Neural Computing and Applications
https://doi.org/10.1007/s00521-021-06206-0

S.I. : DEEP NEURO-FUZZY ANALYTICS IN SMART ECOSYSTEMS

Dynamic traffic forecasting and fuzzy-based optimized admission control in federated 5G-open RAN networks

Abida Perveen[1] · Raouf Abozariba[1] · Mohammad Patwary[2] · Adel Aneiba[1]

Received: 20 October 2020 / Accepted: 8 June 2021

Abstract
Providing connectivity to high-density traffic demand is one of the key promises of future wireless networks. The open radio access network (O-RAN) is one of the critical drivers ensuring such connectivity in heterogeneous networks. Despite intense interest from researchers in this domain, key challenges remain to ensure efficient network resource allocation and utilization. This paper proposes a dynamic traffic forecasting scheme to predict future traffic demand in federated O-RAN. Utilizing information on user demand and network capacity, we propose a fully reconfigurable admission control framework via fuzzy-logic optimization. We also perform detailed analysis on several parameters (user satisfaction level, utilization gain, and fairness) over benchmarks from various papers. The results show that the proposed forecasting and fuzzy-logic-based admission control framework significantly enhances fairness and provides guaranteed quality of experience without sacrificing resource utilization. Moreover, we have proven that the proposed framework can accommodate a large number of devices connected simultaneously in the federated O-RAN.

Keywords Open radio access network (O-RAN) · Traffic forecasting · Fuzzy-logic · Admission control · Resource allocation · Quality of experience (QoE)

1 Introduction

The adaptive nature of fifth-generation (5G) wireless network architecture offers a significant opportunity to enhance system capacity and provide more efficient radio resource utilization. Adaptivity is achieved partly by recent advances in network function virtualization, network slicing, and the coexistence of multiple radio access technologies (RAT) [1]. One of the principal incentives behind redesigning cellular networks is to serve a plethora of devices with different requirements. There are numerous techniques for enhancing system capacity (e.g., the use of ultra-dense small cell distribution, millimeter waves (mmWave), new radio (NR), and intelligent cognitive radio) [2–7]. However, there is little emphasis in the literature on the interpolation of current standards with existing ones to guarantee a seamless multi-operator orchestration. Such a seamless ecosystem is required to provide enhanced quality of experience (QoE) and efficient radio resource utilization [8, 9].

The coexistence of various heterogeneous wireless networks provides better performance by accumulating system capacity, supporting higher data rates, and reducing latency and packet loss. From the implementation perspective, operators often use existing network infrastructure to serve traditional voice and Web browsing applications as they offer satisfactory service performance. Also, network operators expect that whenever users are outside 5G coverage, available legacy RATs are needed to provide a seamless service to end users [10, 11]. Such coexistence is becoming the dominant feature of the current and next

✉ Abida Perveen
abida.perveen@mail.bcu.ac.uk

Raouf Abozariba
raouf.abozariba@bcu.ac.uk

Mohammad Patwary
patwary@wlv.ac.uk

Adel Aneiba
adel.aneiba@bcu.ac.uk

[1] School of Computing and Digital Technology, Birmingham City University, Birmingham, UK

[2] School of Mathematics and Computer Science, University of Wolverhampton, Wolverhampton, UK

Published online: 18 June 2021

generation of cellular networking. In this regard, the novel concept of the open radio access network (O-RAN) could be considered a complementary option to the new 5G RATs [12, 13].

Network selection using the tenant-based approach is a common technique to provide heterogeneous RAT services [1, 14–17]. Selecting the best RAT from an available set adds latency, leading to network congestion, particularly when the number of devices requesting access to the network is extensive. Furthermore, inefficient admission control saturates the network, impractical for providing real-time services with stringent QoE demand [6, 18]. Furthermore, lack of optimal network resource utilization and fairness impact the network's QoS. In such scenarios, the choice of technique for selecting the best access network from multiple heterogeneous RATs by the user remains an open problem. We propose a forecasting and admission control (FAC) framework to support the presented federated O-RAN. This framework builds on dynamic traffic demand and augmented by particle filtering, followed by a network selection method using NSGA-II fuzzy-logic optimization to address the challenges mentioned above.

This work focuses on an efficient tenant-aware network configuration in the federated O-RAN, an extended version of the well-known O-RAN architecture, to achieve this autonomously. A federation controller is added to the O-RAN architecture and acts as a switch to select the optimal network based on various networks and user statistics, including network bandwidth, required data rate, latency sensitivity of the requested service, packet loss, priority, and signal strength. The contributions in this paper are (1) a novel federation framework for tenant-aware network configuration which features a dynamic demand-estimation scheme embedded with fuzzy-logic-based optimization for the optimal network selection, and (2) two algorithms in which we establish a multivariate service allocation priority factor and admission queue and build a service profile used for service monitoring. (3) We analytically compare to several state-of-the-art methods for admission control and resource allocation schemes and show how FAC is more efficient and provides higher QoE.

The remainder of the paper is organized as follows. Related work in correspondence with the proposed work is described in detail and also summarized in Table 1 in Sect. 2. The system model considered in the scope of the proposed work, its design, and statistics are explained in detail along with a summary of key symbols used throughout this paper are listed in Table 2 in Sect. 3. Section 4 presents the proposed forecasting and admission control scheme and its respective algorithms. Section 5 introduces the evaluation schemes. We showed our results with a detailed comparison in Table 3 in Sect. 6. The conclusion is presented in Sect. 7.

2 Related work

5G promises on-demand service deployment, support of service heterogeneity, and coordination of multiple access network technologies. Over the last few years, multiple research teams initiated work on the transformation of RANs by integrating various technologies to provide more agile services to end-users. For example, 3GPP TR 38.804 Release 14 describes dual connectivity between 4G LTE and 5G NR. Work by 3GPP in TR 37.900 Release 15 explains multi-RAT deployment architecture. It also identifies relevant scenarios and their radio frequency requirements for implementing the Multi-Standard Radio (MSR) Base Station [9, 15–17, 19]. More recently, open RAN (O-RAN) (aka V-RAN), a novel architecture with interoperability of various networks at its core principle, has been proposed. O-RAN is an emerging technology that incorporates virtualization and intelligence in networks. One such example is the OpenRAN project by Telecom Infra, a software-driven architecture that evolved from the concept of Cloud RAN (C-RAN). It provides a solution based on software-defined networks and openness of general-purpose hardware [20]. Another individual effort is the xRAN forum, an open-source alternative to the traditional RAN architecture, which provides a solution by separating data and control planes of network devices and opening interfaces with intelligence between various RAN's building blocks [21].

In 2018, the C-RAN Alliance and the xRAN forum merged into an open radio access network (O-RAN) to support the evolution of 5G networks and beyond. More than 160 well-known contributors from large and small companies, vendors, academic institutions, and network operators are participating in the standardization of this technology (e.g., Nokia, Hewlett Packard Enterprise, Intel, Vodafone [12]). This multi-vendor, interoperable technology eliminates dependencies and opens the protocols and interfaces between various components to incorporate intelligence into RAN to support different deployment scenarios [13, 22, 23]. This paper continues the trend by introducing a federation layer within an O-RAN architecture to enable dynamic traffic forecasting, efficient admission control, and service monitoring.

In addition to the coexistence of various access technologies, future demand prediction (via efficient forecasting techniques) is among the operators' main challenges. Network operators strive to make resource management and orchestration (MANO) processes highly automated to cope with the volatile demand. To realize this,

Table 1 Table on existing research

Subject	Authors and publications	Description
Support of heterogeneous connectivity	3GPP TR 23.234 (R-12), 38.804 (R-14), and 37.900 (R-15) [9, 17, 19]	The 3rd Generation Partnership Project (3GPP) worked on interworking of cellular network and WLAN, dual connectivity of cellular user with 4G LTE and 5G NR, and Multi-RAT deployment architecture that can be found in releases 12, 14, and 15
	Telecom Infra [20], xRAN forum [21], O-RAN Alliance [12], Open RAN technical report [22, 24, 25] Gavrilovska et al. [13] and Niknam et al. [23]	The open radio access network is a multi-vendor, interoperable product. This intelligently opens the protocols and interfaces between various RAN components to integrate various operators' networks and supports different deployment scenarios with lower cost and time to market
Demand forecasting	Sciancalepore et al. [26, 27], and Raikwar et al. [28]	The authors implemented Holt-Winters theory for long- and short-term traffic forecasting to ensure efficient admission control and resource management in cellular networks
	Tseliou et al. [1], Dudek et al. [29, 30], and Hippert et al. [31]	Monte Carlo-based prediction frameworks are proposed by the authors for on-demand resource allocation in cellular and neural networks
	Narmanlioglu et al. [32], Miao et al. [33], and Zhang et al. [34]	Bayesian techniques are adopted by the authors to predict the number of active users and their distribution within the cellular network for localization and resource allocation over handover
Fuzzy-logic-based network selection and resources allocation	Inaba et al. [35], Bouali et al. [18], Goudarzi et al. [36], and Kaloxylos et al. [37]	The authors implemented the fuzzy-logic approach in their proposed hybrid model for an efficient access network selection among heterogeneous networks
	Khan et al. [38], Zeng et al. [39], silva et al. [40], and Shrimali et al. [41]	The authors adopted fuzzy-logic and multi-criterion optimization schemes, or algorithms such as a genetic algorithm, to propose their framework for resource allocation in 5G cellular and vehicular networks

