IEEE ComSoc IEEE Open Journal of the Communications Society

Received 30 May 2024; revised 2 July 2024; accepted 16 July 2024. Date of publication 19 July 2024; date of current version 6 August 2024.

*Digital Object Identifier 10.1109/OJCOMS.2024.3431286*

# Unlocking O-RAN Potential: How Management Data Analytics Enhances SMO Capabilities?

**MOHAMMAD ASIF HABIBI [1], GIRMA MAMUYE YILMA [2,3],**
**UMBERTO FATTORE [2,4] (Graduate Student Member, IEEE),**
**XAVIER COSTA-PÉREZ [2,5,6] (Senior Member, IEEE), AND HANS D. SCHOTTEN [1,7] (Member, IEEE)**

[1] Division of Wireless Communications and Radio Navigation (WiCoN), Department of Electrical and Computer Engineering (EIT), University of Kaiserslautern (RPTU), 67663 Kaiserslautern, Germany

[2] 6G Networks Research and Development Department, NEC Laboratories Europe, 69115 Heidelberg, Germany

[3] Network Engineering Department, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain

[4] Telematic Engineering Department, Universidad Carlos III de Madrid, 28911 Leganés, Spain

[5] Department of AI Driven Systems, i2CAT Research Center, 08034 Barcelona, Spain

[6] Department of Engineering Sciences, Catalan Institution for Research and Advanced Studies (ICREA), 08034 Barcelona, Spain

[7] Intelligent Networking Research Group, German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany

CORRESPONDING AUTHOR: M. A. HABIBI (e-mail: asif@eit.uni-kl.de)

This work was supported in part by the German Federal Ministry for Education and Research (BMBF) through the Open6GHub Project under Grant 16KISK003K.

**ABSTRACT** The service management and orchestration (SMO) framework plays a major role in the management and orchestration (M&O) of applications, services, and components within the open radio access network (O-RAN) architecture, as defined by the O-RAN Alliance. It comprises the non-real-time RAN intelligence controller (Non-RT RIC) and possesses the capability to incorporate management systems and components from multiple standards development organizations (SDOs), notably the Third Generation Partnership Project (3GPP) and the European Telecommunications Standards Institute (ETSI). This multi-SDO-based SMO framework aims to provide a rich set of M&O services in a coherent and unified manner. Leveraging management data analytics (MDA) to deliver M&O services can enhance SMO capabilities. This article proposes an intelligence-driven approach by integrating MDA into the 3GPP and ETSI management systems within the context of SMO. Furthermore, it introduces the artificial intelligence (AI)/machine learning (ML) Function into the Non-RT RIC, consolidating existing intelligent components and introducing novel ones to enhance the intelligence capabilities of the Non-RT RIC. The article further proposes an architectural solution that unifies and facilitates interoperability among the intelligent systems of the 3GPP, ETSI, and Non-RT RIC – collectively referred to as "three modules" – within SMO. It also presents the end-to-end lifecycle workflow of the AI/ML model across the three modules. Finally, the article outlines key research challenges related to integrating MDA within the SMO framework.

**INDEX TERMS** 3GPP, artificial intelligence, ETSI, management & orchestration, management data analytics, machine learning, Non-RT RIC, O-RAN architecture, O-RAN Alliance, SMO framework.

## I. INTRODUCTORY REMARKS

The OPEN radio access network (O-RAN) architecture represents a transformative approach to designing, deploying, managing, and maintaining the wireless domain of a cellular network [1]. It aims to make the RAN domain more open, intelligent, virtualized, cloud-native, programmable, and interoperable [1], [2]. O-RAN moves away from traditional, often proprietary, and closed RANs [3], typically provided by a single equipment manufacturer with limited flexibility and interoperability [1]. It introduces standardized interfaces and encourages a more competitive, innovative, and diverse supplier ecosystem for the RAN domain.

The open radio access network (O-RAN) architecture consists of several components that work together to create a flexible, cloud-based RAN [1], [4]. The service management and orchestration (SMO) framework is responsible for deploying and operating RAN services, coordinating the RAN components [5]. O-RAN functions include the near-real-time RAN intelligence controller (Near-RT RIC), open centralized unit (O-CU)-control plane, O-CU-user plane, and the open distributed unit (O-DU), which can be hosted on the O-Cloud or customized hardware [1]. The O-Cloud provides the cloud infrastructure that hosts these functions and includes the O-Cloud Notification interface for receiving relevant notifications.

The SMO framework is responsible for integrating the non-real-time RAN intelligence controller (Non-RT RIC) and the management systems of other standards development organizations (SDOs) [1]. These management systems primarily encompass those specified by the Third Generation Partnership Project (3GPP) and the European Telecommunications Standards Institute (ETSI). We refer to the management systems of these two SDOs as the 3GPP-network slicing management system (3GPP-NSMS) and the NFV-management and orchestration (NFV-MANO), respectively. To ensure that the SMO framework can deliver unified management and orchestration (M&O) services for O-RAN components, seamless interoperability among the three modules is crucial. In [5], we tackled the challenges of unification and interoperability among the three modules by introducing a unified and standard-compliant SMO framework. However, our proposal, while acknowledging the necessity, did not address the integration of artificial intelligence (AI)/machine learning (ML)[1] models into SMO.

The integration of AI/ML has revolutionized network operations management and service optimization, offering unparalleled capabilities in analyzing vast amounts of data, predicting network behavior, and automating decision-making processes [6], [7], [8]. AI/ML algorithms have the potential to enhance the operations and network planning of mobile service providers [9]. These technologies enable swift and effective network optimizations, thereby enhancing efficiency. Moreover, they open avenues for generating additional revenue by integrating big data and networking capabilities, enabling the delivery of tailored customer experiences while maintaining stringent assurance standards [6]. In particular, within the context of O-RAN architecture, the incorporation of AI/ML algorithms holds immense potential for enhancing the efficiency, reliability, and intelligence of network M&O functions [8], [10], [11]. By harnessing the power of data analytics, predictive modeling, and automated decision-making, the integration of AI/ML algorithms is aimed at elevating the capabilities of SMO to new heights, enabling it to proactively identify and mitigate network and service issues, optimize resource allocation, and deliver superior quality of service to end-users [10], [12]. Hence, it is crucial to explore the potential advantages of integrating AI/ML models at an early stage, as their incorporation could enhance the performance of the SMO framework [13], leading to more advanced and intelligent M&O capabilities [6].

## A. LITERATURE REVIEW

Recently, the O-RAN Alliance has studied the integration of AI/ML into the Non-RT RIC. In [14], the Alliance explores various use cases, outlining requirements and exploring multiple ML-assisted solutions to enhance the capabilities of the Non-RT RIC. Despite its valuable contributions, this study exhibits two limitations: First, it neglects to define the interactions between the intelligent system of the Non-RT RIC and other intelligent systems residing within SMO. Second, this study is based on an AI/ML architecture for the Non-RT RIC, which, in our opinion, demands substantial revision to enable the effective and unified delivery of ML-assisted solutions throughout the O-RAN architecture. The ETSI Industry Specification Group (ISG) Network Function Virtualization (NFV) has also made significant contributions to the intelligentization and automation of the latest NFV-MANO framework, as evidenced in [15]. In addition, the 3GPP Technical Specification Group (TSG) Service and System Aspects Working Group 5 (SA5) has studied the integration of management data analytics (MDA) into the 3GPP-NSMS in [16]. These two documents, [15], [16], establish a comprehensive foundation for understanding the role of AI/ML technologies in the M&O of services, networks, and their associated management systems, providing valuable insights into their potential benefits and challenges. However, to this end, neither the 3GPP nor the ETSI has yet explored the integration of their corresponding intelligent systems into O-RAN. The intelligent systems of these two SDOs may require significant architectural restructuring to seamlessly integrate into SMO. Moreover, NFV-MANO and 3GPP-NSMS may present incompatible deployment requirements, and their integration may not be compatible with the specifications defined by the O-RAN Alliance, demanding a harmonization of performance, functional, and operational requirements across all three modules.

In addition to the contributions made by the three SDOs, academia and industries have also dedicated substantial efforts to researching the integration of AI/ML capabilities into O-RAN. For example, in [17], the authors employed an intelligent master-actor-based architecture for testing AI/ML models within O-RAN, considering aspects such

[1] The terms "MDA," "intelligence," and "AI/ML models or algorithms" are used interchangeably to refer to the implementation of advanced computational techniques aimed at extracting insights from various management systems, facilitating informed decision-making, and optimizing M&O processes within O-RAN in an intelligent and autonomous manner. Consistently throughout the article, these terms are utilized synonymously to underscore the adoption of data-driven approaches, aligned with the latest frameworks of prominent SDOs (e.g., 3GPP, ETSI, and O-RAN Alliance), to enhance managerial intelligence within SMO.

as performance, decision-making, security, and vulnerabilities. In [12], an AI/ML-based architectural solution is proposed to enable hierarchical and distributed optimization and automation for both Near-RT RIC and Non-RT RIC. Although the paper overviews the intelligent systems of other SDOs (including those of ETSI and 3GPP), their integration into and utilization within the O-RAN architecture is not studied. Another notable contribution in [13] explores the ML models' lifecycle workflow within the Non-RT RIC, illuminating the processes and considerations involved. Reference [18] introduces a network telemetry architecture aimed at supporting end-to-end (E2E) analytics in O-RAN. It showcases an AI/ML workflow for two ML algorithms using the xApps. In [19], the authors investigated the integration of deep learning into O-RAN and outlined the key steps to embed a deep learning algorithm within O-RAN. Finally, in [20], a network intelligence orchestration solution is introduced as an rApp. This solution computes an optimal set of data-driven algorithms along with their execution locations. Its aim is to prevent conflicts between these algorithms, regulate desired parameters, and meet expected time requirements.

### B. MOTIVATION AND RESEARCH CHALLENGES

Despite the significant efforts made by the corresponding SDOs, academia, and industry to enhance the intelligentization and automation of the O-RAN architecture, **a major challenge remains in realizing a unified, interoperable, and multi-SDO-based intelligent system within SMO**. While advancements in the intelligentization of individual module of each SDO have been notable, **the lack of seamless collaboration and interoperability among these intelligent systems impedes the delivery of a coherent and unified approach to the M&O of O-RAN components**. This fragmentation poses **a formidable obstacle to realizing the full potential of AI/ML-driven solutions**, hindering the SMO framework from achieving a holistic and synergistic intelligent system. Furthermore, with the integration of intelligent management systems from different SDOs into SMO, a variety of data types from diverse components are considered as inputs for ML model training. The **interpretation, normalization, and processing of this input data pose complex challenges**, necessitating the design of a tailored intelligent system to manage the ML workflow within the SMO framework. Addressing these research challenges is crucial for unlocking the transformative capabilities of AI/ML algorithms and ensuring their effective integration into the operations of the SMO framework for delivering advanced M&O services in an intelligent, autonomous, and efficient manner. The intelligent automation of M&O services within SMO directly influences the quality of service (QoS), resource allocation, energy consumption, and the overall performance of the O-RAN architecture.

### C. MAJOR CONTRIBUTIONS AND GOALS

In response to the outlined research challenges, this article presents the following major contributions:

- First, **we focus on the integration of MDA into 3GPP-NSMS**. The objective of this integration is to enhance the performance of M&O operations from applications standpoint in conjunction with an open next generation node B (O-gNB) within O-RAN.
- Second, **we integrate MDA into NFV-MANO**. This integration aims to intellegentize operations associated with the M&O of the virtualized and cloudified aspects of an O-gNB, as well as the underlying wireless infrastructure, within the O-RAN architecture.
- Third, **we introduce an architectural solution to support AI/ML Function within the Non-RT RIC**. The AI/ML Function was initially proposed in [5]. In this article, we delve into its capabilities, components, and broader impact on the intelligentization of operations associated with the Non-RT RIC.
- Fourth, **we propose an architecture that unifies the intelligent systems of the three modules and enables interoperability among them** via standard-compliant interfaces. The objective of such a unification is to design a unified, crowd-sourced, interoperable, and multi-SDO-based intelligent system for SMO.
- Fifth, **we present a comprehensive E2E lifecycle workflow for realizing ML models** across the three modules. The goal is to showcase step-by-step operations for the development, deployment, and refinement of ML models across a unified intelligent system proposed for the SMO framework.
- Sixth, **we identify a number of research challenges related to integrating intelligence and MDA within SMO.** These challenges help prioritize areas of focus, guide research efforts, and advance the state-of-the-art, ultimately contributing to improved network performance and decision-making capabilities within SMO.

### D. THE STRUCTURE OF THE ARTICLE

The rest of this article is structured as follows: Sections II and III delve into the integration of MDA into the 3GPP-NSMS and NFV-MANO, respectively. Section IV introduces the AI/ML Function for the Non-RT RIC. Following this, Section V proposes an architectural solution that facilitates interoperability among the intelligent systems of the three modules. Subsequently, in Section VI, we outline the E2E lifecycle workflow of the AI/ML models across the SMO. Additionally, Section VII presents several research challenges associated with integrating AI/ML models that must be addressed in order to achieve a fully intelligent SMO. Finally, in Section VIII, we summarize key points and identify promising directions for future research.

