networked Intelligence
ETSI	European Telecommunications Standards Institute
EU	European Union
FCAPS	Fault, Configuration, Accounting, Performance and Security
FG-ML5G	Focus Group on ML for Future Networks including 5G
FL	Federated Learning
IBN	Intent-Based Networking
ICN	Information-Centric Networking
IDN	Intent-Driven Networking
IETF	Internet Engineering Task Force
IoT	Internet of Things
IP	Internet Protocol
ISG	Industry Specification Group
ISO	International Organization for Standardization
IT	Information Technology
ITU	International Telecommunication Union
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
KDN	Knowledge-Defined Networking
KPI	Key Performance Indicator
LSO	Lifecycle Service Orchestration
LSTM	Long-Short Term Memory
LTE	Long Term Evolution
LVAP	Light Virtual Access Point
LVNF	Light Virtual Network Function
MAB	Multi-Armed Bandit
MAC	Medium Access Control
MANO	Management and Orchestration
MAPE-K	Monitor-Analyze-Plan-Execute over a shared Knowledge
MCS	Modulation Coding Scheme
MDAF	Management Data Analytics Function
MDP	Markov Decision Process
MEC	Multi-Access Edge Computing
MEF	Metro Ethernet Forum
MIB	Management Information Base
ML	Machine Learning
mMTC	massive Machine Type Communications
MMW	Milli-meter Wave
MNO	Mobile Network Operator
MSE	Mean Square Error
NF	Network Function
NFV	Network Function Virtualization
NFVI	Network Function Virtualization Infrastructure
NFVO	NFV Orchestrator
NGNM	Next Generation Mobile Networks
NMA	Network Management Automation
NMS	Network Management System
NN	Neural Network
NRF	NF Repository Function
NSMF	Network Slice Management Function
NSSF	Network Slice Selection Function
NSSMF	Network Slice Subnet Management Function
NWDAF	Network Data Analytics Function
ONAP	Open Network Automation Platform
OpenSig	Open Signalling
OPEX	Operational Expeditures
OSI	Open Systems Interconnection
OS	Operating System
OSM	Open Source MANO
O-RAN	Open Radio Access Network
OSS	Operations and Support System
PBNM	Policy-based Network Management
PCA	Principal Components Analysis
PCF	Policy Control Function
PRB	Physical Resource Block
QCI	QoS Class Index
QML	Quantum Machine Learning
QoE	Quality of Experience
QoS	Quality of Service
QoT	Quality of Trust
RAM	Random Access Memory
RAN	Radio Access Network
RF	Random Forest
RIC	RAN Intelligent Controller
RL	Reinforcement Learning
RMSE	Root Mean Square Error
RNIS	Radio Network Information Service
RNN	Recursive Neural Network
RRH	Remote Radio Head
RRM	Radio Resource Management
RRU	Remote Radio Unit
RSU	Road Side Unit
RSSI	Received Signal Strength Indicator
RU	Radio Unit
SA	Service and System Aspects

Acronym	Definition
SBA	Service-Based Architecture
SDN	Software-Defined Networking
SD-RAN	Software-Defined Radio Access Network
SDO	Standard Development Organization
SGMP	Simple Gateway Management Protocol
SL	Supervised Learning
SLA	Service Level Agreement
SMF	Session Management Function
SMI	Structure of Management Information
SMO	Service Management and Orchestration
SNMP	Simple Network Management Protocol
SNR	Signal to Noise Ratio
SOC	Service Oriented Core
SON	Self Organizing Networking
SVM	Support Vector Machines
TCO	Total Cost of Ownership
TCP	Transmission Control Protocol
TINA	Telecommunications Information Networking Architecture
TL	Transfer Learning
TMF	Tele Management Forum
TMN	Telecommunication Management Network
TN	Transport Network
TSG	Technical Specification Group
UAV	Unmanned Aerial Vehicle
UDM	Unified Data Management
UE	User Equipment
UL	Unsupervised Learning
UPF	User Plane Function
URLLC	Ultra Reliable Low Latency Communications
V2X	Vehicle To Everything
VANET	Vehicular Ad-hoc Network
VIM	Virtual Infrastructure Manager
VNF	Virtualized Network Function
VNFM	Virtualized Network Function Manager
VM	Virtual Machine
XAI	Explainable Artificial Intelligence
WLAN	Wireless Local Area Network
ZOOM	Zero-touch Orchestration, Operations and Management
ZSM	Zero-touch Service Management
ZTC	Zero-touch Commissioning
ZTM	Zero-touch Network Management

II. Network Management Paradigms

In the first generation telecommunication networks, human operators manually managed the circuit-switched communication networks until the likes of software agents, Open Signalling (OpenSig), and ANs, came along to gradually relieve them from the burden of network management. With the advancement to packet-switched 3G and 4G communication networks, network management paradigms, such as Policy-based Network Management (PBNM), Intent-Based Networking (IBN), and SON came to the fore. Subsequently, 5G communication networks fostered the ANM aspect by leveraging on the advancements in Machine Learning (ML) techniques. Finally, to realize 6G communication network requirements, zero-touch closed-loop network management is envisioned as the primary enabler. Below, we review the fundamental network management paradigms.

- *Active Network Management (AN).* It is a network management paradigm that enable packets flowing through the network to contain programming codes that can modify the network operation at run-time. Many modern-day networking paradigms that intend to revolutionize the network design by separating control and data plane, centralized network management, and network programmability stem from the idea of ANs.
- *Policy-based Network Management (PBNM).* It is a network management paradigm where policies are used to configure network elements and services. Policies are often defined as a set of rules that, upon verification of a certain condition, apply a corresponding action. They can be found at different levels of abstraction, from the business level down to the device-specific aspects and configuration. The transition from one level of abstraction to another is often done utilizing external information such as device configuration.
- *Intent-based Network Management (IBN).* An Intent is a declaration of operational goals that a communication network is supposed to deliver, without specifying how actually to achieve them. The overall goal here is to replace the tedious and error-prone work of configuring network devices as well as the reaction to networking issues. As opposed to the policy, the intent defines a high-level operational goal without specifying how it should be achieved. The translation into actual administrative actions is then performed by the network itself without human intervention.
- *Self-organizing Network Management (SON).* It is an automation technology designed to make the planning, management, optimization, and healing of mobile networks (including radio, core, and transport segments) simpler and faster. It enables the automation of selection and execution of network management actions based on pre-defined rules, therefore reducing human involvement. It then interprets events under different contexts to determine their cause-effect relationships. As opposed to ZTM, SON solutions are characterized by a very specific set of inputs and outputs, both of them tightly defined in 3GPP standards, whose parameters are manually configured by the operators.
- *Zero-Touch Network Management (ZTM).* A debate exists within the networking community on zero-touch automation's definition and its relationship with ANM. In this survey, we define ZTM as a fully autonomous network management solution with human oversight, which is also the vision of ZSM [5]. Therefore, ZTM must be able to reason, interpret and provide recommendations for subsequent network automation behavior driven by AI/ML solutions. Additionally, depending on the risk level of the use case, ZTM may also require a human network operator to verify and accept those recommendations before they are executed.
- *Autonomic Network Management (ANM).* It can be defined as the automated decision-making stage where self-configuration, self-monitoring, self-healing, and self-optimization can be achieved without requiring dictation

TABLE II
SUMMARY OF NETWORK MANAGEMENT PARADIGMS.

Network Management Paradigms	Summary
Active Network Management (AN)	Enables packet flowing through the network to contain programming codes that can modify the network operation at run-time.
Policy-based Network Management (PBNM)	Policies are defined as a set of rules that, upon verification of a certain condition, apply a corresponding networking action.
Intent-based Network Management (IBN)	Declaration of operational goals that a communication network is supposed to deliver, without specifying how actually to achieve them.
Self-organizing Network Management (SON)	Enables the automation of selection and execution of network management actions based on pre-defined rules.
Zero-Touch Network Management (ZTM)	Joint ML and networking validation environments that enables both a reduction in the time required for testbed setup and a more powerful and precise validation of ZTM approaches.
Autonomic Network Management (ANM)	Automated decision-making for self-configuration, self-monitoring, self-healing, and self-optimization operations.

of network management rules from other entities/humans and can completely act independently.

A summary of the network management paradigms discussed before can be found in Table II.

III. RELATED WORK AND SCOPE OF THE SURVEY

In this section, we review the most relevant surveys on network management and mobile networking with a particular focus on network automation and ZTM. We classify these works into three categories: (i) ANM and SON management, (ii) ANM for 5G networks; and (iii) ANM for beyond 5G networks. We summarize these surveys in Table III.

A. Surveys on Autonomic and SON Management

Since ANM was launched in 2001 by IBM, many works in the literature have analyzed its features and challenges, such as [6] and [7]. In [6], the authors provide an introduction to the concepts of autonomic computing, outline the motivation, and describe the fundamental research influencing a large proportion of early work on self-management systems. In particular, they describe the control theory-based feedback model that was introduced by IBM for self-adaptive systems, i.e., Monitor-Analyze-Plan-Execute over a shared Knowledge (MAPE-K). N. Samman et al. in [7] focus on automated management paradigms, covering system, user, and application modeling, as well as autonomic monitoring and analysis approaches. Furthermore, they provide a classification of autonomic architectures according to the degree of adaptability, intelligence, awareness, and autonomy.

SON is considered as a particular case of the ANM family for self-operating and self-x functions. The functionalities enabled by the SON paradigm can be introduced at several levels in both the management and the control planes. This specific paradigm received particular attention since 3rd Generation Partnership Project (3GPP) Release 8 of mobile network standards. The authors of [8] provide a detailed survey on autonomic computing and communications in the context of software-driven networks aided by the Software-Defined Networking (SDN) and Network Function Virtualization (NFV) cloud ecosystems. They also discuss various research challenges related to SON and propose potential solutions to self-management mechanisms (e.g., automatic testing, integration, and deployment of virtual network functions) and the associated architectures in software-driven systems. Other surveys and books on SON for 4G and 5G networks can also be found in [9]–[11]. The authors in [12] highlight SON features expected during the 5G evolution, together with the impact on Quality of Service (QoS) and Quality of Experience (QoE). The survey makes an extensive review of wireless evolution of 5G networks and provides a tutorial on the 5G architectural innovations in terms of self-management expected in the network architecture design, including air interface, smart antennas, cloud, and heterogeneous RAN.

However, a SON-enabled network may still be far from a fully autonomous entity as a whole. Furthermore, the objectives given by network administrators may be conflicting, leading to incompatible behaviors and performance degradation. Moreover, since SON functions depend on the algorithms placed on the control loops and the network status, each of them still needs to be manually configured. Finally, a SON system should be able to handle the interaction between functions. ANM requires a step beyond by introducing cognitive capabilities in the SON framework to allow intelligent adaptation to network context and decisions, enabled either by employing statistical learning or AI.

B. Surveys on ANM for 5G Networks

The fifth generation of mobile networks has monopolized research in the broad information and communication technology field in the last decade. 5G is expected to support a wide range of applications and to meet a very diverse set of Key Performance Indicators (KPIs). The 5G development progress has been summarized in a wide number of books, tutorials, surveys, and magazines, covering standardization roadmaps, technical enablers, and several cross-domain aspects such as energy efficiency and security [13]–[17]. In [13], the authors explain how 5G is not just an incremental advance w.r.t. the fourth generation of the mobile network, but it is instead an end-to-end paradigm shift and a significant breakthrough from the RAN to the core network. The survey identifies challenges for both research and standards-related activities. The authors of [14] identify the most significant requirements of the many verticals that arise in the 5G arena, such as autonomous driving. Furthermore, they present an analysis of the stan-

dardization efforts and entities behind the main test cases. A study on the transition towards a more distributed management approach for mobile networks is performed in [15]. This survey analyzes the main existing solutions, including those deployed in real-world prototypes. A comprehensive survey of the expected architectural and technological innovations of 5G is presented in [16]. In addition, the paper surveys the most relevant research projects on 5G in different countries. In [17], the authors first discuss the limitations of 4G networks, and then analyze the features of 5G to address those limitations. The survey also identifies the open challenges and presents a comparative study of the proposed innovations in terms of energy efficiency, network hierarchy, and network types.

The advances in SDN and NFV, the proliferation of new data sources, low-cost storage and computing resources, and advanced ML tools in the cloud, pave the way for easier adoption of ML into network management. Several comprehensive surveys exist concerning different AI techniques on a wide range of applications, i.e., Supervised Learning (SL) [18], Unsupervised Learning (UL) [19], (Deep) Reinforcement Learning (RL) [20], Deep Learning (DL) [21] and Transfer Learning (TL) [22].

AI-based techniques, for network management, have been the topic of several surveys in the Cognitive Network Management (CNM)/ANM domain [23]–[32]. The work in [23] overviews the most relevant AI tools in meeting 5G network requirements and puts special attention on DL-driven network management, covering from data analysis procedures to network state and health prediction and reconfiguration. Moreover, it analyzes the impact of users' location and system nature (i.e., distributed, hierarchical, etc.) on the level of automation that can be inducted. The authors of [24] also highlight the importance of ML and DL to enable an autonomous 5G network, especially concerning its optimization in terms of energy. Likewise, the work in [25] is focused on data-driven proactive 5G networks and highlights the change from reactive to proactive and automated management, the technical enablers to make it a reality, and the main networking areas affected, including network-level traffic prediction, temporal dimension, and metadata analysis, and proactive caching with social concentration prediction. Moreover, this paper analyses future research directions for a full data-driven user-oriented mobile network, such as context granularity, prediction error impact, and a strong need for collaboration with data providers. By contrast, the authors of [26] provide a broad overview of autonomic management regarding mobility operations in complex (and heterogeneous) 5G and future internet scenarios. The paper provides an overview of autonomic management, robustness issues, and key network segments where self-management must be applied according to several autonomic criteria. Finally, the paper in [32] provides an overview of IBN, or Intent-Driven Networking (IDN), covering the standardization bodies and open-source communities promoting this idea.In addition, they also focus on the formal definition of intent-based networking, the enabling technologies, and the main architectural elements in 5G networks.

Specifically related to network management automation, the study in [27] discusses the bottlenecks that are restricting the large-scale deployment of ANM systems and the role that ML can play to realize a cognitive MAPE-K control loop for network management. Moreover, the study describes, in detail, the scope of ML in the Fault, Configuration, Accounting, Performance and Security (FCAPS) management areas while discussing the challenges of using ML for ANM. Concerning CNM, the authors of [28] provide a comprehensive survey of ML algorithms that can be applied in the context of SDN, mainly focusing on traffic classification, routing optimization, QoS and QoE prediction, resource management, and security aspects. In [29], the authors describe a new paradigm called Knowledge-Defined Networking (KDN), which exploits SDN, telemetry, network analytics, and AI. The KDN paradigm operates utilizing a control loop to provide automation, recommendation, optimization, validation, and estimation mechanisms. Moreover, the authors describe relevant use cases to illustrate the KDN's applicability to networking and present experimental results to determine the benefits of using ML. The authors of [30] examine the problem of reliable resource provisioning in a joint edge-cloud NFV ecosystem and present a survey on various ML technologies used to enhance the reliability of distributed applications in heterogeneous networks. They give particular emphasis to workload characterization and prediction, component placement and system consolidation, application elasticity and remediation aspects of NFV. Conversely, the paper in [31] introduces an innovative concept called COgnition-BAsed NETworkS (COBANETS) that leverages unsupervised DL for system-wide learning, optimization, and data representation in an NFV domain.

C. Surveys on ANM for Beyond 5G Networks

Even though the deployment of 5G has just begun, both industry and academia have started to look beyond 5G and 6G networks. As a result, there is a growing number of surveys looking deeper into the ongoing research towards the sixth generation of the mobile network. Some of these surveys intend to provide a bird's-eye view of the evolution of the system architecture from 5G to 5G and beyond networks [33], [34]. The authors of [33] first provide a 5G/6G network system architecture split into access cloud, forward cloud (i.e., 5G switch, and service enabler), and control cloud functionalities (i.e., radio resource management, information center, etc.). Then, they propose several centralized, semi-centralized, and hierarchical management options based on SDN technologies, paying particular attention to interference management. By contrast, in [34] the authors perform a deeper analysis of the main architectural elements that will be part of 6G networks. Furthermore, they propose a taxonomy building on the main technological enablers, use cases, AI schemes, and communication and computing technologies that are expected to play a key role in the 6G domain.

There are several surveys which aim to provide a broad overview of open research challenges for the race towards 6G [35]–[37]. Different from the rest, A. Dogra et al. in [38] compare the features and requirements of 5G and 6G technologies. Furthermore, they study the architecture of a virtualized

6G architecture based on the concept of network slicing through a practical use case. In addition to describing the next-generation network technologies, the work in [39] details the feasibility of Information-Centric Networking (ICN) in 5G and beyond networks, paying particular attention to existing in-network content caching schemes over ICN, as well as the challenges and open issues in the management of 6G networks.

There exist other surveys that make a point for the role that AI and ML will play in the ANM of beyond 5G and 6G networks and their contribution toward fully autonomous network management [40]–[46]. The main argument of the authors of [40] is that the complexity of future 6G networks will reach a point where traditional analytical and numerical simulation approaches will not be feasible anymore. Conversely, the survey in [41] overviews the privacy concerns associated with the inherent inclusion of AI in the 6G network architecture, analyzing the main violation techniques and attacks, as well as the corresponding protection methods based on the type of ML models used. Related to the intelligence topic, the authors of [42] analyze the technical aspects of large, intelligent surfaces, including physical principles and capacity analysis, and present the research directions on practical problems on distributed and centralized topologies. Conversely, the survey in [43] first provides a tutorial on the main AI techniques that will be used in 6G networks and how they can be applied in different aspects such as radio interface, intelligent traffic control, CNM and security. Moreover, this work analyzes the new services that AI will enable in 6G networks, such as holographic applications and brain-computer interactions. The study in [44] covers the potential of ML to support beyond 5G and 6G requirements. Moreover, it reviews the main considerations regarding AI to be taken into account when envisioning ZTM for beyond 5G and 6G networks. Among such considerations, the authors highlight the extreme need for defining and developing 5G infrastructure to generate research datasets (involving security and privacy issues), the non-existence of a one-fits-all ML model providing absolute best learning and performance in all applications, and interpretability vs. accuracy trade-offs, among others. In addition to the state-of-the-art in ML, the work in [45] makes a step forward by surveying the fundamentals of Quantum Machine Learning (QML), the main existing solutions in QML-assisted communications, as well as future research directions in terms of intelligent proactive caching and Multi-Access Edge Computing (MEC), massive big data analytics, and intelligent cognitive radio and self-sustaining networks, among others. Recently, the IBN paradigm has also been extended beyond 5G and 6G networks. The authors in [46] provide a comprehensive survey on the architectures and key techniques of IBN in both RAN and core network domains for 6G networks. Furthermore, they discuss various industrial efforts of intent-based networks to solve the problems of open data platforms and automated network operations.

D. Our Survey

The objective of this paper is to provide a complete survey of ZTM and ANM techniques in the mobile networking area. In pursuit of this goal, we aim to discuss why ZTM and ANM are suitable for solving mobile networking problems (Sec. IV), what are the most recent applications of ANM concepts in the ZTM domain (Sec. VI), and what are the most important and promising directions worthy of further investigation (Sec. VII). Moreover, we cover the different technological enablers required to realize the ZTM and ANM vision including programmatic control, virtualization, closed-loop control, AI/ML, and open-source initiatives (Sec. V).

It is our standpoint that the literature surveyed in this section only partially addressed these questions. The only other works which are in principle similar to ours are [47]–[50]. The paper in [47] is a magazine aiming at discussing mainly the challenges and research directions in the ZTM domain. On the other hand, the survey in [48] focuses only on security aspects. Conversely, the work in [50] introduces a functional architecture and requirements for ZTM, and reviews the main elements enabling this vision, namely policy-based automation, intents, ML, and RL algorithms. Similar to ours, this survey addresses ZSM standardization issues. However, the rest of the paper is focused only on cross-domain ZSM approaches and security aspects. The most similar to this work is [49], in which the authors provide an overview of the ZSM vision, including the review of the ETSI ZSM and the ETSI ENI architectures, as well as an analysis of the contributions provided by European projects in terms of ZSM. The core of the paper provides a high-level set of 7 ML algorithms (i.e., logistic regression, random forest, neural networks, Support Vector Machines (SVM), naive Bayes, k-nearest neighbor, and RL), and links them with four network management functions, namely RAN management, resource management, flow inspection, and multi-domain management. In contrast to our paper, the work in [49] does not cover the main technical enablers of ZTM, lacks an in-depth literature review, and only examines in a high-level manner a subset of the network management areas and strategies toward a ZTM approach. Moreover, the rest of the related survey papers are fully oriented to a specific area of ZTM (e.g., security, main challenges, etc.) and do not examine control loops and methods for B5G and 6G networks. To the best of our knowledge, this is the first survey looking specifically at ZTM and ANM solutions in the field of management of 5G and 6G networks, providing a tutorialistic vision of the network management paradigms, covering network management functions and requirements from all network segments, and reviewing the main technical enablers to make ZTM and ANM a reality in future networks.

IV. Zero-touch Network and Service Management

This section introduces the concept and architectural principles of the ETSI ZSM ISG, given that it is considered as the reference solution for the realization of ZTM-powered systems. In addition to this, other related standardization groups are also presented to provide a more complete overview of current efforts in the standardization arena to boost the adoption of ZTM.

TABLE III
EXISTING SURVEYS IN THE ROAD TO ZERO-TOUCH NETWORK MANAGEMENT.

Year	Mgmt. Paradigm	Ref.	Scope
2008	ANM	[6]	Definition of autonomic computing, and discussion on self-management properties and degrees of autonomicity.
2009	ANM	[7]	Comparison of traditional vs. autonomic management. Classification of the areas, (e.g, monitoring and analysis), and main architectures in ANM.
2010	SON	[10]	Overview of self-configuration and self-optimization procedures in LTE for SON, focusing on challenges from new radio nodes without dedicated backhaul.
2014	ANM	[13]	Breakthroughs brought by 5G w.r.t. 4G networks, including the involvement of the first automation processes in core and RAN.
2015	ANM	[15] [16]	Evolution of the 5G architecture, and discussion on emerging technologies for management, from SON to ANM, covering spectrum sharing, mmwaves, mobility management, and cloud.
		[31]	Vision on cognitive networks, comprising the enablers of a functional architecture (e.g., SDN, NFV, and system-wide policies handling various data flows), highlighting the role of deep learning.
2016	SON	[12]	Comprehensive survey on 5G networks, giving major importance to SON in the paradigm shift enabled by smart antennas, virtualization, and greater agility and adaptability, especially on QoE provisioning and quality management.
	ANM	[17]	Limitations and challenges of 5G, e.g., self-healing infrastructures and flexibility. Review of multi-tier CNM-based architectures aiming for interference and energy consumption minimization.
2017	SON	[8]	Discussion on self-management and self-optimization of networks in 4G, e.g., network function testing and integration, SON actuators, and standardization. Autonomic operations and security functions are also considered.
	ANM	[29]	Review of knowledge-defined networking for 4G enabled by ML, including a control loop examining automation, recommendation and optimization functions.
2018	SON	[11]	Evolution of SON and its architectures in 3GPP (e.g., centralized/distributed SON), and their role in softwarized networks from 4G, including management and use cases through ML.
	ANM	[27]	Review of ML for network management in 5G, putting special attention on FCAPS operations and the relationship with the ML techniques used in each management area.
		[28]	Survey on the ML techniques most widely applied in programmatic and autonomic network management enabled by the SDN architecture in a centralized manner.
2019	ANM	[30] [23] [26]	Overview of AI and ML tools being embedded in the 5G network architecture (with a special focus on deep learning algorithms) to enable CNM in various areas (e.g., data analysis procedures, health prediction, handover and self-reconfiguration).
		[33]	Analysis of network resource and interference management systems in SDN-enabled intelligent 5G/6G networks.
		[35] [36] [40]	Driving applications, requirements, and promising techniques for evolving 6G systems, including massive analytics, highly scalable ML operations, and ML and big data assisted ANM operations.
		[44]	Overview of considerations to embed ML as a native part of B5G networks classified by main learning types (SL, UL, RL) and the management areas where they present the greatest performance.
		[45]	Survey of deployment options of Quantum ML for 6G and CNM, covering aspects such as intelligent and proactive caching for user-centric deployments, full harmonization in the manipulation of wireless networks.
2020	IBN	[32] [46]	Discussion on the evolution of IDN working groups, industry groups and standardization efforts, and the relationship with ZTM. Review of enabling technologies and architectures in 5G and B5G networks.
	ANM	[24]	Vision of the role of ML, and especially DL, in enabling ANM. Particular attention is provided to energy saving and optimization for B5G networks.
		[25] [34] [43]	Data-driven and user-oriented approaches to enable a complete ANM concept, together with the technical enablers required in terms of traffic prediction, metadata and social analysis, and security for heterogeneous B5G systems.
		[37] [38]	Analysis of the 6G vision, technical requirements and role of AI as a transformative foundation for ANM on a fully virtualized architecture with ML and network slicing as main contributors.
		[39]	Review of the concept of ICN and their inclusion in the architecture and management of B5G networks. A taxonomy in vertical (centralized, distributed, collaborative) and horizontal (from edge to cloud) management is proposed.
		[41]	Review of privacy issues arising from the automatized management employing AI in 6G networks, including attack surfaces for ML tools.
		[42]	Review of the physical aspects and challenges when considering the automation and management of future mobile and wireless networks using large intelligent surfaces.
	ZTM	[47]	Overview of the research projects and existing standardization related to ZSM and ZTM. A preliminary architectural vision and its limitations in terms of AI are identified.
	ZTM	[48]	Analysis of potential security threats related to the ZSM architecture and its main technological enablers (e.g., closed loops and AI/ML-based attacks), also identifying potential mitigation mechanisms.
2021	ZTM	[49]	Overview of the ZSM vision, related standards (e.g., ETSI ZSM and ETSI ENI), and relevant European projects. Broad ML classification linked with network management areas (e.g., resource management, multi-domain management, etc.).
2022	ZTM	[50]	Overview of ZSM standards, including the description and requirements of a ZSM functional architecture. Focus on cross-domain lifecycle management and security aspects of ZSM.

A. Concept and Architecture

The ever-increasing complexity of next-generation wireless and mobile networks imposes a very challenging scenario for traditional human-centric management systems. The need to cope with the unprecedented flexibility and dynamic adaptation introduced by 5G/6G networks has triggered the transition toward ZTM solutions. The ZTM vision implies exploiting the automation capabilities of autonomously operated and self-adapting networks to improve agility in deployment and maintenance operations, with the consequent reduction of costs.

Motivated by this reality, in 2017 ETSI founded ZSM ISG, which introduces the vision of ZTM-driven networks and services. The overarching goal of ETSI ZSM is to provide a framework that enables *full end-to-end automation of network and service management* in a multi-vendor environment [51]. More specifically, the ZSM framework envisions the following operational processes and tasks automatically executed without human intervention.

a) *Planning and design*, to accelerate the creation of tailored networks and services that meet the needs of subscribers.
b) *Delivery*, to enable on-demand delivery of services while ensuring requirements fulfillment.
c) *Deployment*, to improve network and resource utilization and expedite the service roll-out.
d) *Provisioning*, to reduce manual configuration errors and automate service assurance.
e) *Monitoring and optimization*, to effectively avoid service degradation and ensure quick network recovery.

The ZSM reference architecture, presented in [5] and depicted in Fig. 2, proposes a modular, scalable, and extensible Service-Based Architecture (SBA) based on the decomposition of complex management services and management functions into fundamental building blocks that are integrated following a consistent set of composition and interoperation patterns.

Separation of concerns is one of the fundamental principles behind ZTM and requires management domains to be defined around administrative, geographical, or technological boundaries. For cross-domain management and coordination, an *End-to-End (E2E) Service Management Domain* is introduced as a special management domain responsible for the services that span across multiple management domains. Every management domain contains the following building blocks:

- *Management Functions*, which provide a set of capabilities by exposing and/or consuming a set of endpoints.
- *Data Services*, which enable data sharing, persistence, and authorized access management across services within management domains, avoiding the need for management functions to handle their own data persistence.
- *Domain Integration Fabric*, which facilitates the exposure of service capabilities beyond domain boundaries while controlling the access to exposed management services.

Complementing this, the *Cross-domain Integration Fabric* is responsible for the visibility and the accessing of cross-domain services endpoints (i.e., between the management domains and the E2E service management domain). This entity is also in charge of the communication between management functions and the ZSM framework consumers. Similarly, the *Cross-domain Data Services* are included to provide data persistence across management domains while also allowing processing tasks to run on the stored data as a way to achieve end-to-end global optimization.