Table 2 Key symbols and definitions

Symbols & Definitions	
$\mathcal{U}$	Set of tenants in the network
$\mathcal{M}$	Set of MNOs
$\mathcal{V}$	Set of MVNOs
$\mathcal{S}$	Set of services
$\mathcal{S}\_Op$	List of service operators on $\mathcal{S}$
$\mathcal{N}$	Set of resources
τ	Tenant's forecasted demand
$d_{(n)}$	Aggregate nth resource demand
$R_{(n)}$	Aggregate nth resource allocation
$\mathcal{R}\_Op$	Service operator's available resources
τ_γ, τ_h	Acceptable tenant resource bounds
$\mathcal{Q}_{(\gamma)}, \mathcal{Q}_{(h)}$	Expected tenant QoE bounds
$B_{(\gamma)}, B_{(h)}$	Service network resource bounds
$\mathcal{S}_{\mathcal{Q}_\gamma}, \mathcal{S}_{\mathcal{Q}_h}$	Network's guaranteed QoS bounds

Sciancalepore et al. proposed a Holt-Winters theory-based mobile traffic forecasting and slice scheduling approach for admission control in 5G networks [26]. Two low-complexity algorithms were developed into a geometric knapsack problem to ensure optimal slice admission and better QoE. The authors further proposed enhancements to their work for traffic and user mobility analysis on guaranteed and best-effort traffic. The signaling-based network slicing broker is used for capacity forecasting of the cellular network [27]. Raikwar et al. proposed the concept of using the Holt-Winters method for predicting vehicular traffic demand in long- and short-term traffic windows for an efficient transportation management system [28]. Tseliou et al. proposed a multi-tenant slicing capacity framework for on-demand resource allocation in LTE networks. The authors integrated the capacity broker into the 3GPP architecture for extracting variations in traffic patterns. The technique improves traffic forecasting based on the Monte Carlo method [1]. For short-term forecasting, a few other network models using the Monte Carlo approach can be found in [29–31]. Narmanlioglu et al. investigated a Bayesian technique for predicting active users in the LTE network to ensure efficient resource allocation [32]. Miao et al. proposed a multi-spatiotemporal framework for

Table 3 Summary of comparisons of average efficiency between the proposed work and existing methods

Evaluation Parameters	Approaches	Efficiency
User satisfaction level	Forecasting model and availability of multiple heterogeneous networks (O-RAN) ensure optimal resource allocation to the tenants in the proposed work	Average efficiency 93% from $U = 50$ to $U = 300$
	Online auction on available resources and greedy approaches are applied for resource allocation in [48]	Average efficiency approximately 73%, and 43% from $U = 50$ to $U = 300$
Network resources utilization	In the proposed work, forecasting modifier ($\mathcal{P}\epsilon$) and multi-variate priority factor (φ) ensures optimum admission control	Approximately 90% average efficiency from $U = 5$ to $U = 20$
	Mobile traffic forecasting [26] and reinforcement learning [27] approaches for tenants admission control are applied	Average efficiency approximately 27%, and 59% from $U = 5$ to $U = 20$
Resources allocation fairness	In the proposed work, fuzzy-logic-based network selection, QoE-based admission control and resource allocation approaches are applied	Average efficiency approximately 99.5% from $U = 315$ and $U = 330$
	Bankruptcy game approach is applied for admission control and resource allocation in [49]	Average efficiency approximately 99% from $U = 315$ and $U = 330$

forecasting cellular user traffic [33]. Another forecasting framework based on the Bayesian model and Markov chain Monte Carlo (MCMC) techniques is proposed in [34] to predict the spatiotemporal information of traffic distribution in a cellular network. Holt-Winters and Bayesian are basic exponential smoothing techniques. These techniques are simple, yet work well over short time series in a linear system using prior assumptions about the user. Monte Carlo-based forecasting techniques make predictions according to data from previous instances [42]. However, these approaches are not suitable for forecasting demand in dense networks with limited or no prior information about user demand and network capacity. In this paper, we implement a hybrid Monte Carlo-based particle filtering technique to predict traffic demand. The hybrid Monte Carlo-based particle filtering technology has proved its superiority over other approaches, due to its non-dependency on previous data samples in nonlinear and non-Gaussian systems, and its multimodal processing capability makes it suitable for a wide range of communication applications.

Fuzzy-logic optimization proved to be among the most effective approaches in coping with the lack of information and associated uncertainties surrounding users. A significant amount of research is available on fuzzy-logic optimization. Bouali et al. proposed a novel, context-aware, user-driven framework for network selection. The authors implemented fuzzy multiple-attribute decision-making (MADM) approach to select the best RAT of the network's defined policies [18]. Goudarzi et al. proposed a hybrid model that uses a multi-point algorithm in heterogeneous networks for the most suitable RAT selection. This scheme implements biogeography-based optimization (BBO) on RAT selection probabilities obtained from a Markov decision process (MDP) [36]. Kaloxylos et al. applied a fuzzy-logic approach for efficient RAT selection mechanism between (H)eNBs and Wi-Fi APs, addressing static and low-mobility users [37]. The authors of [35] proposed a fuzzy call admission control scheme for wireless multimedia networks. Khan et al. [38] proposed a hybrid fuzzy-logic-based genetic algorithm (H-FLGA) for resource allocation in 5G-driven VANETs. Zeng et al. [39] proposed a fuzzy-logic-based multi-criterion scheme to investigate a coordinated NOMA system for resource allocation in 5G. Their proposed resource allocation algorithms serve channel-gain-based subchannel allocation (SCG-SA), low-complexity, fuzzy-logic user-ranking-order-based joint resource allocation (FLURO-JRA). In [40], Silva et al. proposed a self-tuning approach based on fuzzy-logic for resource allocation during handover in small, dense cells. This approach compared the received signal level with a proposed fuzzy-logic-based threshold derived from the user's velocity, signal power, and channel quality. This way, the number of handovers was reduced, and a lower failed handover ratio was achieved. Shrimali et al. [41] proposed a weighted-sum-based, multi-objective optimization framework for resource allocation in cloud infrastructure. This framework applied the fuzzy-logic approach to generate coefficients of the defined objectives. These coefficients used by the genetic algorithm to generate Pareto optimal solutions. The existing research is considering few application requirements or network statistics for best network selection among two heterogeneous RATs and resource allocation. However, today's network is more complex and dynamic due to multiple and heterogeneous application requirements and network statistics, and especially in the case of the coexistence of various heterogeneous RATs such as in O-RAN. In such scenarios, Information about uncertainty on application requirements is essential for fuzzy-logic operations. This is

because higher uncertainty leads to inefficient network selection, overload, and agreed QoE degradation [43]. The proposed framework applies the Non-Dominated Sorting Genetic Algorithm II, coupled with the fuzzy-logic approach to provide an optimal network selection based on user forecasted demand, network capacity and fitness policies of the fuzzy-logic model to reduce the effects of uncertainty (see Fig. 3).

3 System design

The management of resource allocation to support a massive amount of heterogeneous traffic flow in proportion to network capacity is still an open issue [44]. In this section, a system model, known as federated O-RAN (FORAN), is presented, which is similar to the O-RAN architecture, as given in [12]. A federation controller is added to the O-RAN referenced architecture, which acts as a switch to select the optimal network based on various network and user demand statistics as discussed in the detail in the following subsections.

3.1 Federation logical architecture with O-RAN

O-RAN is a unified architecture designed with openness and intelligence and built by disaggregating three main components of the traditional RAN: Radio Unit (RU), Distributed Unit (DU), and Centralized Control Unit (CU), via intelligently decoupling the virtualized software and hardware functionalities [24, 25]. Establishing open-standard protocols and interfaces between hardware and software eliminates vendor dependency on conventional networks. O-RAN facilitates a wide range of services by transforming existing business models into a new paradigm or launching new business models with a shorter time to market and lower cost [12].

O-RAN consists of four functional building blocks: (1) *Orchestration and Automation*, (2) *RAN Intelligent Controller (RIC) near-real Time*, (3) *Multi-RAT Control Unit protocol stack*, and (4) *Distributed Unit* and *Remote Radio Unit* (RRU) [24], as shown in Fig. 1. Contrary to the traditional RAN, the O-RAN non-real-time controller (with latency $>$ 1s) and near-real-time (near-RT) controller (with latency $<$ 1s) are decoupled from each other and placed as an isolated layer, connected via A1 interface. Orchestration and automation function is responsible for non-real-time services, network design, configuration, and policy management. The function also analyzes the RAN traffic and models the training data for run-time executions by RIC near-real-time functionality. It also contains the RAN database, the trained model, and an intelligent radio resource management unit, which provides a robust and reliable execution platform for third-party applications.