## II. INTEGRATING MDA INTO 3GPP-NSMS WITHIN THE CONTEXT OF SMO FRAMEWORK

In this section, we discuss the management functions (MFs) of the 3GPP-NSMS. We elaborate on how these MFs can be utilized within the SMO framework to manage the O-RAN components and orchestrate their required resources. Moreover, we explore the integration of the MDA into the MFs of the 3GPP-NSMS. The aim of this integration is to enhance the operations of the 3GPP-NSMS within SMO.

### *A. 3GPP-NSMS FOR THE M&O OF O-RAN*

The 3GPP TSG SA5 is among the various working groups within the 3GPP ecosystem, with a specific focus on the M&O aspects of public networks. This group therefore defines an architectural framework for the M&O of 3GPP networks [21], aimed at addressing a long list of desired use cases and requirements for the fifth-generation (5G) and beyond communications systems [22]. Notably, among the use cases considered in [22], managing network slices within 5G networks is highlighted. This capability enables 3GPP networks to deliver diverse services by utilizing dedicated slices of a public network, each tailored to different service requirements and featuring ad-hoc performance monitoring and configuration [23]. Such services may include, but are not limited to, vehicle-to-everything (V2X), massive Internet of Things (IoT), and enhanced mobile broadband (eMBB) [23]. The 3GPP TSG SA5 specifications also provide an example of a network slice as a service (NSaaS) provided to customers (i.e., the owners of private networks), demonstrating how different network slice instances can encompass a set of dedicated and/or shared network functions (NFs) over a common telecommunications infrastructure.

In the architectural framework defined by the 3GPP TSG SA5 [21], two MFs become relevant for the M&O of O-gNBs within O-RAN: the network function management function (NFMF) and the network slice subnet management function (NSSMF). The NFMF is responsible for the lifecycle management and fault, configuration, accounting, performance, security (FCAPS) of NFs within a network [22]. In O-RAN, the NFMF can manage each distinct NF of an O-gNB, such as the O-CU, O-DU, and open radio unit (O-RU), with dedicated NFMFs for each [24]. The key responsibilities of the NFMF with respect to the O-gNB encompass provisioning (i.e., setting up and configuring NFs to meet requirements), fault management (i.e., detecting and solving issues), and performance management (i.e., monitoring and analyzing performances) [21].

On the contrary, the NSSMF concentrates on the management and FCAPS of a network slice subnet [22], assuming responsibilities like slice subnet creation and configuration (i.e., defining and deploying slice subnets according to the customer requirements), resource allocation (i.e., allocating resources for the slice subnets to meet the requirements), monitoring (i.e., to check if specific service level agreements are respected), and adaptation (e.g., in case of changing demands over time). Within the context of O-RAN, the NSSMF is responsible for the M&O of an entire O-gNB, managing the O-RAN slice subnet and its required resources. As depicted in Fig. 1, multiple NFMFs manage distinct NFs for the same slice subnet, which can then be managed by a single NSSMF managing that slice subnet. Furthermore, the NSSMF can interact with other 3GPP MFs, such as the network slice management function (NSMF), to realize an E2E service M&O within the 3GPP TSG SA5 framework [21] and beyond (i.e., transport network M&O).

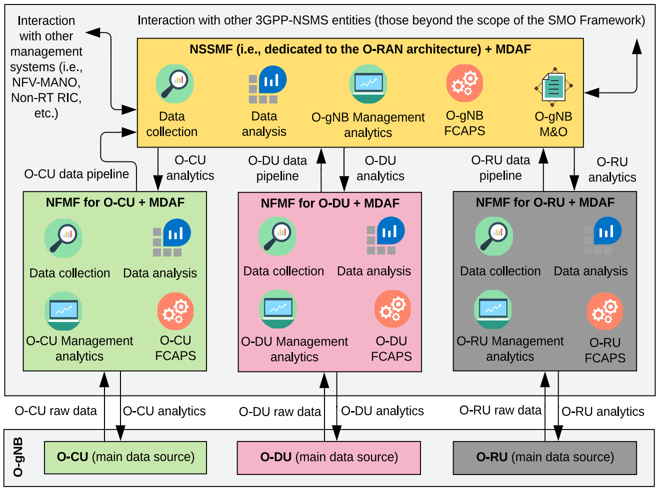

FIGURE 1. Integrating MDAS into the 3GPP-NSMS within the SMO framework.

The NSSMF, NFMF, and NSMF collectively form the 3GPP-NSMS [25], which, in turn, can be regarded as part of the SMO framework within O-RAN [5]. Hence, the NSSMF and NFMF may be anchored within SMO. The inclusion of the NSMF within the scope of SMO is variable and may or may not be considered [5]. We consider the NSMF beyond the scope of the SMO framework in this article. Moreover, to comprehensively and efficiently manage a network slice subnet, encompassing NFs and O-RAN aspects, the 3GPP-NSMS must interact (i.e., produce and consume M&O services) with NFV-MANO and with components of the O-RAN Alliance, notable the Non-RT RIC, within SMO.

### *B. INTEGRATING MDA INTO 3GPP-NSMS*

Within the research and standardization communities, there is widespread acknowledgment that the performance of the 3GPP-NSMS, whether within or beyond the SMO framework, can be significantly enhanced through the application of automation techniques, data-driven solutions, and MDA-assisted mechanisms [16], [24]. Furthermore, it is also acknowledged that employing MDA also improves the capabilities of the 3GPP-NSMS during its interaction with other systems (both management and non-management), such as the NFV-MANO and the Non-RT RIC.

The MDA, conceived as the endeavor to derive meaningful insights from unprocessed management data, originated within the management domain of a cellular network [16]. The MDA System is also standardized by the 3GPP TSG SA5 [16]. Its aim is to autonomously and intelligently address specific optimization challenges related to the M&O of networks and services. The MDA System provides management data analytics service (MDAS), comprising statistical data from the past or predictive insights into the future. The determination of this depends on whether the requested timeframe is situated in the past or the future, respectively [26]. The output reports generated by the MDAS might incorporate a confidence parameter ranging between 0 and 100 [26]. This parameter communicates the level of certainty associated with the prediction made, potentially influenced by factors such as the volume of data utilized for prediction generation, the age of the AI/ML model, and other relevant considerations.

Within SMO, the MDA System is used to gather management data from diverse management systems and components, train dedicated ML models, and generate insightful analytics. A pivotal obstacle within the MDA System lies in effectively managing the vast volume of management data stemming from the SMO framework and transforming it into valuable insights. This involves understanding data structures and relationships, extracting actionable knowledge, and facilitating well-informed decision-making [16]. The overarching goal, anticipated to materialize in SMO, is to facilitate a highly autonomous management plane for O-RAN, which is capable of self-configuration, self-monitoring, self-healing, and self-optimization with minimal intervention from human users and/or human administrators [16], [27].

To integrate analytics into the functionalities of the NSSMF and NFMF within SMO, the 3GPP-NSMS can leverage the MDA System [16]. The MDA System can collect management data from NSSMF and NFMF, analyze the collected data, and provide MDA back to the corresponding functions [28]. To achieve this capability within the 3GPP-NSMS, the MDA System can employ the management data analytics functions (MDAFs) [16], as illustrated in Figure 1. The MDAFs can be integrated into or deployed alongside the NSSMF and NFMF via standard-compliant interfaces. In either case, each NSSMF and NFMF determines which essential management data needs to be exposed to the MDA System, taking into account privacy and security concerns. In turn, the MDA System must specify the necessary output required by the NSSMF and NFMF to support the corresponding use case within the 3GPP-NSMS.

When designing the MDA System for the 3GPP-NSMS, the unique characteristics of wireless communication channels within O-RAN must be taken into consideration [24]. It specifically means that the inherent unpredictability linked to factors like traffic conditions and radio-wave propagation, as well as the dynamic nature resulting from interference conditions and user mobility, necessitates thorough attention during data collection and processing, as well as leveraging analytics derived from such data. For example, several supervised ML models operate under the fundamental assumption that the data characteristics utilized for the design of the MDA System of the 3GPP-NSMS may remain consistent post-deployment. Yet, during the deployment phase, this crucial assumption is frequently breached, primarily due to the non-stationarity of the cellular network sites. Consequently, the accuracy of the ML model, deployed within the 3GPP-NSMS, may endure substantial performance degradation. Given the unique characteristics of wireless systems, it is crucial that the MDA System within O-RAN incorporates methods to guarantee resilience against non-stationarity.

The MDAS within the 3GPP-NSMS can be produced and consumed in centralized or distributed manners [26], [29]. In a centralized intelligence paradigm, the NSSMF collects management data and that of its respective NFMFs. Next, it provides the collected data to its associated MDAF. The MDAF cleans and analyzes the collected data, trains the respective AI/ML model, and delivers analytics to the NSSMF and NFMFs. In a distributed intelligence paradigm, the NFMFs collect data and provide it to their corresponding MDAFs. Each local MDAF trains its local model and delivers analytics to the corresponding NFMF. The customization of MDAFs for the NSSMF and NFMF can be such that the MDAF for the NSSMF may possess greater analytical capabilities compared to the MDAF for the NFMF.

It is also worth noting that each MDAF and its consumer(s) interact with each other using `Analytic ID`. This ID indicates the nature of information collected by the MDAF from data sources (i.e., NSSMF and NFMF in the case of 3GPP-NSMS) and the type of information furnished in an analytics report to the MDAF consumer.

Fig. 1 shows the integration of the 3GPP-NSMS with MDA System within the SMO framework. Different NFMFs manage various NFs of an O-gNB: one for the O-CU, one for the O-DU, and one for the O-RU. Each NFMF is integrated with a MDAF, tasked with collecting and analyzing local analytics pertaining to the NFs managed by the respective NFMF. These analytics from each NFMF are aggregated at a higher level in the NSSMF, providing an overarching view of all NFs and NFMFs. The NSSMF can collect and analyze data through its associated MDAF. The MDAF co-located in the NSSMF can therefore combine the results of the lower-level MDAFs. It then provides analytics to both its corresponding NSSMF and also to local MDAFs.

Finally, it is critical for the MDA System of 3GPP-NSMS to generate and utilize AI/ML services exchanged with the intelligent systems of the management systems belonging to other SDOs within SMO. This approach facilitates the creation of a unified intelligent architecture for SMO, regardless of whether its MDAS paradigm is centralized or distributed.

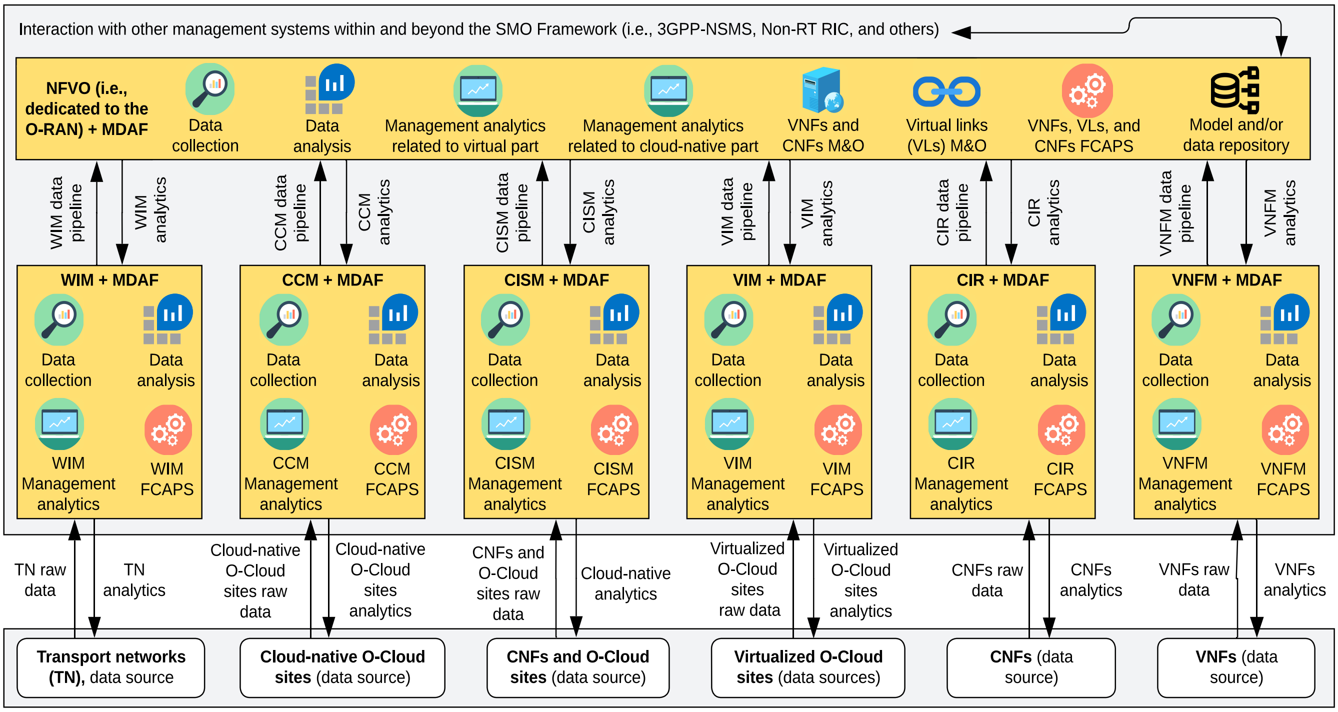

FIGURE 2. Integrating MDAS into the NFV-MANO within SMO aims to intelligentize the M&O of CNFs and VNFs through intelligent processes.