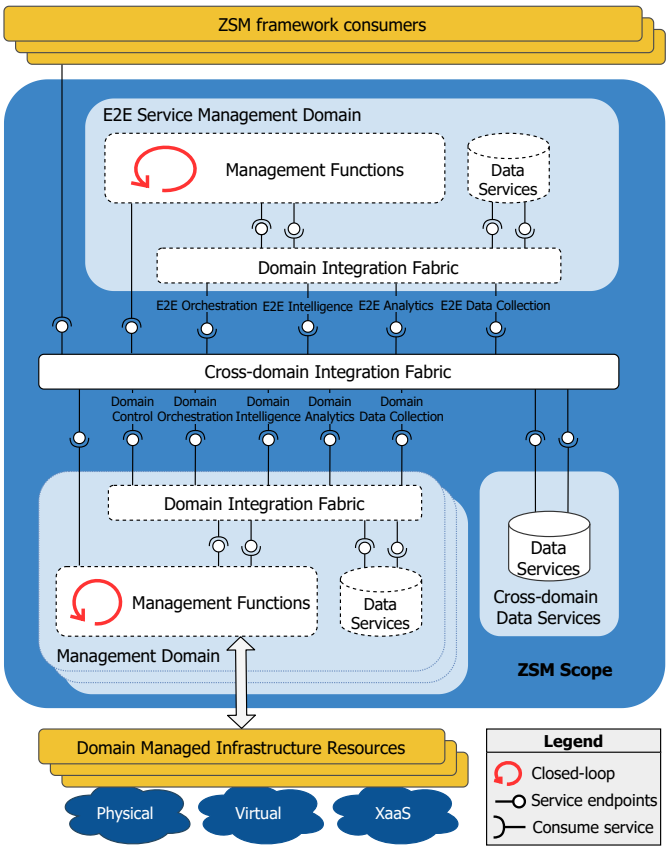


Fig. 2. ZSM reference architecture [5].

As one of its major architecture principles, closed-loop operations are supported at the management domain level and the E2E service management domain level to allow management functions involved in the loop to adapt the behavior of the managed entities [52].

The ZSM architecture follows a model-driven approach, in which managed entities are defined in terms of flexible and reusable information models that specify the entity's attributes and supported operations. This way, managed resources or services can be described independently of their implementation, avoiding vendor-locking and allowing the system to be domain and technology agnostic.

Overall, the ZSM holistic management framework focuses on automation techniques to enable full automation of E2E service management over different technology and administrative domains. This is supported through the use of stateless management functions that inter-work across various domain-specific network management systems.

B. Other Related Standards

The need for leveraging autonomous and automated network and service management solutions is not only gaining a lot of interest in the research community, but also driving notable

standardization efforts within several Standard Development Organizations (SDOs). In addition to ETSI ZSM, other standardization groups are also working towards the consolidation of networks and services automation. Next, the most relevant standardization approaches, as well as their interlink with ZSM in terms of complementarity and overlapping, are outlined [53].

The **ETSI Experiential Networked Intelligence (ENI)** ISG focuses on the design of a CNM architecture based on AI/ML techniques and context-aware policies to realize an effective and adaptive service delivery experience [54]. By enabling agile service optimization based on changing user requirements, service context, and business goals, ENI aims to improve the full management cycle of 5G networks (i.e., provision, operation, and assurance), with particular emphasis on known 5G challenges, namely slice management and resource orchestration. Designed to be agnostic from the managed system specifics, the proposed architecture, shown in Fig. 3, acts as an external functional block and interacts with every layer of the assisted Management and Orchestration (MANO) system to (i) retrieve all the available information (e.g., network status, associated metrics, faults, and errors); and (ii) generate outcomes in the form of recommendations and/or commands (according to the desired operational mode) [55]. To achieve this, the ETSI ENI system architecture consists of three main functional blocks, namely: input processor, analyzer, and output generator. Moreover, an API broker is optionally considered to ensure the integration and format translation between the ENI system and heterogeneous managed entities, which include applications, end-users, orchestrators, and network infrastructure, among others. Compared to ZSM, both architectures offer a framework supporting the use of analytics techniques to provide intelligent management. However, in contrast to ZSM, the scope of ENI focuses on a single administrative domain, and therefore it assumes a centralized system for data collection, processing, and policy generation. As stated in [53], ZSM can leverage the outputs of ENI in terms of AI/ML algorithms, intent policies and Service Level Agreement (SLA) management to foster service assurance and service intelligence capabilities in cross-domain scenarios.

In **3GPP, the Service and System Aspects (SA)** Technical Specification Group (TSG), responsible for specifying the overall architecture (SA2) and service management capabilities (SA5), is committed to enhancing telecom networks operation with a strong focus on network automation. The 5G SBA defined by **3GPP SA2** helps to achieve service modularity and reusability while providing guidelines to improve deployment agility, network slicing flexibility, and Network Function (NF) reliability [56]. In particular, the Network Data Analytics Function (NWDAF), defined as part of the 5G core [57], assists other NFs with slice-specific data collection and analytics services based on a request/subscription model. With this module, consuming NFs is endowed with powerful real-time AI/ML-driven analytics concerning several use cases (such as anomaly detection, network load performance computation, and future load prediction) to perform a zero-touch, dynamic, and proactive network management [58]. In parallel, **3GPP SA5** further investigates solutions for better support

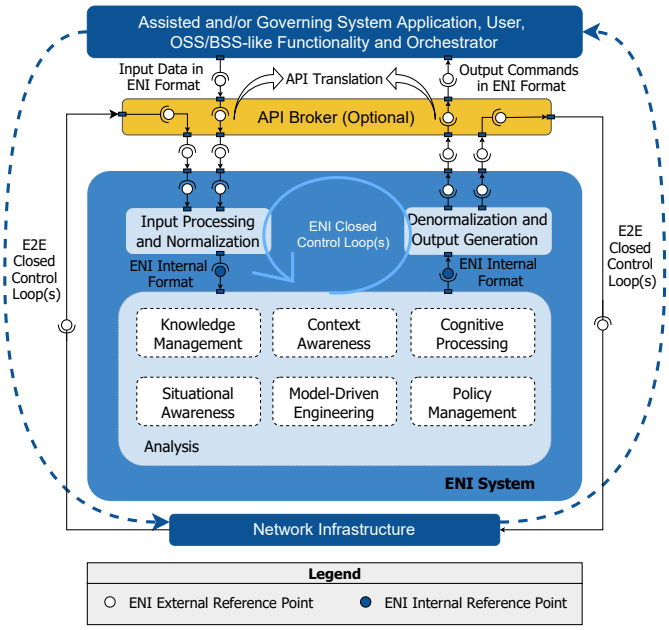


Fig. 3. ENI reference architecture [55].

automation on performance management [59], fault management [60], configuration management [61], network policy management [62], intent-driven management [63], SON [64], and network slice management and orchestration [65]. While 3GPP specifications are intrinsically meant for cellular (core and radio access) networks, ZSM scope assumes generic management systems not tied to any particular network type. Nonetheless, both frameworks are fully aligned in the proposal of SBA architectures to leverage the potential of key features, such as modularity, deployment agility, independence, and reusability of services, as important means to better support automation and high-reliability principles.

Moving from single-domain management solutions, in the arena of service management and orchestration across technology and/or administration domains, there are other organizations relevant to the work in ZSM. Such is the case of the **Tele Management Forum (TMF) Zero-touch Orchestration, Operations and Management (ZOOM)** project that aims to provide a generic and open management solution for virtualized networks and services through automated provisioning, configuration, and assurance [66]. The ultimate ZOOM's vision is to define the principles and guidelines for the new generation of service provider support systems to increase service delivery agility and management effectiveness as main drivers for new business opportunities. The main guiding principles of the proposed management architecture include near real-time requests execution without human intervention, open standard Application Programming Interfaces (APIs), closed-loop control, and end-to-end management [67]. In particular, the support for open standard interfaces is a key feature that identifies and strengthens the adoption path of both TMF and ZSM solutions. In particular, the TMF suite, defining more than 50 API specifications, is supported by most

Tier-1 operators for seamless end-to-end service management underpinned by standards-based interoperability principles.

The **Metro Ethernet Forum (MEF) 3.0 Transformational Global Services Framework** aims to define, deliver, and certify agile and assured communication services orchestrated across a global ecosystem of automated networks [68]. The MEF 3.0 service family includes dynamic Optical Transport, Carrier Ethernet, Internet Protocol (IP), and other virtualized services, all of them orchestrated over programmable networks using the open Lifecycle Service Orchestration (LSO) [69]. MEF LSO specifications enable full-service lifecycle automation for coordinated management and control across multiple technology boundaries within a provider network, as well as across multiple provider domains responsible for delivering an end-to-end orchestrated service. In particular, within the LSO reference architecture, Sonata APIs are used to support zero-touch/automated business-to-business interactions between multiple service providers [70]. In addition to addressing the automation challenge of network and service management across multiple technological and/or administrative domains, MEF LSO and ZSM are both aligned in the pursuit of architecture simplification. To achieve this, the proposed architectures allow the composition and use of individual components as needed, reducing the complexity of deployment blueprints and ensuring scalability.

V. Enablers for Zero-Touch Management

The concepts and architectural components presented in the previous section demand a set of technologies to make ZTM a reality for 5G and beyond networks. This section provides an analysis of these potential technological enablers, including programmatic control loops and management, virtualization and orchestration, data analytics and closed-loop control, AI techniques and AI playgrounds in the networking domain.

A. Programmatic Control and Management

In this subsection, we identify the common aspects concerning programmatic control and management of wireless and mobile networks. Moreover, we also provide some pointers toward popular toolkits and open-source platforms in this domain.

Programmatic control of wireless and mobile networks requires identifying how network resources are exposed (and represented) to software modules written by developers, and how software modules can affect the network state. Although OpenFlow [71] provides a practical forwarding abstraction for packet-switched networks, it has been argued that programming current networks using OpenFlow is equivalent to programming applications in assembler, i.e., the interface is too low-level and exposes too many implementation details to the programmers. As a result, we have witnessed a plethora of efforts aimed at providing developers with higher-level interfaces to their SDN [72]–[78].

More in general, the fundamental SDN principles call for a consolidation of network management functions at a *logically* centralized controller, where a user-defined control application can operate on the global network view maintained by the controller itself. Such concepts have also been brought to the mobile networking domain under the generic term Software-Defined Radio Access Network (SD-RAN). Fig. 4 depicts a reference SD-RAN architecture. At the bottom of the figure, we find the radio access segment composed of various radio access elements. These can range from Wi-Fi Access Points (APs) to 4G/5G eNodeBs/gNodeBs. Event-driven simulators such as ns3 or mininet can also be used for development and debugging purposes. A southbound interface enables communication between the RAN and the actual SD-RAN controller. The latter is where the intelligence of the network resides. It typically consists of a virtualization layer in charge of abstracting the control applications from the underlying radio access technology as well as enabling network slicing, a radio information base maintaining a global view of the network, and then an application layer where the various control and management applications are found. The task performed by such applications can range from radio resource management to mobility management. A northbound interface enables communication with the Orchestration and Automation layer, where entities, such as Business Support System (BSS)/Operations and Support System (OSS) and Network Management System (NMS), reside.

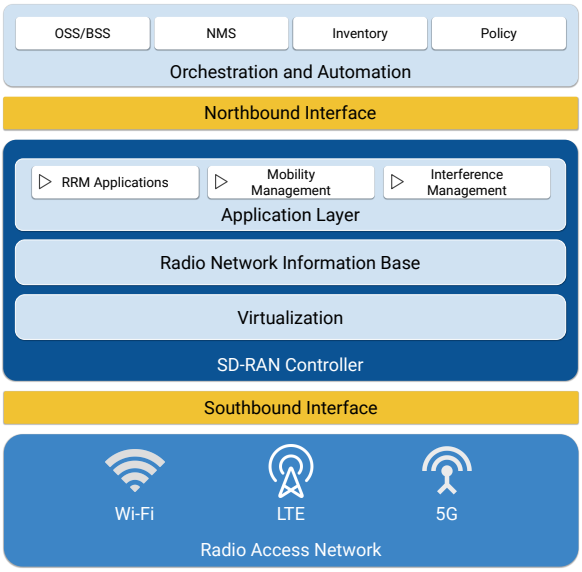


Fig. 4. SD-RAN reference architecture.

In [79], the authors draw a clear line between network control and network management. The former (control) deals with fast timescale operations executed by the elements at the edges of the network, such as scheduling and transmission rate adaptation. The latter (management) is in charge of checking whether the operating conditions for a certain policy are still met and, if not, of re-configuring or replacing the policy. For example, a given scheduling algorithm could be optimized for a uniform distribution of clients across the sectors of a mobile cell. This work resulted in the *5G-EmPOWER* platform [80], which has been used for several resource allocation problems in wireless and mobile networks [81]–[86].

In principle, the work presented by the authors of [87] is similar to [79]. In their paper, the authors introduce the Light Virtual Access Point (LVAP) abstraction. This programming primitive allows network developers to treat a handover in an 802.11-based network as if it was a virtual machine migration. The resulting system, named *Odin*, has then been used by several other researchers for their work [83], [88]–[91].

Other works that conceptually fall in this category are CoAP [92], AeroFlux [93], and Lego-Fi [94].

CoAP [92] is a vendor-neutral cloud-based centralized framework that can be used to configure, co-ordinate and manage individual home APs using an open API implemented over the OpenFlow protocol. As opposed to the proprietary cloud-based AP management solutions, which do not attempt to enable cooperation among neighboring APs, the CoAP API can be effectively implemented by AP vendors and leveraged by a third-party controller to supervise the operation of all the APs in a given building. In their evaluation, the authors show how by using CoAP it is possible to deliver a significant reduction in airtime utilization in a group of AP (due to reduced contention).

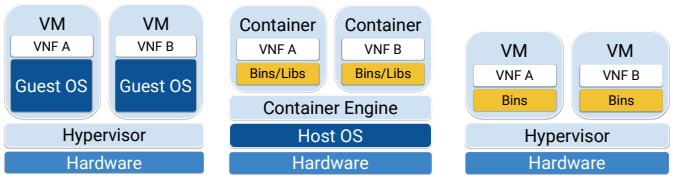
AeroFlux [93] is an SDN-based solution providing large enterprise and carrier Wi-Fi deployments with low-latency programmatic control of fine-grained WiFi-specific transmission settings. The main innovation provided by AeroFlux lies in its two-tier design, which handles non-time critical, global events at a centralized controller and time-critical, local events at a near-sighted controller co-located with the AP.

Lego-Fi [94] aims at functionally decomposing a Wi-Fi AP into smaller functional blocks that are then allocated and composed dynamically, e.g., during a failover or a scale-out. To achieve this goal, Lego-Fi introduces the concept of Light Virtual Network Function (LVNF), a programming abstraction allowing network programmers to implement complex services by composing several elementary packet processing blocks, (i.e., the LVNFs), into a more complex packet processing pipeline. In their paper, the authors show how this solution can be applied to several relevant scenarios, such as packet deduplication, probabilistic load balancing, and mobility.

The complexity of current wireless and mobile networks has been acknowledged recently also in the SDN domain. The growing consensus is that future Software-Defined Wireless and Mobile Networks must follow an AI-native approach. In [81], the authors introduce the *aiOS* platform precisely intending to enable autonomous management of Software-Define Wireless Local Area Networks (WLANs). The proposal is very well aligned with similar efforts from International Telecommunication Union - Telecommunication Standardization Sectors (ITU-Ts) and from the Open Radio Access Networks (O-RANs) Alliance. The latter, in particular, has released the blueprints for a hierarchical SDN platform for mobile networks fully supporting and complementary to standards promoted by 3GPP and standardization organizations.

B. Virtualization and Orchestration

Virtualization is a technique widely used in Information Technology (IT), and specifically in the cloud domain, to create isolated logical components on top of a standard and shared physical infrastructure. Virtualization makes system design more flexible, agile, and efficient compared to the traditional fixed and hardware-based approaches, and it has also found great applicability in mobile networks [95]. In particular, with the emergence of NFV as a major enabler for 5G networks, which employs virtualization to decouple NFs from the hardware executing them. Virtual Machines (VMs), containers, and unikernels are the three leading virtualization technologies for deployment of NFs as Virtualized Network Functions (VNFs).



(a) VMs (b) Containers (c) Unikernels

Fig. 5. High-level comparison of virtualization technologies.

As illustrated in Fig. 5, VMs are emulated computer systems providing isolated guest Operating Systems (OSs) on top of a shared underlying physical infrastructure that can perform all the functionalities of a real physical computer. Provisioning VMs requires specialized hardware and software techniques to be performed by a hypervisor, a thin software system responsible for the instantiation, coordination, and isolation of guest OSs running on top of shared hardware. The hypervisor provides an abstraction view of the underlying hardware, by which multiple isolated VMs can run on top of a shared physical server without interfering with each other [96]. Containerization is another virtualization technology that is very lightweight, highly portable, and has a very low startup time, making them a fascinating application deployment technique. These benefits can be gained through a shared underlying kernel for all containers. The shared OS kernel among the containers running on the same machine is considered one of the major drawbacks of containers, resulting in some security issues [97]. Unikernels are another virtualization technology that relies on the Library Operating System (LibOS) [98] concept for constructing lightweight, single-purpose, and secure VMs. Unikernels embed NF/application and required OS libraries into a single compiled entity, and eliminate all the OS services, features, and libraries that are not necessary for running an NF/application. Given the limited OS libraries, the attack surface shrinks significantly, leading to better security levels. This technology allows running only one NF/application inside each unikernel instance and cannot be expanded after construction [99].

Regardless of the underlying virtualization technology employed for VNF deployment, NFV technology introduces a new way of service creation, deployment, and management [100]. However, realizing an NFV-based mobile network necessitates having a platform able to model and manage services in an automated manner. Automatically managing, scheduling, and allocating resources to maintain certain SLAs

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2022.3212586

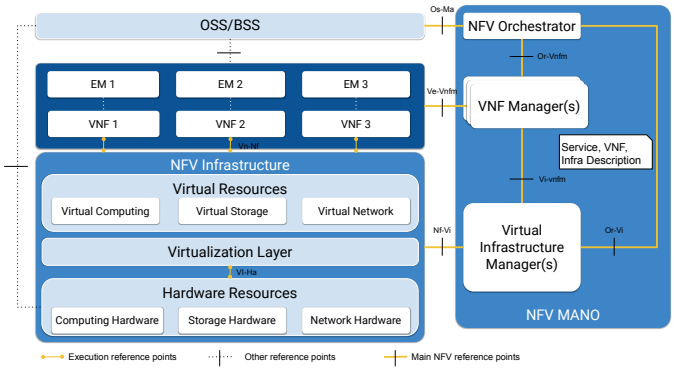


Fig. 6. NFV MANO reference architectural framework.

and ensuring efficient utilization of resources is commonly referred to as *orchestration*. Ensuring a highly available framework that can provide E2E service provisioning requires an efficient system to manage, orchestrate, monitor, and control system components. In this regard, a significant effort from standardization bodies has been devoted to developing an NFV management and orchestration framework. ETSI has realized the high potential of NFV and allocated an ISG to develop an NFV framework called NFV MANO [101].

In the ETSI NFV MANO framework, shown in Fig. 6, VNFs are software implementations of legacy NFs that run on top of the Network Function Virtualization Infrastructure (NFVI), which by itself comprises a virtualization layer that provides virtual resources from a pool of physical resources to make VNFs run on a logically isolated shared infrastructure. NFV MANO is the entity responsible for making sure that services requested by the users are up and running. It includes three modules: Virtual Infrastructure Manager (VIM), NFV Orchestrator (NFVO), and VNF manager. The VIM constitutes the functionalities for controlling and managing the interaction between VNFs with the physical and virtualized resources. The NFVO is responsible for orchestration and management of physical and software components and realizing services on NFVI. Finally, the VNF manager is in charge of the lifecycle management of VNFs. A detailed description of the framework and the reference points can be found in [101].

Recently, a considerable effort from standardization bodies, including ETSI, Next Generation Mobile Networks (NGNM), MEF, Internet Engineering Task Force (IETF), and International Telecommunication Union (ITU) has been devoted to specifying standards and solutions to define the reference architectures, the interfaces, and the protocols for an orchestration platform. Different open source and commercial orchestration solutions have been proposed lately. These solutions can be distinguished by the network environment they can operate on, the number of services they support, implemented NFV MANO functional blocks, and operating domain. Open Source MANO (OSM) [102], Cloudify [103], Open Baton [104], and Open Network Automation Platform (ONAP) [105] are examples of open source orchestration solutions. For a more detailed overview of orchestration solutions, we refer the reader to [100].

C. Data Analytics and Closed-loop Control

Data analytics have been of great importance since the early days of networking. Data analytics refers to the continuous process of extracting data, analyzing, and performing some actions based on the analyzed data aiming to improve system performance. AI/ML approaches for data analytics have been used in 4G networks for network management, anomaly detection, data monetization, and so on. Compared to 4G, 5G is much more complex. Apart from that, previous reactive approaches for data analytics are very limited in terms of the short-term advantages they can achieve. Therefore, data analytics plays a pivotal role, and ML can be leveraged as an excellent proactive tool that can assist Mobile Network Operators (MNOs) in improving their productivity. The increasing complexity in handling a huge volume of data from networking systems has led to the definition of platforms known as Artificial Intelligence for IT Operations (AIOps) [106], which cover the operations to perform data collection, storage, and advanced analytics in an IT context [107], [108].

To achieve the vision of ZTM in both multi-vendor and multi-domain networks, it is important to standardize the interaction between various components of data analytics and closed-loop control. This allows for the efficient creation, execution, and governance of single or multiple closed loops within end-to-end networks and contributes to the adoption of ZTM by the MNOs. Therefore, various SDOs, such as ETSI, 3GPP and O-RAN, are defining standardized functions/services and the corresponding interfaces for data analytics and closed-loop control.

The ETSI ZSM framework supports many data analytics services, covering both single and end-to-end management domains [5]. These management services are categorized into Data Collection, Analytics, and Intelligence services. The domain data collection services monitor the managed entities, consume managed services and provide live performance and fault data to support closed-loop automation, which needs to be able to verify how the network reacts to changes. As part of the closed loops at the management level, domain data collection services interact with domain analytics and domain intelligence services, but also with domain orchestration services and domain control services as needed to trigger actions or changes to the managed entities of the management domain. The domain analytics services provide domain-specific insights and generate domain-specific predictions based on data collected by domain data collection services and data collected by other domains or stored in data services. The domain intelligence services are responsible for driving intelligent closed-loop automation in a domain by supporting variable degrees of automated decision-making and human oversight, with fully autonomous management being the final stage. Intelligence services can be categorized as follows: Decision support, Decision-making, and Action planning. Decision support services enable decision-making via technologies such as AI, ML, and knowledge management. Decision-making is based on information provided by ZTM services defined as part of domain data collection, domain analytics, and domain data services. Action planning defines orchestration/control actions

to be executed by ZTM services defined as part of domain orchestration and domain control.

3GPP specifications support several ZTM enablers such as NWDAF, Management Data Analytics Function (MDAF) and Closed Loop Communication Service Assurance (COSLA).

The NWDAF, first defined in 3GPP TS 23.501 [109], is an entity in the 5G core that collects information about other network functions in the 5G core, e.g., User Plane Function (UPF), Access and Mobility Management Function (AMF), and Session Management Function (SMF), and applies real-time analysis and predictive decisions to help MNOs to manage the network proactively. NWDAF is equipped with an ML engine, incorporates standard interfaces to the 5G SBA, and uses the publish/subscriber method for retrieving the information from the network functions and updating them with the results from the data analysis. In the literature, several works study the capabilities of NWDAF to facilitate the introduction of ML-enabled management and propose network architectures on the road toward ZTM and self-healing systems [110]–[112]. The authors of [110] propose a 5G architecture for vertical industries highlighting an AI-based data analytics module, responsible for the overall management, control, and configuration of the network leveraging NWDAF for the optimization of the 5G core NFs following the modular system architecture of ETSI ENI [113]. Nonetheless, no validation or performance results are provided. The work in [111] focuses on the use of NWDAF to incorporate in the network several AI techniques (linear regression, logistic regression, extreme gradient boosting, Recursive Neural Networks (RNNs), and Long-Short Term Memory (LSTM)) to predict abnormal User Equipment (UE) behavior and network load performance in a specific area.

MDAF, defined in 3GPP TR 28.809 [114], is an entity that provides the capability of processing and analyzing the raw data related to network and service events and status to provide analytics reports (including recommended actions), and enables the necessary actions for network and service operations. The MDAF, together with AI and ML techniques, brings intelligence and automation to network service management and orchestration. These service operations are depicted together with the functionalities enabled by NWDAF in Fig. 7.

Finally, COSLA defined in 3GPP TS 28.535 [115] and depicted in Fig. 8, is a management control loop for communication service assurance that consists of monitoring, analytics, decision, and execution. The adjustment of the resources used for the communication service is completed by the continuous iteration of the steps in a management control loop. The closed control loop for the communication service is deployed in the preparation phase and takes effect during the preparation phase and operation phase of the lifecycle management.

The O-RAN architecture defines two layers of control: the near-realtime RAN Intelligent Controllers (RICs) and the non-real-time RICs. The former operates at the timescale of 100 ms while the latter operates at the level of seconds. Both of them can leverage AI/ML solutions for their operation; however, the non-realtime RIC is envisioned to perform offline network optimization and possibly ML model training, while the near-realtime RIC must execute the inference engine and

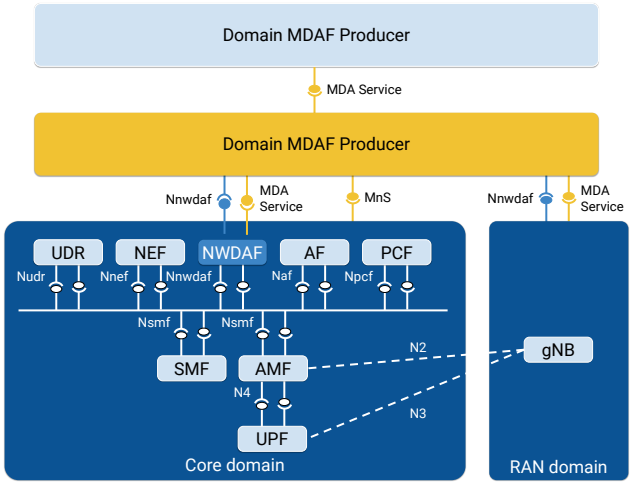


Fig. 7. Layout of the interaction between the NWDAF and the MDA Function and Service for data collection and analysis.

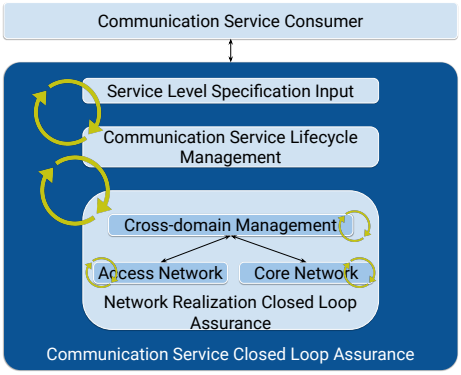


Fig. 8. Layout of the Closed Loop Communication Service Assurance (COSLA) architecture.

perform online optimization. Both layers are expected to gather network data at different spatial and time granularity. A detailed introduction to O-RAN can be found in [116].

The O-RAN architecture introduces three control loops to allow leveraging analytics and data-driven approaches, including advanced AI tools for improved resource management [117]. These control loops, shown in Fig. 9, are placed at different levels, attending to the latency requirements of the operations. The first control loop focuses on the non-real-time RIC and handles management operations via de O1 interface, such as beamforming configurations, requiring a big amount of data. It also supports the deployment, (re)training, and inference of AI/ML-based applications to detect and predict anomalies on a UE level in the near-real-time RIC via the A1 interface. The second control loop is centered on the near-real-time RIC and enables programmatic control operations on the O-Central Units (CUs) and O-Distributed Units (DUs). Moreover, it provides near-realtime monitoring of traffic and radio conditions in the near-real-time RIC and via the E2 and A1 interfaces. Finally, the third control loop enables real-time tasks directly on the O-DUs, e.g., Radio Resource Management (RRM) configuration tasks.

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2022.3212586

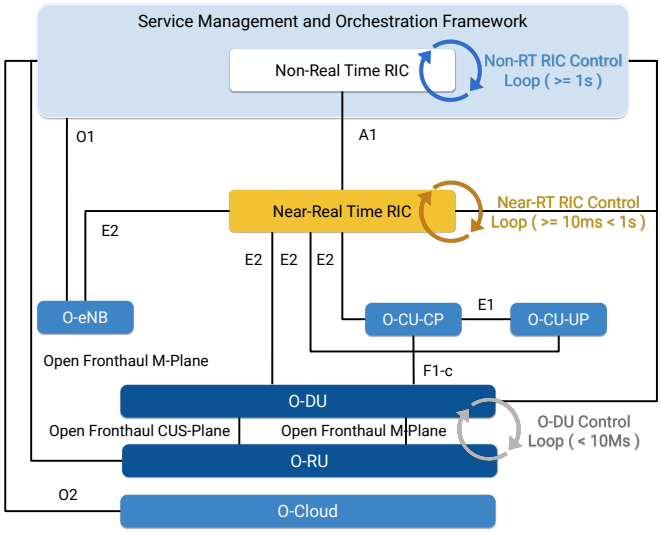


Fig. 9. Control loops in the O-RAN architecture.

The O-RAN technical report in [118] discusses the mapping of AI/ML-related functionalities into the aforementioned loops. Furthermore, it provides an exhaustive analysis of the location of the training and inference of the various ML model types (i.e., SL, UL and RL) according to computational complexity, availability, the quantity of data exchanged, and response time requirements from the model.

D. AI techniques for ZTM

The level of automation expected from ZTM in 5G and in beyond 5G networks requires incorporating innovative end-to-end approaches comprising spectrum sharing, resource scheduling, service orchestration, security, and trust, among others. To achieve this disruptive change, AI has been acknowledged as one of the main enablers in scenarios where the whole networking system is expected to self-manage, self-adapt, and self-react to disruptions and changes with minimal intervention. For a high-level introduction to the most significant AI/ML approaches, we point the reader to [23].