The proposed framework's main purpose is to utilize network resources more efficiently via optimal network selection and users' admission. In this framework, the federation layer has three major functions managed by the federation controller: (1) *Demand and Capacity Analyzer* (DCA), (2) *Network Selection and Configuration Function* (NSCF), and (3) *QoS/QoE and Traffic Flow Monitoring* (QTFM). The DCA analyzes incoming traffic demand in order to generate forecasts. It also analyzes the available network capacity and resources for network selection and resource allocation. The DCA contains the mobile virtual network operator (MVNO) resource inventory: its services, content, and billing information. Given the network availability and forecasted demand, the NSCF selects an optimal network for the user at a particular instance via fuzzy-logic optimization for configuration and resource allocation. Network selection is based on the suitability factor, as defined by the policies on multiple decision objectives, to ensure the network QoS and user QoE continue to meet the agreed level. We then send the service requests to the selected MVNO, choose a gateway, and establish an end-to-end connection for management and data transmission. Next, we continuously monitor the admitted traffic to ensure efficient resource utilization and that the granted QoE are within guaranteed bounds. If the user's QoE degrades and resources are over/underutilized, the analyzer will be triggered by the QTFM to modify the predictions after observing the difference between forecasted and actual demand.

A multi-RAT CU protocol stack is installed on the virtualization platform to process the heterogeneous wireless generation protocols. DU and RRU functions are responsible for baseband and radio frequency (RF) processing. These functional units are linked to the O-RAN via the E2 interface [24]. This novel, vendor-neutral architecture has a huge potential for virtual industries to quickly deploy and upgrade their network architecture in various deployment scenarios and geographies.

3.2 E2E customized network configuration via FORAN

Researchers and industry professionals envisage that the O-RAN will be an essential part of wireless networks because compared to traditional RAN, the neutral architecture of the O-RAN can efficiently accommodate the rising heterogeneous services demand [22]. Opening protocols and interfaces among various building blocks of the access network reduce the operator's and vendor's dependency on conventional network deployment and operations. Therefore, the coexistence of various heterogeneous

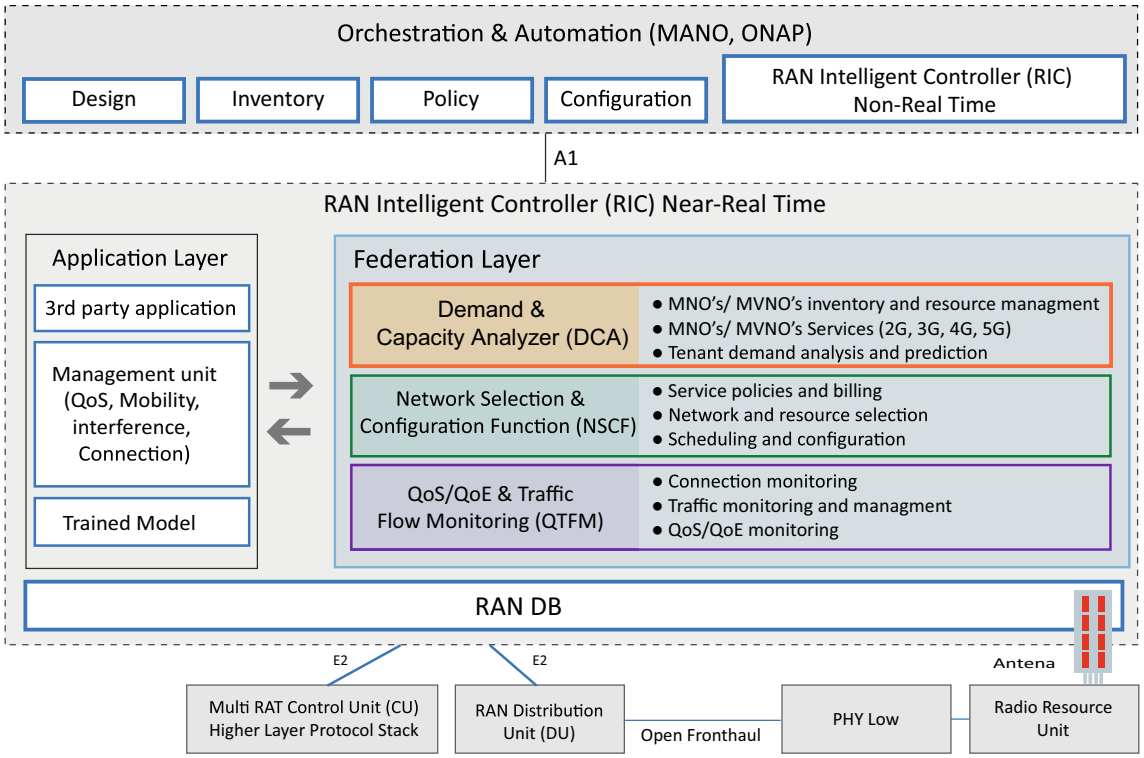

Fig. 1 FORAN architecture illustrating network operations and management functions

technologies in O-RAN will facilitate an automated vendor network that reduces operational cost and enhances network performance [13].

The device sends a control signal to the O-RAN for an E2E customized network configuration when it is turned on, as shown in Fig. 2. The `Service Request` and `Registration Request` in the control signal are received by the CU of the respective virtual access node (i.e., *virtual Base Transceiver Station* (vBTS), *virtual Node B* (vNB), *virtual evolved Node B* (veNB), *virtual Next Generation Node B* (vgNB)). The CU sends the control signal to the unified real-time or non-real-time controller based on the requested service sensitivity. The virtual access node requests user information from the corresponding *Home Location Register* (HLR) and *Unified Data Management* (UDM) in the core network, which then sends back the subscription user data and confirms if the tenant is authorized for service from the network. If authorization and authentication are successful, the user's request will be sent to the federation controller in the real-time O-RAN controller for admission control and resource allocation by the NSCF. However, the user's admission will go through

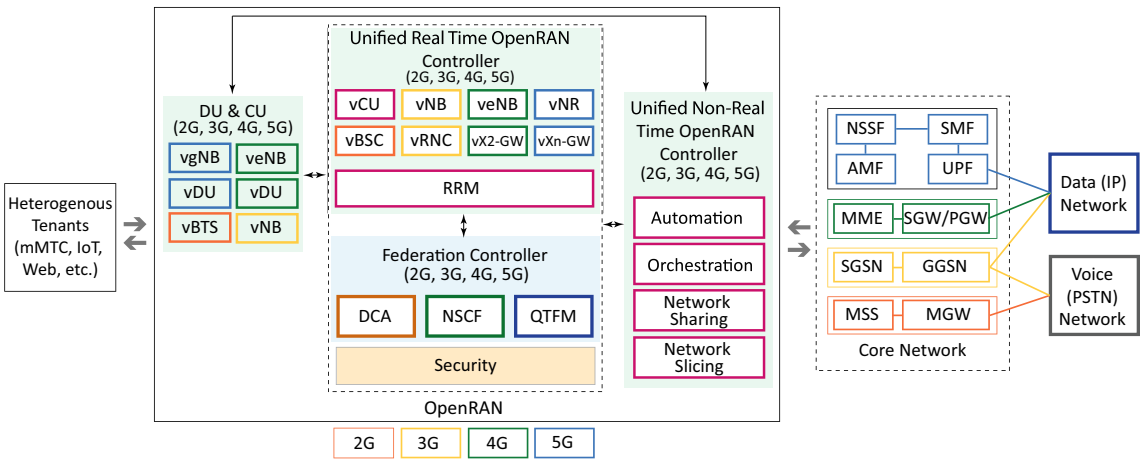

Fig. 2 FORAN architecture illustrating the network functions for the customized network configuration

the conventional process in the non-real-time O-RAN controller. Based on the user's preferences and demanded service network statistics, the NSCF selects an optimal network from the service operator list for service provisioning. The optimal networks in the list are in order based on the forecasts generated by the DCA. After that, the federation controller sends the service request and the session ID to the corresponding network functions in the core to perform a customized network configuration. An example of such functions is *Mobile Switching Station* (MSS), *Serving GPRS Support Node* (SGSN), *Mobility Management Entity* (MME), and *Access and Mobility Management Function* (AMF). The ID contains the network function instance address, where the NAS message terminates [45]. Finally, the request forwards to the respective core entities (i.e., *Media Gateway* (MGW), *Gateway GPRS Support Node* (GGSN), *Serving Gateway/ Packet Gateway* (SGW/PGW), *Session Management Function* (SMF), *user plane function* (UPF)) for gateway selection, E2E session establishment for data or voice communication via DU, and service management and monitoring via QTFM.