## III. LEVERAGING MDA IN NFV-MANO WITHIN THE CONTEXT OF SMO FRAMEWORK

In this section, we examine the framework of the NFV-MANO and its major building blocks. We discuss its integration into the SMO framework for the M&O of the virtualized and cloud-native aspects of O-RAN components and resources. In addition, we explore the integration of MDA into NFV-MANO. The goal is to enhance the functionalities of NFV-MANO within the SMO framework.

### A. NFV-MANO FOR THE M&O OF O-RAN

The ETSI NFV is one of the ISGs in the European standardization body. Founded in 2012 to address the increasing need for flexible, scalable, and efficient network infrastructure through virtualization, its scope has since expanded to encompass containerization as an advanced deployment option [30]. Recently, it has broadened its focus to include the physical management of network and computing infrastructure, supporting initiatives such as RAN virtualization [31]. Today, the ISG NFV remains a pivotal reference point for industry, with its standards continuing to be widely utilized and, consequently, remaining highly relevant.

The NFV-MANO framework, introduced by the ETSI ISG NFV, initially comprised three primary functional blocks (FBs), which are integral for the M&O aspects of NFV resources and services [32]. These blocks include: (a) the network function virtualization orchestrator (NFVO), tasked with the lifecycle management of VNFs and the allocation of network functions virtualization infrastructure (NFVI) resources; (b) the virtualized infrastructure manager (VIM), responsible for managing NFVI resources, including compute, storage, and networking; and (c) the virtual network function manager (VNFM), which manages the lifecycle and FCAPS of individual VNF instances, encompassing tasks such as instantiation, scaling, and termination.

In Release 4 [15], the ETSI ISG NFV introduced several MFs in addition to the existing FBs to the NFV-MANO framework. The latest architecture of the NFV-MANO is illustrated in Figure 2. The newly introduced MFs consist of container infrastructure service management (CISM), container image registry (CIR), container infrastructure service cluster management (CCM), and wide area network infrastructure manager (WIM) [5], [15]. The CISM is responsible for managing container-based applications and infrastructure services, including container management, container cluster orchestration, and management of CNFs. The CIR manages the secure and efficient lifecycle of container images, encompassing tasks such as image storage, retrieval, modification, and deletion. The CCM provides tools and procedures for managing the lifecycle of container infrastructure service (CIS) clusters, covering deployment, scaling, and monitoring. Lastly, the WIM is tasked with managing virtual links among O-Cloud sites and between O-Cloud sites and cellular network sites in the underlying NFVI. This MF facilitates the E2E M&O of transport network and resources.

Both the FBs and MFs within the NFV-MANO framework play a key role in the M&O of VNFs and CNFs [5], as well as their required resource. They can be considered part of the

SMO framework within O-RAN. They are also accountable for the M&O of the virtualized and cloudified resources within the underlying wireless network infrastructure. In addition, the NFV-MANO, particularly the NFVO, within the SMO framework must interact with the management systems of other SDOs to enable SMO to deliver unified M&O services [5]. These management systems from other SDOs may encompass the 3GPP-NSMS and the Non-RT RIC [5].

### B. INTEGRATING MDA INTO NFV-MANO

In the pursuit of achieving unprecedented operational agility and efficiency within SMO, it is crucial to imbue NFV-MANO with automation and intelligent solutions. With this objective in mind, the ETSI ISG NFV identified four deployment areas (options) in Release 4 to facilitate autonomous management and integrate intelligence into NFV-MANO [15]. These deployment areas encompass intent-based management for network services, MDA-assisted management, autonomous container infrastructure management, and hierarchical closed-loop automation (CLA). As of this writing, the ETSI ISG NFV has not specified the integration of intelligence into the NFV-MANO within the aforementioned areas. However, there is anticipation that the ISG will introduce two new enhancements within the NFV-MANO in the near future: intent-based management solutions and MDA-assisted mechanisms. While these two solutions have been briefly examined within the context of NFV-MANO in [15], the ETSI ISG NFV has not yet undertaken an in-depth exploration of the precise specifications encompassing the potential interfaces, procedures, and functionalities of the aforementioned solutions.

The intent-based management solution enables a third party to send an intent, utilizing standard-compliant interfaces, to the NFV-MANO. The NFV-MANO framework will be required to fulfill the specified intent. The third party can be a management service provider, a Non-RT RIC, a 3GPP-NSMS, etc. The MDA System gathers network and service management data from NFV-MANO, processes the collected data, trains the selected AI/ML model, and subsequently provides management analytics to the corresponding MDA consumers. The MDA-assisted solution is claimed to enhance the performance of interactions among the FBs and CNFs MFs. Additionally, it aims to optimize tasks associated with M&O and FCAPS within the NFV-MANO, as referenced in [15], [16]. The intent-based and MDA-assisted management solutions can be implemented individually, or in specific scenarios, they can be deployed side by side to enhance the capabilities of the NFV-MANO with respect to automation and intelligentization. In this article, our focus is solely on the integration of the MDA System within the internal processes of NFV-MANO.

The goal of integrating MDA into NFV-MANO is to achieve unprecedented operational agility and efficiency [33]. Hence, we consider MDA to be part of the SMO framework. According to [15], MDA serves as a key tool for diagnosing ongoing issues affecting the performance, health, or behavior of network services. Furthermore, it enables the prediction of potential issues, such as potential failures or performance degradation, thereby facilitating proactive management and maintenance.

Moreover, in [34], the recommendations previously determined were addressed by the ISG NFV to derive service and interface requirements for the MDA. Based on this, [34] introduced the MDAF within NFV-MANO, equipped with AI/ML models aimed at enhancing decision-making processes in automation related to network and service M&O. Another crucial aspect introduced is the MDA-1 interface, which facilitates communication between a MDA consumer (e.g., NFVO) and the MDA producer (i.e., MDAF). This interface enables the consumer to invoke data analytics operations provided by the producer in a subscription/notification approach, wherein the consumer subscribes to specific data analytics, and the producer delivers results when available. It is important to note that the MDA consumer may be an internal NFV entity such as the NFVO, but also other systems external to NFV-MANO. In the scope of this paper, these external management systems to the NFV-MANO, consuming MDA services, may be the Non-RT RIC or the 3GPP-NSMS.

With the commencement of Release 5 [35], ETSI ISG NFV continued to specify alternative applications of the MDAS, aiming for more automated and integrated data analytic use cases. This initiative holds the potential for increased integration with external frameworks (both management and network). However, despite the progress made in Release 5, ISG NFV has not yet achieved full integration with external frameworks from other SDOs, such as the intelligent systems of the 3GPP-NSMS and Non-RT RIC. As described in [5], such integration is crucial for establishing a unified and standard-compliant SMO framework.

The MDA System can integrate MDAFs into NFV-MANO through two distinct scenarios. One possible scenario is to create FBs and MFs that are inherently compatible with MDAS. An alternative scenario is to create MDAFs as distinct entities and link them to their corresponding FBs and MFs via standard-compliant interfaces, such as MDA-1 interface. In both scenarios, the MDA System collects network and service management data from NFV-MANO, processes the gathered data, trains the AI/ML model, and then furnishes MDAS back to the respective FBs and MFs.

The MDAS within the NFV-MANO can be provisioned in either a centralized or distributed manner.

- In a centralized approach, the NFVO acts as the central data aggregator, collecting data from all the FBs and MFs. It then provides this aggregated data to its customized MDAF, which is responsible for cleaning, analyzing, and training ML models. The MDAF then delivers the generated analytics back to the NFVO, which utilizes them to enhance the M&O capabilities of all FBs and MFs within NFV-MANO. As previously mentioned, the MDA-1 interface facilitates communication between NFVO and its MDAF.

- In a distributed approach, each FB and MF is directly connected to its corresponding MDAF through standard-compliant interfaces, such as the MDA-1 interface. This decentralized approach allows each FB and MF to independently collect, process, and analyze their own data, leading to more localized and targeted analytics. Subsequently, the MDAFs provide their analytics directly to the corresponding FBs and MFs, enabling them to make well-informed decisions.

Regardless of whether the NFV-MANO adopts a centralized or distributed approach, it must have the capability to generate and utilize AI/ML services exchanged with the intelligent systems of management systems belonging to other SDOs within the SMO framework. This capability to exchange data and analytics among various intelligent systems fosters a unified architectural solution for the intelligentization and automation of SMO. The NFV-MANO can facilitate this interaction via the NFVO and through standard-compliant interfaces, as shown in Figure 2.

In Figure 2, the integration of MDA capabilities into NFV-MANO is illustrated. Each yellow box shown in the figure represents a FB or MF, responsible for managing different resources and services of the NFV infrastructure, as indicated at the bottom line of the figure, e.g., WIM for managing the underlying transport network (TN), VNFM for managing the VNFs. Each FB and MF can be linked with a local MDAF, performing local data collection and analysis of the respective FB and MF and therefore providing local analytics. This matches to the distributed approach mentioned earlier. All the FBs and MFs, as outlined in [32], interact through standardized interfaces with the NFVO. The NFVO, in turn, can be integrated with its customized MDAF, enabling the collection and analysis of statistics at the orchestration level and integrating inputs from local MDAFs. Finally, at the top of Fig. 2, an interface is shown, assumed to allow interaction with other systems, such as the 3GPP-NSMS and the Non-RT RIC. This interface serves the purpose of enabling interaction among the intelligent systems of the three modules. More details on such integration are further described in the continuation of this paper.

## IV. ENHANCING NON-RT RIC CAPABILITIES WITH AI/ML-ASSISTED SOLUTIONS

In [5], we extensively examined the architectural design of the Non-RT RIC. The proposed design incorporates a variety of functions, with three specifically pivotal for fostering its intelligentization and automation capabilities. These functions include the AI/ML Monitoring, External AI/ML Termination, and AI/ML Workflow. We introduced their incorporation into a unified function, collectively referred to as the AI/ML Function. This newly proposed function must have the capability to encompass not only the mentioned three functions (i.e., those introduced by the O-RAN Alliance) but also any additional ones related to AI/ML services in the future. Its primary responsibility is to automate and infuse intelligence into the operations of the Non-RT RIC.

In this section, our attention centers on the AI/ML Function. Here, we explore the architecture underpinning this critical function, delving into both its established functionalities and the introduction of innovative capabilities and components. These enhancements are designed to ensure alignment with the diverse requirements of AI/ML model training, hosting, execution, and monitoring within Non-RT RIC.

Figure 3 portrays an illustrative depiction of the architectural framework proposed for the AI/ML Function. This framework is crafted to enhance the ability of the Non-RT RIC to make data-driven decisions, optimize the performance of SMO, and ensure the efficient management of managing objects and managed objects within the O-RAN architecture. By offering real-time insights into Non-RT RIC behavior and AI/ML model performance, the proposed architectural framework plays a pivotal role in realizing the vision of intelligent and automated network and service management within the SMO framework. Through this discussion, we aim to elucidate the different parts, components, interactions, and objectives of the proposed AI/ML Function in a comprehensive manner.

Figure 3 illustrates the composition of the AI/ML Function architecture, which consists of four parts: (a) Management Data Collection and Processing; (b) AI/ML Model Preparation and Management; (c) Compulsory and Optional Components; and (d) External Integration Points. The Management Data Collection and Processing part encompasses tasks associated with collecting, processing, storing, and managing management data. It also identifies the sources from which management data needs to be collected. The AI/ML Model Preparation and Management part assumes responsibility for various tasks like training, testing, validating, hosting, management, and monitoring of the AI/ML model. The Compulsory and Optional Components part outlines various elements that may be incorporated into the AI/ML Function. These components can be either mandatory (compulsory) or discretionary (optional), depending on the specific requirements and objectives of the AI/ML Function. The External Integration Points part plays a critical role in facilitating the connection and interoperability of the AI/ML Function with other external AI/ML service-producing components, both within and outside the SMO framework. This part enables data/model exchange, communication, and collaboration, essential for harnessing the full potential of AI/ML capabilities.

In the subsequent subsections, we provide a detailed breakdown of the components and functionalities within each of the aforementioned parts, offering a comprehensive view of the AI/ML Function and its enhanced capabilities and features within the Non-RT RIC architecture.