There is no doubt that AI in ZTM and ANM is a double-edged sword. While AI is an unquestionable enabler, it is also very hard to find simple, trustable, explainable, and accurate models that perfectly satisfy the requirements of all the network operations and segments involved in ZTM. For this reason, big research efforts have been devoted to analyzing how the multi-disciplinary techniques of AI, ranging from ML to optimization theory and meta-heuristics, can cover the specific necessities of networking, for instance, in terms of granularity, timing, or data-awareness [119]–[121].

ML algorithms can be broadly classified into three main categories according to their nature, namely SL, UL, and RL. Other complementary categories to these, such as (Deep) Federated Learning (FL) and TL can also be distinguished in the literature. Nonetheless, for a more detailed description of ML taxonomies, we refer the reader to [23], [122]. Below, we provide an overview of how specific ML algorithms from all broad categories are suited to address specific needs in ZTM [119]–[121], [123], [124].

The authors of [119] divide AI functionality into four modules, namely sensing, mining, prediction, and reasoning, and review the techniques found under each category. Moreover, specific examples of networking problems are provided for each AI module, showing the evolution from 4G to intelligent 5G networks. For instance, looking at SL, logistic regression and SVM are categorized under sensing tasks and are a valuable option for anomaly detection. Likewise, mining modules reflect provisioning mechanisms (e.g., bandwidth) and cover decision trees and unsupervised methods such as replicator Neural Networks (NNs) and one-class SVM. The prediction module is dedicated to trend forecasting and comprises techniques, such as RNNs and LSTM. Finally, reasoning tasks refer to the parameter re-configuration for service adaptation and contemplate mainly RL and TL techniques. The work in [120] discusses a similar idea to [119]. Moreover, this work studies the autonomous 5G traffic control from the perspective of SL, UL, and RL, showing the different stages in training and deployment, as well as the strengths and drawbacks of each method. In [121], AI is put as the central element of the intelligent wireless network architecture and the advantages and limitations of the most used AI techniques for intelligent resource management are summarized, including SL (e.g., NNs and SVM), UL (e.g., k-means and Principal Components Analysis (PCA)), RL, DL (e.g., Deep Neural Networks (DNNs)) and Deep Reinforcement Learning (DRL). In particular, the effectiveness of an RL algorithm called deep Q-learning for the joint resource allocation problem covering communication aspects, caching model, and computing model is studied. Finally, the works in [123], [124] argue the role of Explainable Artificial Intelligence (XAI) for 5G and 6G networks to enable ZTM and define it as the ability of AI to clarify itself in human-understandable ways. Given the level of autonomy expected from ZTM, XAI aims to implement algorithms that can be understood and trusted in a technology-agnostic manner. Furthermore, the authors of [123] provide a mapping between the most used XAI approaches to improve explainability and the problem domains (e.g., resource management optimization) and discuss the tradeoff between explainability and performance gain of different AI techniques.

Fig. 10 depicts the AI taxonomy discussed above by classifying the various AI inference tasks according to their nature and networking management areas and entities, together with the most relevant approaches used for each of them.

E. Playgrounds for Building and Testing AI-based Solutions in the Networking Domain

Although AI has proved to fit the requirements of modern network management, there are still issues that slow down the path toward a full ZTM architecture. On the one hand, many networking researchers lack the knowledge and skills to build powerful AI pipelines and integrate them into a data-driven architecture [125]. On the other hand, despite having experience in ML toolkits, there is a clear lack of construction and validation environments specific for networking that enables

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2022.3212586

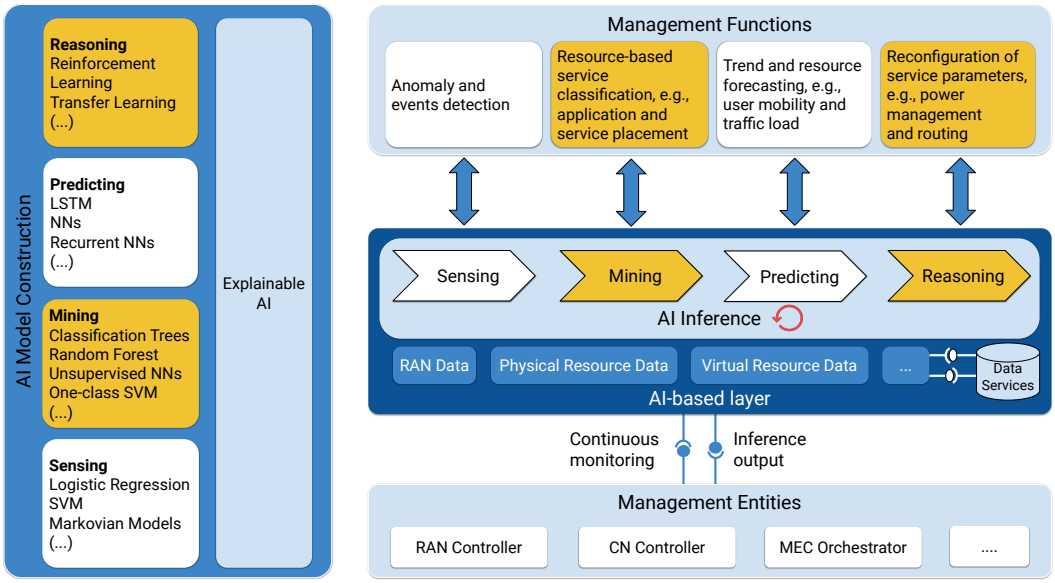


Fig. 10. Classification of AI techniques for ZTM attending to network management functions and requirements.

both a reduction in the time required for testbed setup and a more powerful and precise validation process, allowing to isolate networking from AI issues.

To give a solution to the problem above, the authors of [126] present a playground for AI in networking building on OpenAI Gym. The work, called ns3-gym, embeds the OpenAI Gym framework into the widely-used ns3 simulator to enable fast prototyping of RL-based algorithms (using, for instance, Keras or Tensorflow) and benchmarking. Moreover, emulation environments are also supported. The workflow is demonstrated on an IEEE 802.11-based network to select free-of-interference time slots. Another toolkit called Optical RL-Gym has similar objectives in optical networks [127], assisting in the implementation of RL-based solutions and their validation. Other frameworks, such as Acumos, which is promoted by the Linux Foundation [128], harmonize diverse existing AI solutions (e.g., SciKit Learn, Tensorflow, and H2O) and facilitate their interoperability and the isolation of model training processes over the same dataset without requiring special knowledge on ML. However, this is a general-purpose solution, not specifically designed to solve networking problems. Therefore, after the ML algorithms have been properly assessed via simulation, their validation on a realistic scenario is necessary before they are deployed on production networks.

Concerning experimental testbeds, the work in [129] describes an open-source testbed that can be customized with different network topologies for testing AI solutions in 5G, supporting slicing and service orchestration, and that can be used without specialized hardware. The working model of the testbed is showcased through two use cases, namely RAN slicing, and VNF placement, leveraging DNNs. Although other open-source testbeds can be found, the support for integration with AI pipelines is very limited [130], [131].

F. Summary of Enablers for Zero-touch Management

In Table IV, we provide a summary of the key technology enablers for zero-touch management described for the convenience of the reader.

VI. Network Automation Solutions

Given the dynamic nature of beyond 5G and 6G networks, with changing conditions along the time and space dimensions, static network management solutions are doomed to provide suboptimal performance. In addition, their increasing complexity and the advanced mission-critical services to be supported with stringent performance requirements makes the management of beyond 5G and 6G networks a daunting challenge. As a result, the automation of beyond 5G and 6G networks shall encompass all segments of the network (i.e., radio access, core, cloud and edge, and end-to-end) to guarantee holistic and end-to-end optimized ANM operation. This section overviews the literature on zero-touch management for the main network segments, paying particular attention to the automation and AI/ML technique employed in each of them. Each subsection surveys the latest state-of-the-art contributions across different network segments. The final subsection is devoted to end-to-end approaches. It is worth noting how the topics under each segment have been selected due to their closeness to ZTM and ANM concepts and principles.

A. Radio Access Network

Data-driven and intelligent solutions aim to introduce a degree of autonomy in the networked systems that only require human intervention to oversee or verify the processes. At RAN level, the main areas in which the ZTM vision is exhibiting more outstanding capabilities can be identified as (i) radio resource allocation; (ii) energy-aware radio resource scheduling; (iii) operation, management, and self-deployment; (iv) RAN slicing; and (v) programmatic and distributed control

TABLE IV
SUMMARY OF ENABLERS FOR ZERO-TOUCH MANAGEMENT.

Zero-touch Management Enablers	Summary
Programmatic Control and Management	Programmatic control of wireless and mobile networks identifying how network resources are exposed (and represented) to software modules written by developers, and how such software modules are leveraged to enforce management operations.
Virtualization and Orchestration	Virtualization creates isolated logical components on top of shared physical infrastructure. Orchestration automatically manages, schedules, and allocates resources to maintain SLAs and ensures efficient utilization of resources.
Data Analysis and Closed Loop Control	Data analytics services, covering both single and end-to-end management domains, are categorized into data Collection, analytics, and intelligence services. The domain data collection services monitor the managed entities, consume managed services and provide live performance and fault data to support closed-loop automation, which needs to be able to verify how the network reacts to changes.
AI Techniques for ZTM	AI techniques for scenarios where the whole networking system is expected to self-manage, self-adapt, and self-react to disruptions and changes with minimal intervention. It comprises four modules (sensing, mining, prediction, and reasoning) leveraging on SL, UL, RL, FL and TL AI techniques.
Playgrounds for Building and Testing	Joint ML and networking validation environments that enables both a reduction in the time required for testbed setup and a more powerful and precise validation of ZTM approaches.

loops. More specific capabilities and their relation to the cellular and Wi-Fi protocol stacks are depicted in Fig. 11 and will be discussed throughout this section. In addition to this, Table V conveys the related work associated with the various ZTM topics in radio access networks, grouped by the specific automation technique used. Note that we decided to include also Wi-Fi in the list of surveyed technologies in that it will play an increasingly more important role in 5G and beyond networks, especially due to the work carried out in 3GPP on Multipath TCP [175].

Radio Resource Allocation. In the current literature, there exists a plethora of solutions introducing AI to adapt the allocation of radio resources to the evolving network conditions and user demands [132]–[141]. The main strategies include dynamic rate and channel adaptation, modulation and coding schemes, and power control, among others, to support the growth of the demands in the radio spectrum while ensuring low latency and ultra-high reliability. However, they all have in common the aim to characterize in a (nearly) real-time fashion the communication and channel performance and use it as a basis for various resource allocation purposes.

The authors of [132] formulate a resource allocation problem as a contextual Multi-Armed Bandit (MAB) for mmWaves. In this work, the context is based on data rate requirements, harvested energy and available channel, the arm of the model regards modulation, coding scheme, and transmit power, and the reward is based on a packet success indicator. The paper demonstrates the ability of the online model proposed to deal with the fast-changing nature of mmWaves by exploiting unimodality to reduce the number of arms employed and to use past context information to make better decisions in the path from CNM to ZTM. Similar approaches can also be found in [133], [134]. In comparison with these, the work in [132] exploits both the unimodality and context of the arms, thus reducing the range of context, convergence time, and performance significantly. All these works prove the ability of RL approaches for modeling the rapidly-varying conditions of wireless resources in 4G, 5G, and Wi-Fi networks and better structure resource allocation strategies.

The work in [135] proposes a DRL algorithm to optimize resource allocation and bitrate selection in 5G networks with the ultimate goal of improving video QoE. The problem is formulated as a Markov decision process using a multi-agent actor-critic DRL model and leveraging an extra LSTM-based NN to predict changes in the channel quality of the wireless clients over time. The performance analysis via simulation demonstrates the effectiveness w.r.t. the related work to allocate network resources and avoid drastic variability in the quality perceived. A distributed version of DRL is also used by the authors of [139], who propose a resource allocation scheme pursuing low latency, improved fairness, and greater resiliency for the control and management of resilient microgrids that leverage small cell networks. The results showcase greater network agility and self-organization w.r.t. state of the art, while higher throughput and fairness are achieved. The resource allocation approach in [141] also exploits DRL for the autonomous management of network slicing engines regarding bandwidth allocation according to QoS requirements and IP shuffling. Additionally, this paper presents enhanced resiliency of the RL agents by introducing an anomaly detection mechanism with a memory-based technique, avoiding attackers manipulating the training information. A combination of LSTM and Convolutional Neural Networks (CNNs) are used in [137] for the same purpose of looking at the prediction of the Channel State Information (CSI). Different from the rest, an offline-online training mechanism is presented to diminish the lack of stability of the CSI in such a manner that the neural networks are built first offline and then progressively updated with recently measured CSI. The solution is validated via simulation against an offline Artificial Neural Network (ANN) approach in several scenarios, such as free space, outdoor, and within a building, showing the greater accuracy obtained when updating the NNs' weights. Different from the rest, the work in [138] advocates for more interpretable ML approaches, such as Random Forest (RF) and rule-based models, to improve the use of resources in Wi-Fi networks by optimizing traffic packet size length. The models are built taking as a basis the channel

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2022.3212586

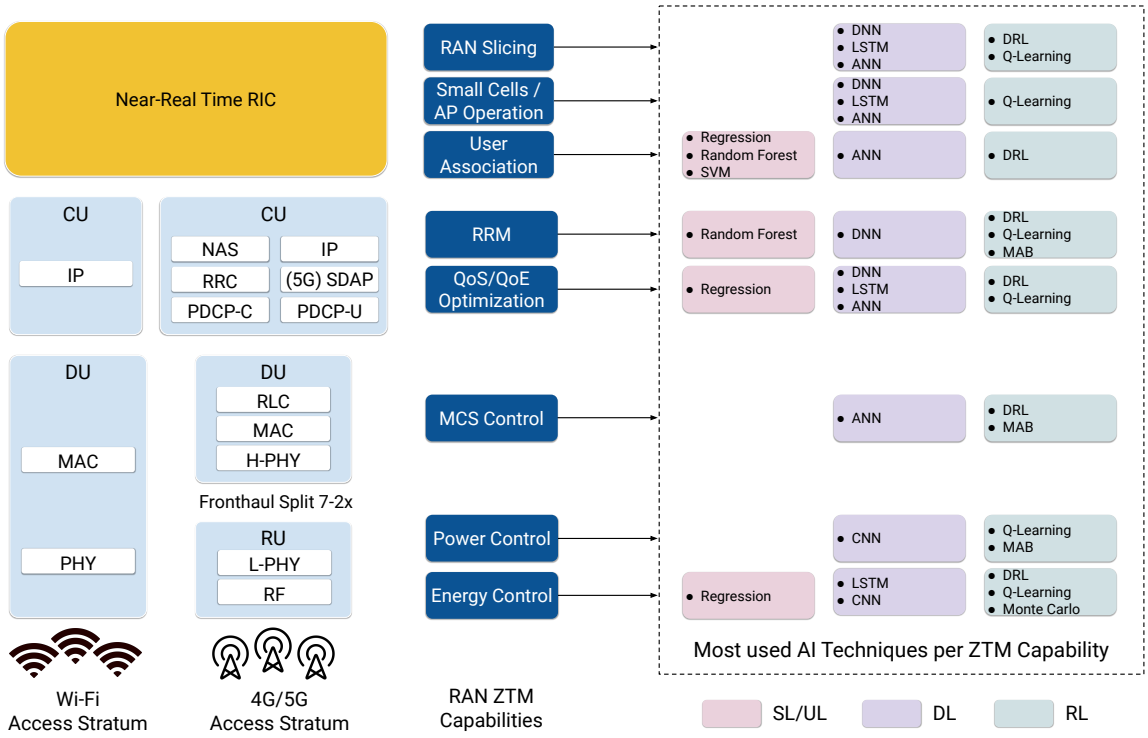


Fig. 11. Main AI techniques used to fulfill ZTM capabilities for cellular and Wi-Fi stacks.

occupancy and the channel status of the users. Experimental results on an open-source testbed [80] demonstrate the high accuracy of random forest models and the ability to tackle problems with high bias.

Resource management is also studied for multicast applications in several works [136], [140]. The authors of [136] propose an LSTM model to construct a dynamic traffic model of multicast services for TV multimedia applications in 5G. Then, taking this forecasting as a reference, a DRL model solves the resource allocation problem for multicast and unicast traffic considering energy efficiency aspects through a DNN. The solution is compared against Auto-Regressive Integrated Moving Average (ARIMA) and Holt-Winters models, demonstrating that although it decreases with a high number of concurrent users, the presented model provides higher precision and accuracy. A similar vision is presented in [140], which introduces a DRL model for unicast-multicast resource allocation that pursues the reduction of the energy consumption in the Remote Radio Heads (RRHs). The solution is validated in a small scenario via simulation and compared with other non-ML approaches, such as fully unicast transmission and simple access decision.

Energy-aware Resource Scheduling. In addition to high data rates and large coverage capacity, enhanced energy consumption is essential in 5G RANs. In this respect, making the best use of the resources at the radio level becomes indispensable in ZTM. Consequently, several approaches for radio resource optimization and QoS assurance also pursue the objective of maximizing energy efficiency. It is worth mentioning that given the non-linear nature of the problem, most of the existing works rely on RL and DRL to provide an efficient solution in real-time. Nonetheless, a few DL-based models can be also found in the literature [142]–[149].

In terms of QoS assurance, the authors of [142] use several ML models trained with traffic and energy production patterns in operational Long Term Evolution (LTE) networks. The paper presents zero-touch resource management strategies of the considered RAN portions to deliver a trade-off between energy consumption and QoS provided by the resources allocated. The analysis of the ML models built (i.e., linear regressor, ANN with different layers/neurons compositions and LSTM) demonstrates the greater ability of the simplest ones to deliver a high accuracy while reducing computational cost and simplifying training data collection. RL algorithms have also been the option adopted in several HetNets deployments. The work in [143] presents a Q-learning solution to manage power allocation on a multi-agent network. The proposed reward function allows for reducing interference and keeping the capacity of the cells beyond a certain threshold, while maximizing user fairness. The model is evaluated via simulation with prior works using the same predictive model, but is not designed for a dense network [144] and does not ensure a minimum cell capacity in the reward function. By contrast, the authors of [145] are in favor of CNNs to maximize energy efficiency in HetNets by provisioning decisions on subchannel and power allocation. The training data is generated via simulation to build a CNN whose input layer collects the channel gains from each user, subchannel, and base station tuple. Numerical analyses demonstrate similar results as analytical methods while reducing computational time. Nevertheless, the performance of the CNN is not compared against other related works or learning types. This is also a topic of discussion in vehicular

TABLE V
RAN AUTOMATION SOLUTIONS.

Topics	Automation technique	Ref.
Radio Resource Allocation	MAB	[132]
		[133]
		[134]
	Multi-agent DRL, LSTM	[135]
		[136]
	LSTM, CNN	[137]
	Random forest	[138]
	DRL	[139]
		[140]
		[141]
Energy-aware Resource Scheduling	Linear regressor, ANN, LSTM	[142]
	Q-learning	[143]
		[144]
	CNN	[145]
	CNN, DNN, LSTM	[146]
	LSTM	[147]
	Monte Carlo RL	[148]
	DRL	[149]
Operation, Management and Self-deployment	NN	[150]
	Deep Q-learning	[151]
	LSTM	[152]
		[153]
		[154]
	SVM	[155]
		[156]
	ANN, decision trees, random forest	[157]
RAN Slicing	ANN	[158]
	Q-learning	[159]
		[160]
	Deep Q-learning	[161]
	LSTM	[162]
		[163]
	RL	[164]
		[165]
	DRL	[166]
	DNN	[167]
	Blockchain	[168]
		[169]
Programmatic Control Loops	DRL	[170]
	ANR	[171]
	Random Forest	[172]
	Multi-agent team learning	[173]
	CNN	[174]

networks. The paper in [146] uses a mininet-wifi-based setup to deploy a Vehicular Ad-hoc Network (VANET) and employs a set of algorithms such as CNNs, DNNs, and LSTM to adjust the bandwidth of the transmission queues to satisfy QoS requirements. Similar to [142], the authors of [146] evaluate the performance and accuracy of various ML approaches in the path toward zero-touch network management, showing, in this case, the ability of LSTM to provide the highest accuracy at the cost of a very reduced increase in reaction time.

Driven by the increasing data needs, the concept of cloud and fog RAN has been defined to form a centralized architecture composed of the Base-Band Unit (BBU) pool, the Remote Radio Units (RRUs) and the fronthaul. This architecture enables the centralization of the BBU computational and storage pools at the central office, while the RRUs are distributed, thus bringing larger cost-efficient deployments. In this respect, the work in [147] seeks efficient resource and energy utilization in Fog RANs while satisfying the low-latency requirements of Internet of Things (IoT) applications employing an RL algorithm based on the Monte Carlo method. The reward function is based on user utility, which is characterized by the Signal to Noise Ratio (SNR) through the Shannon channel capacity. Moreover, the reward considers a penalty to encourage less idle time of small cells and improve energy efficiency. Similar to this, in [148] a two-step DRL model is introduced for dynamic resource allocation in Cloud-Radio Access Network (C-RAN) minimizing, by contrast, power consumption, and satisfying user demands. The proposed agent first determines the set of RRHs to be turned on/off to reduce the search space, and then derives the optimal resource allocation based on beamforming. The DRL model, validated via simulation, is composed of an offline DNN (that estimates states and actions) and an online deep Q-learning algorithm that chooses the action with the highest reward. No prior related work attempted to apply DRL to dynamic resource allocation in cloud RANs. Finally, the authors of [149] present a C-RAN architecture aiming to automatically configure and affiliate RAN units to edge and clouds nodes (including Radio Unit (RU), DU, and CU interdependencies) following the Zero-touch Commissioning (ZTC) model in both a distributed and a centralized approach to meet the sustainability goals expected in 6G. Modules for resource discovery as well as automated helm charts creation and deployment launching are introduced to make ZTC possible. In contrast to the works discussed above, the architecture is, in this case, hosted on a testbed to evaluate deployment time, infrastructure health, and service health, and provides service discovery features, which are absent from the rest of the papers.

Operation, Management and Self-deployment. AI-based RAN management solutions can be found from both academia and industry perspectives, dealing with aspects related to ZTM management of small cells and APs, including activation and user association, self-deployment, and optimized configuration [150]–[157].

Regarding user association management, the paper in [150] presents a DL-based algorithm for dense networks aiming to intelligently associate UEs to macro and small cells. The proposed NN is composed of four convolution blocks and is trained using a synthetic dataset. The dataset considers features such as the throughput of the cells and the current association matrix. The model is validated via simulation against a genetic algorithm, showing how the computation time of the DL model scales with the number of layers and improves with hardware acceleration. A deep Q-network strategy to solve this same problem is also sought in HetNets in [151]. Similar to [147], the QoS requirements are mapped to channel SNR levels using the Shannon capacity, which is used to derive the UE's utility and the reward function reflecting if the choice made is enough to guarantee the minimum QoS. As opposed to the previous work, the approach in [151] uses a double deep Q-network approach to achieve its goal. The accuracy of the

model is validated via simulation vs. other RL methods, i.e., Multi-Agent Q-learning and multi-agent RL in mobile and static scenarios, comparing different neurons constructions, learning rates, and optimization strategies, showing faster convergence speed and better generalization ability than the baseline models.

Closely related to user association approaches are the solutions that employ ML to predict and optimize handover processes [152], [153]. The work in [152] builds a stacked residual NN to predict traffic demands in cellular base stations and cells, and incorporates fine-grained handover information. This information is captured on a graph at the RAN based on the handover frequencies of the cells, and then it is fed to the model to account for the traffic characteristics. The proposed model is compared with state-of-the-art solutions using historical averages such as ARIMA and LSTM models, showing that, although the Root Mean Square Error (RMSE) and the Mean Square Error (MSE) increase with the period ahead of the prediction, the model can reduce the handover frequency. Different from the rest, LSTM is exploited in [154] for multi-task learning (also covering incremental learning) pursuing handover management, and initial Modulation Coding Scheme (MCS) selection in 5G networks with the ultimate aim of optimizing operational efficiency and autonomy in the RAN management. Similar works, such as ABRAHAM [153], can also be found in Wi-Fi networks. This paper presents a proactive ML-based handover algorithm building on [80] aiming to introduce zero-touch approaches able to optimize network-wide load while preserving QoS. The authors take as input several features, such as users' Received Signal Strength Indicator (RSSI) and throughput, association matrix, APs' load and location of users and APs, and use them to build a three-layer LSTM model to first predict expected RSSI for all AP-user pairs and then condense the output on possible handover decisions. The experimental analysis demonstrates the low MSE, and improved latency and throughput of the algorithm in different traffic and mobility scenarios in comparison with the standard and the maximum RSSI handover algorithms.

Improved agility in resource-constrained environments and user experience are also the motivations of multiple studies found in Wi-Fi networks, which intend to facilitate self-deployment in indoor environments building on semi-supervised algorithms [155] and SVM [156]. Both works present an AI-driven approach leveraging sensing and perception to collect measurements from user devices (e.g., RSSI, transmission bit rate, received/transmitted data, etc.) that are then used to explore optimal self-deployments. In particular, in [156], the aforementioned measurements are used in the context of cognitive networking to estimate performance indicators that are then fed on a semi-supervised SVM to create location heat-map. Conversely, the authors of [157] present a set of classification and regression models to predict the utilization of the APs in a Wi-Fi network and proactively activate/deactivate the APs to reduce the energy consumed. Similar to [138], the dataset is collected using off-the-shelf APs running OpenWRT. In particular, algorithms, such as multi-label ANNs, decision trees, and RFs are used in this work. The experimental evaluation studies the seasonality and the importance of using models able to reflect the time-series nature of the problem, reporting a greater accuracy and adaptability to changes from the RF model, as it is also the conclusion of [138].

RAN Slicing. Given the fluctuating nature of the radio resources and the need to deal with multiplexing gains and isolation, RAN slicing can be even more challenging than Core Network (CN) and Transport Network (TN) slicing. As a matter of fact, a huge body of literature has appeared in recent years to provide adaptive radio resource management to satisfy the SLAs of the RAN slices [158]–[169].

RAN slice orchestrators have been taken into account in several solutions as an entity that is able to manage RAN resources, orchestrate slice lifecycle management, and be the entry point for service requirements and slices' SLA specifications [158]–[160]. In this respect, the authors of [158] envision a cellular network managed by a RAN controller and supporting three types of slices (i.e., best effort, constant bit rate, and minimum bit rate) with distinct and adjustable parameters, such as minimum bitrate and admission rate. On this basis, the paper presents a scheduler based on a 3-layer ANN that takes as input network conditions (i.e., SNR and arrival rate of the slices) and control parameters (i.e., slice scheduler weights and resource ratio) to optimize the slice parameters and supervise the admission control. The performance evaluation does not comprise another reference technique but shows the ability of the model to adapt control parameters to respect slices' SLAs regardless of the network load. In [159], a RAN controller is also in charge of allocating Physical Resource Blocks (PRBs) to slices based on the directives, in this case, provided by a DRL algorithm. In particular, the paper introduces a distributed version of a Q-learning algorithm (a Dueling Network first presented in [160]) composed of a learner, multiple actors (that allows distributively collecting various environmental states), and replay memory. This algorithm performs autonomous PRBs allocation attending to slice requirements and SLA, e.g., a high-throughput slice may require a periodic number of PRBs, while a low-latency slice could advocate for fewer PRBs but allocated at the time at which the request arrives. The simulation-based evaluation compares the performance of the model in comparison with the work in [176] (using a Deep Q Learning (DQN) model) and with other hard slicing strategies. The results show a 20% improvement in the slice requirements satisfaction w.r.t the remaining approaches. However, it does not give clear light to the scalability of the method.

Virtualization and preemption techniques enable a greater elasticity in slice management by seamlessly allowing the sharing of unused resources based on long-term demand predictions [161]–[164], [166]. The work in [161] presents a two-level framework for dynamic radio resource virtualization and slice allocation in 5G aligned with the ZTM concepts. The problem is formulated as a Markov Decision Process (MDP) and leverages deep Q-learning using QoS utility as a reward function to allow allocating the unused resources from a set of slices to the ones that may need extra resources until it is predicted that the slice may experience congestion. The scheme is evaluated via simulation w.r.t. various state-of-the-art solutions not leveraging ML techniques [177], [178],

showing a greater slice requirement satisfaction while minimizing the resources used. The same objective is pursued by the authors of [162], which present a modified LSTM model, called X-LSTM, trained on a custom-designed experimental LTE testbed to improve prediction accuracy tens of seconds ahead. The QoS required by the slices is emulated using the QoS Class Index (QCI). The model introduces notions from ARIMA and the X-11 statistical method, allowing the decomposition of time series data into seasonal data patterns, and uses multiple LSTMs for different time scales. Experimental results demonstrate the ability to reduce the amount of over-provisioned PRBs and SLA violations w.r.t. ARIMA and LSTM. This idea is improved in [164], which keeps the identification of services through the QCI but also enables multi-class slices. In particular, a slice broker aligned with 3GPP for advanced RAN sharing is presented for admission control and traffic forecasting. The RL model presented builds on predefined SLAs of the tenants, traffic usage patterns, and user distribution. The performance evaluation via simulation demonstrates the effectiveness of the solution as the number of requests and system load increase for regular and irregular network slice requests. A similar concept for vehicular networks is proposed by the authors of [163], where a Road Side Unit (RSU) virtualization layer enables the functions of slice management and coordination. Similar to [147], the QoS requirements of the slices are related to SNR values using the Shannon capacity. On this basis, an LSTM-based allocation algorithm is proposed for RAN slicing aiming to first enable traffic prediction and then use it to match vehicles' mobility and minimize system delay. A dataset from a real operational network is used for training and numerical validation, demonstrating the ability to significantly reduce the total delay. Building on the OpenAI Gym toolkit, a set of DRL models is designed by the authors of [166] to enable AI-driven zero-touch network slicing in C-RAN scenarios to reconfigure compute processes while minimizing VNF instantiation cost. The work is evaluated together with [179] and [180], reducing average delay and energy consumption for the same slice admission ratios w.r.t the other DRL benchmarks.