3.3 Network description

In the proposed work, a cellular network is considered with a set $\mathcal{U}$ consisting of U number of tenants, indexed by $\mathcal{U} = \{1, 2, \ldots, U\}$, and a set $\mathcal{M}$ of M number of mobile network operators (MNOs), indexed by $\mathcal{M} = \{1, 2, \ldots, M\}$, We assume that each MNO supports a set $\mathcal{V} = \{1, 2, \ldots, V\}$ of V mobile virtual network operators (MVNOs), each of which has a set $\mathcal{N} = \{1, 2, \ldots, N\}$ of N equal number of similar resources. Let us assume that each v, where $v \in \mathcal{V}$ is independent and different from the other v associated with the same MNO with respect to the capacity of the resources, guaranteed QoS, and billing. This information is stored in the inventory matrix in the repository of DCA by the federation controller and symbolized as $\mathbf{P}$,

$$\mathbf{P} = \begin{bmatrix} \mathbf{v}_{11} & \mathbf{v}_{12} & \mathbf{v}_{13} & \cdots & \mathbf{v}_{1V} \\ \mathbf{v}_{21} & \mathbf{v}_{22} & \mathbf{v}_{23} & \cdots & \mathbf{v}_{2V} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{v}_{M1} & \mathbf{v}_{M2} & \mathbf{v}_{M3} & \cdots & \mathbf{v}_{MV} \end{bmatrix}, \tag{1}$$

where the resource vector $\mathbf{v}_{ij\,(1\times N)}$, $i \in \mathcal{M}$ and $j \in \mathcal{V}$. Each operator provides some services, as represented by the service set $\mathcal{S} = \{g_2, g_3, g_4, g_5\}$ from 2G, 3G, 4G, and 5G. These services represent a row entry in the mask matrix $\mathbf{G}$ and

$$\mathbf{G} = \begin{bmatrix} g_2\mathbf{v}_{11} & g_2\mathbf{v}_{12} & g_2\mathbf{v}_{13} & \cdots & g_2\mathbf{v}_{MV} \\ g_3\mathbf{v}_{11} & g_3\mathbf{v}_{12} & g_3\mathbf{v}_{13} & \cdots & g_3\mathbf{v}_{MV} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ g_5\mathbf{v}_{11} & g_5\mathbf{v}_{12} & g_5\mathbf{v}_{13} & \cdots & g_5\mathbf{v}_{MV} \end{bmatrix}. \tag{2}$$

If the operator supports that particular service g_ι, where $g_\iota \in \mathcal{S}$ and $\iota = \{2, 3, 4, 5\}$, then the entity is represented by 1, otherwise by zero in the mask $\mathbf{G}$. Operators supporting 5G have an additional feature, "slicing," for latency-sensitive or critical applications from the tenant or for tenants with frequently varying demand. Each 5G slicing operator is assumed to have S number of slices with homogeneous and heterogeneous, capacity. In that case, $\mathbf{v}_{ij}$ is $S \times N$ dimension matrix. This is because each slice has the potential of elasticity with regard to the capacity of the network to support a massive number of connection requests. The DCA holds this matrix for the provisioning of services to the tenants over forecasting and optimal admission control.

4 Proposed forecasting and admission control framework

In this section, we propose a forecasting and admission control framework. This framework applies the aforementioned sampling-based forecasting technique to obtain optimal network selection by the key entities of the federation controller. The federation controller consists of three key entities: the Demand and Capacity Analyzer (DCA), the Network selection and configuration function (NSCF), and QoS/QoE and Traffic Flow management (QTFM). The systematic diagram of the proposed *Forecasting and Admission Control* (FAC) framework processed by these entities is illustrated in Fig. 3 and discussed in detail in the following subsections.

4.1 Forecasting and demand characterization

Traffic analysis is an essential part of traffic engineering and is the fastest method to gain insights into the nature of future service demand [33]. More accurate forecasting leads to maximizing the network QoS, and resource utilization, as well as QoE. In the proposed work, *SequentialMonteCarlo* (SMC)-based particle filtering is used for future demand forecasting in the wireless network. Through this, the DCA isolates tenants with their specific demand from each other based on the service requirement from the heterogeneous network and then observes the actual demand using the particle filter for future demand forecasting, as shown in Fig 3.

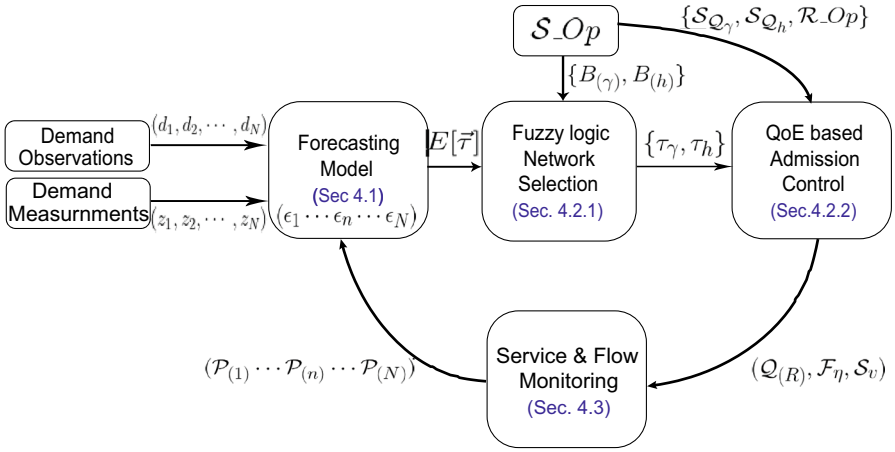

Fig. 3 Systematic diagram of the proposed forecasting and admission control (FAC) framework

During observation, whenever the uth tenant requires access to the network for provisioning of its service s, where $u \in \mathcal{U}$ and $s \in \mathcal{S}$, it issues a request denoted as $\mathbf{d}_{un} = [d_{u1}, d_{u2}, \ldots, d_{un}]$, where $\mathbf{d}_{un} \in \mathcal{D}$. The vector $\mathbf{d}_{un}$ consists of tenant-demand-specific characteristics such as physical resources, latency, service holding time, priority, and cost revenue. The federation controller has to assess that request with respect to tenant-specific application characteristics and populate the respective coefficient as a row entry in the demand matrix $\mathcal{D}$, as illustrated in Eq. (3).

$$\mathcal{D} = \begin{bmatrix} d_{11} & d_{12} & d_{13} & \cdots & d_{1N} \\ d_{21} & d_{22} & d_{23} & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ d_{U1} & d_{U2} & d_{U3} & \cdots & d_{UN} \end{bmatrix}. \tag{3}$$

After resources are allocated to the tenant from a particular network, the respective coefficients are populated as a row entry in the allocation matrix $\mathcal{R}$. Thus, after resources are allocated to the tenant u for the service s, the vector of the allocated resources can be indexed by $\mathbf{r}_{un} = [r_{u1}, r_{u2}, \ldots, r_{un}]$, where $\mathbf{r}_{un} \in \mathcal{R}$ and

$$\mathcal{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1N} \\ r_{21} & r_{22} & r_{23} & \cdots & r_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_{U1} & r_{U2} & r_{U3} & \cdots & r_{UN} \end{bmatrix}. \tag{4}$$

Initially, due to periodicity, the uth tenants nth resource demand, symbolized as $d_{(n)}$, have to be analyzed using a probability density function (PDF) over the observation time window, denoted by t_{ow} where $t_{ow} = t - t_{\text{obs}}$, as shown in the equation below:

$$d_{(n)} = \int_{t-t_{\text{obs}}}^{t} f(d_{n,t})\, d(d_{n,t}), \tag{5}$$

where $d_{n,t} \in \mathcal{D}$, and $d_{(n)}$ is known by the ground truth demand shown in Fig. 4. This is obtained from the initial prior Gaussian distribution function. Now, the measured demand, symbolized as $z_{(n)}$, over the ground truth of the previous interval, along with the measurement noise covariance, Υ, in the system, can be obtained from

$$z_{(n)} = \frac{d_{(n,t-1)}}{20} + \Upsilon. \tag{6}$$

In view of the network capacity and required services, the observation of the resource allocation, denoted as R, to the tenants over observation time window (from t to t_{obs}) with PDF can be written as

$$R_{(n)} = \int_{t-t_{\text{obs}}}^{t} f(r_{n,t})\, d(r_{n,t}), \tag{7}$$

where $r_{n,t} \in \mathcal{R}$. $R_{(n)}$ is the observed nth resource allocation and $R_{(n)} \leq d_{(n)}$. Therefore, the initially acquired tenant QoE, symbolized as $\mathcal{Q}_{(n)}$ for nth resource, can be obtained from