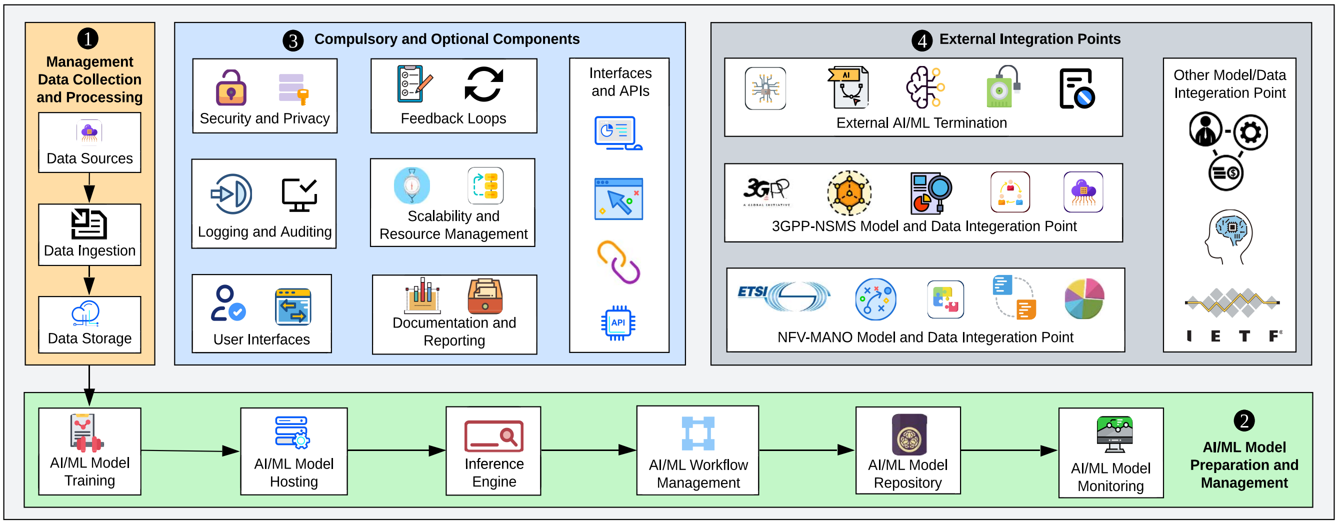
1
Management Data Collection and Processing
Data Sources
Data Ingestion
Data Storage
3 Compulsory and Optional Components
Security and Privacy
Feedback Loops
Interfaces and APIs
Logging and Auditing
Scalability and Resource Management
User Interfaces
Documentation and Reporting
4 External Integration Points
External AI/ML Termination
3GPP-NSMS Model and Data Integeration Point
NFV-MANO Model and Data Integeration Point
Other Model/Data Integeration Point
IETF
AI/ML Model Training
AI/ML Model Hosting
Inference Engine
AI/ML Workflow Management
AI/ML Model Repository
AI/ML Model Monitoring
2 AI/ML Model Preparation and Management

FIGURE 3. The proposed architecture for the AI/ML Function embedded within the Non-RT RIC. This diagram portrays its integral elements, encompassing crucial stages such as data collection and ingestion, model training and management, external reference points, and several optional and mandatory functionalities.

### A. MANAGEMENT DATA COLLECTION AND PROCESSING

**Data Sources** are the foundational elements of the AI/ML Function, encompassing various management data producers both within and beyond the Non-RT RIC architecture. These sources include the components of the three modules, which generate a wealth of data related to network and service performance. They may also include components beyond the three modules. Their primary role is to identify and collect a diverse range of management data types [36], including network telemetry, user behavior, and service performance metrics, from various sources within and beyond SMO.

Data Sources serve as the initial data touchpoints, providing management data – whether raw or cleansed – to fuel subsequent AI/ML processes. To fulfill this role effectively, Data Sources must ensure data quality, preprocessing, and validation, accommodating various data formats and sources [36]. They play a critical role in maintaining scalability, data privacy, real-time capabilities, and seamless integration within the AI/ML Function. The efficiency of Data Sources directly impacts the quality and effectiveness of AI-driven insights and decision-making processes [36], making them integral components of the Non-RT RIC‘s intelligent system.

**Data Ingestion** serves as the crucial bridge connecting Data Sources to subsequent AI/ML processes within the AI/ML Function. Its primary role is to efficiently receive, preprocess, and structure raw data collected from various sources. Data Ingestion ensures data quality, cleanliness, and compatibility by performing tasks such as validation, format conversion, and organization [36], [37]. It plays a pivotal role in preparing data for AI/ML model training, real-time inference, and monitoring. This contribution enhances the reliability and effectiveness of data-driven insights and decision-making within the AI/ML Function.

**Data Storage** serves as the foundation for the management and persistent storage of data. It retains historical and real-time data collected from diverse sources [38]. These components are designed to be highly scalable, accommodating the ever-growing volumes of data generated within SMO. They ensure data accessibility for various AI/ML processes, including model training, real-time analytics, and performance monitoring [36]. Security measures, including data encryption and access controls, are integral to safeguard data integrity and confidentiality. Effective data lifecycle management, encompassing retention policies and archiving, optimizes storage resources while maintaining data relevance. By fulfilling these objectives, Data Storage components support the Non-RT RIC in making data-driven decisions, facilitating real-time analysis, and preserving historical data as a valuable resource for trend analysis, compliance, and informed decision-making.

### B. AI/ML MODEL PREPARATION AND MANAGEMENT

**AI/ML Model Training** is responsible for the training and refinement of AI/ML models, designed to process, analyze, and extract insights from the management data collected by Data Sources and stored within Data Storage. The primary objective is to craft AI/ML models that effectively capture patterns, make predictions, and provide recommendations based on the data’s characteristics [38]. This entails processes like feature engineering and hyperparameter tuning to enhance model accuracy. The quality of training data is paramount, necessitating data preprocessing and validation for optimal results. Rigorous testing against real-time and accurate data ensures the reliability and performance of the AI/ML models. Once trained, these models are deployed to undertake real-time analytics, anomaly detection, and predictive tasks. The AI/ML Model Training component stands as a linchpin in the realization of data-driven decision-making within the AI/ML Function, influencing the precision and effectiveness of its insights and actions across the Non-RT RIC [36], [38].

**AI/ML Model Hosting** focuses on the hosting, deployment, and real-time execution of trained AI/ML models. Once models are crafted and validated, this component ensures their integration into the AI/ML Function, making them available for inference. Its core objectives encompass deploying models, enabling real-time analysis by accepting data inputs, processing them through the models, and delivering swift predictions or recommendations. In this component, scalability becomes paramount to accommodate varying workloads and dynamic network environments [36], while efficient resource utilization and low-latency execution are critical for timely decision-making [38]. AI/ML Model Hosting also orchestrates version management, ensuring compatibility and facilitating updates. Through these endeavors, it serves as the vital conduit that empowers the Non-RT RIC's intelligent system, delivering actionable insights and enabling data-driven decision-making.

**Inference Engine** plays a critical role in the AI/ML Function as the real-time execution powerhouse. Its objective is the execution of trained AI/ML models, enabling timely, responsive, and data-driven decision-making [38]. This component accepts input data from diverse sources, preprocesses it as per model requirements, and guides it through the deployed models to produce instant predictions or recommendations [36]. Similar to AI/ML Model Hosting, achieving low latency is a top priority, ensuring that insights are swiftly delivered for real-time monitoring and decision-making. Furthermore, scalability is key, with the Inference Engine designed to accommodate concurrent requests and dynamic network environments. Compatibility with different model versions and resource optimization further enhance its capabilities. Through these efforts, the Inference Engine stands as a key, enabling the Non-RT RIC's intelligent system to translate data into actionable insights and support informed decision-making [36].

**AI/ML Workflow Management** is an essential element within the AI/ML Function. The O-RAN Alliance included this component as an internal feature of the Non-RT RIC [5]. We utilize this function as a component of the AI/ML Function. It aims to effectively manage, orchestrate, automate, and optimize the E2E lifecycle of the AI/ML workflow, which is crucial for improving the performance of the AI/ML model. The key functionalities of this component involve designing workflow, handling the lifecycle of an AI/ML model, scalability, integration, and compliance. These aspects collectively strive to improve the efficiency and reliability of SMO by leveraging AI/ML algorithms. To implement the functionalities listed above, it is necessary for the AI/ML Workflow Management to provide interoperability across the various components of the AI/ML Function within the Non-RT RIC.

**AI/ML Model Repository** serves as a centralized platform for the systematic storage, version control, and management of AI/ML assets, with a specific focus on models, datasets, configurations, and metadata [37]. Its core objectives encompass asset management, collaboration facilitation, reproducibility assurance, security provision, scalability support, and comprehensive documentation. Its key features include meticulous storage combined with version control, metadata management, robust access control, streamlined model deployment, efficient asset discovery, tight integration with AI/ML workflows, data versioning capabilities, and audit trail maintenance [37]. By fulfilling these roles, the AI/ML Model Repository fosters collaboration among AI/ML engineers, ensures experiment reproducibility, simplifies model deployment processes, and provides a secure, well-organized repository for AI/ML assets, ultimately supporting the development, deployment, refinement, and management of AI/ML-assisted solutions within the Non-RT RIC architecture [37].

**AI/ML Model Monitoring** monitors the operations of the AI/ML Function and ensures its optimal performance. This component plays a diverse role, starting with the meticulous monitoring of every component inside the AI/ML Function, ensuring their well-being and operational effectiveness. It records AI/ML Function events through logging and auditing, enabling anomaly detection and troubleshooting. By efficiently controlling and maintaining the various components of the AI/ML Function, this component ensures the reliability, stability, and responsiveness of the Non-RT RIC, ultimately supporting data-driven decision-making.

### C. COMPULSORY AND OPTIONAL COMPONENTS

**Interfaces and application programming interfaces (APIs)** orchestrate seamless communication and interaction among the components of the AI/ML Function. The objective of APIs and open Interfaces is to establish channels that enable the smooth exchange of data and commands [33]. This enables the uninterrupted flow of information from Data Source components to trained AI/ML models, through the inference engine, and across different aspects of the AI/ML Function. In addition to internal connections, they also offer standardized interfaces and APIs to external components, allowing external platforms, management systems, and orchestration frameworks to easily connect and utilize the capabilities of the AI/ML Function. These components are designed with scalability, extensibility, and performance optimization in mind, adhering to industry standards and best practices. Through this framework, Interfaces and APIs empower the Non-RT RIC‘s intelligent system to seamlessly connect, share, and act upon data-driven insights, fostering agile and informed decision-making.

**Security and Privacy** assume the role of vigilant sentinels, standing guard over the sanctity of data and operations within the AI/ML Function. Their mandate encompasses a multi-faceted mission, beginning with the protection of data's confidentiality, integrity, and availability. Through encryption, access controls, and data segmentation, they shroud sensitive information in a protective cloak. Access control mechanisms ensure that only trusted individuals and systems gain entry, reinforced by rigorous authentication and authorization processes [33]. Audit trails and logs diligently record every system interaction, serving as watchful eyes that monitor activities and detect anomalies. Compliance with data privacy regulations and standards is non-negotiable, with stringent privacy safeguards in place. A well-honed incident response protocol stands ready to thwart any security breaches. Regulatory compliance, data minimization, security testing, and privacy impact assessments all fall within their purview [33]. In summary, Security and Privacy components play a crucial role in maintaining trust by ensuring that users, administrators, and stakeholders have their data protected, their privacy maintained, and the integrity of the AI/ML Function preserved.

**Logging and Auditing** act as careful record keepers and attentive guardians within the AI/ML Function. They capture and record a rich spectrum of system events, from user interactions and configuration changes to data accesses and security-related incidents. These records serve as the memory of the AI/ML Function, enabling retrospective analysis, anomaly detection, and forensic investigations. Data retention policies ensure that historical records remain accessible for as long as required, supporting compliance verification with data privacy regulations. Integrity and access control measures safeguard the sanctity of log and audit data, preventing tampering or unauthorized access. In the crucible of log aggregation, data from various system facets converges for centralized analysis. Lastly, Logging and Auditing are the custodians of transparency, accountability, and security, nurturing a culture of vigilance and data-driven insights within the AI/ML Function.

**User Interfaces** are facilitating communication between administrators and the components of the AI/ML Function. These interfaces serve a diverse range of functions, providing users with user-friendly access to the AI/ML Function while delivering information, analysis, and management options in visually understandable formats. Operators have the ability to effortlessly engage with the AI/ML Function, utilizing configuration options and initiating actions. The utilization of real-time dashboards and alerts serves to provide users with up-to-date information, while the implementation of role-based access restrictions guarantees that users only view the data that is related to their specific duties. The primary focus in interface design is on usability, which aims to facilitate the efficient completion of tasks while minimizing the need for extensive learning. The consideration of scalability ensures that interfaces maintain their responsiveness and functionality when the AI/ML Function undergoes scaling. To sum up, User Interfaces serve as the portal to the AI/ML Function, rendering complex data and operations accessible, comprehensible, and responsive to the human touch, empowering operators and administrators to make informed decisions.