Slice admission control is studied from a revenue point of view by several works [165], [167]. The authors of [165] introduce an RL-based slice admission algorithm for 5G that aims to maximize infrastructure providers' revenue (i.e., accepting as many high revenue requests as possible) while minimizing penalty (i.e., rejecting the slices whose revenue is expected lower than its rejected penalty or that would cause service degradation). Although the work focuses on RAN slicing, the architecture considers a cross-domain orchestrator covering radio, transport, and cloud that processes admission requests for two types of slices (including data, such as holding time, service priority, resources required, etc.). The RL algorithm also takes as input features the monitoring information from the orchestrator and uses the loss of revenue as a reward function to estimate the probability of accepting a given request. The evaluation presented compares the proposed approach with a static and a threshold-based heuristic, showing a remarkable profit increase. Similarly, the approach introduced in [167] builds on a DNN to incorporate Operational Expeditures (OPEX) control by integrating constraints on OPEX violation rates in the training dataset that can act as hyperparameters of the model. Moreover, input features, such as aggregated throughput, users per cell, and occupied PRBs, have been considered to build a four-layer DNN considering three types of slices (enhanced Mobile Broad Band (eMBB), social media and browsing) with different traffic demands. The functional validation via simulation shows the resources' evolution over time together with the OPEX distribution without adding any comparative strategy.

Finally, the role of blockchain in the accuracy of 5G network slice brokers has also been studied to satisfy all users' diversified requirements and enable more efficient collaborative and business ecosystems [168], [169]. This role is acknowledged by the work in [168], which exploits blockchain to build a network slice broker that reduces service creation time through a smart contract system. The difference w.r.t. other works resides in the fact that the broker takes care of on-demand resource allocation, admission control based on traffic forecasting and RAN scheduling, and managing slices that are characterized by QCI (in terms of priority, packet delay, and packet loss) and associated billing. Conversely, the solution presented in [169] aims to take a step beyond the previous works by not only proposing a consensus protocol for decentralized network slice systems but also improving the system performance of the services in a federated environment able to work autonomously. To this end, the protocol determines the same sequence of slice trading transactions for each miner before generating each block to ensure block generation efficiency and consensus robustness. The paper analyses the performance with respect to a basic Proof-of-Majority consensus protocol, showing a 50% latency reduction and 30% throughput improvement on the same slice trading transactions. Nevertheless, the proposed method experiences considerable scalability issues.

Programmatic and Distributed Control Loops. The Programmatic control of the network is essential to enable automation and zero-touch management. In this regard, the O-RAN initiative is one of the main representatives in building several distributed control loops based on open interfaces and modularization, and one of the major contributors to their standardization. With the definition of the O-RAN architecture and the AI/ML control loops presented in Section V-C, the challenges related to making the open, virtualized, and data-driven vision a reality has become a topic of discussion in both academia and industry [170]–[174].

The authors of [170] investigate the limitation of the current O-RAN specification and analyze the deployment of data-driven policies at different control loops on a wireless network emulator named Colosseum [181], which enables the generation of massive datasets to fulfill this purpose. In particular, the work aims to show the feasibility of O-RAN control loops by implementing a DRL agent to select scheduling policies on RAN slices (i.e., round-robin, proportionally fair, and waterfalling) running as a xApp on the near-real-time RIC. The experimental scenario comprises base stations and UEs implemented through srsLTE, while the traffic of the various slices, i.e., eMBB, Ultra Reliable Low Latency

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2022.3212586

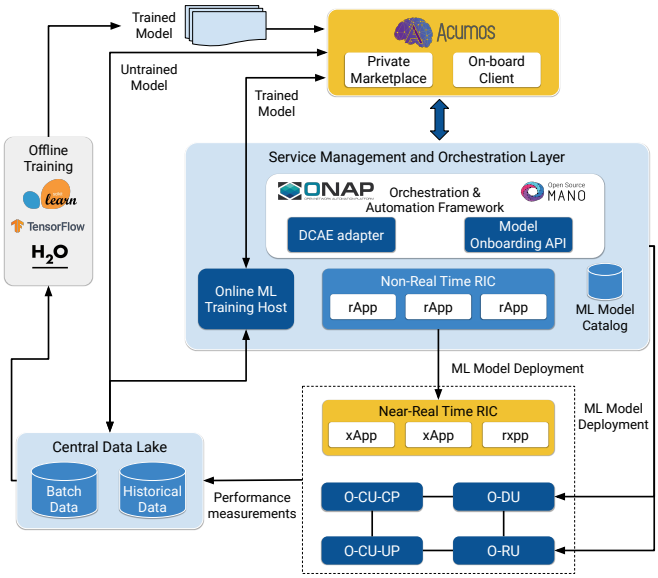


Fig. 12. O-RAN ML Model Lifecycle Implementation Example with ONAP and Acumos.

Communications (URLLC) and massive Machine Type Communications (mMTC), is generated with a version of Coliseum implementing the near-real-time RIC control loop. Moreover, a DRL agent per each slice/base station runs in parallel in several xApps. The comparison with the static scheduling strategies demonstrates considerable gains in spectral efficiency, although it does not cover the computational cost of the xApps at the various control loops and the overhead on the E2 interface from the distributed nodes.

The lifecycle management of ML models is a key part of the implementation of control loops and ZTM. In this regard, O-RAN has proposed an example of such an implementation in the AI/ML specification, as depicted in Fig. 12. In this case, the Service Management and Orchestration (SMO) framework (e.g., ONAP [182], OSM [102], etc.) provides the capabilities to host the ML models packaged by Acumos [128] acting as ML designer catalog. Note that the ML models can be built offline or online (either through the O-RAN ML training host or through the ML designer in Acumos powered by widely-known toolkits, e.g., sci-kit-learn, TensorFlow, etc.). The training datasets are collected from the central data lake, which is filled with information from the near-real-time RIC, the O-CU and O-DU. Once the models are trained, Acumos stores them in the private marketplace and prepares their onboarding autonomously. In the case of ONAP, given its wide integration as SMO of the O-RAN architecture, it contains in its last release the Data Collection, Analytics and Events (DCAE) adapter, which assists the onboarding of ML models from catalogs and their publication as rApps (containerized by Acumos) with the associated metadata. Then, the Non-real-time RIC deploys them on the near-real-time RIC and the O-DUs control loops via the O1 interface. By contrast, the incorporation of these functionalities is currently being investigated in OSM. Based on this architecture, the paper in [171] complements [170] by analyzing tools that facilitate the deployment of AI/ML-enabled control loops in O-RAN and presents a framework that assists in the development, tracing, and debugging tasks. The work also discusses in a general way the open points in the native introduction of AI in the O-RAN control loops regarding scalability, hardware acceleration, and online training. However, the discussion does not cover how the aforementioned features affect the various control loop levels and the distributed architecture.

The work presented in [172] introduces an ML-based approach for Automatic Neighbour Relation (ANR) optimization that leverages the ability of the Non-real-time RIC and the near-real-time RIC to run rApps and xApps in their control loops, respectively. The proposed workflow allows fetching telemetry data from the 5G RAN nodes through the O1 interface of the O-RAN architecture and making it available at the SMO, where it is used to train an ML model. This model, built using random forest techniques, is deployed as an rApp to predict the cells to be marked as permitted/prohibited for handover, and uses this input to generate a new policy that is passed to an xApp on the near-real-time RIC and enforced on the RAN nodes.

The authors of [173] propose a multi-agent team learning for O-RAN architectures to study the range and AI payload limitations. These agents, called AI-enabled NF in the paper, are distributed along with the system according to their delay budgets in the control loops managed, e.g., at the CU or the DU, and must interact with each other to exchange AI payloads. Conversely, the medium used for backhaul determines the range, e.g., fiber optic would have a propagation delay of around 5 μs/km, while it would be in the order of 50 μs/km for Ethernet cables. Direct open issues from this analysis, such as convergence guarantee in decentralized learning and team size, are also discussed.

Finally, the work in [174] presents a framework for anomaly detection and root cause analysis for distributed architectures enabling programmatic control loops. The solution, named Soda, builds on a CNN encoder that takes the cell status telemetry over time and historical anomaly scores, to predict whether the cell status will lead to an abnormal point. The model is deployed at the SMO via de non-real-time RIC control loop of an O-RAN architecture on a small-scale testbed and is functionally validated for both 4G and 5G networks in several locations. The results include the functional verification of the closed-loop and report a 15x computational efficiency w.r.t. standard approaches.

Lessons Learned. AI/ML processes are key elements in enabling ZTM at the RAN. Programmatic control loops and open interfaces together with the trends in ML (especially in scenarios with a lack of data), are opening the door for autonomous processes related, but not limited, to resource management, demand-driven power allocation, and user association.

The main conclusions to be highlighted regarding the adoption of ZTM at RAN level are as follows:

- *Standardization efforts.* While the work from standardization entities (e.g., 3GPP, O-RAN, ETSI) is simplifying the adoption of ZTM automation among vendors and operators, as well as their interoperability, there are still

crucial gaps across the specifications covering the various network segments. For that reason, discussions on their alignment on the road to fully automated and interoperable operations across network segments are mandatory.

- *Interaction between closed loops at the access network.* The coordination across and inside control loops requires well-defined open APIs that allow control loops to cooperate across various domains and network segments, e.g., edge nodes and transport networks. This has been proven crucial by several research works at the RAN (especially based on O-RAN) to provide broader and fine-grained decisions affecting control loops managing different functionalities. By contrast, decisions in control loops from one segment can also impact other ones (e.g., transport), which is still an open subject for ZTM, and a key research aspect for ZTM in 6G.
- *Efficient telemetry management mechanisms.* The incorporation of these mechanisms is a must to enable ZTM in 6G networks. The lack of advancement on this point can lead to drops in both accuracy and performance, as well as an increase in the training time of the aforementioned closed-loop and closed-feedback loop processes. Moreover, enhance telemetry and exchange systems will also allow control loops to be aware of the whole context of the system, instead of creating data silos, where the resulting automation processes may generate glitches.
- *Differences in the automation techniques adopted.* While the state of the art in programmatic control loops explores several automation techniques (i.e., SL, RL, and DL), the remaining approaches related to ZTM, are more inclined toward RL. In the former case, this choice is given by the novelty of the topic and the need for understanding how well-different methods can fulfill the needs of these new AI operations. By contrast, the remaining topics, especially the most recent papers, usually adopt several forms of RL-based solutions, especially motivated by the difficulties of having access to updated datasets. The ability of RL to adapt to more complex and dynamic network systems as in 6G, is especially suitable for the constantly changing nature of the RAN.

B. Distributed Core

The path towards automation in the mobile core has been widely investigated by the research community in recent years. In this direction, the introduction of a service-based 5G core architecture by 3GPP, entirely relying on NFs, has further extended the possibilities to exploit autonomous mechanisms in the management and orchestration of this network segment [209]. This subsection is divided into three parts, as the main areas that are driving ZTM-related research outcomes, namely: (i) core network slicing; (ii) scalability of 5G core NFs; and (iii) service automation for 5G core. The automation techniques considered for each one of the surveyed papers within each topic, including data-driven analytics functionalities (ranging from traditional optimization and heuristic algorithms to AI/ML based solutions) together with control/orchestration functionalities for automated operability (ranging from cloud-native engines to blockchain-powered frameworks) are summarized in Table VI.

TABLE VI
DISTRIBUTED CORE AUTOMATION SOLUTIONS.

Topics	Automation technique	Ref.
Core Network Slicing	Online Optimization	[183]
	Policy-based Algorithms	[184]
	Lasso Regression	[185]
		[186]
	DNN	[187]
		[188]
	5G core OSM-based MANO	[189]
	DNN and Random Forest	[190]
	Big Data Analytics and Policy-based Algorithms	[191]
		[192]
	NN, LSTM and Support Vector Regression (SVR)	[193]
Scalability of 5G Core Network Functions	MADM Algorithm	[194]
	Control Theory Algorithm	[195]
	Policy-based Decision Making	[196]
		[197]
	Q-Learning with Gaussian Processes	[198]
	CNN, LSTM and K-means	[199]
	DNN and LSTM	[200]
		[201]
Service Automation for the 5G Core	ARIMA and Facebook Prophet	[202]
	ONAP-based Policy Creation	[182]
	Cloud-native compliant Core NFs	[203]
		[204]
		[205]
	SPE plus sidecar based NFs design	[206]
	Context and Execution split of NFs	[207]
	Blockchain-based Data Storage	[208]

Core Network Slicing. Being a fundamental driver for the adoption of 5G-and-beyond mobile technologies, the lifecycle management of network slices [210], [211] represents a key aspect to consider and as such has received tremendous research attention [183]–[193].

In this vein, authors in [183] revise traditional algorithmic methodologies to efficiently tackle the challenges imposed by network slice provisioning modeled as online optimization problems (i.e., deciding how to efficiently allocate, manage, and control the slice resources in real-time). The outcomes of this work corroborate the need for accurate monitoring and prediction tools, efficient resource allocation mechanisms, and flexible network orchestration techniques to correctly deal with time-varying network requirements in the ZTM-driven path for dynamic network slicing.

The work in [184] addresses the case of local 5G operators as a disruptive model for industrial Internet environments, serving the communication needs of use cases confined inside factory premises. Three architectural options are analyzed for core network slices: (i) Pure Local Architecture; (ii) MNO Architecture; and (iii) Hybrid Architecture, which contemplates the different alternatives for placing core NFs as a case of 5G Public Network Integrated Non-Public Networks (PNI-NPNs) [212]. For the hybrid environment, the authors provide a predictive placement scheme based on the history data that

distributes NFs among the local 5G operator and the MNO intending to reduce the latency as well as the number of local NFs. Simulation results exhibit the performance in terms of latency of the three architectural options and evaluate the suitability of analyzing past usage data to achieve a trade-off of low latency with minimal NFs deployed locally for the hybrid scheme. This paper represents a very well-placed use case in the context of ZTM to enhance automation capabilities for planning, design, and deployment of virtualized and distributed core network functions.

With a special focus on the automation of 5G core network slicing, the work in [185] identifies the main functionalities in the slice lifecycle management (i.e., design, construction, deployment, operation, control, and management) in which ZTM principles have a major impact. To illustrate the possibilities of integrating intelligent management techniques for network slice operation, the authors provide an overview of relevant ML techniques that can be of applicability to enable their autonomous execution. This revision is retaken by the work in [186], which also introduces a dynamic resource adaptation framework. The proposed framework, based on analyzing continuous telemetry data about workload, performance, and resource utilization, performs one of three main control actions, namely: resource arbitration, network function migration, and network slice reconfiguration. Experiments are conducted using actual demands data collected from a web server to evaluate the workload prediction accuracy obtained by the lasso regression model, outperforming the least-squares method in terms of prediction error.

Motivated by the aim of reducing operational expenses as a result of the trade-off between SLA violations and network design over-provisioning, [187] proposes a DL architecture designed for network slices' capacity forecasting. Unlike traditional mobile traffic predictors, and in contrast to the work in [186], the ultimate goal of this solution is to anticipate not only the expected traffic demand but the required resource capacity that needs to be allocated by the slice provider in advance, as a way to facilitate and accelerate network self-planning, as envisioned by ZTM. Through empirical evaluations, the effectiveness of the proposed solution is validated against state-of-the-art predictors, also depicting a suitable configuration to balance the cost incurred by service orchestration at different network segments (i.e., C-RAN, MEC, and core datacenters).

The authors of [188] introduce a zero-touch framework for network slicing running dedicated AI solutions, where the benefits associated with monetary cost, capacity allocation, and SLA violations are reviewed. Several DNNs are incorporated in the framework to achieve long-term and short-term capacity forecasting and show the advantage of a zero-touch network slicing paradigm over dedicated and shared slices. The performance of the framework is assessed via experiments with real-world mobile data traffic in terms of incurred management fees. Observed behavior outperforms other state-of-the-art forecasting solutions, such as [187], since in addition to resource over-provisioning and demands satisfaction; the proposed approach also takes into account costs associated with resource instantiation and reconfiguration.

In [189], the authors demonstrate the operational capabilities of a zero-touch framework for 5G core network slicing management and orchestration featuring closed-loop automation that performs monitoring, selection, and lifecycle control (deployment, configuration, activation, and deactivation) of slices. To do so, a 3GPP compliant 5G core testbed is presented including slice control functions (i.e., Network Slice Management Function (NSMF) and Network Slice Selection Function (NSSF)). In this way, the automatic functioning of various NFs is enabled via on-demand triggering of auto scale-out/in of the slices by complimenting MANO capabilities. The functionalities of the systems are validated through the instantiation (using OSM as NFVO) of three slices, namely the common slice with the 5G core control plane functions; an eMBB slice with the UPF and a video stream server; and an URLLC slice with the UPF and an Intelligent Transport Service (ITS) application.

In order to address the selection of a specific slice instance to serve the incoming user service requests, the work in [190] leverages DL techniques for the selection of network slices to handle incoming user requests and for the prediction of future traffic demand, aimed to reserve resources for the selected slice in advance. The developed model, designated as *DeepSlice*, is a data-driven smart decision-making engine for network resource adaptation without the need for human intervention. By training the model with network KPIs, connection requests from different device types are mapped to pre-defined slice categories (i.e. eMBB, URLLC and mMTC). Similar to the concept of dedicated and shared slices in [188], in this work, the incoming traffic is allocated to the selected slice or a master slice acting as backup in case of slice failure or resource exhaustion. Performance evaluations are dedicated to validating the accuracy of the proposed model to predict the right slice category (around 95% for unknown device types), and the selection of the master slice to meet load balancing and network availability constraints.

In the same context, the authors of [191] have proposed a self-regulating framework, depicted in Fig. 13, that uses a closed-loop automation mechanism, a key enabler for zero-touch automation, with three main stages, namely slice monitoring, slice analytics, and slice selection. A distinctive aspect of this work, compared to the aforementioned research papers, is that the authors implement the proposed framework in the corresponding network entities of the 5G core. In particular, slice monitoring is implemented at the NSMF, while slice analytics and slice selection are part of the NSSF. In the proposed framework, the Slice Selection Algorithms module is in charge of implementing the selection scheme and performing the decision of the most suitable slice instance among a set of candidates provided by the analytics module. To ensure efficient use of resources dedicated to the deployed slices while reducing the response times incurred during the slice commissioning phase, the NSMF dynamically controls the activation/deactivation of selected slices on a need and timely basis, as proposed in the authors' previous work [192]. Using synthetic traffic data for eMBB, URLLC, and mMTC load patterns, the framework performance is evaluated considering three basic slice selection strategies (most loaded, less loaded,

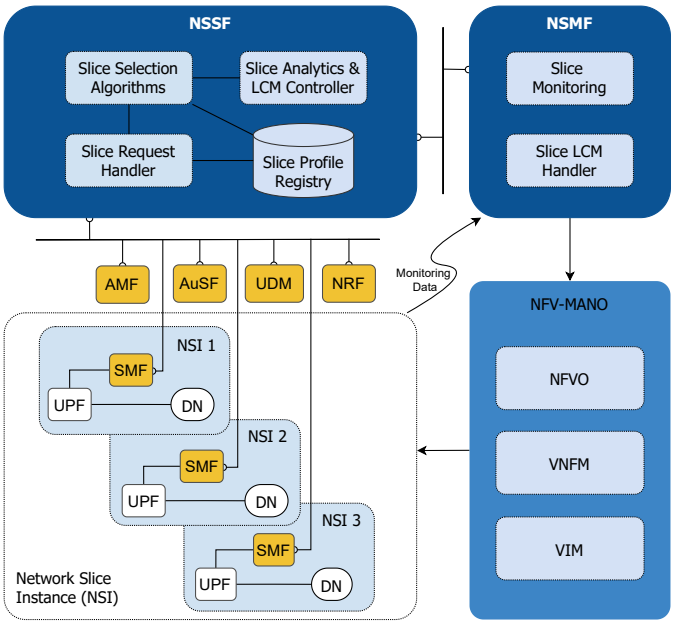


Fig. 13. Network slice monitoring, analytics and selection.

and random), showing in the first case a higher potential for slice de-commissioning.

Finally, in [193], several forecasting methods are analyzed to predict slices' congestion probability. Similar to [191], such a predictor framework is integrated into a 5G core architecture, acting as an implementation of the NWDAF, in conjunction with the session and policy control functions, to achieve ZTM goals that enable the system to act autonomously and dynamically manage the resources needed for multiple slices. Two scenarios are considered for performance evaluation: in the first one, limited resources are allocated among multiple slices maximizing the network utilization and minimizing the degraded network sessions; in the second one, the goal is to prevent network congestion and minimize the use of resources per slice. Obtained results show that the utilization of accurate predictions in the network allows better performance in terms of session loss and congestion reduction, validating the effectiveness of integrating forecasting techniques for intelligent network management.

Scalability of 5G Core NFs. Crucial to the consolidation of mobile network management automation is the capability of dynamic scaling 5G core NFs. In this regard, several works have proposed reactive (posterior) and proactive (prior) scaling mechanisms that rely on the NFV capabilities for dynamically (re-)dimensioning virtualized functions, either by instantiating (horizontal scaling) or reassigning resources (vertical scaling) to VNFs on the fly [194]–[201].

The work in [194] considers, in addition to system resource usage (in terms of Central Processing Unit (CPU) and Random Access Memory (RAM) usage) typically addressed by other related works, real-time users' QoE feedback as criteria in a Multi-Attribute Decision Making (MADM) algorithm to decide whether to scale up/down cloud-based mobile core resources. To achieve the envisioned VNF resource elasticity automatically and without any human intervention, additional functional entities are integrated into the MANO architecture. Such components are dedicated to constantly monitoring the resource consumption and service quality and to allocating resources accordingly to VNFs at instantiation as well as during their run-time. The proposed modules are implemented and evaluated in an experimental testbed for a video streaming service, analyzing the system responsiveness to effectively handle vertical/horizontal scaling up/down actions (e.g., within 15s) based on usage and perceived quality index thresholds.

In [195], the authors propose an algorithm based on control theory for scheduling incoming users to attach requests and scaling of AMF instances. The proposed control model aims to redirect incoming requests to ensure the optimal AMF load while reducing unnecessary latency and the number of blocked arrivals. Upon reaching the threshold level in terms of the number of attaching requests received by AMF, the system is dynamically re-dimensioned (scaled in/out) in order to guarantee at every moment the exact number of deployed AMF instances that automatically satisfy the arrival rate. The performance of the proposed model is evaluated by considering several scenarios with different load patterns and outperforming another scheduling model used as a benchmark in terms of stability, load balancing among active AMF instances, and blocked arrivals reduction.

Other similar threshold-based mechanisms have been proposed in the arena of virtualized functions scaling [196], [197]. However, such static schemes suffer from a lack of flexibility, rely entirely on the accuracy of pre-computed threshold levels, and are likely to become unstable and exhibit transitory oscillating behavior in the presence of highly dynamic ecosystems.

Alternatively, the use of adaptive scaling techniques overcomes the need for fixed boundaries' definition. Unlike previous works, the solution in [198] opts for using a RL-based scaling strategy that learns from dynamic environments to manage performance variations following a reward/penalty model. The proposed solution, combining Q-Learning with Gaussian Processes-based system modeling to improve the scaling policy, was proved more accurate than approaches based on static threshold rules, with results that avoid performance target violations (e.g., mean response time below 1 ms) while avoiding transitory oscillations. Despite its benefits compared to static approaches, adaptive strategies remain reactive, meaning that they respond to network events, which implies potential performance degradation and requests loss during the time needed for the scaling process.

To solve the drawbacks of the abovementioned reactive approaches, ML-based mechanisms, and in particular traffic forecasting techniques, have been identified as very promising solutions in the path towards ZTM-enabled networks capable of anticipating the execution of needed scaling operations. Following this goal, in [199], the authors present a combined strategy that uses a CNN to extract traffic patterns and an LSTM to predict traffic evolution. In order to accelerate the training and avoid over-fitting, K-means clustering with random injection is introduced during the training phase. Obtained simulation results prove a reduction of 50% of the training duration time without affecting the prediction precision.

Hinging on such two main zero-touch enablers (i.e., traffic prediction and automated management capabilities), the works in [200], [201] propose a dynamic scaling mechanism based on traffic forecasting to predict changes in the incoming load and modify the number of deployed AMF instances accordingly. To predict the traffic changes, two different NN are designed and evaluated, namely DNN and LSTM. In the proposed system, the predicted number of required AMF instances is informed to the NFVO in order to proactively conduct the scaling operations before the change of traffic load. Performed simulations corroborate that the proposed predictive scaling solution outperforms static threshold-based reactive approaches in terms of latency to respond to traffic changes and delays required for resource setup.

Service Automation for 5G Core. To unlock the high levels of automation and orchestration envisioned in 5G, novel technologies (such as AI, microservices, and blockchain) that improve the reliability, agility, and flexibility of core networks are rising in importance and as such, receiving significant attention from the research community [182], [202]–[208].

Driven by the idea of exploiting the benefits of ML techniques to enhance the reliability and performance of ZTM-compatible networks, a practical approach to integrating data analytics for better control and management of 5G core networks is provided in [202]. In particular, the designed framework focuses on the use of an analytics module that predicts the system behavior (in terms of resource usage and operational statistics) based on traffic logs collected from the control plane core NFs, enabling the proactive identification of anomalies and the avoidance of potential system failures. For the system evaluation, the authors compare two time-series forecasting models, namely ARIMA and Facebook Prophet, which are integrated with the Open5GCore [213] used as a reference 5G testbed toolkit. Prediction results, in terms of CPU and memory usage, were overall less error-prone when using the Prophet model.

Data-driven management applications, as a crucial recommendation for enhanced network management capabilities, are also advocated in [182] in order to automate the access discovery and selection process. In this work, the authors emphasize the obsolescence of current manual approaches for the discovery and selection of access points, traditionally performed by operators through the establishment of static policies. Instead, the paper presents an ONAP-based Service Oriented Core (SOC) that allows dynamic prediction, policy creation, and self-learning based on current and historic service level KPI data collected from multiple access and core network nodes. For validating the proposal, an experimental testbed is conformed by the ONAP-orchestrated SOC together with 5G access nodes, and real and non-real-time traffic traces are used for different service scenarios. The accuracy of policies dynamically allocated to selected access nodes is verified against legacy networks, and related research works, showing promising results in terms of (i) reducing the average RAN energy consumption; (ii) increasing the number of supported IoT devices; and (iii) increasing the admission percentage of new vehicles without additional latency.

Regarding the principles of modularity, programmability, and reconfigurability expected in the 5G core, stateless microservices arise as very well-suited solutions to realize the SBA design and to optimize management and orchestration functionalities of networking systems towards the ZTM vision. In this regard, recent works have started to discuss the implicit advantages of this implementation approach, together with the challenges to be addressed and some solutions to overcome them. In particular, the paper in [203] examines the role of AI-driven solutions for the placement, resource allocation, and scaling of a microservice-based stateless 5G core. In addition to discussing the main challenges to address in the context of the aforementioned orchestration operations, introduced by functionally-decomposed microservice-based implementations, the authors provide an experimental evaluation of the impact of latency among components, scaling strategy, and scaling sequence on performance. The authors claim that AI-driven scaling algorithms ought to be embedded into the orchestration of 5G core networks to achieve effective closed-loop solutions that can help to meet the stringent and dynamic requirements of future network slices.

The adoption of cloud-native 5G core network functions is also contemplated in [204]. In this paper, the authors leverage open platforms (such as Kubernetes) for the zero-touch provisioning and closed-loop orchestration of network slices with the aid of AIOps. In particular, operational analytics functionalities (integrated as part of NWDAF and MDAF core functions) are employed for the identification of anomalies and root-cause analysis. Subsequently, the identified failure (i.e., specific failure type together with impacted component) is notified to the orchestrator (via publish-subscribe message queue) for performing actions required for automated remediation. The proposed flow is demonstrated considering as affected metric the session registration time which, after the corrective scaling action, returns to normal condition according to the corresponding Service Level Objective (SLO) threshold.

In the same context, the work in [205] proposes Kube5G, a platform focused on building, packaging, and automating the lifecycle of cloud-native compliant 5G NFs in a cloudified environment, aligned with the ZTM needs of an agile, scalable, and on-demand managed systems. For enabling the lifecycle management of 5G network services in runtime, the authors develop a custom resource definition (called Kube5G-Operator) that automates both basic (e.g., installation, configuration, and monitoring) and advanced (e.g., reconfiguration, update, backup, failover, restore) operations. Experimental results of a prototype implementation are compared against traditional bare-metal and docker deployments demonstrating roll-out efficiency and failure recovery capabilities with provisioning times of less than two minutes and reconfiguration times of less than a minute.