$$\mathcal{Q}_{(n)} = \frac{R_{(n)}}{d_{(n)}} \leq 1. \tag{8}$$

The forecasting model $f(\cdot)$ uses the particle filter estimates, $\hat{\tau}$, to forecast tenant demand, symbolized as τ, over $d_{(n)}$. Hence, the overall estimates over the forecasting window t_w ($t+1$ to $t+f$), as shown in Fig. 4, are obtained from

$$E[\tau_{(n)}] = \int_{t+1}^{t+f} f(\hat{\tau}_{n,t})\, d\hat{\tau}_{n,t}, \tag{9}$$

where

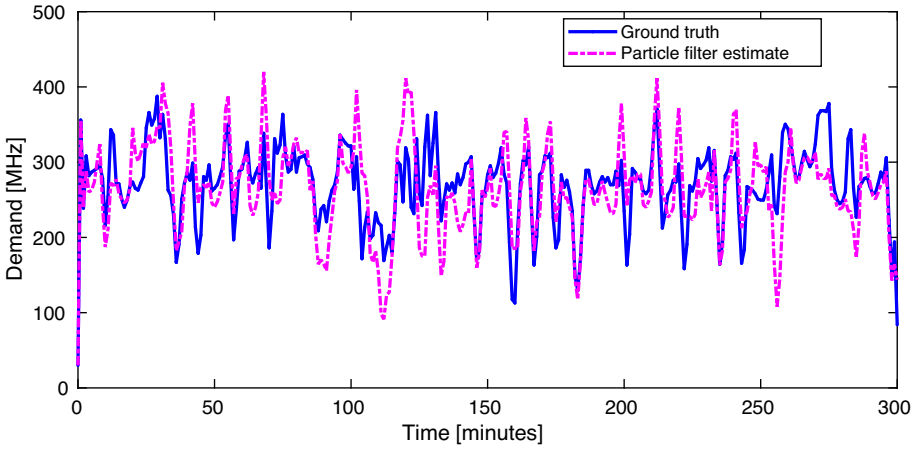

Fig. 4 Long-term resource demand forecasting; $U = 1$ with $E[u_i] = 100$ users, $|\mathcal{N}| = 1$. $t_w = 300$ minutes, and $\tau_h = [10, 100]$ MHz

$$f(\hat{\tau}_{n,t}) = \frac{\hat{\tau}_{t-1}}{2} + \frac{25\hat{\tau}_t}{1+\hat{\tau}_t^2} + 8\cos(1.2(t-1)) + \epsilon_{(n)}\,, \qquad (10)$$

where $\frac{\hat{\tau}_{t-1}}{2} + \frac{25\hat{\tau}_t}{1+\hat{\tau}_t^2} + 8\cos(1.2(t-1))$ gives the estimates over the posterior probability distribution function, and $\epsilon_{(n)}$ is the noise covariance in the system. Over the forecasting window t_w, the uth tenant overall forecasted demand, symbolized as $E[\tau]$, is obtained by $E[\tau_{(n)}]$ over set $\mathcal{N}$, where $|\mathcal{N}| \geq 1$. Now, the updated measured estimated value, symbolized as $\hat{z}_{(n)}$, can be obtained from

$$\hat{z}_{(n)} = \frac{f(\hat{\tau}_{n,t})}{20}\,. \qquad (11)$$

Based on the estimated measured value, $\hat{z}_{(n)}$, from the particle filter and the measured value from demand observation, $z_{(n)}$, the computed error, also known as noise covariance, in the system can be written as

$$\epsilon_{(n)} = \sqrt{\frac{1}{t_w^2}\sum_{t-t+1}^{t+f}(z_{(n,t)} - \hat{z}_{(n,t)})^2}\,. \qquad (12)$$

The mean squared error determines the covariance of the observation and estimates. The $\epsilon_{(n)}$ updates until errors in the prediction converge, where $z_{(n)} - \hat{z}_{(n)} \approx 0$.

4.2 Fuzzy-logic- and QoE-based admission control

In the federation approach, the key objective is to provide an efficient admission control based on the tenants forecasted demand, which ensures the enhanced network QoS and provisioning of desired tenants QoE. Therefore, a *Non-Dominated Sorting Genetic Algorithm II* (NSGA-II) is considered for the optimum network selection. This is a multi-objective elitist method of quickly obtaining non-dominated and optimal solutions that uses an explicit diversity preserving mechanism [46]. The given optimization approach still considers tenant QoE and throughput maximization for network selection to acquire fair resource allocation and maximum utilization, as will be explained in the following subsections.

4.2.1 Fuzzy-logic-based network selection

In this work, the federation controller deploys the fuzzy-logic based NSGA-II framework for the tenants' optimal network selection, as shown in algorithm 1. The uth tenant-forecasted demand, $E[\tau]$, on set $\mathcal{N}$ and service-specific network characteristics are provided to the FLC as inputs for selection. These characteristics include required data rate, latency, packet loss, available bandwidth, and billing information. Given the tenant's forecasted demand, the lowest and highest guaranteed resource bounds, symbolized as $\{B_{(\gamma)}, B_{(h)}\}$, of the corresponding service network from $\mathcal{S}\_Op$ are selected. The tenant-acceptable resource demand among $\{B_{(\gamma)}, B_{(h)}\}$ are represented as a genome of size ρ for generation of the initial population, which is denoted as P_0. The selection criteria are required to evaluate the fitness of the resource demand characteristics from the initial population following the objective function (i.e., desired tenant QoE and throughput maximization). The selection process goes through several iterations, also called generations (or gen), until they converge to a global optimum. After crossover of the parent and mutation of the child population, the most suitable statistics among the service guaranteed network resource bounds are selected with respect to their fitness for the defined objectives. The selected resource demand statistics (τ_γ, τ_h) are placed in the tenant forecasted demand matrix, symbolized as $\mathbf{Ten}_{u(\rho\times k)}$ where $k = 2|\mathcal{N}|$, in descending order of the tenant $\mathcal{Q}$ and η for provisioning of the service from the selected network from $\mathcal{S}\_Op$. This is to present the corresponding selected network resources with guaranteed QoS to the tenant for customized network configuration.

Algorithm 1: NSGA II based Optimization for Network Selection

Input: Forecasted demand ($\boldsymbol{\tau}$), number of generations (gen), population size (ρ), evaluation objectives ($\mathcal{Q}$, and η), network resource bounds ($B_{(\gamma)}$, $B_{(h)}$) and $B_{(\gamma)} < \tau \leq B_{(h)}$.
Output: Optimized $\mathbf{Ten}_u$ w.r.t. $\mathcal{Q}$ and η.
begin
  $P_0(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_\rho)$ = select $\boldsymbol{\tau}$ from resource bounds $(B_{(\gamma)}, B_{(h)})$ of networks from $\mathcal{S}\_Op$.
  $F_o(\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_\rho)$ = evaluate objective $(P_0(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_\rho))$.
  Sort P_0 w.r.t. F_0.
  for $i = 1 \rightarrow gen$ **do**
    $P_{i,parent}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_\rho)$ = select $(P_{i-1}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_\rho))$.
    $P_{i,child}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_{\frac{\rho}{2}})$ = crossover $(P_{i,parent}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_\rho))$.
    $P_{i,child}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_\rho) = P_{i,child}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_{\frac{\rho}{2}})$+ mutation $(P_{i,child}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_{\frac{\rho}{2}}))$.
    $F_{i,child}(\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_\rho)$ = evaluate objective $(P_{i,child}(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_\rho))$.
    $P_i(\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \ldots, \boldsymbol{\tau}_\rho)$ = sort $(P_{i,parent} + P_{i,child})$ w.r.t. F_i and select optimal ρ solutions.
  $\mathbf{Ten}_u = P$.
  sort $\mathcal{S}\_Op$ w.r.t. $\mathbf{Ten}_u$.

In addition to providing the guaranteed QoS, the uth tenant will now be admitted to a particular service network belonging to $\mathcal{S}\_Op$, subject to the availability of the resources. However, simultaneous access of various tenants to the service network creates competition. This can lead to network congestion. Thus, a priority-based admission queue is generated to ensure efficient admission control. A service allocation priority factor of the tenant, denoted as φ, can be computed as follows:

$$\varphi_u = f(\lambda_u, \psi_u, \kappa_u, h_u), \tag{13}$$

where λ_u represents the frequency of the tenant u requests, ψ_u represents the generated revenue, κ_u represents the requested service type across default classification, and h_u represents the n resource utilization history of the tenant u, where $n \in \mathcal{N}$ and $|\mathcal{N}|$ is considered to be 1 for simplicity. The admission priority list is generated in descending order of allocation factor in the tenant set $\mathcal{U}$. The tenant with the highest allocation factor is served first among all tenants by the network.