**Scalability and Resource Management** are key components that orchestrate computational resources to maintain system harmony and responsiveness. The objective of these components is to effectively distribute and orchestrate resources, including the compute, memory, and storage, with accuracy and flexibility. These computing resources are employed for training, validating, and testing the AI/ML models within the Non-RT RIC. The optimal distribution of these resources is crucial for enhancing the speed and effectiveness of training [33]. The components are responsible for implementing auto-scaling techniques that adaptively allocate resources based on fluctuating workloads. These strategies are designed to efficiently handle both high-demand periods and low-activity periods by either scaling up resources or conserving them accordingly.

Load balancing techniques can be effectively utilized to achieve an equitable distribution of data and workloads across available resources, thereby preventing congestion and ensuring optimal system performance. Continuous resource monitoring maintains a constant state of vigilance, promptly alerting operators when thresholds are reached or surpassed. Capacity planning is a strategic process that involves forecasting and preparing for future resource needs in order to facilitate system expansion. Resource optimization strategies seek to minimize waste and enhance efficiency. The deployment of these techniques can ensure that the AI/ML Function architecture is capable of scaling and orchestrating harmoniously to meet the dynamic performance requirements.

**Feedback Loops** act as the cognitive foundation of the AI/ML Function, facilitating an ongoing exchange of information between its components. These loops extract data-driven insights, detect anomalies, and orchestrate adaptive responses, ensuring that the AI/ML Function remains agile, responsive, and perpetually improving. By integrating real-time data and historical knowledge, the Feedback Loops unravel hidden patterns and deviations, sounding the alarm when discrepancies are spotted and proactively averting potential disruptions. These loops drive the engine of continuous improvement, channeling feedback into the refinement of AI/ML models, system configurations, and resource allocation strategies. By utilizing predictive foresight, they enhance the resilience of the system by anticipating and proactively resolving forthcoming trends and challenges. In the context of dynamic data landscapes, Feedback Loops monitor and optimize the performance of the AI/ML Function. These loops possess the essential qualities of vigilance, insight, adaptability, and automation, enabling them to effectively coordinate and enhance operational efficiency.

**Documentation and Reporting** are responsible for recording and communicating information. They rigorously document the progress of the AI/ML Function and convey its insights to the components (within or beyond the Non-RT RIC) interested in understanding its operations. Documentation and Reporting diligently capture and store log data, performance metrics, and noteworthy events. This practice aims to create a comprehensive historical narrative that sheds light on the behavior and evolution of the AI/ML Function. Equipped with an abundance of data, these components generate notifications, dashboards, visual representations, and reports that transform unprocessed data into actionable insights.

The purpose of these reports extends beyond mere documentation, as they serve as coherent narratives that effectively convey the state of the AI/ML Function, including its overall condition, any deviations from normalcy, and prevailing patterns, to administrators and operators for making decisions. By employing real-time monitoring, configurable reporting, and adherence to compliance standards, these components play a crucial role in maintaining transparency, accountability, and the ability to make well-informed decisions and continuously improve the AI/ML Function. In essence, Documentation and Reporting are the chroniclers and interpreters of the AI/ML Function. They provide guidance and insights that contribute to achieving operational excellence and fostering data-driven knowledge and well-informed decision-making.

### D. EXTERNAL INTEGRATION POINTS

**External Integration Points** connect the AI/ML Function with the AI/ML (or management data) producers anchored within and beyond SMO, facilitating access to a wide range of data, services, and opportunities for cooperation. The AI/ML producers encompass components such as the 3GPP-NSMS MDAFs, the NFV-MANO MDAFs, and other producers of data or models. The major role of the External Integration Points is the efficient acquisition of data from external sources, the sharing of AI/ML models and insights with external services, and the capability to receive event notifications from the wider ecosystem. These integration points also facilitate data enrichment, which involves incorporating information from external sources to enhance the contextual comprehension of AI/ML analysis. To achieve interoperability, it is envisaged that the External Integration Points will engage in communication using standardized language and protocols while also implementing rigorous security measures and access controls. External Integration Points enable the AI/ML Function to surpass its limitations by handling various data formats, protocols, and compliance requirements. This allows for the integration of external information, enhances the analytical capabilities of the system, and facilitates data-driven intelligence and collaboration.

## V. UNIFICATION AND INTEROPERABILITY AMONG THE INTELLIGENT SYSTEMS OF THE THREE MODULES

To facilitate seamless integration and the effective implementation of intelligent-driven solutions within the SMO framework, it is crucial for the distinct components of the intelligent systems across the three modules to collaborate in a unified fashion. To realize this unification, we propose an architectural solution, as illustrated in Figure 4. The proposal leverages two logical functions, also referred to as terminations, namely the NSSMF Termination and the NFVO Termination. It also utilizes their associated interfaces, namely the NSSMF_NonRTRIC and NFVO_NonRTRIC. The aim of leveraging the two terminations and their associated interfaces is to unify the three modules and facilitate interoperability among them. We conceptualized and detailed the proposed terminations and interfaces in [5]. In this article, specifically in this section, we explore the complexities involved in fostering seamless interoperability among components that both generate and consume AI/ML services across the three modules. To further enhance the capabilities of the three modules, the two terminations and their associated interfaces shall expand their support to encompass functionalities and aspects related to AI/ML service exchange and delivery. This extension aims to complement the fundamental features and characteristics of the terminations and interfaces, as addressed in [5].

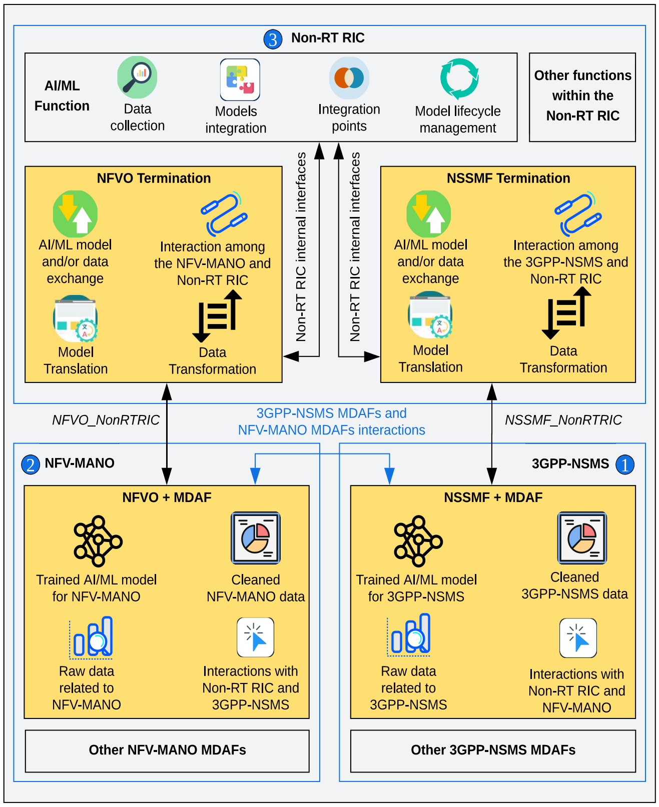
3 Non-RT RIC
AI/ML Function
Data collection
Models integration
Integration points
Model lifecycle management
Other functions within the Non-RT RIC
NFVO Termination
AI/ML model and/or data exchange
Interaction among the NFV-MANO and Non-RT RIC
Model Translation
Data Transformation
Non-RT RIC internal interfaces
Non-RT RIC internal interfaces
NSSMF Termination
AI/ML model and/or data exchange
Interaction among the 3GPP-NSMS and Non-RT RIC
Model Translation
Data Transformation
NFVO_NonRTRIC
3GPP-NSMS MDAFs and NFV-MANO MDAFs interactions
NSSMF_NonRTRIC
2 NFV-MANO
NFVO + MDAF
Trained AI/ML model for NFV-MANO
Cleaned NFV-MANO data
Raw data related to NFV-MANO
Interactions with Non-RT RIC and 3GPP-NSMS
Other NFV-MANO MDAFs
3GPP-NSMS 1
NSSMF + MDAF
Trained AI/ML model for 3GPP-NSMS
Cleaned 3GPP-NSMS data
Raw data related to 3GPP-NSMS
Interactions with Non-RT RIC and NFV-MANO
Other 3GPP-NSMS MDAFs

FIGURE 4. Proposal for seamless unification and interoperability among the intelligent systems of the three modules within SMO.

The NFVO Termination and NSSMF Termination serve as pivotal components for the exchange of data (cleaned and raw) and AI/ML models between the MDAFs of the NFVO and NSSMF and the AI/ML Function of the Non-RT RIC. To achieve effective data or model exchange, Data Transformation and Model Translation are crucial features of the two terminations, as shown in Figure 4. Data Transformation acts as a data wrangler, ensuring that the data formats and structures between modules align. It performs data cleaning, normalization, encoding, and other necessary operations to harmonize the data flow. On the other hand, Model Translation acts as a linguistic interpreter, translating the models developed in one module into a language understandable by the other module. It encapsulates models as serialized representations, converts algorithms to compatible formats, and ensures parameter compatibility. By seamlessly transforming data and translating models, both terminations enable the seamless exchange of information and expertise between the intelligent systems of the three modules, enhancing the overall performance and capabilities of SMO. Furthermore, the two terminations also establish communication with the AI/ML Function of the Non-RT RIC via internal interfaces. This interaction enables seamless data and model exchange among the three modules within the SMO framework. Further details on the definitions and scope of both terminations, in relation to supporting AI/ML service exchange within the SMO framework, are provided in Table 1.

The `NFVO_NonRTRIC` and `NSSMF_NonRTRIC` interfaces enable communication between the MDAFs of NFVO and NSSMF and the NFVO Termination and NSSMF Termination, respectively. Beyond their core functionalities described in [5], the two interfaces can additionally enable signaling, synchronization, and transfer of data and/or models between the intelligent systems of the three modules. They can handle high-throughput data transfer, ensuring data accuracy and reliability, even in the presence of network disruptions. Furthermore, they facilitate authentication, authorization, and data encryption, safeguarding sensitive information from unauthorized access. By implementing these functionalities, the two interfaces ensure a secure, reliable, and efficient communication channel between the three intelligent systems. Table 2 provides additional information regarding the scope and implications of the two interfaces in supporting AI/ML service exchanges between the two proposed terminations.

Moreover, the MDAFs for the NSSMF and NFVO can have additional capabilities beyond what is shown in Figures 1 and 2. These supplementary capabilities would enable the two MDAFs to consume and produce AI/ML services both within and outside SMO. These capabilities include, but are not limited to, archiving and exchanging raw data, cleansed data, and trained models. The MDA Systems of the 3GPP-NSMS and NFV-MANO execute these tasks in collaboration with the Non-RT RIC through the defined interaction points depicted in Figure 4. The interaction points not only facilitate the creation and provisioning of AI/ML models and data for the Non-RT RIC but also enable NFV-MANO and 3GPP-NSMS to retrieve and extract AI/ML models or data from the Non-RT RIC as required. This capability empowers the Non-RT RIC to influence the decision-making of the MDA Systems in 3GPP-NSMS and NFV-MANO. Furthermore, it grants both MDA Systems access to a wealth of data and models from the AI/ML Function and R1 services within the Non-RT RIC, further bolstering their model development, deployment, and enhancement capabilities.

Introducing this multi-SDOs, unified, interoperable, and crowd-sourced intelligent system across the three modules represents a significant shift within the SMO framework, transcending siloed limitations and embracing a holistic perspective of the O-RAN architecture. It fosters seamless communication and collaboration among the intelligent systems of the three modules, breaking down data silos and enabling the exchange of knowledge and insights. This interconnectedness empowers SMO with the ability to make well-informed decisions grounded in a wider and more comprehensive understanding of the network's dynamics in an intelligent manner. The crowd-sourcing aspect further strengthens the SMO framework's capabilities by tapping into the collective intelligence of a vast network of components. By incorporating data and insights from diverse sources, the intelligent system of each module can continuously refine its AI/ML models, adapt to evolving network conditions, and provide personalized recommendations to optimize user experiences. In conclusion, the proposed unification and interoperability among the three intelligent systems represent a groundbreaking advancement in the M&O of O-RAN components and interfaces. This transformative approach paves the way for proactive issue identification, intelligent solution deployment, and optimized network performance and resource utilization, propelling O-RAN into a new era of efficiency, resilience, intelligence, automation, and innovation.