Considering the enhancements related to service framework aspects proposed in Release 16 [56], the authors of [206] went a step further, compared to related works, and introduce NoStack, an SBA design based on two major components: service management plane and sidecar. In particular, each NF instance is comprised of a Service Processing Entity (SPE) for service logic operations and a sidecar for service com-

munication aspects. By decoupling service-specific logic from service framework common functionalities, NoStack improves the reliability of inter-service communication to better support automation of NFs service management and to improve service agility. Experimental evaluation is dedicated to analyzing the impact of the sidecar approach on the performance of inter-service communication by varying the request size, the number of clients, network latency, and packet loss. Results from conducted experiments have shown that although the sidecar model has some influence on SBA performance, it represents an appropriate mechanism to balance performance, system flexibility, reliability, and inter-communication efficiency.

In [207], the authors propose a novel framework to dynamically re-orchestrate a virtualized mobile network by leveraging the principles of the split between the context and execution engine (c/e split) of NFs. In this way, running NFs can be reallocated on the fly by only moving the function context towards a different location, where a new execution engine is instantiated, to support flexible and rapid network management and orchestration. The proposed scheme is implemented, including basic 5G core functionalities with seamless relocation capabilities. Major implementation novelties of this paper, compared to existing research works, include re-orchestrable functions compliant with the proposed c/e split paradigm and the support of new interfaces between VNFs and the MANO domain for (i) context extraction/injection; and (ii) VNF performance monitoring. The authors evaluate their framework implementation in an experimental testbed, over which a performance analysis is conducted. Obtained results validate the ability of the proposed approach to reduce relocation delays, compared to standard VIM-managed live migration without any perceptible service disruption, allowing to seamlessly modify the Service Function Chains (SFC) of a running slice to partially relocate context, to improve perceived latency due to VNF relocation closer to end-users, and to dynamically re-assign radio resources to UEs on-demand.

Extensions to 5G core SBA are also introduced in [208] for better management and coordination of decentralized functions suitable for cross-domain ZTM. The distinct feature of this paper is the application of blockchain at the 5G mobile core packaged as a standard NF. In essence, the authors introduce CoNTe, a blockchain system intended to complement the cellular core with a decentralized and immutable data record. Designed as an agnostic storage service, CoNTe is exposed to be consumed by any NF of the 5G core. Following the evolution path towards microservices, the proposed mechanism is meant to be deployed as a cloud-native application and orchestrated by a MANO-compatible Container Orchestration Engine (COE). Together with the implementation of the proposed system, the authors also present the performance and scalability results in terms of total block throughput and network efficiency for different rates of block contribution and distances among nodes.

Lessons Learned. The introduction of ZTM-compliant systems unlocks significant opportunities for core network management enhancement. These approaches allow the management of the SBA to be automated with the help of AI/ML-powered orchestration workflows. As can be seen from the surveyed proposals, the combination of such appealing mechanisms brings substantial benefits in different aspects of the core network management process, including network slicing-related tasks (e.g., selection, arbitration, and reconfiguration); provisioning operations of core NFs (e.g., placement, migration, and scalability); extensions for automated service response (e.g., context and execution split, inter-service communication and decentralized data record); among others.

Based on the aforementioned literature review, major points regarding the adoption of ZTM for core networks can be briefly outlined as follows:

- *Intelligent and dynamic management frameworks.* These frameworks are becoming a reality for core networks thanks to data-driven and closed-loop control strategies aimed to constantly ensure and preserve performance and latency requirements. In particular, traffic and capacity forecasting methods are commonly considered in many of the revised works for intelligent network management.
- *Programmable orchestration and reconfiguration techniques.* Aligned with the aforementioned advances, diverse programmable orchestration and reconfiguration techniques are currently being exploited to further improve the reliability, agility, and flexibility of cloud-native core networks, tackling the inherent complexity of distributed scenarios, especially in what regards ZTM. Worth to mention that some of the revised works stand above by presenting experimental cloud-native testbeds and prototypes formed by open-source core networks and orchestration tools.
- *Continuous gathering and processing of monitoring information.* While the modular, programmable, reconfigurable, and micro services-oriented SBA gives the ability to seamlessly integrate monitoring and analytics engines (e.g., the NWDAF module) with the rest of the control plane NFs and to conduct the exchange of performance data within it, the continuous gathering and processing of monitoring information from different sources remain as a fundamental challenge to enable fully self-operating communications systems. In this regard, in the surveyed research, there is a lack of assessment of potential limitations in terms of scalability and computational complexity when dealing with a large amount of heterogeneous monitoring data.
- *Interoperability and integration in mind.* Finally, it is important that ZTM-compliant systems, whether they follow SBA patterns or not, are implemented with the principles of interoperability and integration in mind. While most of the works are focused on the enhancement of core network management automation via the introduction of AI/ML-powered engines, only scarce literature takes into account this aspect by relying for instance on event-driven approaches and microservices-based design as predominant means of integration.

C. Distributed Edge

Throughout this section, we shed light on some of the most compelling research topics at the network edge, where applying a ZTM approach can lead to better network management

TABLE VII
DISTRIBUTED EDGE AUTOMATION SOLUTIONS.

Topics	Automation technique	Ref.
Task Offloading and Resource Allocation	Reinforcement Learning	[214]
		[215]
		[216]
		[217]
	Deep Reinforcement Learning	[218]
		[219]
		[220]
		[221]
		[222]
		[223]
	Deep Neural Networks	[224]
	Adaptive Ant Optimization	[225]
	Long Short-term Memory (LSTM)	[226]
	Support Vector Machine (SVM)	[227]
	Reinforcement learning and Neural Networks	[228]
	Deep Deterministic Policy Gradient (DDPG)	[229]
Proactive Scaling	Deep Neural Networks and LSTM	[230]
	Neural Networks	[231]
	Neural Networks and Federated Learning	[232]
	LSTM	[233]
Proactive Caching	Transfer Learning	[234]
	Reinforcement Learning	[235]
		[236]
	Random Forest and XGBoost	[237]
	Distributed Learning	[238]
	LSTM and Ensemble Learning	[239]
	Reinforcement Learning and Neural Networks	[240]

decisions. These research areas include (i) task offloading and resource allocation, (ii) proactive scaling, and (iii) proactive caching. Table VII.

Task Offloading and Resource Allocation. Task offloading and resource allocation are classes of problems in which decisions should be made about the location of executing users' tasks and the number of resources allocated to perform such tasks. Several options/locations can be considered to offload users' tasks; specifically, user tasks can be executed locally in the user device, be offloaded to edge servers in proximity or aggregation points, or be executed in a central cloud far from the task source. After performing and deciding on offloading tasks, it becomes essential to determine the number of resources (computing, transport) allocated to perform a specific task. Fig. 14 illustrates an example of a scenario in which MEC is used to execute a user's task, which cannot be processed locally due to a lack of resources. As can be comprehended from the figure, the user's task undertakes multiple steps in order to be ready for the intelligence engine to decide whether to perform the task locally or offload it to a MEC server in proximity [241].

The problem of task offloading and resource allocation become vitally important and challenging research topics when it comes to the near/far edge nodes that are equipped with

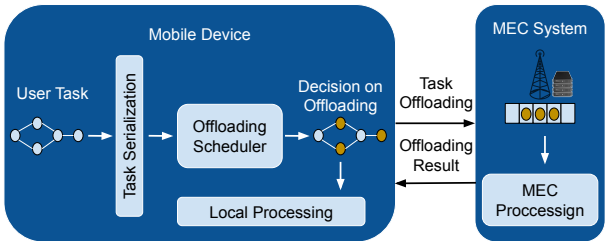


Fig. 14. Task Offloading Process.

an extremely limited and costly set of resources [242]. This necessitates a proper and efficient allocation and utilization of resources to have a performant edge system. Although task offloading and resource allocation problems can be studied independently, there is a sizeable body of literature that addresses these problems jointly [214]–[229]. Thus, in this section, we review some research works in these areas.

Wang et al. [214] study the problem of task offloading and resource allocation in a MEC-enabled mobile network in which MEC hosts are collocated with base stations. A multi-stack RL algorithm is proposed to jointly solve the problems of task offloading, spectrum, and transmission power allocation. The main objective of the work is to optimize both computational and transmission delay for all the users in the network. They assume that each base station records the history of the allocation schemes and user information. This information later will be fed to the learning model to improve the convergence speed and learning efficiency. This study acknowledges the need for continuous monitoring and optimization of the network to meet the users' demands and optimize the network performance, as is foreseen by ZTM. Similarly, Wang et al. [218] employ a DRL algorithm to minimize the average service time and balance the resource usage among MEC servers. The proposed solution takes computing resources and route selection jointly into account to respond to the mutative environment of MEC. SDN technology is applied to the MEC architecture, and a DRL unit is placed inside the SDN controller to manage the routing of requests. This study dedicates special attention to the lifecycle management of services, to which ZTM principles have notable attention. Aiming at tackling the security, reliability, and irreversibility of data in the MEC, along with the proper utilization of scarce resources at the edge, Fen et al. [224] present a DNN-based model. In particular, in the proposed method, blockchain technology guarantees data security and privacy issues in the MEC system. Moreover, an optimization model is proposed, which jointly maximizes the computation rate of the MEC system and transaction throughput of the blockchain system. The study in [217] presents an RL-based state-action-reward-state-action approach for the joint problems of task offloading and resource allocation in MEC systems. A hierarchical mobile network, consisting of near edge, adjacent edge, and remote cloud is considered, with different computing resources and the cost associated with the resources. The main objective of the work is to optimize the system cost, which includes the energy and computation time delay. The proposed method is evaluated against RL-based Q learning and demonstrated

superior performance. This work highlights the importance of continuous optimization, which is envisioned by ZTM.

Regarding the fact that long-distance transmissions hinder the realization of stringent latency applications that demand sub-millisecond latency, the authors of [215] propose an offloading and resource allocation strategy in the IoT scenario in which mobile devices offload some of their tasks to the MEC hosts that are placed in the proximity of the end device. The problem of resource allocation among MEC and IoT devices is tackled using RL and a belief-learning-based double auction mechanism. The resource management problem for MEC and IoT devices is formulated as a double auction game. The experience-weighted attraction algorithm is proposed to search the Nash equilibrium behind each participant. Different from this study, which only considers the offloading tasks with stringent latency requirements, the work in [219] proposes a framework employing a DRL-based model to enhance the performance of applications for the scenario of smart cities taking into account the caching, computing, and networking resources. This study concentrates on the continuous optimization of different aspects of the network, which has been considered one of the main pillars of the path toward ZTM. Unlike the studies mentioned above, Rathore et al. [243] employ Hesitant Fuzzy for the offloading decision based on the state of the resources and the security level requested by the user. The main idea behind applying fuzzy logic to the problem is due to the scenario's multi-variable requirements and specifications.

Motivated by the Vehicle To Everything (V2X) scenario, in which the frequent movement of vehicles results in imbalanced task offloading over the MEC hosts, the authors of [244] present a Self-Imitation Learning (SIL)-based DRL algorithm for the problem of task offloading. A central server is connected to the MEC system is considered, where the learning model runs. The study's main objective is to minimize energy consumption at the MEC servers while simultaneously guaranteeing the tasks' delay requirements. The authors of [226] focus on one of the main aspects of the ZTM (i.e., prediction) by using ML techniques, such as CNN and LSTM, to predict the state of the energy at the MEC hosts. The idea of the study is to minimize energy consumption by correctly predicting the state of the energy on the MEC hosts and accordingly distributing the load on the applications hosted at the edge. The model treats the problem as a multivariate time series forecasting problem, and it predicts the energy state for the multiple time steps ahead. CNN and LSTM models are combined to take advantage of both models for this time series prediction. Having the objective to minimize energy and cost consumption, the authors of [220] propose a DRL-based optimization framework to jointly tackle the problem of task offloading and resource allocation in the scenario of MEC-enabled mobile networks. Similarly, the authors of [221] propose a DRL-based approach to minimize delay and energy consumption in MEC systems for the problems of task offloading and resource allocation. An SVM-based model has been proposed in [227] to optimize energy consumption at the MEC nodes by adjusting user association, task, and power allocation schemes at the same time. Aiming to optimize energy consumption by the MEC hosts in the scenario of mMTC, the authors of [223] propose an offloading method based on DRL that can decide on the location where an incoming task is to be offloaded.

The study in [228] investigates the problem of task offloading for the scenario of dependent tasks. Specifically, they consider applications composed of multiple tasks in which the tasks are dependent on each other, which makes the problem of task offloading more difficult to solve. An off-policy RL method is proposed to offload the tasks to the edge servers intelligently. The proposed RL method uses Sequence-to-Sequence (S2S) NNs. The results acquired from the simulation experiments show that the method improves the latency and energy consumption compared to the existing works. Zhang et al. [229] study the problem of task offloading for a scenario in which users are mobile, and MEC hosts' status can change dynamically. User mobility can cause very dynamic communication conditions and increase the load on some host spot MEC servers; it induces fluctuation in the latency of the task completion and inefficient usage of MEC servers. In this regard, a Deep Deterministic Policy Gradient (DDPG)-based task offloading approach is proposed, capable of improving computing resource utilization and simultaneously meeting the latency requirement of mobile users. In this approach, the tasks are divided into two parts and then possibly offloaded to different servers to be executed in parallel, aiming to distribute the load among the MEC servers and decrease the latency of the task execution. The experimental results demonstrate noticeable performance improvements in latency for latency-stringent tasks. Finally, a task offloading and trajectory control approach is proposed in [216] for MEC-assisted Unmanned Aerial Vehicle (UAV) using DRL techniques. Despite the significant advantages of MEC-equipped UAVs for scenarios, such as sports game areas or natural disasters, which become hotspots temporally, they are extremely constrained in terms of computing and energy resources. Employing traditional optimization techniques for such a constrained scenario cannot be efficient in terms of energy since both trajectory of the UAVs and the task offloaded should be carried out at the same time. In this regard, the authors of [216] propose a DRL-based approach to optimize UAV trajectory controlling and task ratio offloading jointly. The results demonstrate improvements in maximizing system stability, energy consumption, and computation latency.

Aiming at improving task execution probability, the authors of [222] propose a learning-based prediction method for self-resilient task offloading at the edge. The problem of redundant task offloading has been modeled mathematically. Considering the time complexity of the mathematical model, a DRL-based method has been proposed, able to perform task offloading based on the application, network, and telemetry features in a shorter time scale. The evaluations have been performed in a realistic testbed, and the results show a significant improvement concerning task execution probability. In another study [225], an intelligent task offloading mechanism is proposed to transfer intensive tasks to picocells equipped with computing power. Moreover, a framework is also proposed to work as a mesh middle layer to provide seamless and high

resilience communication. The Adaptive Ant Optimization algorithm is developed for backhaul routing conflicts.

Proactive Scaling. A sizable body of research has been conducted on using time series forecasting techniques to predict traffic loads and perform scaling operations to respond to the demand variations [230]–[233].

In predictive scaling, mostly prediction techniques (a key enabler for zero-touch automation), such as time series, are used to learn a model that can anticipate future demands for performing scaling operations based on the scaling decisions made in the past. A comprehensive survey on applying ML algorithms for reliable resource provisioning in edge-cloud computing is presented in [30]. The work is specifically focused on workload characterization and prediction, component placement, and system consolidation, and application elasticity and remediation. The authors of [230] employ DNN and LSTM algorithms to predict the number of VNFs required to satisfy the demand. Network traffic load fluctuation is considered the main feature for predicting the number of required VNFs. The proposed method exhibits superior performance in terms of accuracy compared to other ML methods applied for the same problem. The study only focuses on load prediction on VNFs and does not propose any scaling strategy to react to prediction outcomes. Similarly, Subramanya et al. [231] propose an NN-based model to predict the number of VNFs and automatically scale the VNFs to meet certain traffic demands. The proposed method is trained using the traffic traces collected over a real-operator commercial network, and it demonstrates high performance in terms of accuracy, precision, recall, and F-measure. Consequently, combinatorial optimization techniques are employed for embedding the SFCs into the substrate network. The objective of the SFC placement method is to minimize the overall end-to-end latency from all users to their respective SFCs. Motivated by the scenario of multi-domain orchestration in 5G networks, the same authors propose in [232] two ML-based models to predict the number of VNFs needed to serve a certain traffic demand. More specifically, centralized and distributed DL methods are used for VNF prediction. The applicability of the proposed approach is shown using an AI-driven Kubernetes orchestration prototype. The studies mentioned above give a particular attention to the traffic prediction and automation of the lifecycle management of services in the 5G networks, which manifest their adjustability with the ZTM principles. Finally, Tao et al. in [233] study the problem of VNF scaling based on traffic fluctuation in the network. A proactive resource allocation method is proposed to predict users' demands and perform resource allocation accordingly. The prediction model is based on LSTM to predict users' demands. After acquiring the prediction results, a cost-minimization model is proposed to optimize communication and computing resources and perform scaling if needed. Regarding the time complexity of the optimization model, a Markov approximation method is also proposed that can approximate the VNF scaling problem and reach a near-optimal solution in a shorter time scale.

Proactive Caching. Caching in mobile networks can be implemented both in the core and in the RAN. Caching in the core can be implemented through Content Delivery Network (CDN) and bring advantages, such as ease of management, scalability, and high cache hit ratio. Contrary to that, caching in the RAN improves the QoE for consumers, reduces traffic, and can alleviate the load on backhaul by up to 35% [245]. Recently, with the evolution of MEC, caching in the RAN can be performed more efficiently. An example of such a service is the Radio Network Information Service (RNIS), which provides information about the radio context and helps make caching decisions. The study in [246] proves that caching in the RAN can reduce the time taken to start a video by up to 30%. It is clear by now that caching in mobile networks is a trade-off between network bandwidth usage and storage usage. Despite the advancements in caching in the RAN, still, the storage capacity at the RAN is limited and shared among many applications. Therefore, there is a high demand for devising intelligent methods able to cache the video contents properly [247], [248]. In this regard, a plethora of research work [234]–[240] has been conducted on devising novel and automated solutions to come up with strategies that can use cache resources at the edge efficiently and, at the same time, improve users' perceived QoE.

Tingting et al. [234] propose an ML-based approach for proactive caching of video content at the edge of mobile networks. The study's main objective is to minimize the transmission over the backhaul by prefetching the content to the edge of the network. Transfer learning is used to predict content popularity. Following the predictions of the TL-based model, combinatorial optimization techniques are used to optimize cache content placement in the network. Regarding the NP-hardness of the content placement problem, a greedy approach algorithm is also proposed to reach a near-optimal solution in a shorter timescale. This work predicts the popularity of the video content to prefetch them to the network edge, but prefetching to the edge, where there is a small portion of users in the coverage area of the base station, might lead to inefficient utilization of resources at the edge. Different from the work mentioned above, Wei et al. [235] address the problem of cooperative content caching for a MEC-enabled mobile network scenario using RL. The idea of the work is to minimize downloading latency for the users, and in this regard, a weighted reduction of downloading latency is considered as the caching reward. The model is trained using the historical data acquired from a real-world network. MEC servers use Q-learning to learn how they can improve the caching decisions. After that, the problem of content caching is modeled as a multi-agent, multi-armed bandit problem to maximize the total expected accumulated caching reward over a long-term horizon. The authors show a significant improvement in reducing downloading latency and cache-hit ratio compared to the state-of-the-art works. The study presents a closed-loop automation mechanism, which is one of the main enablers for ZTM. Behravesh et al. in [237] present an ML-based approach for proactive video content caching at the edge of mobile networks. RF and Gradient Boosting Tree (XGBoost) methods are used to predict video segment requests, bitrate of the video segments, and base station association of the users. Following the prediction results, combinatorial optimization techniques are used for

the video content placement problem at the MEC servers. This work presents an end-to-end approach toward proactive caching and management of networks, which is considered a key enabler for zero-touch automation. A cache management algorithm using combinatorial optimization techniques is presented in [236] with the objective of efficient cache usage in the MEC systems. To tackle the scalability of the proposed solution, a RL-based algorithm is proposed that can reach near-optimal solutions in a much shorter timescale. Unlike the study in [236], which does not take users' information into account, the authors of [238] present two proactive caching strategies using DL to predict users' content demand. While the first solution depends on a central server to collect data and train the model, the second approach trains the models in a distributed manner.

The authors of [239] study the problem of content caching at the edge, aiming at improving latency and energy jointly. To achieve the mentioned objectives, they try to maximize the amount of content served from the edge. The proposed method is based on LSTM local learning and ensemble-based meta-learning. LSTM is used as a local training model in the first stage of the work to predict attributes of the demands in each demographic user group. Then regression ensemble methods are used to make a unified caching strategy out of multiple demographic user preferences. The simulation results demonstrate a significant improvement in the cache hit ratio compared to the state-of-the-art works. Similarly, Jiang et al. [240] propose an actor-critic-based RL model. This content caching problem is modeled as a Markov decision process and is used to minimize caching costs and downloading latency. After that, in order to reduce the computation complexity of the model, a branching NN is proposed to approximate the policy function in the actor part.

Lessons Learned. Edge computing has been proven to be a key enabler for 5G and 6G networks in realizing many futuristic use cases with stringent latency demands. Despite the many advantages offered by edge computing, there are many challenges such as dynamicity, heterogeneity, scarcity, and cost of resources at the edge that demand thorough attention. Thus, network management at the edge needs to be highly agile and efficient to grasp the full potential of the resources at the edge. In order to attain such a level of agility, speed, and efficiency, network management, including life cycle management of service and resources, must be fully automated.

Recently, a sizeable body of research has been conducted on adopting a ZTM approach to tackle network management problems using ML techniques at the edge. This section overviewed some of the most recent research works that take a ZTM approach toward solving problems, such as task offloading and resource allocation, proactive scaling, and proactive caching at the network edge. The studies presented demonstrated satisfactory improvement in the network management at the edge by utilizing ML techniques. The lessons learned from this section can be summarized in the below points.

- *Continuous optimization and life-cycle management*. AI, and in particular ML, is considered the main technique allowing ZSM principles, such as continuous optimization, life-cycle management, and closed-loop automation. However, the research studies surveyed in this section do not consider the practical implementation difficulties and overheads of such approaches, which is critically important as continuous optimization leads to huge overhead of control data transmission in the network and requires timely responses to network dynamism.
- *Continuous monitoring and collection of telemetry data.* Closed-loop systems at the edge need to continuously monitor the network, collect telemetry data, and accordingly adjust network management decisions as foreseen by the ZSM, which still needs to be explored profoundly. Despite the usefulness of telemetry data in the network for proper decision makings, the research studies surveyed in this sections do not consider the technical complexities of the such approaches. Moreover, scalability of such methods is not assessed in these studies.
- *Designing intent-driven systems*. The ultimate goal of management automation is to enable largely autonomous networks to be managed using high-level intents. Therefore, future studies need to address this challenge by designing intent-driven systems in order to build a complete ZTM-compliant system. However, the surveyed works in this section do not introduce fully intend-based systems, which requires further studies to be conducted.

D. End-to-end Aspects

This section covers three main topics within the end-to-end aspects of ZTM: (i) End-to-End Network Slicing; (ii) AI-based security of 5G/6G networks; and (iii) Security of AI-based solutions in 5G/6G networks. Table VIII conveys the related work surveyed associated with the end-to-end aspects of ZTM topics, grouped by the specific automation technique used.

End-to-End Network Slicing. MNOs need a programmable solution that allows them to support multiple independent tenants on the same physical infrastructure. Therefore, 5G and beyond networks must support end-to-end network resource allocation using the concept of network slicing which, together with AI, can accelerate the 5G network performance, as shown in many research activities [190], [220], [249]–[257].

The high-level reference architecture for cognitive end-to-end network slicing is presented in Fig. 15. When an MNO receives a request to create network slices from various verticals, the functions at the service and network management layers (i.e., Communication Service Management Function (CSMF), NSMF and domain-specific Network Slice Subnet Management Function (NSSMF)) translate the slice level requirements into service and resource level requirements for access, transport and core network domains, and instantiate end-to-end slices covering multiple domains. These management functions can include cognitive analytics to monitor the slices and proactively adapt their requirements to meet the required SLA. Additionally, ETSI ZSM003 specification [272] proposes an example ZSM architecture deployment for network slicing management. The components of the ZSM E2E network slicing architecture include an E2E Service Management Domain as a provider of the E2E network slicing related management services and an E2E Service Management

TABLE VIII
END-TO-END AUTOMATION SOLUTIONS.

Topics	Automation technique	Ref.
End-to-End Network Slicing	Deep Neural Networks	[190]
		[249]
		[250]
		[251]
	Deep Reinforcement Learning	[220]
	Blockchain	[252]
	Generative Adversarial Networks	[253]
	Federated Learning	[254]
	MAPE	[255]
	Decentralized Marketplace	[256]
		[257]
AI-based Security for 5G/6G Networks	Deep Neural Networks	[258]
	Gaussian Mixture Model	[259]
	Multi-objective Genetic Algorithm	[260]
	Federated Learning	[261]
		[262]
	Differential Privacy	[263]
	Anomaly Detection	[264]
		[265]
		[266]
Security of AI-based Solutions	Supervised Learning, Unsupervised Learning and Reinforcement Learning	[267]
		[268]
	Deep Learning	[269]
	Deep Learning and Blockchain	[270]
	Federated Learning and Blockchain	[271]

Domain as also a consumer of the management services provided by the access network, transport network and/or core network Management Domain(s). The Access Network Management Domain is a provider of the access network slice subnet-related management, the core network Management Domain is a provider of the core network slice subnet-related management services, and the transport network Management Domain is a provider of the transport network slice subnet-related management services.

The authors of [250] present a unified framework that has generality to integrate AI to conduct intelligent tasks for all network aspects, ranging from radio channels to signal processing, from resource allocation to network slicing orchestration, from local control to end-to-end optimization, which is in line with the principles of ZTM. The concept of intelligent slicing was introduced with the flexibility to instantiate, deploy, scale, reconfigure, and transfer AI functional modules on demand. Such slices can be deployed in an arbitrary network entity, facilitating solving a problem through the selection of the best algorithm optimized explicitly for this problem. Additionally, two example slices, i.e., neural network-based Multiple Input Multiple Output (MIMO) channel prediction and security anomaly detection in industrial networks, were illustrated to demonstrate the proposed framework.

In contrast to [250], which leverages the concept of intelligent slices to conduct different tasks with the flexibility to accommodate arbitrary AI algorithms, in [249], the authors propose an AI-based framework for network slice management by introducing AI in the different phases of the slice lifecycle to achieve the goal of ZTM. They propose practical deep learning architectures to solve very complex networking problems by considering three different case studies: scheduling of slice traffic at RAN, resource allocation to slices in the network core, and admission control of new slices. They conclude that AI has a clear potential to become a cardinal technology for future-generation zero-touch mobile networks and illustrate the typical high gain that one can expect from integrating AI in network slicing.

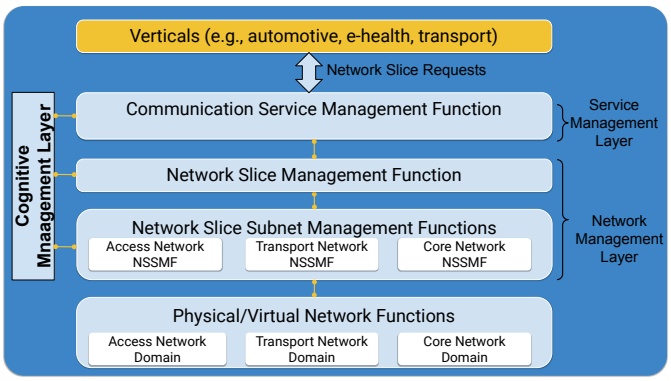


Fig. 15. Cognitive end-to-end network slice management.

The authors of [249] also propose in [190] a novel data analytics tool, named DeepCog, for cognitive management of resources in 5G networks. DeepCog leverages deep learning architectures to forecast the capacity needed to accommodate future traffic demands within individual network slices by considering the tradeoff between resource overprovisioning and service request violations. They also perform comparative evaluations using real-world data, proving that DeepCog's tight integration of machine learning into resource orchestration allows for substantial (50% or above) reduction of operating expenses concerning resource allocation solutions based on state-of-the-art mobile traffic predictors.

The research mentioned above works mainly focuses on leveraging SL techniques for optimizing end-to-end network slicing operations. In contrast, the authors in [220] propose applying Q-learning and DRL schemes, which remove data preparation effort like in SL, to solve resource allocation problems in network slicing scenarios (both radio resource slicing and core network slicing). Moreover, they perform extensive simulations to demonstrate that the proposed schemes achieve a significant reduction in the sum cost compared to other baselines. Finally, they also discuss all the possible challenges in applying RL for optimized network slicing to achieve zero-touch environment.

The aforementioned research works do not address the problem of end-to-end isolation of network slices, which facilitates customizing each slice based on their service requirements in a secure manner. The work in [251] presents a NN-based network slicing resilient model called Secure5G to proactively identify and mitigate threats from incoming connections before they infest the 5G network. The Secure5G model quarantines the threats to ensure that end-to-end security from the device to the core and external networks is guaranteed. They also

claim that the designed model will enable the MNOs to sell network slicing as-a-service with high security and reliability, which is crucial to ensure safe ZTM.

Unlike previous solutions that deal with single-domain slicing challenges, the work in [252] proposes a zero-touch framework for the automated service assurance of cross-domain network slices, based on the use of Distributed Ledger Technologies (DLT) and AI-driven closed-loop automation techniques. Underpinned by a decentralized marketplace for infrastructure resources and network services sharing among multiple providers [257], this work employs trained AI prediction models to forecast violations of network slices' SLA [256], and consequently, it triggers proactive mitigation actions in the form of slice extensions over infrastructure resources available in the marketplace (which are programmatically enabled via smart contracts) together with the corresponding orchestration workflows. Using an experimental prototype, the authors also evaluate the accuracy of the proposed approach for service demand short-term predictions and validate the ability of the framework to respond with preventive scaling actions timely.