4.2.2 QoE-based admission control

Each network processes admission requests in terms of QoE constraints. A higher acquired tenant satisfaction level represents the optimal admission control, provisioning of better QoE, and resource utilization. The acquired QoE, $\mathcal{Q}_{(R)}$, by tenant u will not go beyond the expected QoE bounds, $\{\mathcal{Q}_{(\gamma)}, \mathcal{Q}_{(h)}\}$, where $R \leq \tau_h \leq \tau$. Therefore, the highest demanded QoE, denoted as $\mathcal{Q}_{(h)}$, by the forecasted demand τ and the acquired QoE, symbolized as $\mathcal{Q}_{(R)}$, due to acquired resources R of the tenant u are determined as follows:

$$\mathcal{Q}_{(h)} = f(\tau_h, \beta_\tau, \iota_\tau, \varphi_u), \tag{14}$$

$$\mathcal{Q}_{(R)} = f(R, \beta_R, \iota_R, \varphi_u), \tag{15}$$

where β, ι, and φ are the acceptable user-application-specific packet loss, latency sensitivity, and priority, respectively. To simplify, these measures are normalized to zero or 1 for the summation in $f(\cdot)$. Likewise, during peak hours the tenant served with the least-expected QoE, denoted as $\mathcal{Q}_{(\gamma)}$, due to softness in its QoE demand, such that $\gamma == R$. Thus, $\mathcal{Q}_{(\gamma)}$ can be written as

$$\mathcal{Q}_{(\gamma)} = f(\tau_\gamma, \beta_\gamma, \iota_\gamma, \varphi_u). \tag{16}$$

The uth tenant service request arrives in order with respect to φ from the prioritized admission queue, symbolized as

$$\mathcal{A}\_List = \{u_1(\mathcal{Q}_{(\gamma)}, \mathcal{Q}_{(h)}), u_2(\mathcal{Q}_{(\gamma)}, \mathcal{Q}_{(h)}), \ldots\}. \tag{17}$$

This is to access the optimal network from the service operator list, $\mathcal{S}\_Op$, which is in order according to tenant preferences for the customized network configuration. The NSCF assesses each tenant's desired $\mathcal{Q}$ across the network guaranteed QoS bounds, denoted as $\mathcal{S}_{\mathcal{Q}_\gamma}$ and $\mathcal{S}_{\mathcal{Q}_h}$, as shown in Algorithm 2. $\mathcal{S}_{\mathcal{Q}_\gamma}$ and $\mathcal{S}_{\mathcal{Q}_h}$ are the lowest and highest service network QoS bounds, respectively. After satisfying the available network QoS, the tenant τ resource demand (either guaranteed or demanded) will be checked against network capacity for resource allocation. The tenant's resource demand should be less than the serving network resource capacity. Thus, the tenant-acquired QoE, resource utilization and overall network throughput are obtained to compute the resource allocation fairness on set $\mathcal{U}$. The fairness of resource allocation of a particular service operator network is entered into the network service

profile, Ψ_List, along with the tenant-achieved $\mathcal{Q}_{(R)}$. This is for the federation controller to examine the fairness of resource allocation and user satisfaction level from the serving network. In the case of selected network resource unavailability or unsatisfied QoS bounds for the tenant, the next network from $\mathcal{S}\_Op$ will be examined by the NSCF for admission control. After admission, the tenant-acquired $\mathcal{Q}$ will be monitored to ensure efficient network performance. In the case of violation of QoS/QoE bounds, the user will be dropped from the serving network and reassessed with higher priority by the NSCF. To summarize, by optimizing the forecasted demand and service network statistics, a customized network is selected, and resources are allocated with guaranteed QoS bounds to ensure efficient resource utilization and tenant-acquired QoE.

Algorithm 2: QoE based Admission Control

Input: Service operator list ($\mathcal{S}\_Op$), tenant forecasted demand ($\{\tau_\gamma, \tau_h\}$), and $\{\tau_\gamma, \tau_h\} \in \mathbf{Ten}_u$, admission queue ($\mathcal{A}\_List$).
Output: $\Psi\_List = \{u_1(\mathcal{Q}_{(R)}, \mathcal{F}_\eta, \mathcal{S}_v), \ldots\}$.
for $i = 1 \rightarrow \mathcal{A}\_List.length$ **do**
  for $(j = 1 \rightarrow \mathcal{S}\_Op.length)$ **do**
    Select $\{\mathcal{S}_{\mathcal{Q}_\gamma}, \mathcal{S}_{\mathcal{Q}_h}\}$ bounds of $\mathcal{S}\_Op(j)$.
    if $(\mathcal{S}_{\mathcal{Q}_\gamma} < \mathcal{Q}(\tau_h(i)) \leq \mathcal{S}_{\mathcal{Q}_h}) \&\& (\mathcal{S}_{\mathcal{Q}_\gamma} \leq \mathcal{Q}(\tau_\gamma(i)) < \mathcal{S}_{\mathcal{Q}_h})$ **then**
      $\mathcal{R}\_Op =$ assign $\mathcal{S}\_Op(j)$ operator resources.
      if $(\tau_h(i) \leq \mathcal{R}\_Op) || (\tau_\gamma(i) > \mathcal{R}\_Op)$ **then**
        Allocate resources via $\mathcal{R}\_Op = \mathcal{R}\_Op - R(i)$.
        Obtain tenant acquired QoE via $\mathcal{Q}_{(R)} = \frac{R(i)}{\tau_h(i)}$.
        Obtain resources utilization via $\mathbf{U}_i(R(i))$.
        Compute throughput via $\eta_i = p_{R(i)} p_{\iota(i)}$.
        Update $\mathcal{F}_\eta$ by including ith tenant.
        $\mathcal{S}_v =$ save $\mathcal{S}\_Op(j)$.
        $\Psi\_List = u_i(\mathcal{Q}_{(R)}, \mathcal{F}_\eta, \mathcal{S}_v)$.
      else
        Check $j + 1 \in \mathcal{S}\_Op$ for tenant resource allocation.
    else
      Check $j + 1 \in \mathcal{S}\_Op$ against tenant QoE demand.

4.3 Service and flow monitoring

Once the decision has been made to admit the tenant to a particular network, the network is configured with guaranteed QoS for the respective service provisioning requested by the tenant. The E2E service flow should also be monitored to ensure the tenant's acquired QoE does not degrade during service from the operational network, and that traffic flow is proportional to the capacity. Therefore, we propose a QoS/QoE and traffic flow monitoring system (QTFM) that pursues the following goals: 1) monitor flow to ensure the tenant's acquired QoE is within guaranteed bounds, where the least-expected QoE is denoted as $\mathcal{Q}_{(\gamma)}$ and the highest achievable QoE is denoted as $\mathcal{Q}_{(h)}$, 2) providing feedback to the analyzer for modification of the forecasted demand in proportion to the actual demand, and utilization, as shown in Fig. 3 and described in detail in the following subsections.

4.3.1 QoS/QoE monitoring

For network operators, continuous service monitoring is essential to ensure that network QoS and tenant QoE remain above the agreed least guaranteed bound, where violation in provisioning of agreed QoE and QoS can occur. Therefore, network QoS is monitored for tenant service duration by the QTFM to ensure the tenant's acquired QoE is within expected bounds, as shown in Algorithm 3. In the case of violation of QoS/QoE bounds, the tenant will be dropped from Ψ_List and added to the admission queue, $\mathcal{A}\_List$ with higher priority as compensation. Now, the tenant will be reassessed by the NSCF, along with the change in QoE statistics. The QTFM will also trigger the forecasting model to take appropriate action to enhance the overall network's QoS and tenant's QoE. The network service profile, Ψ_List, will also be updated to maintain the network service inventory by the federation controller.

Algorithm 3: Service and Flow Monitoring

Input: Tenant QoE ($\mathcal{Q}_{R_u}$), serving network QoS ($\mathcal{S}_{\mathcal{Q}_{(h,u)}}$), service operator list ($\mathcal{S}\_Op$), service profile (Ψ_List).
Output: Updated Ψ_List.
if ($\mathcal{Q}_{R_u} > \mathcal{S}_{\mathcal{Q}_{(h,u)}}$) **then**
  $\Psi\_List = \Psi\_List - u(\mathcal{Q}_{R_u}, \mathcal{F}_\eta, \mathcal{S}_v)$.
  $\varphi_u =$ increase φ_u.
  $\mathcal{A}\_List = \mathcal{A}\_List + u(\mathcal{Q}_\gamma, \mathcal{Q}_h)$.
  Compute Algo. 2.
  Update $\mathcal{P}$ via (18).
  Update τ via (19).
else
  $u \in \Psi\_List$