## VI. E2E LIFECYCLE WORKFLOW OF THE AI/ML MODEL ACROSS THE THREE INTELLIGENT SYSTEMS

Building upon the contributions from the previous sections, we present an E2E workflow for the lifecycle of a ML model across the three intelligent systems in this section. The proposed E2E workflow provides a well-structured plan for ML development within the SMO framework, fostering efficient and effective deployment of trained models. It promotes collaboration and knowledge sharing among the three intelligent systems, ensuring that ML models are enriched by collective insights and expertise. Furthermore, the workflow streamlines seamless integration and exchange of data, enabling the continuous refinement and optimization of ML models based on real-time management and network data. The holistic approach presented in this section aims to maximize the intelligence and adaptability of the SMO framework by leveraging the strengths of each module and fostering a synergistic relationship among them.

The comprehensive and integrated workflow for the development, deployment, and optimization of ML models across the three modules within SMO is illustrated in Figure 5. This workflow is structured into several distinct phases,

TABLE 1. The definitions and scope of the two proposed functions that enable interoperability among the intelligent systems of the three modules.

| Termination Name | Scope Regarding Support for AI/ML Capabilities |
|---|---|
| NSSMF Termination | Building upon its main features (described in [5]), the NSSMF Termination must also support advanced features related to AI/ML and automation within SMO. The newly introduced enhancements aim to leverage the capabilities of AI/ML to improve efficiency, accuracy, security, and performance when the intelligent systems of the Non-RT RIC (i.e., the AI/ML Function) and 3GPP-NSMS (i.e., the MDA System) interact via this termination. These mainly include the exchange of data (both raw and cleansed) as well as AI/ML model between the MDAF of the NSSMF and the AI/ML Function of the Non-RT RIC. In the exchange processes, AI/ML plays a key role by enabling intelligent routing and load balancing through predictive analytics. These analytics anticipate data traffic patterns and adjust routing strategies dynamically to balance the load across two modules, ensuring efficient and timely data/model exchanges. Additionally, real-time monitoring and feedback mechanisms continuously track the performance of data/model exchanges, identifying bottlenecks and inefficiencies. They utilize AI/ML algorithms to adapt and improve future exchanges based on past performance. Finally, AI-driven error handling and recovery processes employ predictive maintenance to foresee potential failures during model/data exchanges. This enables preemptive measures to avoid disruptions and automated recovery protocols to ensure minimal downtime and data loss in case of issues. |
| NFVO Termination | In addition to the core functionalities outlined in [5] for this termination, it must also support the exchange of AI/ML capabilities between the NFV-MANO and the Non-RT RIC. These capabilities primarily involve the exchange of data (both raw and cleansed) and trained AI/ML models between the AI/ML Function of the Non-RT RIC and the MDAF of the NFVO. This entails robust communication protocols to ensure reliable transmission of diverse data formats, ranging from structured datasets to complex ML models. Moreover, the NFVO Termination boasts comprehensive authentication and authorization mechanisms to safeguard the integrity and confidentiality of the exchanged data and models. Furthermore, it is endowed with scalability features capable of accommodating varying workloads and evolving data requirements over time. Within the NFVO Termination, real-time feedback and monitoring of data and model exchange facilitate the identification of bottlenecks and inefficiencies by both modules. The monitoring mechanism can utilize AI/ML algorithms to adjust and enhance forthcoming data and model exchanges based on insights from previous performance. Finally, during the exchange process, AI/ML takes center stage by facilitating intelligent routing and load balancing through predictive analytics. These analytics forecast data traffic patterns and dynamically adjust routing strategies to evenly distribute the load between the two modules, ensuring the smooth and timely exchange of data/models. |

each of which is represented visually by an individual box. The components of each module undertake a set of specific ML operations throughout each phase of the workflow. The output of each phase is transmitted to the subsequent phase through a dedicated step, as denoted in Figure 5. These phases or steps can be optional or mandatory. The inclusion or exclusion of specific phases and steps within the proposed workflow is determined by the discretion of the network operator and AI/ML model deployment requirements.

In the first phase, data sources and relevant management data are identified and gathered within each module through corresponding interfaces, as shown in Figure 5. For example, the MDA System of the 3GPP-NSMS collects management data from NSSMF, NFMFs, O-gNB NFs, and user equipments (UEs). Likewise, the MDA System of the NFV-MANO and the AI/ML Function of the Non-RT RIC acquire relevant data from their respective sources. Subsequently, the collected data can be archived in a designated data repository within each module (step 1). This repository serves as a central hub for managing and accessing the vast amount of data generated by each module. Following collection, the collected data undergoes thorough storage, management, and organization to ensure seamless access, analysis, and sharing (step 2). Next, the selected data undergoes rigorous cleansing, ingestion, fusion, and preparation for training (step 3). The refined and cleaned data is then meticulously analyzed and prepared to be utilized to train and develop the desired ML models (step 4).

Within each module, the model training host, such as a MDAF responsible for model training, decides whether to utilize an existing model or initiate the training of a new model. The selection of an appropriate ML model is performed in accordance with the requirements of the use case that demands a ML-assisted solution. Upon selecting a ML model, it undergoes training, testing, and validation using the prepared data as a foundation. Once the model has successfully completed training and validation, the training host will communicate the trained model's availability to the model management phase (e.g., any MDAF responsible for managing ML models), initiating the process of certification, onboarding, and deployment (step 5), as shown in Figure 5.

For seamless deployment of the trained ML model across each module, the training host may provide model compilation instructions to the designated compiling host (step 6). The compiling host, if utilized, can be a MDAF or AI/ML Function capability that transforms a trained model into an optimized executable format for deployment across SMO. This information will be subsequently communicated to the model management phase during the deployment stage (step 7), as illustrated in Figure 5. The model management phase will coordinate the deployment of the compiled model within each module, ensuring seamless integration

TABLE 2. **Detailed definitions and descriptions of the two interfaces used to connect the two proposed logical terminations (namely NFVO Termination and NSSMF Termination) within the Non-RT RIC to the intelligent systems of NFV-MANO and 3GPP-NSMS, emphasizing their support for AI/ML capabilities.**

| Endpoints | Interface Name | Scope Regarding Support for AI/ML Capabilities |
|---|---|---|
| MDAF of NSSMF – AI/ML Function | `NSSMF_NonRTRIC` | Initially, the `NSSMF_NonRTRIC` interface is used to connect the NSSMF and NSSMF Termination, aiming to enable interoperability between the 3GPP-NSMS and the Non-RT RIC within the SMO framework. In addition to its fundamental features, as discussed in [5], this interface must also support the exchange of data (both raw and cleansed) and AI/ML models between the AI/ML Function within the Non-RT RIC and the MDAF of the NSSMF. To ensure seamless data/model exchange, the `NSSMF_NonRTRIC` interface should incorporate several key functionalities. Firstly, it should enable efficient data transfer mechanisms, supporting various data formats and protocols. It should also facilitate model serialization and deserialization to ensure AI/ML models can be transmitted in a compact and standardized format. Furthermore, the `NSSMF_NonRTRIC` interface should provide secure authentication and authorization to protect data/model integrity and privacy. Version control features are essential for tracking changes in models and datasets, ensuring compatibility and reproducibility. Moreover, it must support asynchronous communication and include error handling and logging mechanisms to diagnose and resolve issues promptly. Finally, the `NSSMF_NonRTRIC` interface should include monitoring and reporting capabilities to track performance metrics and data flow, ensuring smooth and efficient operations between the intelligent systems of the two modules. |
| MDAF of NFVO – AI/ML Function | `NFVO_NonRTRIC` | Originally, the `NFVO_NonRTRIC` interface facilitates seamless interoperability between the NFV-MANO and Non-RT RIC by linking the NFVO with the NFVO Termination within the SMO framework. Besides its core functionalities outlined in [5], the `NFVO_NonRTRIC` interface must also enable the transfer of both raw and processed data, as well as AI/ML models, between the AI/ML Function within the Non-RT RIC and the MDAF of the NFVO. For a smooth exchange of data and AI/ML models, the interface should integrate several essential functionalities. First and foremost, this interface should facilitate efficient data transmission mechanisms, accommodating diverse data formats and protocols. Additionally, it should streamline model serialization and deserialization processes to guarantee compact and standardized model transmission. Moreover, the `NFVO_NonRTRIC` interface ought to furnish robust authentication and authorization mechanisms to safeguard the integrity and privacy of data and models. Version control functionalities are indispensable for monitoring alterations in models and datasets, thereby ensuring compatibility and reproducibility. Furthermore, it must facilitate asynchronous communication and incorporate error handling and logging mechanisms to swiftly diagnose and resolve issues. Lastly, the `NFVO_NonRTRIC` interface should encompass monitoring and reporting functionalities to monitor performance metrics and data flow, guaranteeing seamless and efficient operations between the intelligent systems of the NFV-MANO and Non-RT RIC. |

and operation. Furthermore, during the model management phase, the compiled model can be stored in a centralized model repository for future retrieval and deployment. The repository functions as a centralized hub where models can be accessed and managed at any stage of their lifecycle.

At this stage, the compiled model transitions from a dormant state into active operation at the ML inference host (step 8) within each module. An inference host can be a MDAF or an AI/ML Function that serves as the platform for the trained model to perform inference tasks, which may include both online learning and model execution. This specifically implies that the MDAFs within 3GPP-NSMS, the MDAFs within NFV-MANO, and the AI/ML Function within Non-RT RIC initiate the deployment of their trained models within their respective domain-specific contexts. While in operation, the inference host can engage in online training by continuously acquiring real-time management data from data sources, enabling the ML model to adapt and improve its performance over time (step 9), as shown in Figure 5. It seamlessly integrates online data with an offline-trained model to dynamically augment the model's capabilities during operation. This continuous learning process ensures that the model remains adaptive and optimized to handle real-world changes and complexities.

The inference host has the capability to exchange model or data within each module. This can be facilitated using standard-complaint components and interfaces. For example, the NSSMF Termination and `NSSMF_NonRTRIC` facilitate model or data sharing between the MDAF of NSSMF and the AI/ML Function within the Non-RT RIC. This capability enables each module to generate and consume data or models from other modules, promoting seamless interoperability among the intelligent systems within the three modules. Utilizing the output of a ML model (step 10), the inference host will proactively notify the relevant actors to undertake the necessary actions (step 11) towards the subject of action, as illustrated in Figure 5. The subject of action can either be a NF or a MF that requires optimization through a

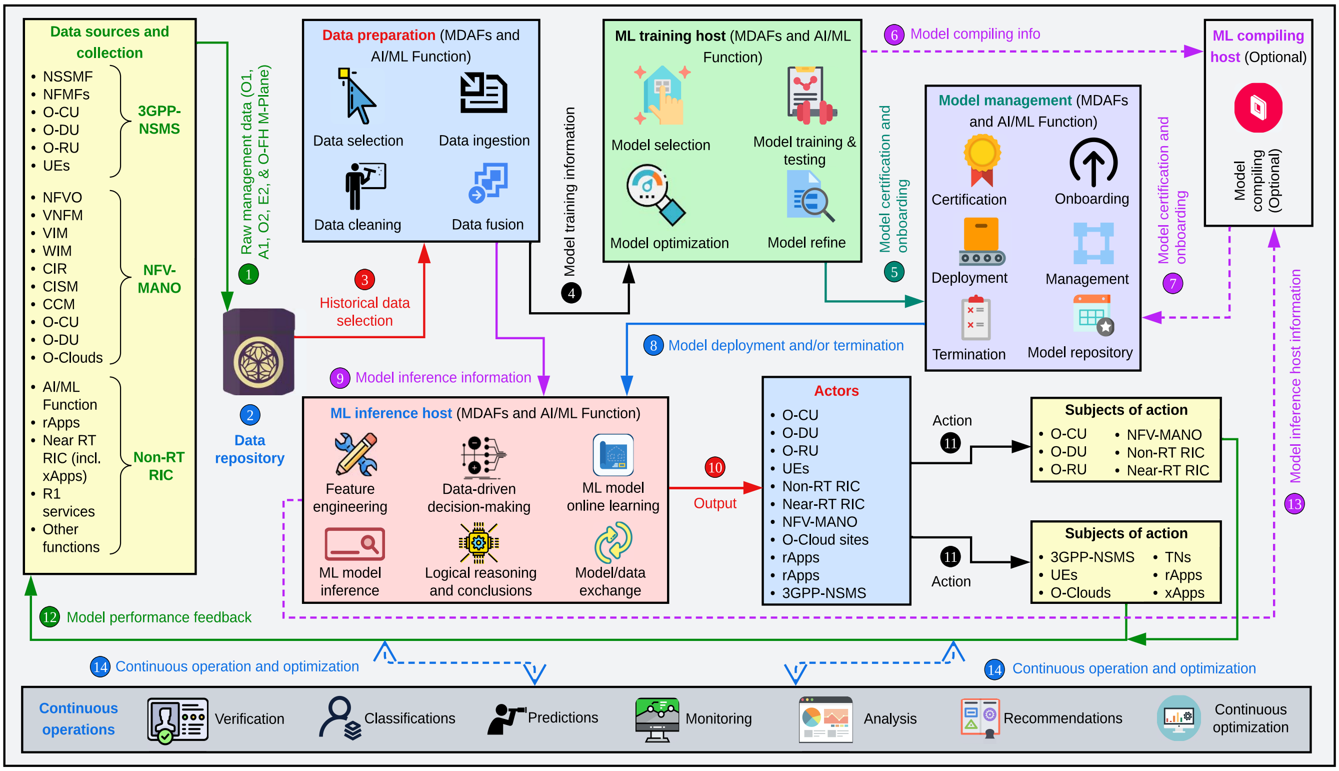

FIGURE 5. E2E workflow for data collection and ML model lifecycle management across the intelligent systems of the three modules within SMO.