In [253], the authors present an intent-based network slicing framework that can efficiently slice and control the RAN and core network resources. The system allows a user to provide high-level information in the form of a network slice intent, and in return, the proposed system deploys and configures the requested resources. Moreover, they propose applying Generative Adversarial NNs to manage network resources, and also evaluate the proposed framework by creating several network slices, which illustrates the performance improvement w.r.t. bandwidth and latency.

The authors of [254] propose an architectural and artificial intelligence-based algorithm to achieve energy efficiency in 6G networks along with the analysis of their trade-offs. The authors then introduce a novel statistical federated learning-based analytic engine for zero-touch 6G massive network slicing. This engine performs slice-level resource prediction by learning offline while respecting some preset long-term service level agreement constraints defined in terms of the empirical cumulative distribution function and the percentile statistics, and hence it uses a new proxy-Lagrangian two-player strategy to solve the local non-convex federated learning task without settling for surrogates only. This guaranteed 20× lower service level agreement violation rate with respect to the federated averaging scheme while achieving more than 10× energy efficiency gain compared to a centralized deep learning algorithm, which paves the way to sustainable massive network slicing.

The authors of [255] propose introducing a Network Intelligence Orchestration layer for effective end-to-end coordination of Network Intelligence instances deployed across the whole mobile network infrastructure. In particular, they first outline requirements and specifications for Network Intelligence design that stems from data management, control timescales, and network technology characteristics. Based on such analysis, they derive initial design principles of the Network Intelligence Orchestration layer, focusing on (i) proposals for the interaction loop between Network Intelligence instances and the Orchestrator, and (ii) a unified representation of Network Intelligence algorithms based on an extended MAPE-K model.

AI-based Security for 5G/6G Networks. 5G/6G networks will introduce new security and privacy challenges largely due to the adoption of new technology enablers such as SDN, NFV, MEC, and network slicing to support diverse 5G/6G network service requirements. Therefore, providing security and protecting privacy has become the primary concern in 5G/6G networks since risks can have enormous consequences for mission-critical applications and as such has received a significant research attention [258]–[266].

The TSG Service and System Aspects of 3GPP has dedicated a working group on security-focused on defining requirements and specifying the architectures and protocols for security and privacy in 5G/6G systems [273]. A comprehensive survey detailing various security and privacy threats, their defense mechanisms, recent advancements, and future directions is presented for both 5G and 6G networks in [274].

The authors in [258] provide their vision of 6G by describing various scenarios, use cases, and their corresponding requirements. They expect that 6G will be an autonomous system with human-like consciousness and intelligence (main enabler for ZTM), which will provide various ways to communicate and interact with smart terminals such as fingers, voice, eyes, and brainwaves. Consequently, the functionalities of the physical layer in the RAN protocol stack need to be redesigned to enhance their performance. Therefore, they propose leveraging DNN for designing and enhancing one or more physical layer functionalities such as signal classification, channel decoding, channel estimation, and signal detection. Additionally, they also mention that each of the data-link layers in the RAN protocol stack can leverage ML to enhance their security during data transmission, especially for IoT scenarios. Two works ([259] and [260]) leverage ML to enhance physical layer security. In [259], the authors analyze the benefits of using SL for enhancing the security of the physical layer authentication process. The receiver needs to check for previous packets received from a transmitter to examine whether the respective channel estimation matches the previous ones before authenticating the transmitter. They use a Gaussian Mixture Model to perform channel estimation and evaluate its performance for different feature dimensions. The authors of [260] present an ML-based antenna design scheme to ensure the security of the IoT communication system, which delivers secure directional communication from the relay tag to the reader by consolidating the patch antenna with a log-periodic dual-dipole antenna.

The authors of [261] provide a vision for distributed and edge-native AI in 6G networks, which would play a key role in the multi-vendor and multi-domain zero-touch automation environment. They describe how AI can improve edge security by enabling personalized, shareable, location-aware security systems adjusted to each user context with fine-grained control. Furthermore, they mention that distributed AI will improve user privacy, increase control of consent management and data ownership (data integrity), and increase overall trust by adopting distributed ledgers with AI. Similar to [261], the authors of [262] anticipate that in 6G networks, edge computing plays a crucial role in realizing new network services, and therefore, propose to leverage distributed and

federated AI techniques in which edge nodes need not send their data to a centralized location, thus enhancing network security, privacy, and trustworthiness. However, as mentioned by the authors of [263], the impact of data correlation in some ML algorithms can lead to an increase in privacy leaks. Therefore, to address the problem of data privacy, the authors present various algorithms based on differential privacy (for both centralized learning and distributed learning approaches) that could be suitable for solving some 5G/6G networks' privacy issues.

Industry 4.0 promises to offer increased automation and inturn productivity but also introduces new security risks to critical industrial control systems from unsecured devices and machines. Therefore, anomaly detection is with no doubt a major concern when it comes to systems controlled by AI [264]–[266] to accomplish safe ZTM. The authors of [264] propose a network-centric, behavior-learning-based, anomaly detection approach for securing vulnerable environments. They demonstrate that the predictability of TCP traffic from IoT devices can be exploited to detect different types of DDoS attacks in real-time, using UL. With a small set of features, their ML classifier separates normal and anomalous traffic, allowing the use of SDN-based mechanisms for blocking attack traffic close to the source. With a similar aim to [264], the work in [265] examines Apache Spot, which is a widely known ML-based platform for anomaly detection in natural language processing. In this particular work, Apache Spot is extended to the networking domain for improved SDN/NFV threat detection and mitigation measures by deducing a probabilistic model for the behavior of each detected IP. The authors of [266] mention that outages and situations leading to congestion in a cell pose a severe hazard to the network. High false alarms and inadequate accuracy are the major limitations of modern approaches for the anomaly-outage and sudden hype in traffic activity that may result in congestion-detection in mobile cellular networks. Therefore, they apply DL for the detection of anomalies mentioned above, and also support the MEC paradigm in which core network computations are divided across the cellular infrastructure among different MEC servers (co-located with base stations) to relieve the CN. Each server monitors user activities of multiple cells and utilizes L-layer feed-forward deep neural network fueled by a real call detail record dataset for anomaly detection.

Security of AI-based Solutions. Most 5G/6G applications depend on AI algorithms due to their recent advances in uncovering hidden patterns from multidimensional data. Unfortunately, AI also presents several adversarial threats such as evasion, poisoning, extraction, and inference that needs to be addressed before jeopardizing the security of future networks. Evasion attacks involve carefully perturbing the input samples during test time to have them misclassified. Poisoning is adversarial contamination of training data. ML systems can be retrained utilizing the operational data collected during deployment. An attacker may poison the operational data by injecting malicious samples that consequently disrupt retraining. Extraction attacks intend to duplicate an ML model via query access to a target model. Inference attacks determine if a set of data samples were used in the ML model's training. Predictably, AI/ML-enabled automation systems, such as network and service management systems based on the ZSM framework, can fail because of adversarial attacks on the ML model and data. Therefore, security threat analysis and corresponding countermeasures for AI/ML-related services of the ZSM framework were investigated in ETSI group report ZSM10 [275].

In this regard, several recent research works have focused on addressing the security of AI-based solutions in communication networks [267]–[271]. The authors of [267] emphasize the importance of AI in enhancing security mechanisms in 5G and beyond networks. On the other hand, they also discuss the AI security risks or vulnerabilities leveraged by malicious actors that come along with the benefits of AI. They present three types of attacks (poisoning, evasion, and model API-based attacks) on ML models (i.e., SL, UL, and RL) deployed in 5G/6G networks and recommend potential defense mechanisms to safeguard the networks. They also clarify that, despite the merits of the recommended defense mechanisms, each has its limitations, and none of them can constitute an all-in-one solution for addressing all AI threats.

In [268], the authors present a cautionary view on using ML in 5G/6G by highlighting the adversarial dimension spanning multiple types of ML (SL, UL, or RL) and support this through several case studies. They also examine various approaches to mitigate such adversarial ML attacks by evaluating ML models' robustness and calling attention to ML-oriented research issues in 5G/6G. Conversely, the work in [269] uses white-box and black-box evasion attacks to generate adversarial examples in the spatio-temporal trajectory data. The authors estimate the travel time of a path in the 5G network, as an example, which is crucial for applications such as route planning, route navigation, and flow monitoring. Experimental results show that the adversarial examples can successfully attack the DL model, thus invalidating the model itself.

The authors of [270] present a blockchain-based secure DL called BlockDeepNet to support secure collaborative DL in IoT. In BlockDeepNet, they propose to perform collaborative DL at the device level to overcome privacy leaks and obtain sufficient data for DL, whereas blockchain is applied to ensure collaborative DL confidentiality and integrity in IoT. The experimental evaluation proved that BlockDeepNet could achieve higher accuracy for DL with tolerable computational and latency overhead of blockchain operations. The paper in [271] presents a comprehensive intelligent, and secure data analytics framework named Block5GIntell by converging Blockchain and AI for 5G networks to support decentralized, distributed, and immutable smart applications. A study is presented for the use case of energy-saving covering performance and security issues to evaluate the proposed framework and further discuss the research challenges of using Blockchain and AI for 5G in security and performance issues.

Predictive network maintenance. Hardware and software failures and malfunctioning are critical issues faced by network operators. Networks require continuous monitoring to minimize the number of inevitable failures and comply with SLA. Moreover, operators have to ensure that their infrastructure provides enough capacity for their subscribers. As data usage

is continuously rising, existing network infrastructure might become insufficient as services and user usage evolves.

Thus, learning about likely faults as early as possible and keeping enough network capacity before the subscribers experience any deficiencies are critical functionalities for network operators. Given the depth of this topic, it is much larger operation time granularity (as compared to zero-touch network management), and the very different technical solutions applied, we do not address this topic in this survey and defer it as future work to be addressed in a dedicated survey.

Lessons Learned. The introduction of ZTM-compliant systems unlocks significant opportunities in RAN, edge, core, and cloud domains as well as across the control plane, user plane, and management plane to optimize network operations in the mobile networks. As can be seen from the surveyed research works, intelligent and dynamic end-to-end network and service management frameworks are becoming a reality thanks to AI/ML strategies aimed to improve the offered service performance constantly. With such a rapid end-to-end mobile network transformation comes the risks, threats, and vulnerabilities in the networks that the adversaries may exploit. Therefore, the research community has also leveraged AI/ML algorithms to design, model, and automate efficient security protocols against a wide range of threats to ensure end-to-end network security. While the AI/ML-driven intelligent and autonomous end-to-end network and service management frameworks offer the opportunity to realize the emerging critical use cases such as autonomous vehicles, industrial automation, etc., there exists a fundamental challenge of AI/ML trustworthiness that needs to be addressed to realize the vision of ZTM.

In summary, based on the aforementioned literature review, the major points regarding the adoption of ZTM for end-to-end networks enhancement are as follows:

- *Transparency, human agency and oversight*. ZTM-driven end-to-end networks must be able to explain the technical processes and decisions made by them to the telecom stakeholders, and human oversight mechanisms must be supported through governance mechanisms. However, most of the surveyed research considers traditional black-box AI/ML models such as deep neural networks, deep reinforcement learning for end-to-end network enhancement, which are not explainable to human stakeholders.
- *Technical robustness and safety*. ZTM-driven end-to-end networks must be resilient to adversarial attacks and have safeguards that enable fallback plans in case of issues. In the surveyed research work, there is a lack of assessment on the potential attack vectors for AI/ML models deployed in 5G/6G networks due to limitations on the considered 5G/6G use cases. Therefore, more efforts are needed from the research and industrial community to analyze various deployment options for AI/ML models in end-to-end mobile networks focusing on multi-vendor, multi-domain and multi-operator use cases.
- *Privacy and data governance*. ZTM-driven end-to-end networks must protect the privacy of sensitive data, ensure the integrity of data and standardize data protocols governing data access for ZTM systems to be put in place. In particular, most of the surveyed research proposes the use of FL with blockchain to protect the privacy and integrity of sensitive data in multi-vendor and multi-domain mobile networks. However, Homomorphic Encryption based ML allows analytical computations to be performed directly on the encrypted data, but more effort is needed from the research community to address computational and scaling challenges associated with it.

VII. Future Challenges and Research Perspectives

Also fuelled by the recent advances in the AI and the ML domains [23], zero-touch, and in general, ANM are receiving increasing attention from the research community. Without the goal of being exhaustive, in this section, we discuss a set of important research challenges that need to be addressed in the future in the ZTM domain.

A. Datasets and Labeling for Zero-Touch Management

As mentioned earlier, one of the principal goals of ZSM and ZTM is to provide a reference architecture for end-to-end management of future networks and services. This means to make a management environment in which all the operational processes, from the planning to the provisioning and monitoring, can be done automatically with full automation, ideally. Achieving such a goal can be attainable by leveraging AI and, in particular, machine learning tools to reach full automation of the network management process [267].

The success of ML models is highly reliant on the availability of rich and representative datasets. Currently, there is an immense volume of public network datasets out there, and it is growing immensely with the roll-out of 5G, which makes billions of devices able to connect to the network [276]. Despite the availability of a large volume of datasets, many data-related issues still need to be tackled for them to apply to real-world scenarios.

One of the main issues is the lack of *representative* datasets, which gets scarce when it comes to real-world scenarios. A dataset is not representative if it does not demonstrate the whole complexity and dynamicity of the network. Apart from that, the network domain is unique compared to many other domains in how different operators design, configure and deploy their networks. The *un-uniformity* is another issue in which datasets generated by one operator cannot be used for another operator. Additionally, mobile networks are very dynamic in the way that data generated at a specific duration of time might not be valid for future use cases due to changes in the network. Besides, network data sets suffer from common well-known data problems such as lack of labels, noise, sparsity, and privacy issues, which limits the range of ML algorithms that can be applied [125], [276].

As mentioned, the roll-out of 5G networks is at an early stage, so still, the lack of 5G specific data is a challenge regarding applying ML-based techniques for ZTM. Although there are efforts to generate 5G-specific data sets [267], most of them are either collected in non-realistic scenarios or do not represent the multi-dimensionality and dynamicity of the

environment. In this regard, there is much ongoing research to tackle these challenges. For example, the authors of [277] propose a framework called Emerge that extends the ideas from the NoMoNoise platform [278] to label networking data in a cost-effective fashion supporting privacy-preserving collaborations assuming the availability of good quality datasets. The ability of the framework is demonstrated by training an LSTM using the extracted labels.

In summary, AI/ML techniques are the main technological enablers towards reaching full network automation, as defined by ZTM. To exploit the full potential of ML methods, the common and domain-specific dataset-related problems should be treated carefully.

B. Explainable Zero-Touch Management

Taking into consideration the requirements of ZTM in which data-driven decisions require to be human-understandable, the vision on AI of the European Commission focuses on excellence, trust, and transparency [279], [280] and will have a fundamental role in shaping 6G networks and its support for what is starting to be known as *Quality of Trust (QoT)* [124]. As stated in the work in [112], XAI is an essential concept to be considered in ZTM since otherwise human operators would be unable to verify the actions issued by ML models. Consequently, a huge body of literature studies the issues of XAI in 5G and beyond 5G networks [123], [124], [281].

XAI defines the ability to provide reasoning behind the decisions made by AI systems to make them more trustworthy. This is especially relevant in services, such as remote surgery, where network management decisions could highly impact human lives. XAI allows networking experts to verify the decisions made by the ML models (e.g., traffic routing and prioritization), understand the input that drove them, and approve their actions following a human-in-the-loop model. This is even more important when various ML models interact with each other in the network. Discerning the ML model responsible for a network fault could be difficult if they are not explainable, since no information would be exposed about the reasoning followed on a certain action. Nevertheless, making XAI a reality in the ZTM paradigm involves several challenges, especially regarding the methods to generate intelligible ML models for networking experts and metrics to measure AI explainability in ZTM closed-loops.

Nowadays, the methods to obtain explainable models are covered by three main currents of thought. The first one advocates the use of inherently interpretable ML models, such as RF, that can apply to regression and classification problems and are suitable for the majority of networking tasks (e.g., networks operated by a single MNO/domain, and scenarios where sensible training data does not have to be transferred across different sites). The second stream of thinking supports the use of statistical procedures to describe the features on which a prediction was based [282]. The last trend promotes the design of surrogate/post-hoc models, aiming to derive an equivalent rule-based model that can provide hints on values determining a decision. In other words, this line of research pursues the placement of an intermediate component that can explain the possible options to be chosen when making a decision (e.g., link reconfiguration) and the cost associated with each (e.g., in terms of latency).

The level of explainability can be defined by interpretability and fidelity. However, measuring this level requires of dedicated metrics, which may be hard to assess due to the lack of ground truth. Metrics of different natures have been proposed to quantify XAI [283]. On the one hand, human-grounded evaluations test the ability of the resulting explanations to assist humans in completing tasks. Focusing on ZTM, the networking experts would receive explanations on the reasons from ML models to perform certain actions in the network. Based on that, the metrics measure qualitative aspects, such as the impact of the decisions in the system, their usefulness, and the trust/reliance of the audience. Nonetheless, this approach depends on the subject and needs a wide audience to clearly analyze the results. On the other hand, functionality-grounded evaluations rely on formal definitions and more quantitative methods, such as the depth of a decision tree from a post-hoc method. In ZTM, these metrics would be more appropriate when networking experts must verify the decisions made by the AI systems. By contrast, if the experts leverage AI assistance to make a decision, the first method would be a more suitable tool and reduce the complexity given by the amount of data and the number of AI pipelines that could be involved in a single task. A good example would be applications and service migration, where various ML models in charge of RAN allocation, VIM management, service orchestrator, etc., would be involved. However, regardless of the type of explainability metrics, further research is required to include it as a native part of beyond 5G and 6G systems.

C. Trustworthy Zero-Touch Management

The original vision for 5G/6G can be achieved only if the system is capable of managing complex chains of end-to-end services, built on assets from different resource providers, with a deep use of resource virtualization and sharing spanning computing, transport, radio, and spectrum resources. The long-term vision for pervasive computing, connectivity, and corresponding services in 5G/6G calls for the flexible, on-demand integration of ubiquitous computing, storage, and network resources, transparently crossing the borders of administrative domains. In this regard, ZTM is poised to catalyze the progressive introduction of 5G elements in operational infrastructures, which is facing the challenge of streamlining the multiple siloed management frameworks specifically tuned for different technologies, aiming at a uniform end-to-end management of the 5G services. While many advances were made in recent years in this regard, there are no fully automated operations systems for 5G addressing the challenges of sharing operational data and establishment of trust across multiple domains required for ZTM.

One approach to ZTM is driven by the idea of sharing operational data across the whole system in a logically centralized data reservoir (a.k.a. Data Lake) so that multiple asynchronous management components may act upon this shared data pool towards optimizing a target set of KPIs. Normally, the multi-stakeholder nature of typical 5G environments often prevents

parties from openly sharing information about resource availability, operational health, performance capacities, and service agreements. In addition, existing approaches tend to separate the provider and the service layers into different subsystems, which are often operated by different entities, making it hard to streamline the operations across the provider-service boundary.

One way to address these challenges is to leverage two major innovations concerning current 5G architectures, including a Distributed Ledger and an Operational Data Lake, as cornerstone components of the zero-touch management solution that ensures interoperability of the management system on the border between the multi-party Single Domain Layer and the unified Inter-domain Layer, as well as on the border between the Inter-domain Layer and the Evolved 5G Service Layer [284]. The 5G Operational Data Lake component serves as a logically centralized reservoir of all the operational data, channeled by management services of the Inter-domain Layer on behalf of domain-specific management services running in every domain of the Single Domain Layer. The 5G Operational Data Lake provides APIs for adding data, processing it (in place), retrieving it for analytical processes, etc., following the AIOps approach. These APIs can be invoked by the management components in the Inter-domain Layer and by the service components in the Evolved 5G Service Layer without incurring any unneeded coupling between the data providers and the data consumers. The 5G Permissioned Distributed Ledger component ensures the required interoperability by providing data governance, multi-party trust, and accounting for data usage by different participating parties.

D. Summary

ZTM will play a key role in future mobile systems and, in particular, in 5G and 6G networks. In this article, we have provided a comprehensive survey of the research literature focusing on both ANM and ZTM and wireless and mobile networking. We have concluded the article by discussing some open research challenges. We target this paper toward researchers and practitioners interested in learning the most recent trends in ZTM for the mobile network, with the hope that this can become a valuable guide for their work.

ACKNOWLEDGEMENTS

This work has been performed in the framework of the European Union's Horizon 2020 projects AI@EDGE and 5GZORRO co-funded by the EU under grant agreement No 101015922 and No 871533, respectively. The authors would also like to acknowledge CERCA Programme / Generalitat de Catalunya for sponsoring part of this work. This work has been also supported by the EU "NextGenerationEU/PRTR", MCIN and AEI (Spain) under project IJC2020-043058-I, and by the ONOFRE-3 PID2020-112675RB-C43 grant funded by MCIN/AEI/10.13039/501100011033.

REFERENCES

[1] T. Saydam and T. Magedanz, "From Networks and Network Management into Service and Service Management," *Journal of Network and Systems Management*, vol. 4, no. 4, pp. 345–348, 1996.

[2] D. L. Tennenhouse, J. M. Smith, W. D. Sincoskie, D. J. Wetherall, and G. J. Minden, "A Survey of Active Network Research," *IEEE Communications Magazine*, vol. 35, no. 1, pp. 80–86, 1997.

[3] 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, 2017.

[4] Z. Movahedi, M. Ayari, R. Langar, and G. Pujolle, "A Survey of Autonomic Network Architectures and Evaluation Criteria," *IEEE Communications Surveys Tutorials*, vol. 14, no. 2, pp. 464–490, 2012.

[5] ETSI, "Zero-touch Network and Service Management (ZSM); Reference Architecture," European Telecommunications Standards Institute, Group Specification (GS) ZSM 002, Aug. 2019, version 1.1.1.

[6] M. C. Huebscher and J. A. McCann, "A Survey of Autonomic Computing—degrees, Models, and Applications," *ACM Computing Surveys*, vol. 40, no. 3, pp. 1–28, 2008.

[7] N. Samaan and A. Karmouch, "Towards Autonomic Network Management: an Analysis of Current and Future Research Directions," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 3, pp. 22–36, 2009.

[8] Z. Zhao, E. Schiller, E. Kalogeiton, T. Braun, B. Stiller, M. T. Garip, J. Joy, M. Gerla, N. Akhtar, and I. Matta, "Autonomic Communications in Software-Driven Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2431–2445, 2017.

[9] S. Hämäläinen, H. Sanneck, and C. Sartori, *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency*. Wiley, 2012.

[10] H. Hu, J. Zhang, X. Zheng, Y. Yang, and P. Wu, "Self-configuration and Self-optimization for LTE Networks," *IEEE Communications Magazine*, vol. 48, no. 2, pp. 94–100, 2010.

[11] J. Moysen and L. Giupponi, "From 4G to 5G: Self-organized Network Management Meets Machine Learning," *Computer Communications*, vol. 129, pp. 248–268, 2018.

[12] M. Agiwal, A. Roy, and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[13] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[14] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A Survey on 5G Usage Scenarios and Traffic Models," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.

[15] F. Giust, L. Cominardi, and C. J. Bernardos, "Distributed Mobility Management for Future 5G Networks: Overview and Analysis of Existing Approaches," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 142–149, Jan. 2015.

[16] A. Gupta and R. K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[17] N. Panwar, S. Sharma, and A. K. Singh, "A Survey on 5G: The Next Generation of Mobile Communication," *Physical Communication*, vol. 18, pp. 64–84, 2016.

[18] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised Machine Learning: A Review of Classification Techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, pp. 3–24, 2007.

[19] R. Xu and D. Wunsch, "Survey of Clustering Algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[20] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y. Liang, and D. I. Kim, "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[21] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. Iyengar, "A Survey on Deep Learning: Algorithms, Techniques, and Applications," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–36, 2018.

[22] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[23] C. Zhang, P. Patras, and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2224–2287, 2019.

[24] A. Mughees, M. Tahir, M. A. Sheikh, and A. Ahad, "Towards Energy Efficient 5G Networks Using Machine Learning: Taxonomy, Research Challenges, and Future Research Directions," *IEEE Access*, vol. 8, pp. 187 498–187 522, 2020.

[25] B. Ma, W. Guo, and J. Zhang, "A Survey of Online Data-Driven Proactive 5G Network Optimisation Using Machine Learning," *IEEE Access*, vol. 8, pp. 35 606–35 637, 2020.

[26] A. Stamou, N. Dimitriou, K. Kontovasilis, and S. Papavassiliou, "Autonomic Handover Management for Heterogeneous Networks in a Future Internet Context: A Survey," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 4, pp. 3274–3297, 2019.

[27] S. Ayoubi, N. Limam, M. A. Salahuddin, N. Shahriar, R. Boutaba, F. Estrada-Solano, and O. M. Caicedo, "Machine Learning for Cognitive Network Management," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 158–165, 2018.

[28] J. Xie, F. R. Yu, T. Huang, R. Xie, J. Liu, C. Wang, and Y. Liu, "A Survey of Machine Learning Techniques Applied to Software Defined Networking (SDN): Research Issues and Challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 393–430, 2018.

[29] A. Mestres, A. Rodriguez-Natal, J. Carner, P. Barlet-Ros, E. Alarcón, M. Solé, V. Muntés-Mulero, D. Meyer, S. Barkai, M. J. Hibbett *et al.*, "Knowledge-defined Networking," *ACM SIGCOMM Computer Communication Review*, vol. 47, no. 3, pp. 2–10, 2017.

[30] T. L. Duc, R. Leiva, P. Casari, and P. Ostberg, "Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey," *ACM Computing Surveys*, vol. 52, no. 5, 2019.

[31] M. Zorzi, A. Zanella, A. Testolin, M. D. F. De Grazia, and M. Zorzi, "Cognition-Based Networks: A New Perspective on Network Optimization Using Learning and Distributed Intelligence," *IEEE Access*, vol. 3, pp. 1512–1530, 2015.

[32] L. Pang, C. Yang, D. Chen, Y. Song, and M. Guizani, "A Survey on Intent-Driven Networks," *IEEE Access*, vol. 8, pp. 22 862–22 873, 2020.

[33] Qingyue Long and Yanliang Chen and H. Zhang and Xianfu Lei, "Software Defined 5G and 6G Networks: a Survey," *Mobile Networks and Applications*, pp. 1–21, 2019.

[34] L. U. Khan, I. Yaqoob, M. Imran, Z. Han, and C. S. Hong, "6G Wireless Systems: A Vision, Architectural Elements, and Future Directions," *IEEE Access*, vol. 8, pp. 147 029–147 044, 2020.

[35] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[36] P. Yang, Y. Xiao, M. Xiao, and S. Li, "6G Wireless Communications: Vision and Potential Techniques," *IEEE Network*, vol. 33, no. 4, pp. 70–75, 2019.

[37] L. Bariah, L. Mohjazi, S. Muhaidat, P. C. Sofotasios, G. K. Kurt, H. Yanikomeroglu, and O. A. Dobre, "A Prospective Look: Key Enabling Technologies, Applications and Open Research Topics in 6G Networks," *IEEE Access*, vol. 8, pp. 174 792–174 820, 2020.

[38] A. Dogra, R. K. Jha, and S. Jain, "A Survey on Beyond 5G Network with the Advent of 6G: Architecture and Emerging Technologies," *IEEE Access*, pp. 1–1, 2020.

[39] O. Serhane, K. Yahyaoui, B. Nour, and H. Moungla, "A Survey of ICN Content Naming and In-network Caching in 5G and Beyond Networks," *IEEE Internet of Things Journal*, pp. 1–1, 2020.

[40] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The Roadmap to 6G: AI Empowered Wireless Networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[41] Y. Sun, J. Liu, J. Wang, Y. Cao, and N. Kato, "When Machine Learning Meets Privacy in 6G: A Survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 4, pp. 2694–2724, 2020.

[42] R. Alghamdi, R. Alhadrami, D. Alhothali, H. Almorad, A. Faisal, S. Helal, R. Shalabi, R. Asfour, N. Hammad, A. Shams, N. Saeed, H. Dahrouj, T. Y. Al-Naffouri, and M. S. Alouini, "Intelligent Surfaces for 6G Wireless Networks: A Survey of Optimization and Performance Analysis Techniques," *IEEE Access*, vol. 8, pp. 202 795–202 818, 2020.

[43] S. Zhang and D. Zhu, "Towards Artificial Intelligence Enabled 6G: State of the Art, Challenges, and Opportunities," *Computer Networks*, vol. 183, p. 107556, 2020.

[44] M. E. Morocho-Cayamcela, H. Lee, and W. Lim, "Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions," *IEEE Access*, vol. 7, pp. 137 184–137 206, 2019.

[45] S. J. Nawaz, S. K. Sharma, S. Wyne, M. N. Patwary, and M. Asaduzzaman, "Quantum Machine Learning for 6G Communication Networks: State-of-the-Art and Vision for the Future," *IEEE Access*, vol. 7, pp. 46 317–46 350, 2019.

[46] Y. Wei, M. Peng, and Y. Liu, "Intent-based Networks for 6G: Insights and Challenges," *Digital Communications and Networks*, vol. 6, no. 3, pp. 270–280, 2020.