4.3.2 Forecasted service demand monitoring

Inefficiency in the forecasting process might over-/under-utilize the network resources, leading to inappropriate tenant admission to the network. This would result in a violation of the agreed QoS/QoE, due to poor network QoS and tenant QoE [26]. Taking into account the above-mentioned issue, a monitoring procedure is designed to consistently monitor the forecasted and actual demand. This keeps track of the number of violations such as inefficient resource utilization, huge forecasting error or variance, agreement violation and poor QoE. For future forecasting optimization, QTFM provides feedback to the forecasting model, DCA, to update forecasted estimates using the penalty history function. This is symbolized as $\mathcal{P}$ and obtained on t_w, such as

$$\mathcal{P}_{(n)} = \exp\left(\frac{p_{(n)}}{\sum_{u\in\mathcal{U}} a_{(u,n)}}\right), \tag{18}$$

where $n \in \mathcal{N}$, $p_{(n)} = 1$ indicates the penalty due to QoE violation on resource demand n of the uth tenant, otherwise zero. The admission indicator $a = 1$ for the uth tenant due to the acquired resource n from the subscribed operator, respectively. On the given number of penalties for resource demand n, the forecasted demand will be updated for future services. Equation (10) is updated by the forecasting modifier, (denoted as $\mathcal{P}\epsilon$) is defined as

$$f(\hat{\hat{\tau}}_{n,t}) = \frac{\hat{\tau}_{t-1}}{2} + \frac{25\,\hat{\tau}_t}{1+\hat{\tau}_t^2} + 8\cos\left(1.2\,(t-1)\right) + \mathcal{P}_{(n)}\,\epsilon_{(n)}\,, \tag{19}$$

where $\frac{\hat{\tau}_{t-1}}{2} + \frac{25\hat{\tau}_t}{1+\hat{\tau}_t^2} + 8\cos\left(1.2(t-1)\right)$ gives the estimates for the Posterior probability distribution function. ϵ_n is the noise covariance in the system for adjusting estimates according to actual demand. Unlike the conservative setting of the existing forecasting techniques (Holt-Winters, Bayesian, and Monte Carlo), the penalty function dynamically updates the system, where no agreed QoE and QoS violation could occur. This is due to the adaptability of the service and flow monitoring feature, which obtains the effective demand from the forecasted information to release inefficient resources for better utilization, permitting the network operator to accommodate more users.

5 Performance evaluation measures

We define the evaluation parameters in this section. In the context of evaluating the proposed framework, the chosen performance metrics set out to be the assessment of resources and network utilization, resource allocation fairness among the tenants as well as their associated user satisfaction level, as discussed in the following subsections. The performance evaluation framework is aligned with the existing work from the literature for comparison.

5.1 Resources utilization

To capture tenant admission by the vth network operator at a given time t, a_u is introduced as a binary indicator that takes the value 1 if the tenant is admitted to the network, subject to the resources and service availability. Otherwise the value of zero is taken. After successful admission to the network, the nth resource assigned to the tenant u from the service operator resource pool is defined as

$$R_u = a_u\, R_{u,n}\,, \tag{20}$$

where $R_u \leq \tau_h$. Now the aggregate resources assigned to the tenant set $\mathcal{U}$ is obtained from

$$\sum_{u\in\mathcal{U}} a_u\, R_{u,n} \leq \mathcal{R}\_Op\,. \tag{21}$$

Aggregate resources should not exceed the network capacity. The uth tenant utility (symbolized as $\mathbf{U}_u$) w.r.t. R_u is computed by

$$\mathbf{U}_u(R_u) = \alpha\, e^{\omega q}\,, \tag{22}$$

where ω is the difference between the achieved and desired resources, q and α represent the utility function slope and

utility function curve slope. Now, the virtual operator v network utility can be given by

$$\mathbf{U}_v = \sum_{u \in \mathcal{U}} \mathbf{U}_u (R_u) . \tag{23}$$

Subsequently, on the basis of virtual network operator utility, $\mathbf{U}_v$, the mean network utility, represented as $\mathbf{U}_{Net}$ on set $\mathcal{M}$ and $\mathcal{V}$ can be obtained from

$$\mathbf{U}_{Net} = \frac{1}{MV} \sum_{i=1}^{M} \sum_{j=1}^{V} \mathbf{U}_{v_{ij}} . \tag{24}$$

5.2 Resource allocation fairness

Maximum resource utilization as well as tenants' acquired throughput is crucial to attain the objective of revenue maximization. Higher acquired throughput indicates a higher fairness of resource allocation and better tenant QoE on admission [47]. Admission control fairness, $\mathcal{F}_{\mathcal{A}}$, is obtained by

$$\mathcal{F}_{\mathcal{A}} = \frac{\left(\sum_{u \in \mathcal{U}} a_u\right)^2}{U \times \sum_{u \in \mathcal{U}} \left(a_u\right)^2} , \tag{25}$$

where $a \in \{0, 1\}$ subject to the availability of service and resources by the subscribed operator v. Similarly, on admission resource allocation is also a key factor to be considered. This determines the acquired throughput on the probability of resource utilization (p_R) within the given latency constraints (p_ι) at the massive number of tenant demand, which is computed as:

$$\eta_u = p_R \, p_\iota . \tag{26}$$

Significantly, the fairness factor in resource allocation can be achieved as follows:

$$\mathcal{F}_\eta = \frac{\left(\sum_{u \in \mathcal{U}} \eta_u\right)^2}{U \times \sum_{u \in \mathcal{U}} (\eta_u)^2} . \tag{27}$$

6 Simulation results

A simulation model developed in MATLAB to evaluate the performance of the proposed framework. A virtual network is constructed with a set of varied system parameters to support four distinguished services. The number of tenants associated with the virtual network is considered to be $U = [5, 330]$ for a heterogeneous service provisioning. The average number of users associated with the tenant u is $E[u_i] = 100$, as considered in [26] and [27]. Significantly, $\beta = [10^{-2}, 10^{-7}]$, $\iota = [10, 200]$ ms, $\varphi = [1, 5]$, $\tau_h = [10, 100]$ MHz, and $\mathcal{R}\_Op = 500$ are the considered ranges of tenant-service-specific packet loss, latency sensitivity, priority, desired resource demand, and available operator resources for each service belong to $\mathcal{S}$, respectively. The overall demand is normalized to 0 or 1 for simplicity.

Figures 5, 6, 7, 8, and 9 show the performance of the proposed dynamic traffic forecasting and fuzzy-based admission control in the context of user satisfaction level, resource allocation fairness, and utilization gain. The results are compared with existing schemes in the literature and summarized in Table 3. We compare to mobile traffic forecasting (MTF) [26], reinforcement learning (RL-NSB) [27], online auction (ORAN) and greedy algorithm [48], and bankruptcy game (BG) [49], based resource allocation and admission control schemes.

6.1 Impact of forecasting

Demand forecasting is an essential part of traffic engineering and network management. It helps the operator to plan network and resource allocation to ensure better user QoE and network QoS. Therefore, the proposed framework admitting tenants to a corresponding network based on their forecasted demand. The impact of forecasting on tenant-acquired QoE, resource allocation, user satisfaction level, and load distribution is illustrated in Figs. 5, 6 and 7 .

6.1.1 Tenant's QoE and fairness

Figure 5 shows the proposed framework's performance concerning a tenant's perceived QoE and resource allocation fairness for forecasted guaranteed resource bounds ($\tau_\gamma = 0.8$ and $\tau_h = 1$). It can be observed that the trend of resource allocation fairness is more than 97% over the entire range of U on τ_γ and τ_h demand. Fairness of allocation begins to rise with more tenant's arrival onto the network for the provisioning of their requested services. This rise is due to resource allocation between their guaranteed and desired demand approximately close to their actual demand originating by the forecasting modifier's convergence. Hence, the relative gain in acquired resource allocation fairness by τ_γ over τ_h is 0.5% at $U = 150$. This change in gain is noticeable in the case of a fully loaded network to keep tenant admission rejection from the network as low as possible. The proposed approach gives a good estimation of future demand due to self-healing of the forecasted demand by the continuous monitoring of tenant's QoE, and network's QoS. Thus, efficient demand forecasting and monitoring result in more appropriate network selection and admission control for the tenants.