ML-assisted solution. Within the proposed workflow, an actor can be any entity that houses a ML-driven solution, leverages the output of ML model inference, and executes a specified action.

Up to this point, the ML model has been trained and deployed within each module and is also accessible for exchange with the other modules. To further enhance the performance of the deployed model, the workflow continuously collects performance reports from the actors. Figure 5 shows that each module anticipates actors to provide feedback or report on the performance of the ML model (step 12). Furthermore, the inference host also provides the compiling host with reports on the performance of the model (step 13). These performance reports are then analyzed and utilized to potentially retrain, refine, optimize, update, re-select, or, in the worst-case scenario, terminate the model for execution in their respective phases (see model training and model management phases).

Finally, the workflow encompasses the continuous operation phase (step 14), as depicted in Figure 5. This phase incorporates a range of features for the continuous enhancement of ML models throughout their E2E lifecycle. During this phase, the ML model and its output undergo verification, monitoring, analysis, and optimization. Recommendations, predictions, and classifications are dynamically provided to the intelligent system of each module based on real-time requirements or predetermined time intervals.

## VII. KEY RESEARCH CHALLENGES

This section identifies key research challenges linked to the integration of intelligence into the SMO framework. Addressing these challenges necessitates substantial interdisciplinary research efforts and collaborative partnerships among various SDOs and stakeholders. This focus is crucial to meet the growing needs of developing and deploying an intelligent SMO. Ultimately, this will unlock the full potential of intelligent M&O for O-RAN components, facilitating the support of a diverse array of use cases and applications.

**Data Security and Privacy:** Ensuring secure and efficient management of the data pipeline while adhering to data governance regulations, such as the General Data Protection Regulation (GDPR) in the European Union or the California Consumer Privacy Act (CCPA) in California, United States, is essential within the three intelligent systems. Without a secure and protected data pipeline, achieving a scalable, reliable, trustworthy, and unified intelligent system for SMO is unattainable. This becomes particularly critical if a centralized intelligent approach is adopted for training a ML algorithm. Furthermore, research is required to develop novel techniques for access control (both internal and external to the SMO framework), to manage data anonymization, to mitigate model inversion attacks, and to address data poisoning attacks customized to SMO. These efforts are essential to mitigate risks and ensure responsible data collection and model training practices. Distributed learning methods, such as federated learning, can be a promising approach that allows training models on distributed datasets without sharing the raw data itself. Integrating federated learning techniques could represent optimal solutions to safeguard data privacy and security during the collection and processing of sensitive management data from various internal and external data sources, as well as during the development and deployment of ML models within SMO.

**Model Development and Deployment Scenarios:** The SMO may support diverse use cases with the assistance of AI/ML models. However, developing and deploying AI/ML models within the Non-RT RIC may not consistently yield the desired performance. Meeting the demands of some use cases and applications might necessitate the partial or full involvement of the Non-RT RIC in AI/ML model training and inference. This implies that the decision of where to train and where to perform inference for the AI/ML model within SMO must be made on a case-by-case basis. For example, in certain scenarios, the AI/ML model may be trained within the Non-RT RIC and then transferred to the 3GPP-NSMS for inference. In other scenarios, both the training and inference processes may be conducted within the 3GPP-NSMS, while the Non-RT RIC orchestrates this and other models from other intelligent systems. Hence, there is a necessity to explore different scenarios for AI/ML model development and deployment within the three intelligent systems of SMO.

**Communication Rounds:** Whether the AI/ML model training paradigm is centralized or distributed, the AI/ML Function plays a central role in the intelligentization and automation of SMO within our proposed unified architectural solution. This means that the 3GPP-NSMS and NFV-MANO can transmit either raw management data or their customized trained AI/ML models to the AI/ML Function to train a centralized model (in the case of centralized learning) or a global model (in the case of federated learning), respectively. Subsequently, the AI/ML Function processes the imported data or models and provides the trained centralized model or global model back to the NFV-MANO and 3GPP-NSMS for inference. The exchange of models and/or data between the intelligent systems within the SMO framework is directly linked to the amount of required computing and communication resources. Considering these aspects, there is a need for an in-depth exploration of novel methods for optimizing the number of communication rounds between the intelligent systems and enhancing the accuracy of the training of desired AI/ML models. There is also a need to discover an optimal solution for balancing the trade-off between computing and communication resource utilization and the number of communication rounds between the intelligent systems. By optimizing local training and raw data processing within the SMO framework, one can indirectly manage resource consumption on devices and network bandwidth usage.

**Resource Constraints and Scalability:** The integration of AI/ML models into SMO within the resource-constrained environment of the network edge presents a significant challenge. O-RAN components residing at the network edge often operate with limited power and processing capabilities, necessitating careful consideration of resource constraints. Deploying complex AI/ML models directly on these components may exceed their processing and power capabilities, thus requiring the development of lightweight models or federated learning approaches optimized for edge deployment. Moreover, the immense volume of data generated within the O-RAN architecture demands efficient storage and transfer mechanisms. Addressing this challenge involves exploring techniques such as data compression, selective aggregation, and in-network processing to mitigate storage and bandwidth limitations at the edge. Research efforts are needed to design and implement solutions that balance the computational demands of AI/ML algorithms with the constrained resources available at the network edge, ensuring efficient utilization and scalability of SMO.

**Explainability and Transparency:** The opacity and complexity inherent in AI/ML models present significant challenges in understanding their decision-making processes, particularly in critical network management and service orchestration tasks within SMO. It is crucial to enhance the explainability, interpretability, and transparency of these models within SMO to facilitate human comprehension, trust, and accountability. Achieving this involves addressing key issues such as model interpretability, feature importance, and decision rationale through the development of explainable AI/ML techniques, visualization tools, and model debugging frameworks. Furthermore, promoting transparency throughout the AI/ML model lifecycle management, including development, deployment, and governance processes, is essential for fostering stakeholder trust, ensuring regulatory compliance, and upholding ethical accountability standards. This necessitates robust mechanisms for documenting and communicating the methodologies, assumptions, and limitations underlying AI/ML models, thereby enabling informed decision-making and mitigating potential risks associated with their use in network and service management.

**Data Quality and Availability:** The quality, relevance, and availability of training data are paramount for effective AI/ML integration within the SMO framework. However, the SMO is expected to generate and collect vast volumes of heterogeneous and noisy data from various management and network components and interfaces. This data ranges from management data (such as configuration data and monitoring data) to network data (such as performance metrics and user behavior patterns). Addressing data quality issues such as missing values, outliers, and data incompleteness requires robust data preprocessing techniques, anomaly detection algorithms, and data augmentation strategies. Moreover, ensuring timely access to high-quality, labeled training data for AI/ML model development within SMO remains a persistent challenge, especially in dynamic network environments characterized by rapid changes and variability. Hence, collaborative efforts among the three intelligent systems are crucial to effectively address these challenges, leveraging multi-SDO and multi-vendor generated data to harness the full potential of ML models within SMO.

**Model Governance and Lifecycle Management:** Ensuring proper AI/ML model governance and lifecycle management is crucial for the automation and intelligentization of SMO. This entails establishing robust procedures and mechanisms for AI/ML model development, deployment, monitoring, refinement, and retirement. Given the critical role these models play in network management, it is essential to address issues related to model versioning, reproducibility, and regulatory compliance. Additionally, managing the lifecycle of AI/ML models involves continuous monitoring for performance degradation, concept drift, and model decay, especially in dynamic network environments. Implementing strategies for model explainability, interpretability, and accountability is also paramount to enhance trust and facilitate human oversight. Furthermore, integrating governance frameworks that adhere to industry standards and regulatory requirements is necessary to mitigate risks associated with biased or unreliable AI/ML models. Finally, establishing transparent and auditable processes for model selection, validation, and validation is essential to ensure the reliability and robustness of AI/ML-driven decision-making within SMO.

**Human-Machine Collaboration:** Promoting effective human-machine collaboration is essential for leveraging the capabilities of AI/ML models within SMO. This involves designing interfaces and interaction paradigms that facilitate seamless communication and collaboration between human operators and SMO-driven systems. Given the complex and dynamic nature of network management tasks, integrating AI/ML models into the decision-making process requires mechanisms for human oversight, intervention, and feedback. Enhancing user interfaces with intuitive visualizations, real-time dashboards, and natural language processing capabilities can empower operators to interpret model outputs, understand underlying insights, and make informed decisions. Moreover, fostering a culture of trust and transparency is vital to encouraging collaboration between humans and machines. This entails providing explanations for AI/ML-driven recommendations, soliciting user feedback, and incorporating human domain knowledge into model training and validation processes. Additionally, developing adaptive learning systems that can dynamically adjust to user preferences and behaviors can further enhance the effectiveness of human-machine collaboration within SMO. By promoting synergy between human expertise and machine intelligence, operators can harness the full potential of AI/ML models to optimize network operations and drive innovation.

## VIII. CONCLUDING REMARKS AND FUTURE OUTLOOK

In conclusion, the analytics-driven approaches and intelligent solutions outlined in this article present a transformative paradigm shift for enhancing SMO capabilities in O-RAN. By seamlessly integrating MDA and intelligence into three modules, unifying their intelligent systems, and streamlining AI/ML model lifecycles, the SMO can transcend its passive management role and evolve into a proactive intelligence engine. The comprehensive nature of these solutions, as detailed across the various sections, underscores their potential to revolutionize the M&O of O-RAN. This transformation will empower O-RAN to achieve unprecedented levels of performance, reliability, and agility, opening up a new era of wireless connectivity characterized by seamless user experiences, enhanced network efficiency, and accelerated innovation. As the industry moves towards increasingly intelligent and automated systems, the proposed approaches pave the way towards an O-RAN where it operates with unparalleled intelligence, agility, and resilience, setting the stage for a truly transformative wireless ecosystem. Although the integration of AI/ML into SMO has not been thoroughly explored, there are several research challenges that need to be addressed. As we outlined these challenges, we anticipate that exploring and offering optimal solutions to such problems will enhance SMO and transform it into a fully AI-native and data-driven M&O platform within O-RAN.

Looking towards the future, two particularly critical research directions stand out as essential for further advancing the proposed intelligence-driven approaches. First, our goal is to extend the proposed approaches by examining use-case scenarios for AI/ML model training, deployment, and monitoring within SMO. The exploration and validation of multiple scenarios will empower operators to choose the most suitable AI/ML model development and deployment scenario for automating and intelligentizing their SMO frameworks. Second, we aim to delve into the deployment of distributed learning, particularly federated learning techniques, within SMO. We are confident that these approaches have the potential to significantly enhance the speed and accuracy of intelligent decision-making while also facilitating scalable and secure AI/ML model training across the management entities of the three modules. By addressing these research directions, we can unlock the transformative potential of intelligence in O-RAN, paving the way for the next generation of intelligent and open wireless networks.

## APPENDIX

To further illustrate the implementation of proposed logical interfaces and functions, which are detailed in Tables 1 and 2, we have provided Swagger API specifications in https://app.swaggerhub.com/apis/girmayilma/O-RAN-SMO-MDAF-ETSI/1.0.0. The API outlines endpoints for managing `NSSMF_NonRTRIC` entries and `NFVO_NonRTRIC` entries, facilitating operations such as creation, retrieval, update, and deletion. Additionally, the API supports message exchange between NSSMF and Non-RT RIC and NFVO and Non-RT RIC. Descriptions of request parameters and response structures are included, covering request identifiers, service types, QoS profiles, security levels, data transformation capabilities, and AI/ML model management. Moreover, the API includes comprehensive endpoints for data or model exchange, enabling the upload, download, and secure transfer of AI/ML models, as well as efficient data transformation and model translation between the integrated systems.