[47] C. Benzaid and T. Taleb, "AI-Driven Zero Touch Network and Service Management in 5G and Beyond: Challenges and Research Directions," *IEEE Network*, vol. 34, no. 2, pp. 186–194, 2020.

[48] ——, "ZSM Security: Threat Surface and Best Practices," *IEEE Network*, vol. 34, no. 3, pp. 124–133, 2020.

[49] J. Gallego-Madrid, R. Sanchez-Iborra, P. M. Ruiz, and A. F. Skarmeta, "Machine learning-based zero-touch network and service management: A survey," *Digital Communications and Networks*, 2021.

[50] M. Liyanage, Q.-V. Pham, K. Dev, S. Bhattacharya, P. Reddy, T. Gadekallu, and G. Yenduri, "A Survey on Zero Touch Network and Service (ZSM) Management for 5G and Beyond Networks," *Journal of Network and Computer Applications*, 03 2022.

[51] ETSI, "Zero-touch Network and Service Management (ZSM); Terminology for concepts in ZSM," European Telecommunications Standards Institute, Group Specification (GS) ZSM 007, Aug. 2019, version 1.1.1.

[52] ——, "Zero-touch Network and Service Management (ZSM); Means of Automation," European Telecommunications Standards Institute, Group Report (GR) ZSM 005, May 2020, version 1.1.1.

[53] ——, "Zero-touch Network and Service Management (ZSM); Landscape," European Telecommunications Standards Institute, Group Report (GR) ZSM 004, Mar. 2020, version 1.1.1.

[54] ——, "Experiential Networked Intelligence (ENI); Terminology for Main Concepts in ENI," European Telecommunications Standards Institute, Group Report (GR) ENI 004, Oct. 2019, version 2.1.1.

[55] ——, "Experiential Networked Intelligence (ENI); System Architecture," European Telecommunications Standards Institute, Group Specification (GS) ENI 005, Sep. 2019, version 1.1.1.

[56] 3GPP, "Study on Enhancements to the Service-Based Architecture," 3rd Generation Partnership Project (3GPP), Technical Report (TR) TR 23.742, Feb. 2018, version 2.0.0.

[57] ——, "5G System; Network Data Analytics Services," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) TS 29.520, Dec. 2020, version 17.1.0.

[58] ——, "Study of Enablers for Network Automation for 5G," 3rd Generation Partnership Project (3GPP), Technical Report (TR) TR 23.791, Jun. 2019, version 16.2.0.

[59] ——, "Performance Management (PM) for Mobile Networks that Include Virtualized Network Functions," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) TS 28.521, Jun. 2018, version 15.0.0.

[60] ——, "Fault Management (FM) for Mobile Networks that Include Virtualized Network Functions; Procedures," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) TS 28.516, Jun. 2018, version 15.0.0.

[61] ——, "Configuration Management (CM) for Mobile Networks that Include Virtualized Network Functions; Procedures," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) TS 28.511, Aug. 2020, version 16.0.0.

[62] ——, "Policy Management for Network Function Virtualization (NFV) Based Mobile Networks," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) TS 28.311, Dec. 2019, version 1.0.0.

[63] ——, "Study on Scenarios for Intent Driven Management Services for Mobile Networks (Release 16)," 3rd Generation Partnership Project (3GPP), Technical Report (TR) TR 28.812, Dec. 2020, version 17.1.0.

[64] ——, "Study on the Self-Organizing Networks (SON) for 5G Networks (Release 16)," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 28.861, Dec. 2019, version 2.0.0.

[65] ——, "Study on Management and Orchestration of Network Slicing for Next Generation Network," 3rd Generation Partnership Project (3GPP), Technical Report (TR) TR 28.801, Jan. 2018, version 15.1.0.

[66] TM Forum, "Management Platform Blueprint and Application to Hybrid Infrastructure," Tele Management Forum, Technical Report (TR) TR262, Apr. 2018, version 17.5.1.

[67] ——, "User Stories for Hybrid Infrastructure Platform," Tele Management Forum, Technical Report (TR) TR229A, Nov. 2017, version 17.0.1.

[68] MEF Forum, "Lifecycle Service Orchestration (LSO): Reference Architecture and Framework," Metro Ethernet Forum, MEF Specification MEF 55, Mar. 2016.

[69] ——, "MEF Core Model (MCM)," Metro Ethernet Forum, MEF Standard MEF 78.1, Jul. 2020.

[70] ——, "Business and Operational Aspects of Implementing LSO Sonata," Metro Ethernet Forum, MEF White Paper, Feb. 2020.

[71] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, 2008.

[72] T. L. Hinrichs, N. S. Gude, M. Casado, J. C. Mitchell, and S. Shenker, "Practical Declarative Network Management," in *Proc. of ACM WREN*, Barcelona, Spain, 2009.

[73] C. Monsanto, J. Reich, N. Foster, J. Rexford, and D. Walker, "Composing Software-defined Networks," in *Proc. of USENIX NSDI*, Lombard, IL, USA, 2013.

[74] A. Voellmy and P. Hudak, "Nettle: Taking the Sting out of Programming Network Routers," in *Proc. of ACM PADL*, 2011.

[75] N. Foster, R. Harrison, M. J. Freedman, C. Monsanto, J. Rexford, A. Story, and D. Walker, "Frenetic: A Network Programming Language," *ACM SIGPLAN Notices*, vol. 46, no. 9, pp. 279–291, 2011.

[76] A. Voellmy, H. Kim, and N. Feamster, "Procera: A Language for High-level Reactive Network Control," in *Proc. of ACM HotSDN*, Helsinki, Finland, 2012.

[77] A. K. Nayak, A. Reimers, N. Feamster, and R. Clark, "Resonance: Dynamic Access Control for Enterprise Networks," in *Proc. of ACM WREN*, Barcelona, Spain, 2009.

[78] C. Monsanto, N. Foster, R. Harrison, and D. Walker, "A Compiler and Run-time System for Network Programming Languages," in *Proc. of ACM POPL*, Philadelphia, PA, USA, 2012.

[79] R. Riggio, M. K. Marina, J. Schulz-Zander, S. Kuklinski, and T. Rasheed, "Programming Abstractions for Software-Defined Wireless Networks," *IEEE Transactions on Network and Service Management*, vol. 12, no. 2, pp. 146–162, 2015.

[80] E. Coronado, S. N. Khan, and R. Riggio, "5G-EmPOWER: A Software-Defined Networking Platform for 5G Radio Access Networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 715–728, 2019.

[81] E. Coronado, A. Thomas, S. Bayhan, and R. Riggio, "aiOS: An Intelligence Layer for SD-WLANs," in *Proc. of IEEE NOMS*, Budapest, Hungary, 2020.

[82] K. Koutlia, A. Umbert, S. Garcia, and F. Casadevall, "RAN slicing for Multi-tenancy Support in a WLAN Scenario," in *Proc. of IEEE NetSoft*, Bologna, Italy, 2017.

[83] I. Vermeulen, P. Bosch, T. De Schepper, and S. Latre, "DiMob: Scalable and Seamless Mobility in SDN Managed Wireless Networks," in *Proc. of IEEE CNSM*, Tokyo, Japan, 2017.

[84] B. Mafakheri, L. Goratti, R. Abbas, S. Reisenfeld, and R. Riggio, "LTE/Wi-Fi Coordination in Unlicensed Bands: An SD-RAN Approach," in *Proc. of IEEE NetSoft*, Paris, France, 2019.

[85] E. F. Aza and J. P. Urrea, "Implementation of Round-Robin Load Balancing Scheme in a Wireless Software Defined Network," in *Proc. of IEEE COLCOM*, Barranquilla, Colombia, 2019.

[86] D. Harutyunyan, S. Herle, D. Maradin, G. Agapiu, and R. Riggio, "Traffic-aware User Association in Heterogeneous LTE/WiFi Radio Access Networks," in *Proc. of IEEE NOMS*, Taipei, Taiwan, 2018.

[87] L. Suresh, J. Schulz-Zander, R. Merz, A. Feldmann, and T. Vazao, "Towards Programmable Enterprise WLANs with Odin," in *Proc. of ACM HotSDN*, Helsinki, Finland, 2012.

[88] J. Schulz-Zander, L. Suresh, N. Sarrar, A. Feldmann, T. Hühn, and R. Merz, "Programmatic Orchestration of WiFi Networks," in *Proc. of USENIX ATC*, 2014.

[89] R. Riggio, C. Sengul, L. Suresh, J. Schulz–zander, and A. Feldmann, "Thor: Energy programmable WiFi networks," in *Proc. of IEEE INFOCOM WKSHPS*, Turin, Italy, 2013.

[90] Z. Han, T. Lei, Z. Lu, X. Wen, W. Zheng, and L. Guo, "Artificial Intelligence-Based Handoff Management for Dense WLANs: A Deep Reinforcement Learning Approach," *IEEE Access*, vol. 7, pp. 31 688–31 701, 2019.

[91] S. Misra and N. Saha, "Detour: Dynamic Task Offloading in Software-Defined Fog for IoT Applications," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 5, pp. 1159–1166, 2019.

[92] A. Patro and S. Banerjee, "COAP: A Software-defined Approach for Home WLAN Management Through an Open API," in *Proc. of ACM MobiArch*, Maui, HI, USA, 2014.

[93] J. Schulz-Zander, N. Sarrar, and S. Schmid, "Towards a Scalable and Near-sighted Control Plane Architecture for WiFi SDNs," in *Proc. of ACM HotSDN*, Chicago, Illinois, USA, 2014.

[94] J. Schulz-Zander, S. Schmid, J. Kempf, R. Riggio, and A. Feldmann, "LegoFi the Wifi Building Blocks! The Case for a Modular Wifi Architecture," in *Proc. of ACM MobiArch*, New York, NY, USA, 2016.

[95] F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "NFV and SDN - Key Technology Enablers for 5G Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2468–2478, 2017.

[96] R. P. Goldberg, "Architecture of Virtual Machines," in *Proc. of workshop on virtual computer systems*, New York, NY, USA, 1973.

[97] R. Behravesh, E. Coronado, and R. Riggio, "Performance Evaluation on Virtualization Technologies for NFV Deployment in 5G Networks," in *Proc. of IEEE NetSoft*, Paris, France, 2019.

[98] D. R. Engler, M. F. Kaashoek, and J. O'Toole Jr, "Exokernel: An Operating System Architecture for Application-level Resource Management," *ACM SIGOPS Operating Systems Review*, vol. 29, no. 5, pp. 251–266, 1995.

[99] A. Madhavapeddy, R. Mortier, C. Rotsos, D. Scott, B. Singh, T. Gazagnaire, S. Smith, S. Hand, and J. Crowcroft, "Unikernels: Library Operating Systems for the Cloud," *ACM SIGARCH Computer Architecture News*, vol. 41, no. 1, pp. 461–472, 2013.

[100] N. F. S. De Sousa, D. A. L. Perez, R. V. Rosa, M. A. Santos, and C. E. Rothenberg, "Network Service Orchestration: A Survey," *Computer Communications*, vol. 142, pp. 69–94, 2019.

[101] ETSI, "GS NFV 002 V1.2.1," *Network Functions Virtualisation (NFV); Architectural Framework*, 2014.

[102] ——. Open Source MANO. Accessed on 23.09.2022. [Online]. Available: https://osm.etsi.org/

[103] GigaSpaces. Cloudify. Accessed on 23.09.2022. [Online]. Available: https://cloudify.co/

[104] T. Fraunhofer, Berlin. An Open Source Reference Implementation of the ETSI Network Function Virtualization MANO Specification. Accessed on 23.09.2022. [Online]. Available: https://openbaton.github.io/

[105] L. Foundation. ONAP – Open Network AutomationPlatform. Accessed on 23.09.2022. [Online]. Available: http://openbaton.github.io/

[106] A. Masood and A. Hashmi, *AIOps: Predictive Analytics and Machine Learning in Operations*. Apress, 2019, pp. 359–382.

[107] M. Mormul and C. Stach, "A Context Model for Holistic Monitoring and Management of Complex IT Environments," in *Proc. of IEEE PerCom Workshops*, Austin, TX, USA, 2020.

[108] Q. Chen, Z. Zheng, C. Hu, D. Wang, and F. Liu, "On-Edge Multi-Task Transfer Learning: Model and Practice With Data-Driven Task Allocation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 6, pp. 1357–1371, 2020.

[109] 3GPP, "5G; System architecture for the 5G System (5GS); (Release 16)," 3rd Generation Partnership Project (3GPP), Tech. Rep. TS 23.501, Oct. 2020, version 16.6.0.

[110] A. Banchs, D. M. Gutierrez-Estevez, M. Fuentes, M. Boldi, and S. Provvedi, "A 5G Mobile Network Architecture to Support Vertical Industries," *IEEE Communications Magazine*, vol. 57, no. 12, pp. 38–44, 2019.

[111] S. Sevgican, M. Turan, K. Gökarslan, H. B. Yilmaz, and T. Tugcu, "Intelligent Network Data Analytics Function in 5G Cellular Networks Using Machine Learning," *Journal of Communications and Networks*, vol. 22, no. 3, pp. 269–280, 2020.

[112] R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed, and J. C. Zhang, "Artificial Intelligence-Enabled Cellular Networks: A Critical Path to Beyond-5G and 6G," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 212–217, 2020.

[113] ETSI, "Experiential Networked Intelligence (ENI); System Architecture," European Telecommunications Standards Institute, Tech. Rep. ETSI GS ENI 005, Sep. 2019, version 1.1.1.

[114] 3GPP, "Study on enhancement of Management Data Analytics (MDA) (Release 17)," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 28.812, Nov. 2020, version 17.1.0.

[115] ——, "5G; LTE; Management and orchestration; Management services for communication service assurance; Requirements (Release 16)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) TS 28.535, Jan. 2021, version 16.2.0.

[116] "O-RAN Alliance." [Online]. Available: https://www.o-ran.org/

[117] O-RAN, "O-RAN Architecture Description," O-RAN Alliance, Technical Specification, Oct. 2020, version 2.0.

[118] O-RAN Working Group 2, "AI/ML Workflow Description and Requirements," O-RAN Alliance, Technical Report, March 2020, v01.01.

[119] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5G: When Cellular Networks Meet Artificial Intelligence," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 175–183, 2017.

[120] Y. Fu, S. Wang, C. Wang, X. Hong, and S. McLaughlin, "Artificial Intelligence to Manage Network Traffic of 5G Wireless Networks," *IEEE Network*, vol. 32, no. 6, pp. 58–64, 2018.

[121] S. Fu, F. Yang, and Y. Xiao, "AI Inspired Intelligent Resource Management in Future Wireless Network," *IEEE Access*, vol. 8, pp. 22 425–22 433, 2020.

[122] O. Simeone, "A Very Brief Introduction to Machine Learning With Applications to Communication Systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 648–664, 2018.

[123] W. Guo, "Explainable Artificial Intelligence for 6G: Improving Trust between Human and Machine," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 39–45, 2020.

[124] C. Li, W. Guo, S. C. Sun, S. Al-Rubaye, and A. Tsourdos, "Trustworthy Deep Learning in 6G-Enabled Mass Autonomy: From Concept to Quality-of-Trust Key Performance Indicators," *IEEE Vehicular Technology Magazine*, vol. 15, no. 4, pp. 112–121, 2020.

[125] M. Usama, J. Qadir, A. Raza, H. Arif, K. A. Yau, Y. Elkhatib, A. Hussain, and A. Al-Fuqaha, "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges," *IEEE Access*, vol. 7, pp. 65 579–65 615, 2019.

[126] P. Gawłowicz and A. Zubow, "Ns-3 Meets OpenAI Gym: The Playground for Machine Learning in Networking Research," in *Proc. of ACM MSWIM*, Miami Beach, FL, USA, 2019.

[127] C. Natalino and P. Monti, "The Optical RL-Gym: An open-source toolkit for applying reinforcement learning in optical networks," in *Proc. of IEEE ICTON*, Bari, Italy, 2020.

[128] S. Zhao, M. Talasila, G. Jacobson, C. Borcea, S. A. Aftab, and J. F. Murray, "Packaging and Sharing Machine Learning Models via the Acumos AI Open Platform," in *Proc. of IEEE ICMLA*, Orlando, FL, USA, 2018.

[129] C. V. Nahum, L. De Nóvoa Martins Pinto, V. B. Tavares, P. Batista, S. Lins, N. Linder, and A. Klautau, "Testbed for 5G Connected Artificial Intelligence on Virtualized Networks," *IEEE Access*, vol. 8, pp. 223 202–223 213, 2020.

[130] N. Nikaein, C.-Y. Chang, and K. Alexandris, "Mosaic5G: Agile and Flexible Service Platforms for 5G Research," *SIGCOMM Computer Communication Review*, vol. 48, no. 3, 2018.

[131] C. Huang, C. Ho, N. Nikaein, and R. Cheng, "Design and Prototype of A Virtualized 5G Infrastructure Supporting Network Slicing," in *Proc. of IEEE DSP*, Shanghai, China, 2018.

[132] M. A. Qureshi and C. Tekin, "Fast Learning for Dynamic Resource Allocation in AI-Enabled Radio Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 95–110, 2020.

[133] M. Hashemi, A. Sabharwal, C. Emre Koksal, and N. B. Shroff, "Efficient Beam Alignment in Millimeter Wave Systems Using Contextual Bandits," in *Proc. of IEEE INFOCOM*, Honolulu, HI, USA, 2018.

[134] N. Modi, P. Mary, and C. Moy, "QoS Driven Channel Selection Algorithm for Cognitive Radio Network: Multi-User Multi-Armed Bandit Approach," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 1, pp. 49–66, 2017.

[135] J. Chen, Z. Wei, S. Li, and B. Cao, "Artificial Intelligence Aided Joint Bit Rate Selection and Radio Resource Allocation for Adaptive Video Streaming over F-RANs," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 36–43, 2020.

[136] P. Yu, F. Zhou, X. Zhang, X. Qiu, M. Kadoch, and M. Cheriet, "Deep Learning-Based Resource Allocation for 5G Broadband TV Service," *IEEE Transactions on Broadcasting*, vol. 66, no. 4, pp. 800–813, 2020.

[137] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel State Information Prediction for 5G Wireless Communications: A Deep Learning Approach," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 1, pp. 227–236, 2020.

[138] E. Coronado, A. Thomas, and R. Riggio, "Adaptive ML-based Frame Length Optimisation in Enterprise SD-WLANs," *Jorunal of Network and Systems Management*, vol. -, pp. 1–32, 2020.

[139] M. Elsayed, M. Erol-Kantarci, B. Kantarci, L. Wu, and J. Li, "Low-Latency Communications for Community Resilience Microgrids: A Reinforcement Learning Approach," *IEEE Transactions on Smart Grid*, vol. 11, no. 2, pp. 1091–1099, 2020.

[140] X. Zhang, P. Yu, L. Feng, F. Zhou, and W. Li, "A DRL-based Resource Allocation Framework for Multimedia Multicast in 5G Cellular Networks," in *Proc. of IEEE BMSB*, Jeju, Korea (South), 2019.

[141] S. Yoon, J.-H. Cho, D. S. Kim, T. J. Moore, F. Free-Nelson, and H. Lim, "DESOLATER: Deep Reinforcement Learning-Based Resource Allocation and Moving Target Defense Deployment Framework," *IEEE Access*, vol. 9, pp. 70 700–70 714, 2021.

[142] G. Vallero, D. Renga, M. Meo, and M. A. Marsan, "Greener RAN Operation Through Machine Learning," *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 896–908, 2019.

[143] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan, and D. Matolak, "A Machine Learning Approach for Power Allocation in HetNets Considering QoS," in *Proc. of IEEE ICC*, Kansas City, MO, USA, 2018.

[144] H. Saad, A. Mohamed, and T. ElBatt, "Distributed Cooperative Q-Learning for Power Allocation in Cognitive Femtocell Networks," in *Proc. of IEEE VTC Fall*, Quebec City, QC, Canada, 2012.

[145] D. Xu, X. Chen, C. Wu, S. Zhang, S. Xu, and S. Cao, "Energy-Efficient Subchannel and Power Allocation for HetNets Based on Convolutional Neural Network," in *Proc. of IEEE VTC Spring*, Kuala Lumpur, Malaysia, 2019.

[146] S. Khan Tayyaba, H. A. Khattak, A. Almogren, M. A. Shah, I. Ud Din, I. Alkhalifa, and M. Guizani, "5G Vehicular Network Resource Management for Improving Radio Access Through Machine Learning," *IEEE Access*, vol. 8, pp. 6792–6800, 2020.

[147] A. Nassar and Y. Yilmaz, "Resource Allocation in Fog RAN for Heterogeneous IoT Environments Based on Reinforcement Learning," in *Proc. of ICC*, Shanghai, China, 2019.

[148] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A Deep Reinforcement Learning Based Framework for Power-efficient Resource Allocation in Cloud RANs," in *Proc. of ICC*, Paris, France, 2017.

[149] B. Angui, R. Corbel, V. Q. Rodriguez, and E. Stephan, "Towards 6G zero touch networks: The case of automated Cloud-RAN deployments," in *Proc. of IEEE CCNC*, Las Vegas, NV, USA, 2022.

[150] Y. Zhang, L. Xiong, and J. Yu, "Deep Learning Based User Association in Heterogeneous Wireless Networks," *IEEE Access*, vol. 8, pp. 197 439–197 447, 2020.

[151] N. Zhao, Y. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep Reinforcement Learning for User Association and Resource Allocation in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5141–5152, 2019.

[152] S. Zhao, X. Jiang, G. Jacobson, R. Jana, W. L. Hsu, R. Rustamov, M. Talasila, S. A. Aftab, Y. Chen, and C. Borcea, "Cellular Network Traffic Prediction Incorporating Handover: A Graph Convolutional Approach," in *Proc. of IEEE SECON*, Como, Italy, 2020.

[153] E. Zeljković, N. Slamnik-Krieštorac, S. Latré, and J. M. Marquez-Barja, "ABRAHAM: Machine Learning Backed Proactive Handover Algorithm Using SDN," *IEEE Transactions on Network and Service Management*, vol. 16, no. 4, pp. 1522–1536, 2019.

[154] Z. Ali, L. Giupponi, M. Miozzo, and P. Dini, "Multi-Task Learning for Efficient Management of Beyond 5G Radio Access Network Architectures," *IEEE Access*, vol. 9, pp. 158 892–158 907, 2021.

[155] R. Atawia and H. Gacanin, "Self-Deployment of Future Indoor Wi-Fi Networks: An Artificial Intelligence Approach," in *Proc. of IEEE GLOBECOM*, Singapore, Singapore, 2017.

[156] H. Gacanin, E. Perenda, and R. Atawia, "Self-Deployment of Non-Stationary Wireless Systems by Knowledge Management With Artificial Intelligence," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 4, pp. 1004–1018, 2019.

[157] G. H. Apostolo, F. Bernardini, L. C. S. Magalhães, and D. C. Muchaluat-Saade, "A Unified Methodology to Predict Wi-Fi Network Usage in Smart Buildings," *IEEE Access*, vol. 9, pp. 11 455–11 469, 2021.

[158] B. Khodapanah, A. Awada, I. Viering, A. n. Barreto, M. Simsek, and G. Fettweis, "Framework for Slice-Aware Radio Resource Management Utilizing Artificial Neural Networks," *IEEE Access*, vol. 8, pp. 174 972–174 987, 2020.

[159] Y. Abiko, T. Saito, D. Ikeda, K. Ohta, T. Mizuno, and H. Mineno, "Flexible Resource Block Allocation to Multiple Slices for Radio Access Network Slicing Using Deep Reinforcement Learning," *IEEE Access*, vol. 8, pp. 68 183–68 198, 2020.

[160] ——, "Radio Resource Allocation Method for Network Slicing using Deep Reinforcement Learning," in *Proc. of IEEE ICOIN*, Barcelona, Spain, 2020.

[161] G. Sun, Z. T. Gebrekidan, G. O. Boateng, D. Ayepah-Mensah, and W. Jiang, "Dynamic Reservation and Deep Reinforcement Learning Based Autonomous Resource Slicing for Virtualized Radio Access Networks," *IEEE Access*, vol. 7, pp. 45 758–45 772, 2019.

[162] C. Gutterman, E. Grinshpun, S. Sharma, and G. Zussman, "RAN Resource Usage Prediction for a 5G Slice Broker," in *Proc. of ACM MobiHoc*, Catania, Italy, 2019.

[163] D. W. Yaping Cui, Xinyun Huang and H. Zheng, "Machine Learning-Based Resource Allocation Strategy for Network Slicing in Vehicular Networks," *Wireless Communications and Mobile Computing*, vol. 2020, 2020.

[164] V. Sciancalepore, X. Costa-Perez, and A. Banchs, "RL-NSB: Reinforcement Learning-Based 5G Network Slice Broker," *IEEE/ACM Transactions on Networking*, vol. 27, no. 4, 2019.

[165] M. R. Raza, C. Natalino, P. Öhlen, L. Wosinska, and P. Monti, "Reinforcement Learning for Slicing in a 5G Flexible RAN," *Journal of Lightwave Technology*, vol. 37, no. 20, pp. 5161–5169, 2019.

[166] F. Rezazadeh, H. Chergui, L. Alonso, and C. Verikoukis, "Continuous Multi-Objective Zero-Touch Network Slicing via Twin Delayed DDPG

and OpenAI Gym," in *Proc. of IEEE GLOBECOM*, Taipei, Taiwan, 2020.

[167] H. Chergui and C. Verikoukis, "OPEX-Limited 5G RAN Slicing: an Over-Dataset Constrained Deep Learning Approach," in *Proc. of ICC*, Dublin, Ireland, 2020.

[168] J. Backman, S. Yrjölä, K. Valtanen, and O. Mämmelä, "Blockchain Network Slice Broker in 5G: Slice Leasing in Factory of the Future Use Case," in *Proc. of IEEE Internet of Things Business Models, Users, and Networks*, Copenhagen, Denmark, 2017.

[169] W. Lin, X. Xu, L. Qi, X. Zhang, W. Dou, and M. R. Khosravi, "A Proof-of-Majority Consensus Protocol for Blockchain-Enabled Collaboration Infrastructure of 5G Network Slice Brokers," in *Proc. of ACM BSCI*, Taipei, Taiwan, 2020.

[170] L. Bonati, S. D'Oro, M. Polese, S. Basagni, and T. Melodia, "Intelligence and Learning in O-RAN for Data-driven NextG Cellular Networks," in *arXiv Preprint 2012.01263*, 2020.

[171] H. Lee, J. Cha, D. Kwon, M. Jeong, and I. Park, "Hosting AI/ML Workflows on O-RAN RIC Platform," in *Proc. of IEEE GLOBECOM Workshops*, Taipei, Taiwan, 2020.

[172] H. Kumar, V. Sapru, and S. K. Jaisawal, "O-RAN based proactive ANR optimization," in *Proc. of IEEE Globecom Workshops*, Taipei, Taiwan, 2020.

[173] P. E. I. Rivera, S. Mollahasani, and M. Erol-Kantarci, "Multi Agent Team Learning in Disaggregated Virtualized Open Radio Access Networks (O-RAN)," in *arXiv Preprint 2012.04861*, 2021.

[174] Y. Yuan, J. Yang, R. Duan, I. Chih-Lin, and J. Huang, "Anomaly Detection and Root Cause Analysis Enabled by Artificial Intelligence," in *Proc. of IEEE GLOBECOM Workshops*, Taipei, Taiwan, 2020.

[175] 3GPP, "Study on access traffic steering, switch and splitting support in the 5G system architecture (Release 16)," 3rd Generation Partnership Project (3GPP), Technical Report (TR) TR 23.793, Dec. 2018, version 16.0.0.

[176] R. Li, Z. Zhao, Q. Sun, C. I, C. Yang, X. Chen, M. Zhao, and H. Zhang, "Deep Reinforcement Learning for Resource Management in Network Slicing," *IEEE Access*, vol. 6, pp. 74 429–74 441, 2018.

[177] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A Substrate for Virtualizing Wireless Resources in Cellular Networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1333–1346, 2012.

[178] R. Mahindra, M. A. Khojastepour, H. Zhang, and S. Rangarajan, "Radio Access Network sharing in cellular networks," in *Proc. of IEEE ICNP*, Goettingen, Germany, 2013.

[179] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in *Proc. of PMLR*, Stockholm, Sweden, 2018.

[180] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. M. O. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv*, 2015.

[181] D. M. Coleman, K. R. McKeever, M. L. Mohr, L. R. Orraca Rosario, K. E. Parker, and M. L. Plett, "Overview of the Colosseum: The World's Largest Test Bed for Radio Experiments," *Johns Hopkins APL Technical Digest*, vol. 35, no. 1, 2019.

[182] R. Banerji, N. Gupta, S. Kumar, S. Singh, A. Bhat, B. J. R. Sahu, and S. Yoon, "ONAP Based Pro-Active Access Discovery and Selection for 5G Networks," in *Proc. of IEEE WCNCW*, Seoul, Korea (South), 2020.

[183] S. Vassilaras, L. Gkatzikis, N. Liakopoulos, I. N. Stiakogiannakis, M. Qi, L. Shi, L. Liu, M. Debbah, and G. S. Paschos, "The Algorithmic Aspects of Network Slicing," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 112–119, 2017.

[184] Y. Siriwardhana, P. Porambage, M. Ylianttila, and M. Liyanage, "Performance Analysis of Local 5G Operator Architectures for Industrial Internet," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 559–11 575, 2020.

[185] V. P. Kafle, Y. Fukushima, P. Martinez-Julia, and T. Miyazawa, "Consideration On Automation of 5G Network Slicing with Machine Learning," in *Proc. of IEEE ITU K*, Santa Fe, Argentina, 2018.

[186] V. P. Kafle, P. Martinez-Julia, and T. Miyazawa, "Automation of 5G Network Slice Control Functions with Machine Learning," *IEEE Communications Standards Magazine*, vol. 3, no. 3, pp. 54–62, 2019.

[187] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "DeepCog: Cognitive Network Management in Sliced 5G Networks with Deep Learning," in *Proc. of IEEE INFOCOM*, Paris, France, 2019.

[188] ——, "AZTEC: Anticipatory Capacity Allocation for Zero-Touch Network Slicing," in *Proc. of IEEE INFOCOM*, Toronto, ON, Canada, 2020.

[189] S. Vittal, S. Sarkar, P. P S, and A. F. A, "A Zero Touch Emulation Framework for Network Slicing Management in a 5G Core Testbed," in *Proc. of IEEE CNSM*, Izmir, Turkey, 2021.

[190] A. Thantharate, R. Paropkari, V. Walunj, and C. Beard, "DeepSlice: A Deep Learning Approach towards an Efficient and Reliable Network Slicing in 5G Networks," in *Proc. of IEEE UEMCON*, New York, NY, USA, 2019.

[191] M. K. Singh, S. Vittal, and A. Antony Franklin, "SERENS: Self Regulating Network Slicing in 5G for Efficient Resource Utilization," in *Proc. of IEEE 5GWF*, Bangalore, India, 2020.

[192] S. Vittal, M. K. Singh, and A. Antony Franklin, "Adaptive Network Slicing with Multi-Site Deployment in 5G Core Networks," in *Proc. of IEEE NetSoft*, Ghent, Belgium, 2020.

[193] D. Ferreira, A. Reis, C. Senna, and S. Sargento, "A Forecasting Approach to Improve Control and Management for 5G Networks," *IEEE Transactions on Network and Service Management*, pp. 1–12, 2021.

[194] S. Dutta, T. Taleb, and A. Ksentini, "QoE-aware Elasticity Support in Cloud-native 5G Systems," in *Proc. of IEEE ICC*, Kuala Lumpur, Malaysia, 2016.

[195] I. Alawe, Y. Hadjadj-Aoul, A. Ksentini, P. Bertin, and D. Darche, "On the Scalability of 5G Core Network: The AMF Case," in *Proc. of IEEE CCNC*, Las Vegas, NV, USA, 2018.

[196] G. A. Carella, M. Pauls, L. Grebe, and T. Magedanz, "An Extensible Autoscaling Engine (AE) for Software-based Network Functions," in *Proc. of IEEE NFV-SDN*, Palo Alto, CA, USA, 2016.

[197] A. B. Alvi, T. Masood, and U. Mehboob, "Load Based Automatic Scaling in Virtual IP based Multimedia Subsystem," in *Proc. of IEEE CCNC*, Las Vegas, NV, USA, 2017.

[198] C. H. T. Arteaga, F. Risso, and O. M. C. Rendon, "An Adaptive Scaling Mechanism for Managing Performance Variations in Network Functions Virtualization: A Case Study in an NFV-based EPC," in *Proc. of CNSM*, Tokyo, Japan, 2017.

[199] I. Alawe, Y. Hadjadj-Aoul, A. Ksentinit, P. Bertin, C. Viho, and D. Darche, "An Efficient and Lightweight Load Forecasting for Proactive Scaling in 5G Mobile Networks," in *Proc. of IEEE CSCN*, Paris, France, 2018.

[200] I. Alawe, A. Ksentini, Y. Hadjadj-Aoul, and P. Bertin, "Improving Traffic Forecasting for 5G Core Network Scalability: A Machine Learning Approach," *IEEE Network*, vol. 32, no. 6, pp. 42–49, 2018.

[201] I. Alawe, Y. Hadjadj-Aoul, A. Ksentini, P. Bertin, C. Viho, and D. Darche, "Smart Scaling of the 5G Core Network: An RNN-Based Approach," in *Proc. of IEEE GLOBECOM*, Abu Dhabi, United Arab Emirates, 2018.

[202] P. Chakraborty, M. Corici, and T. Magedanz, "A Comparative Study for Time Series Forecasting within Software 5G Networks," in *Proc. of IEEE ICSPCS*, Adelaide, SA, Australia, 2020.

[203] A. Sheoran, S. Fahmy, L. Cao, and P. Sharma, "AI-Driven Provisioning in the 5G Core," *IEEE Internet Computing*, pp. 1–1, 2021.

[204] P. Naik, C. Govindarajan, S. Goel, K. Govindarajan, D. Behl, A. Singh, M. Thomas, U. Mangla, and P. Jayachandran, "Closed-Loop Automation for 5G Slice Assurance," in *Proc. of ACM COMSNETS*, Bengaluru, India, 2022.

[205] O. Arouk and N. Nikaein, "Kube5G: A Cloud-Native 5G Service Platform," in *Proc. of IEEE GLOBECOM*, 2020, pp. 1–6.

[206] J. Lu, L. Xiao, Z. Tian, M. Zhao, and W. Wang, "5G Enhanced Service-based Core Design," in *Proc. of IEEE WOCC*, Beijing, China, 2019.

[207] G. Garcia-Aviles, C. Donato, M. Gramaglia, P. Serrano, and A. Banchs, "ACHO: A Framework for Flexible Re-Orchestration of Virtual Network Functions," *Computer Networks*, vol. 180, p. 107382, 2020.

[208] S. Platt, L. Sanabria-Russo, and M. Oliver, "CoNTe: A Core Network Temporal Blockchain for 5G," *Sensors*, vol. 20, no. 18, 2020.

[209] W. d. S. Coelho, A. Benhamiche, N. Perrot, and S. Secci, "Network Function Mapping: From 3G Entities to 5G Service-Based Functions Decomposition," *IEEE Communications Standards Magazine*, vol. 4, no. 3, pp. 46–52, 2020.

[210] A. Papageorgiou, A. Fernández-Fernández, S. Siddiqui, and G. Carrozzo, "On 5G network slice modelling: Service-, resource-, or deployment-driven?" *Computer Communications*, vol. 149, pp. 232–240, 2020.

[211] A. Fernández-Fernández, C. Colman-Meixner, L. Ochoa-Aday, A. Betzler, H. Khalili, M. S. Siddiqui, G. Carrozzo, S. Figuerola, R. Nejabati, and D. Simeonidou, "Validating a 5G-Enabled Neutral Host Framework in City-Wide Deployments," *Sensors*, vol. 21, no. 23, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/23/8103

[212] T. Cogalan, D. Camps-Mur, J. Gutiérrez, S. Videv, V. Sark, J. Prados-Garzon, J. Ordonez-Lucena, H. Khalili, F. Cañellas, A. Fernández-Fernández, M. Goodarzi, A. Yesilkaya, R. Bian, S. Raju, M. Ghoraishi, H. Haas, O. Adamuz-Hinojosa, A. Garcia, C. Colman-Meixner, A. Mourad, and E. Aumayr, "5G-CLARITY: 5G-Advanced Private Networks Integrating 5GNR, WiFi, and LiFi," *IEEE Communications Magazine*, vol. 60, no. 2, pp. 73–79, 2022.

[213] M. Corici, M. Emmelmann, M. Hauswirth, and T. Magedanz, "Paving the Way for Local and Industrial 5G Networks and testbeds," *ERCIM NEWS*, p. 7, 2019.

[214] S. Wang, M. Chen, X. Liu, C. Yin, S. Cui, and H. V. Poor, "A Machine Learning Approach for Task and Resource Allocation in Mobile Edge Computing Based Networks," *IEEE Internet of Things Journal*, 2020.

[215] Q. Li, H. Yao, T. Mai, C. Jiang, and Y. Zhang, "Reinforcement-Learning-and Belief-Learning-Based Double Auction Mechanism for Edge Computing Resource Allocation," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 5976–5985, 2019.

[216] L. Zhang, Z.-Y. Zhang, L. Min, C. Tang, H.-Y. Zhang, Y.-H. Wang, and P. Cai, "Task Offloading and Trajectory Control for UAV-Assisted Mobile Edge Computing Using Deep Reinforcement Learning," *IEEE Access*, vol. 9, pp. 53 708–53 719, 2021.

[217] T. Alfakih, M. M. Hassan, A. Gumaei, C. Savaglio, and G. Fortino, "Task Offloading and Resource Allocation for Mobile Edge Computing by Deep Reinforcement Learning Based on SARSA," *IEEE Access*, vol. 8, pp. 54 074–54 084, 2020.

[218] J. Wang, L. Zhao, J. Liu, and N. Kato, "Smart Resource Allocation for Mobile Edge Computing: A Deep Reinforcement Learning Approach," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–1, 2019.

[219] Y. He, F. R. Yu, N. Zhao, V. C. Leung, and H. Yin, "Software-Defined Networks with Mobile Edge Computing and Caching for Smart Cities: A Big Data Deep Reinforcement Learning Approach," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 31–37, 2017.

[220] J. Li, H. Gao, T. Lv, and Y. Lu, "Deep Reinforcement Learning Based Computation Offloading and Resource Allocation for MEC," in *Proc. of IEEE WCNC*, Barcelona, Spain, 2018.

[221] L. Huang, X. Feng, C. Zhang, L. Qian, and Y. Wu, "Deep Reinforcement Learning-based Joint Task Offloading and Bandwidth Allocation for Multi-user Mobile Edge Computing," *Digital Communications and Networks*, vol. 5, no. 1, pp. 10–17, 2019.

[222] D. Callegaro, M. Levorato, and F. Restuccia, "SeReMAS: Self-Resilient Mobile Autonomous Systems Through Predictive Edge Computing," in *Proc. of IEEE SECON*, 2021.

[223] I. Khan, X. Tao, G. S. Rahman, W. U. Rehman, and T. Salam, "Advanced Energy-efficient Computation Offloading Using Deep Reinforcement Learning in MTC Edge Computing," *IEEE Access*, vol. 8, pp. 82 867–82 875, 2020.

[224] J. Feng, F. R. Yu, Q. Pei, X. Chu, J. Du, and L. Zhu, "Cooperative Computation Offloading and Resource Allocation for Blockchain-Enabled Mobile-Edge Computing: A Deep Reinforcement Learning Approach," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6214–6228, 2019.

[225] D. A. Sivasakthi and R. Gunasekaran, "QoE-aware Mobile Computation Offloading in Mobile Edge Computing," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 11, p. e6853, 2022.

[226] Y.-J. Ku, S. Sapra, S. Baidya, and S. Dey, "State of Energy Prediction in Renewable Energy-driven Mobile Edge Computing using CNN-LSTM Networks," in *Proc. of IEEE IGESSC*, Long Beach, CA, USA, 2020.

[227] S. Wang, M. Chen, W. Saad, and C. Yin, "Federated Learning for Energy-Efficient Task Computing in Wireless Networks," in *Proc. of IEEE ICC*, Dublin, Ireland, 2020.

[228] J. Wang, J. Hu, G. Min, W. Zhan, A. Zomaya, and N. Georgalas, "Dependent Task Offloading for Edge Computing based on Deep Reinforcement Learning," *IEEE Transactions on Computers*, pp. 1–1, 2021.

[229] H. Zhang, Y. Yang, X. Huang, C. Fang, and P. Zhang, "Ultra-Low Latency Multi-Task Offloading in Mobile Edge Computing," *IEEE Access*, vol. 9, pp. 32 569–32 581, 2021.

[230] Z. Zaman, S. Rahman, and M. Naznin, "Novel Approaches for VNF Requirement Prediction Using DNN and LSTM," in *Proc. of IEEE GLOBECOM*, Waikoloa, HI, USA, 2019.

[231] T. Subramanya, D. Harutyunyan, and R. Riggio, "Machine Learning-driven Service Function Chain Placement and Scaling in MEC-enabled 5G Networks," *Computer Networks*, vol. 166, p. 106980, 2020.

[232] T. Subramanya and R. Riggio, "Centralized and Federated Learning for Predictive VNF Autoscaling in Multi-Domain 5G Networks and Beyond," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 63–78, 2021.

[233] J. Tao, Z. Lu, Y. Chen, J. Wu, P. Yu, and C. Lei, "Adaptive VNF Scaling Approach with Proactive Traffic Prediction in NFV-enabled Clouds," in *ACM Turing Award Celebration Conference-China (ACM TURC 2021)*, 2021, pp. 166–172.

[234] T. Hou, G. Feng, S. Qin, and W. Jiang, "Proactive Content Caching by Exploiting Transfer Learning for Mobile Edge Computing," *International Journal of Communication Systems*, vol. 31, no. 11, p. e3706, 2018.

[235] W. Jiang, G. Feng, S. Qin, and Y. Liu, "Multi-Agent Reinforcement Learning Based Cooperative Content Caching for Mobile Edge Networks," *IEEE Access*, vol. 7, pp. 61 856–61 867, 2019.

[236] Z. Yang, Y. Liu, Y. Chen, and G. Tyson, "Deep reinforcement learning in cache-aided MEC networks," in *Proc. of IEEE ICC*, Shanghai, China, 2019.

[237] R. Behravesh, D. F. Perez-Ramirez, A. Rao, D. Harutyunyan, R. Riggio, and R. Steinert, "ML-Driven DASH Content Pre-Fetching in MEC-Enabled Mobile Networks," in *Proc. of IEEE CNSM*, Izmir, Turkey, 2020.

[238] Y. M. Saputra, D. T. Hoang, D. N. Nguyen, E. Dutkiewicz, D. Niyato, and D. I. Kim, "Distributed Deep Learning at the Edge: A Novel Proactive and Cooperative Caching Framework for Mobile Edge Networks," *IEEE Wireless Communications Letters*, vol. 8, no. 4, pp. 1220–1223, 2019.

[239] Y. Qin, D. Wu, Z. Xu, J. Tian, and Y. Zhang, "Adaptive in-network Collaborative Caching for Enhanced Ensemble Deep Learning at Edge," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[240] W. Jiang, D. Feng, Y. Sun, G. Feng, Z. Wang, and X.-G. Xia, "Proactive Content Caching Based on Actor-Critic Reinforcement Learning for Mobile Edge Networks," *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, 2021.

[241] J. Wang, J. Hu, G. Min, W. Zhan, Q. Ni, and N. Georgalas, "Computation Offloading in Multi-access Edge Computing Using a Deep Sequential Model Based on Reinforcement Learning," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 64–69, 2019.

[242] T. K. Rodrigues, K. Suto, H. Nishiyama, J. Liu, and N. Kato, "Machine Learning Meets Computation and Communication Control in Evolving Edge and Cloud: Challenges and Future Perspective," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 38–67, 2019.

[243] S. Rathore, P. K. Sharma, A. K. Sangaiah, and J. J. Park, "A Hesitant Fuzzy Based Security Approach for Fog and Mobile-Edge Computing," *IEEE Access*, vol. 6, pp. 688–701, 2017.

[244] P. Dong, Z. Ning, R. Ma, X. Wang, X. Hu, and B. Hu, "NOMA-based Energy-efficient Task Scheduling in Vehicular Edge Computing Networks: A Self-imitation Learning-based Approach," *China Communications*, vol. 17, no. 11, pp. 1–11, 2020.

[245] ETSI, "Mobile Edge Computing (MEC); Technical Requirements," European Telecommunications Standards Institute, Group Specification (GS) MEC 002, Mar. 2016, version 1.1.1.

[246] M. Patel, B. Naughton, C. Chan, N. Sprecher, S. Abeta, and A. Neal, "Smart Cells Revolutionize Service Delivery," Intel Corporation, Tech. Rep., 2013, Accessed on 23.09.2022. [Online]. Available: http://www.intel.co.uk/content/dam/www/public/us/en/documents/white-papers/smart-cells-revolutionize-service-delivery.pdf

[247] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. Leung, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 131–139, 2014.

[248] Q. Li, C. Lu, B. Cao, and Q. Zhang, "Caching Resource Management of Mobile Edge Network Based on Stackelberg Game," *Digital Communications and Networks*, vol. 5, no. 1, pp. 18–23, 2019.

[249] D. Bega, M. Gramaglia, A. Garcia-Saavedra, M. Fiore, A. Banchs, and X. Costa-Perez, "Network Slicing Meets Artificial Intelligence: An AI-Based Framework for Slice Management," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 32–38, 2020.

[250] W. Jiang, S. D. Anton, and H. Dieter Schotten, "Intelligence Slicing: A Unified Framework to Integrate Artificial Intelligence into 5G Networks," in *Proc. of IEEE WMNC*, Paris, France, 2019.

[251] A. Thantharate, R. Paropkari, V. Walunj, C. Beard, and P. Kankariya, "Secure5G: A Deep Learning Framework Towards a Secure Network Slicing in 5G and Beyond," in *Proc. of IEEE CCWC*, Las Vegas, NV, USA, 2020.

[252] V. Theodorou, A. Lekidis, T. Bozios, K. Meth, A. Fernández-Fernández, J. Tavlor, P. Diogo, P. Martins, and R. Behravesh, "Blockchain-based Zero Touch Service Assurance in Cross-domain Network Slicing," in *Proc. of IEEE EuCNC/6G Summit*, Porto, Portugal, 2021.

[253] K. Abbas, M. Afaq, T. A. Khan, A. Mehmood, and W.-C. Song, "IB-NSlicing: Intent-Based Network Slicing Framework for 5G Networks using Deep Learning," in *Proc. of IEEE APNOMS*, Daegu, Korea (South), 2020.

[254] H. Chergui, L. Blanco, L. A. Garrido, K. Ramantas, S. Kukliński, A. Ksentini, and C. Verikoukis, "Zero-touch ai-driven distributed management for energy-efficient 6g massive network slicing," *IEEE Network*, vol. 35, no. 6, pp. 43–49, 2021.

[255] M. Camelo, L. Cominardi, M. Gramaglia, M. Fiore, A. Garcia-Saavedra, L. Fuentes, D. De Vleeschauwer, P. Soto-Arenas, N. Slamnik-Krijestorac, J. Ballesteros *et al.*, "Requirements and specifications for the orchestration of network intelligence in 6g," in *2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2022, pp. 1–9.

[256] A. Papageorgiou, A. Fernández-Fernández, L. Ochoa-Aday, M. S. Peláez, and M. S. Siddiqui, "SLA Management Procedures in 5G Slicing-based Systems," in *Proc of IEEE EuCNC*, Portugal, Porto, 2020.

[257] A. Fernández-Fernández, M. De Angelis, P. G. Giardina, J. Taylor, P. Chainho, J. M. J. Valero, L. Ochoa-Aday, D. R. López, G. Carrozzo, and M. S. Siddiqui, "Multi-Party Collaboration in 5G Networks via DLT-Enabled Marketplaces: A Pragmatic Approach," in *Proc. of IEEE EuCNC/6G Summit*, Porto, Portugal, 2021.

[258] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 28–41, 2019.

[259] R. Sattiraju, A. Weinand, and H. D. Schotten, "AI-assisted PHY technologies for 6G and beyond wireless networks," *arXiv preprint arXiv:1908.09523*, 2019.

[260] T. Hong, C. Liu, and M. Kadoch, "Machine Learning Based Antenna Design for Physical Layer Security in Ambient Backscatter Communications," *Wireless Communications and Mobile Computing*, vol. 2019, pp. 4 870 656:1–4 870 656:10, 2019.

[261] L. Lovén, T. Leppänen, E. Peltonen, J. Partala, E. Harjula, P. Porambage, M. Ylianttila, and J. Riekki, "EdgeAI: A Vision for Distributed, Edge-native Artificial Intelligence in Future 6G Networks," in *Proc. of 6G Wireless Summit*, Levi, Finland, 2019.

[262] I. Tomkos, D. Klonidis, E. Pikasis, and S. Theodoridis, "Toward the 6G network era: Opportunities and challenges," *IT Professional*, vol. 22, no. 1, pp. 34–38, 2020.

[263] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, "Differentially Private Data Publishing and Analysis: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1619–1638, 2017.

[264] R. Bhatia, S. Benno, J. Esteban, T. V. Lakshman, and J. Grogan, "Unsupervised Machine Learning for Network-Centric Anomaly Detection in IoT," in *Proc. of ACM Big-DAMA*, Orlando, FL, USA, 2019.

[265] C. M. Mathas, O. E. Segou, G. Xylouris, D. Christinakis, M.-A. Kourtis, C. Vassilakis, and A. Kourtis, "Evaluation of Apache Spot's Machine Learning Capabilities in an SDN/NFV Enabled Environment," in *Proc. of ACM ARES*, Hamburg, Germany, 2018.

[266] B. Hussain, Q. Du, S. Zhang, A. Imran, and M. A. Imran, "Mobile Edge Computing-Based Data-Driven Deep Learning Framework for Anomaly Detection," *IEEE Access*, vol. 7, pp. 137 656–137 667, 2019.

[267] B. Chafika and T. Taleb, "AI for Beyond 5G Networks: A Cyber-Security Defense or Offense Enabler?" *IEEE Network*, vol. 34, no. 6, pp. 140–147, 2020.

[268] M. Usama, R. Mitra, I. Ilahi, J. Qadir, and M. Marina, "Examining Machine Learning for 5G and Beyond through an Adversarial Lens," *IEEE Internet Computing*, vol. 25, no. 2, pp. 26–34, 2021.

[269] J. Qiu, L. Du, Y. Chen, Z. Tian, X. Du, and M. Guizani, "Artificial Intelligence Security in 5G Networks: Adversarial Examples for Estimating a Travel Time Task," *IEEE Vehicular Technology Magazine*, vol. 15, no. 3, pp. 95–100, 2020.

[270] S. Rathore, Y. Pan, and J. H. Park, "BlockDeepNet: a Blockchain-based secure deep learning for IoT network," *Sustainability*, vol. 11, no. 14, p. 3974, 2019.

[271] A. El Azzaoui, S. K. Singh, Y. Pan, and J. H. Park, "Block5GIntell: Blockchain for AI-Enabled 5G Networks," *IEEE Access*, vol. 8, pp. 145 918–145 935, 2020.

[272] ETSI, "Zero-touch network and Service Management (ZSM); End-to-end management and orchestration of network slicing," European Telecommunications Standards Institute, Group Specification (GS) ZSM 003, Jun. 2021, version 1.1.1.

[273] 3GPP. SA3 - Security and Privacy. Accessed on 23.09.2022. [Online]. Available: https://www.3gpp.org/Specifications-groups/sa-plenary/54-sa3-security/

[274] R. Khan, P. Kumar, D. N. K. Jayakody, and M. Liyanage, "A Survey on Security and Privacy of 5G Technologies: Potential Solutions, Recent Advancements, and Future Directions," *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 196–248, 2020.

[275] ETSI, "Zero-touch network and Service Management (ZSM); General Security Aspects Security study ," European Telecommunications Standards Institute, Group Report (GR) ZSM 010, Jul. 2021, version 1.1.1.

[276] R. Boutaba, M. Salahuddin, N. Limam, S. Ayoubi, N. Shahriar, F. Estrada-Solano, and O. C. Rendon, "A Comprehensive Survey on Machine Learning for Networking: Evolution, Applications and Research Opportunities," *Journal of Internet Services and Applications*, vol. 9, pp. 16–99, 2018.

[277] Y. Lavinia, R. Durairajan, R. Rejaie, and W. Willinger, "Challenges in Using ML for Networking Research: How to Label If You Must," in *Proc. of ACM NetAI*, Virtual Event, USA, 2020.

[278] A. Muthukumar and R. Durairajan, "Denoising Internet Delay Measurements using Weak Supervision," in *Proc. of IEEE ICMLA*, Boca Raton, FL, USA, 2019.

[279] European Commission, "Assessment List for Trustworthy AI (ALTAI)," European Commission, Tech. Rep., Jul. 2019, Accessed on 23.09.2022. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

[280] ——, "On Artificial Intelligence - A European approach to excellence and trust," European Commission, White Paper, Feb. 2020, Accessed on 23.09.2022. [Online]. Available: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

[281] 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, 2020.

[282] Y. Chung, T. Kraska, N. Polyzotis, K. H. Tae, and S. E. Whang, "Automated Data Slicing for Model Validation: A Big Data - AI Integration Approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 12, pp. 2284–2296, 2020.

[283] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics," *Electronics*, vol. 10, no. 5, 2021.

[284] G. Carrozzo, M. S. Siddiqui, A. Betzler, J. Bonnet, G. M. Perez, A. Ramos, and T. Subramanya, "AI-driven Zero-touch Operations, Security and Trust in Multi-operator 5G Networks: a Conceptual Architecture," in *Proc. of IEEE EuCNC*, Dubrovnik, Croatia, 2020.

Estefanía Coronado is a Juan de la Cierva Senior Researcher at the University of Castilla-La Mancha (Spain) and a Senior Researcher at Fundació i2CAT (Spain). From 2018 to 2020 she was an Expert Researcher at FBK (Italy). She got her Ph.D. on multimedia content delivery over SD-WLANs using ML techniques from the University of Castilla-La Mancha (Spain), where she got also her M.Sc. degrees in Computer Engineering and Advanced Computer Technologies in 2014 and 2015, respectively. She published around 35 papers in international journals and conferences, and received the IEEE INFOCOM Best Demo Award in 2019. Her research interests include AI-driven network management automation, wireless/mobile networks, network slicing, and SDN/NFV.

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2022.3212586

Rasoul Behravesh is a Researcher in the Smart Networks and Services (SENSE) unit at Fondazione Bruno Kessler (FBK) in Trento, Italy. Currently, his main research focus is on zero-touch network and service management and orchestration in 5G networks. He holds a Ph.D. in Telecommunications from the University of Bologna, Italy. Prior to that, he received his M.Sc. in Computer Networks from QIAU and his B.Sc. degree in Information Technology from the Payam-e-Noor University of Mahabad. Rasoul has published several papers on network management and automation in mobile networks in internationally refereed journals and conferences. His main research interests include mobile networks, AI-driven network management and orchestration, and network softwarization.

Muhammad Shuaib Siddiqui is a senior researcher at i2CAT Foundation where he is also the Area Manager for Software Networks research lab. His current research topics include network automation, SDN/NFV based control, management, & orchestration platforms for 5G, network slicing, and NFV/SDN security. Currently, he is also coordinating the H2020 5GZORRO (5G PPP Phase 3 project). He holds a Ph.D. in Computer Science from Technical University of Catalonia (UPC) (Spain), M.Sc. in Communication Systems (2007) from École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, and B.Sc. in Computer Engineering (2004) from King Fahd University of Petroleum & Minerals (KFUPM), Saudi Arabia. He has given several talks at Mobile World Congress, Smart City Expo World Congress, SDN/NFV World Congress and other events. One of his publications, based on his PhD thesis, received the IEEE Internet Technical Committee (ITC) paper of the year award for 2015.

Tejas Subramanya is a Senior Research Engineer at Nokia Standards in Munich, Germany. His current research interests include cognitive management of future mobile networks and applying machine learning to network automation use cases. He has over 8 years of experience in mobile networks related research and development in different roles in India, Finland, Italy, and Germany. He is a (co-)author of several articles and papers on network management automation for next-generation mobile networks. Tejas received his MS degree in Radio Communications from the Aalto University in Finland and his PhD on cognitive network management and orchestration from the University of Trento in Italy.

Xavier Costa-Pérez is ICREA Research Professor, Scientific Director at the i2Cat Research Center and Head of 5G Networks R&D at NEC Laboratories Europe. His team generates research results which are regularly published at top scientific venues, produces innovations which have received several awards for successful technology transfers, participates in major European Commission R&D collaborative projects and contributes to standardization bodies such as 3GPP, ETSI NFV, ETSI MEC and IETF. He has served on the Organizing Committees of several conferences, published papers of high impact and holds tenths of granted patents. Xavier received his Ph.D. degree in Telecommunications from the Polytechnic University of Catalonia (UPC) in Barcelona and was the recipient of a national award for his Ph.D. thesis.

Adriana Fernández-Fernández is a Senior Researcher at i2CAT Foundation (Spain), mainly focused on NFV, management & orchestration, and network slicing research and innovation. She graduated in Telecommunication and Electronic Engineering; and since 2018, she holds a Ph.D. in Network Engineering from the Universitat Politècnica de Catalunya (UPC). Before joining i2CAT, she worked as a postdoctoral researcher within the BAMPLA Research Group at UPC. She is an author/co-author of over 20 papers in peer reviewed international journals and conference proceedings. She has been awarded several scholarships from the Spanish Government through the Research Training Program (FPI). Her research interests also include communication network modeling and optimization, energy aware routing, SDN and traffic engineering.

Roberto Riggio is an Associate Professor at the Polytechnic University of Marche in Ancona, Italy. He received his PhD from the University of Trento (Italy), after that he was postdoc at University of Florida, Researcher/Chief Scientist at CREATE-NET in Trento (Italy), Head of Unit at FBK in Trento (Italy), Senior 5G Researcher at the i2CAT Foundation in Barcelona (Spain), and Senior Researcher at RISE AB in Stockholm (Sweden). His research interests revolve around optimization and algorithmic problems in networked and distributed systems. His current fields of applications are edge automation platforms, intelligent networks, and serverless computing. Roberto Riggio has published more than 130 papers in internationally refereed journals and conferences. He is a Senior Member of the IEEE.