## ACRONYM LIST

| | |
|---|---|
| 3GPP | Third Generation Partnership Project |
| 3GPP-NSMS | 3GPP-network slicing management system |
| 5G | fifth-generation |
| AI | artificial intelligence |
| API | application programming interface |
| CCM | container infrastructure service cluster management |
| CCPA | California Consumer Privacy Act |
| CIR | container image registry |
| CIS | container infrastructure service |
| CISM | container infrastructure service management |
| CNF | cloud-native network function |
| E2E | end-to-end |
| eMBB | enhanced mobile broadband |
| ETSI | European Telecommunications Standards Institute |
| FB | functional block |
| FCAPS | fault, configuration, accounting, performance, security |
| GDPR | General Data Protection Regulation |
| IoT | Internet of Things |
| ISG | Industry Specification Group |
| M&O | management and orchestration |
| MDA | management data analytics |
| MDAF | management data analytics function |
| MDAS | management data analytics service |
| MF | management function |
| ML | machine learning |
| Near-RT RIC | near-real-time RAN intelligence controller |
| NF | network function |
| NFMF | network function management function |
| NFV | Network Function Virtualization |
| NFVI | network functions virtualization infrastructure |
| NFV-MANO | NFV-management and orchestration |
| NFVO | network function virtualization orchestrator |
| NFVO | network function virtualization orchestrator |
| Non-RT RIC | non-real-time RAN intelligence controller |
| NSaaS | network slice as a service |
| NSMF | network slice management function |
| NSSMF | network slice subnet management function |
| O-CU | open centralized unit |
| O-DU | open distributed unit |
| O-gNB | open next generation node B |
| O-RAN | open radio access network |
| O-RU | open radio unit |
| QoS | quality of service |
| SA5 | Service and System Aspects Working Group 5 |
| SDO | standards developing organization |
| SDO | standards development organization |
| SMO | service management and orchestration |
| TN | transport network |
| TSG | Technical Specification Group |
| UE | user equipment |
| V2X | vehicle-to-everything |
| VIM | virtualized infrastructure manager |
| VL | virtual link |
| VNF | virtual network function |
| VNFM | virtual network function manager |
| WIM | wide area network infrastructure manager |
| CLA | closed-loop automation |

## REFERENCES

[1] K. Alam et al., “A comprehensive overview and survey of O-RAN: Exploring slicing-aware architecture, deployment options, and use cases,” 2024. *arXiv:2405.03555*.

[2] B. Han et al., “DMH-HARQ: Reliable and open latency-constrained wireless transport network,” 2024, *arXiv:2212.03602*.

[3] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, ‘A comprehensive survey of RAN architectures toward 5G mobile communication system,” *IEEE Access*, vol. 7, pp. 70371–70421, 2019.

[4] *WG 1; O-RAN Use Cases Analysis Report; Release 03, V13.00*, O-RAN Alliance, Alfter, Germany, Feb. 2024.

[5] M. A. Habibi, G. Yilma, X. Costa-Perez, and H. D. Schotten, “Unifying 3GPP, ETSI, and O-RAN SMO interfaces: Enabling slice subnets interoperability,” in *Proc. IEEE FNWF*, Baltimore, MD, USA, 2023, pp. 1–8.

[6] K. Samdanis, A. N. Abbou, J. Song, and T. Taleb, “AI/ML service enablers and model maintenance for beyond 5G networks,” *IEEE Netw.*, vol. 37, no. 5, pp. 162–172, Sep. 2023.

[7] M. Q. Hamdan et al., “Recent advances in machine learning for network automation in the O-RAN,” *Sensors*, vol. 23, no. 21, p. 8792, 2023.

[8] P. H. Masur, J. H. Reed, and N. K. Tripathi, “Artificial intelligence in open-radio access network,” *IEEE Aerosp. Electron. Syst. Mag.*, vol. 37, no. 9, pp. 6–15, Sep. 2022.

[9] F. Rezazadeh, H. Chergui, L. Alonso, and C. Verikoukis, “SliceOps: Explainable MLOps for streamlined automation-native 6G networks,” *IEEE Wireless Commun.*, early access, Mar. 19, 2024, doi: 10.1109/MWC.007.2300144.

[10] A. Arnaz, J. Lipman, M. Abolhasan, and M. Hiltunen, “Toward integrating intelligence and programmability in open radio access networks: A comprehensive survey,” *IEEE Access*, vol. 10, pp. 67747–67770, 2022.

[11] X. Lin, L. Kundu, C. Dick, and S. Velayutham, “Embracing AI in 5G-advanced toward 6G: A joint 3GPP and O-RAN perspective,” *IEEE Commun. Mag.*, vol. 7, no. 4, pp. 76–83, Dec. 2023.

[12] Q. Sun, N. Li, I. Chih-Lin, J. Huang, X. Xu, and Y. Xie, “Intelligent RAN automation for 5G and beyond,” *IEEE Wireless Commun.*, vol. 31, no. 1, pp. 94–102, Feb. 2024.

[13] H. Lee, J. Cha, D. Kwon, M. Jeong, and I. Park, “Hosting AI/ML workflows on O-RAN RIC platform,” in *Proc. IEEE Globecom Workshops*, 2020, pp. 1–6.

[14] *WG 2; AI/ML Workflow Description and Requirements, V01.03*, O-RAN Alliance, Alfter, Germany, Oct. 2021.

[15] “Network functions virtualisation (NFV) release 4; management and orchestration; report on enabling autonomous management in NFV-MANO,” ETSI, Sophia Antipolis, France, document ETSI GR NFV-IFA 041, Aug. 2021.

[16] “TSG SA5, Technical specification group services and system aspects; management and orchestration; study on enhancement of management data analytics (MDA); (release 17); V17.0.0,” 3GPP, Sophia Antipolis, France, Rep. TR 28.809, Mar. 2021.

[17] B. Tang, V. K. Shah, V. Marojevic, and J. H. Reed, "AI testing framework for next-G O-RAN networks: Requirements, design, and research opportunities," *IEEE Wireless Commun.*, vol. 30, no. 1, pp. 70–77, Feb. 2023.

[18] A. Giannopoulos et al., "Supporting intelligence in disaggregated open radio access networks: Architectural principles, AI/ML workflow, and use cases," *IEEE Access*, vol. 10, pp. 39580–39595, 2022.

[19] B. Brik, K. Boutiba, and A. Ksentini, "Deep learning for B5G open radio access network: Evolution, survey, case studies, and challenges," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 228–250, 2022.

[20] S. D'Oro, L. Bonati, M. Polese, and T. Melodia, "OrchestRAN: Orchestrating network intelligence in the open RAN," *IEEE Trans. Mobile Comput.*, vol. 23, no. 7, pp. 7952–7968, Jul. 2024.

[21] "3GPP TSG SA5; technical specification group services and system aspects; management and orchestration; architecture framework; (Release 18); V18.1.0," 3GPP, Sophia Antipolis, France, Rep. TS 28.533, Mar. 2024.

[22] "3GPP TSG SA5; technical specification group services and system aspects; management and orchestration; concepts, use cases and requirements (release 18) V18.0.0," 3GPP, Sophia Antipolis, France, Rep. TS 28.530, Dec. 2023.

[23] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[24] M. A. Habibi, F. Z. Yousaf, and H. D. Schotten, "Mapping the VNFs and VLs of a RAN slice onto intelligent PoPs in beyond 5G mobile networks," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 670–704, 2022.

[25] "3GPP TSG SA5; technical specification group services and system aspects; management and orchestration; provisioning; (release 18), V18.5.0," 3GPP, Sophia Antipolis, France, Rep. TS 28.531, Mar. 2024.

[26] M.-A. Garcia-Martin, M. Gramaglia, and P. Serrano, "Network automation and data analytics in 3GPP 5G systems," *IEEE Netw.*, vol. 38, no. 4, pp. 182–189, Jul. 2024.

[27] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: How to empower SON with big data for enabling 5G," *IEEE Netw.*, vol. 28, no. 6, pp. 27–33, Nov./Dec. 2014.

[28] E. Pateromichelakis et al., "End-to-end data analytics framework for 5G architecture," *IEEE Access*, vol. 7, pp. 40295–40312, 2019.

[29] R. Ferrús, O. Sallent, and J. Perez-Romero, "Data analytics architectural framework for smarter radio resource management in 5G radio access networks," *IEEE Commun. Mag.*, vol. 58, no. 5, pp. 98–104, May 2020.

[30] "Network functions virtualisation (NFV) release 4; management and orchestration; or-Vnfm reference point—Interface and information model specification," ETSI, Sophia Antipolis, France, document ETSI GS NFV-IFA 007, Mar. 2023.

[31] "Network functions virtualisation (NFV) Release 4; network functions virtualisation (NFV) release 5; architectural framework; report on NFV support for virtualisation of RAN," ETSI, Sophia Antipolis, France, document ETSI GR NFV-IFA 046, May 2023.

[32] "Network functions Virtualisation (NFV); management and orchestration; report on management and orchestration framework," ETSI, Sophia Antipolis, France, document ETSI GR NFV-MAN 001, Dec. 2021.

[33] M. A. Habibi et al., "Toward an open, intelligent, and end-to-end architectural framework for network slicing in 6G communication systems," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1615–1658, 2023.

[34] "Network functions Virtualisation (NFV) release 4; management and orchestration; management data analytics service interface and information model specification," ETSI, Sophia Antipolis, France, document ETSI GR NFV-IFA 047, Mar. 2023.

[35] "ETSI NFV release 5 kicks off with increased support for cloud-enabled deployment," 2021. [Online]. Available: https://www.etsi.org/newsroom/press-releases/1992-2021-11-etsi-nfv-release-5-kicks-off-with-increased-support-for-cloud-enabled-deployments

[36] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: A big data—AI integration perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1328–1347, Apr. 2021.

[37] M. Liu, C. Zhao, X. Peng, S. Yu, H. Wang, and C. Sha,, "Task-oriented ML/DL library recommendation based on a knowledge graph," *IEEE Trans. Softw. Eng.*, vol. 49, no. 8, pp. 4081–4096, Aug. 2023.

[38] C. Chai, J. Wang, Y. Luo, Z. Niu, and G. Li, "Data management for machine learning: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 4646–4667, May 2023.

**MOHAMMAD ASIF HABIBI** received the B.Sc. degree in telecommunications engineering from Kabul University, Afghanistan, in 2011, and the M.Sc. degree in systems engineering and informatics from the Czech University of Life Sciences, Czech Republic, in 2016. He is currently pursuing the Ph.D. degree with the Division of Wireless Communications and Radio Positioning, University of Kaiserslautern (RPTU), Germany, where he has been working as a Research Fellow since January 2017. From 2011 to 2014, he worked as a Radio Access Network Engineer with Huawei. His main research interests include network management, network slicing, machine learning, and network architecture.

**GIRMA MAMUYE YILMA** received the B.Sc. degree in electrical and computer engineering from Addis Ababa University, Ethiopia, in 2010, and the M.Sc. degree in telecommunications engineering from the University of Trento, Italy, in 2016. He is currently pursuing the Ph.D. degree with the Universitat Politècnica de Catalunya and also works as a Senior Research Engineer with NEC Laboratories Europe. In 2014, he was an Intern with the Advanced Digital Science Laboratories in Illinois, Singapore. As part of his work, he mainly focuses on cloud-native computing and networking, O-RAN security, and telco-clouds in the context of next-generation mobile networks. He is also investigating generative AI for telecommunication services.

**UMBERTO FATTORE** (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in computer engineering from the University of Naples "Federico II," Italy, in 2018. He is currently pursuing the Ph.D. degree with the Universidad Carlos III de Madrid, Spain. Since then, he works with NEC Laboratories Europe GmbH, first as an Early Stage Researcher for the EU SPOTLIGHT Project, focusing on 5G mobile core flow optimization, then as a Researcher for the EU 5G-CARMEN Project or cross-border seamless 5G connectivity. Since 2021, he has been a Standardization Engineer with NEC, mainly involved in the recently established IOWN Global Forum and in the ETSI NFV ISG, but also following several other groups' activities in ETSI, 3GPP, and IETF.

**XAVIER COSTA-PÉREZ** (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications from the Polytechnic University of Catalonia, Barcelona. He is the Head of 6G Networks Research and Development with NEC Laboratories Europe, the Scientific Director of the i2Cat Research and Development Center, and a Research Professor with ICREA. His team contributes to products roadmap evolution as well as to European Commission research and development projects and received several awards for successful technology transfers. In addition, the team contributes to related standardization bodies: 3GPP, ETSI NFV, ETSI MEC, and IETF. He has been a 5GPPP Technology Board Member, served on the Program Committee of several conferences (including IEEE Greencom, WCNC, and INFOCOM), published at top research venues, and holds several patents. He was the recipient of a national award for his Ph.D. thesis. He also serves as an Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING and IEEE TRANSACTIONS ON COMMUNICATIONS.

**HANS D. SCHOTTEN** (Member, IEEE) received the Ph.D. degree from the RWTH Aachen University of Technology, Germany, in 1997. From 1999 to 2003, he worked with Ericsson. From 2003 to 2007, he worked with Qualcomm. He became a Manager of the Research and Development Group, the Research Coordinator with Qualcomm Europe, and the Director for Technical Standards. In 2007, he became a Full Professor with the University of Kaiserslautern (RPTU). In 2012, he became the Scientific Director of the German Research Center for Artificial Intelligence (DFKI) and the Head of the Department for Intelligent Networks. He served as the Dean of the Department of Electrical Engineering, RPTU, from 2013 to 2017. Since 2018, he has been the Chairman of the German Society for Information Technology and a member of the Supervisory Board of the VDE.