The achieved average QoE is also high at the beginning of the acquired result. This is because tenants are acquiring

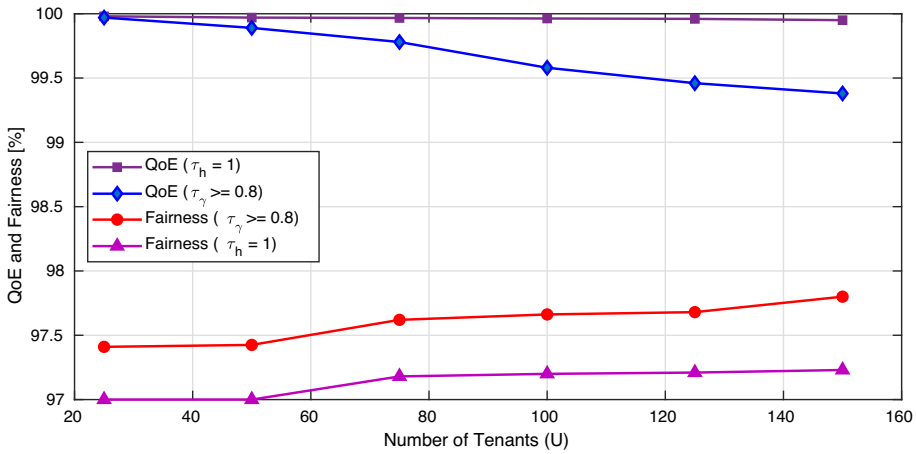

Fig. 5 Computation of QoE and fairness on $U = 150$ number of tenants with $\tau_\gamma \geq 0.8$ and $\tau_h = 1$ forecasted demand

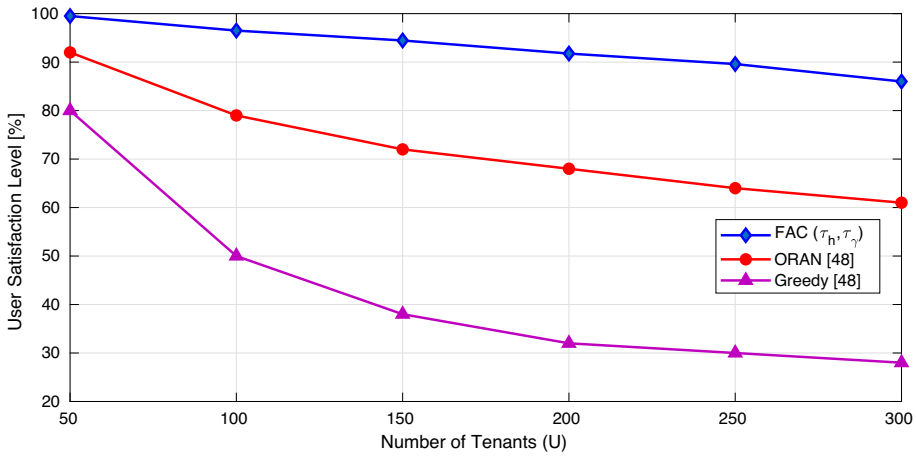

Fig. 6 Comparison of user satisfaction level on varying demand; number of tenants $U = [50, 300]$ with respect to $|\mathcal{S}| = 4$ and $|\mathcal{N}| = 1$

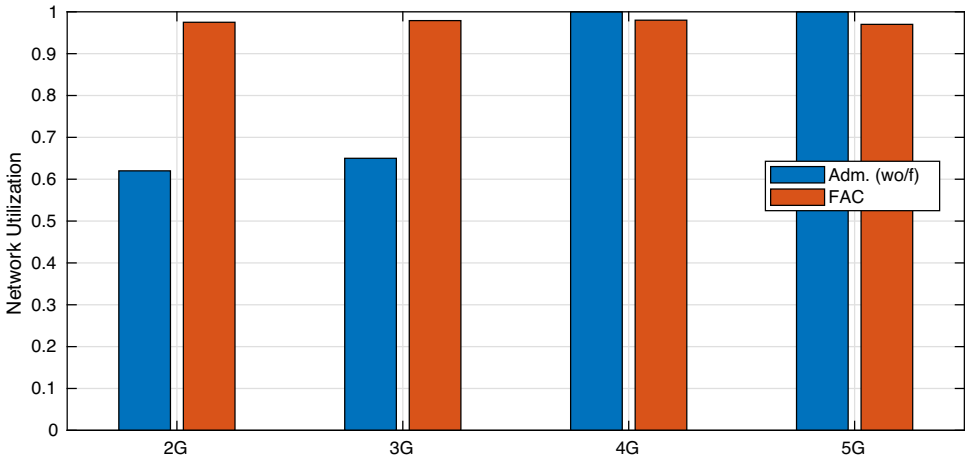

Fig. 7 Average network utilization of the fully loaded wireless network, with and without tenant demand forecasting and admission control on $|\mathcal{S}| = 4$ and $|\mathcal{N}| = 1$

resources for their forecasted demand at $\tau_h = 1$, which might be greater than the actual demand obtained after modification of the forecasted demand. The achieved QoE begins to decline with an increase in the number of tenants. However, it is over 99% on the entire range of U. This is because tenants are acquiring resources between their guaranteed and desired demand approximately close to their actual demand to reduce the number of rejections and improve fairness among tenants. The relative loss in the QoE by τ_γ over τ_h is 0.6% at $U = 150$, which is noticeably

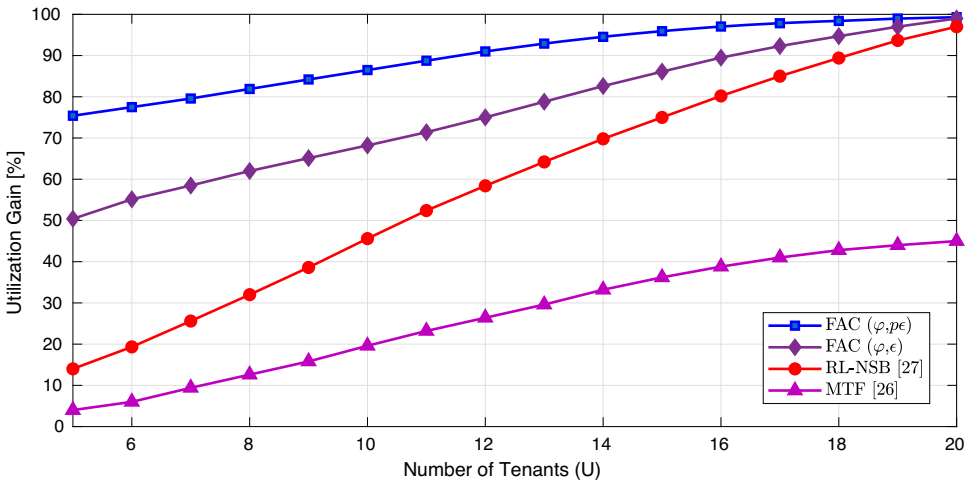

Fig. 8 Computation of network utilization gain by forecasting and legacy approach across $U = 20$ number of tenants along with $E[u_i] = 100$ user each, and $\tau_h = [10, 100]$ MHz

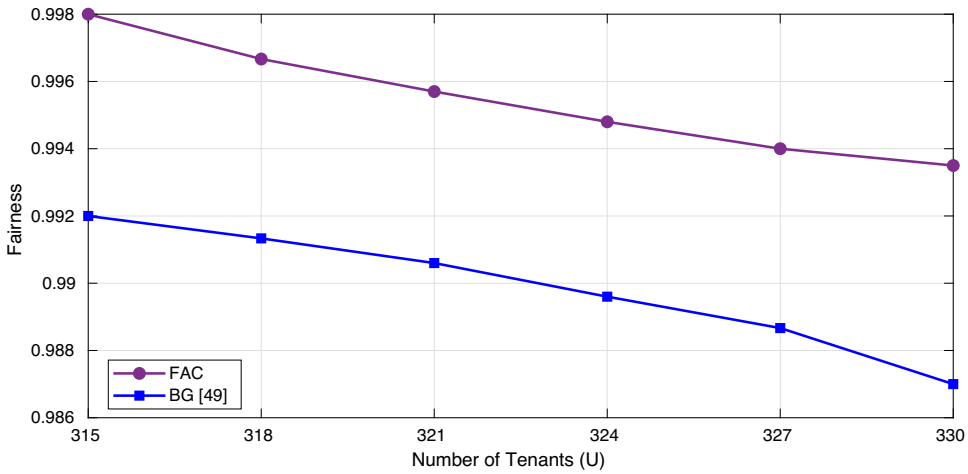

Fig. 9 Resource allocation fairness on varying demand; number of tenants $U = [315, 330]$ with $E[u_i] = 100$ users each on $|\mathcal{S}| = 4$ and $|\mathcal{N}| = 1$

low in the case of a fully loaded network to keep the tenants' admission rejection from the network as low as possible. Hence, the proposed framework is slightly better at managing a massive number of tenant's demand, because of the adaptability of the convergence to actual demand. This framework achieves a higher QoE and fairness of resource allocation on massive connectivity demand.

6.1.2 User satisfaction level on forecasting

Monitoring of various parameters during service provisioning is an added feature of the proposed framework that can impact user satisfaction level. Figure 6 indicates the user satisfaction of 300 tenants with approximately 100 users each. User satisfaction is computed by the accepted user's acquired QoE from guaranteed resource bounds ($\tau_\gamma = 0.8$ and $\tau_h = 1$) over the desired QoE and the average of the total number of received requests. This user satisfaction level reflects the proportion of the accepted users on their desired QoE for service provisioning. The achieved performance of the proposed framework is superior compared to its counterparts that are ORAN and greedy algorithms [48]. The relative gains of the proposed framework at $U = 50$ are 8% with ORAN, and 20% with the greedy algorithm, respectively. The variance between the gain increases when the tenants' number and associated users increase. The relative gains of the proposed framework at $U = 300$ are 25% with ORAN and 58% with the greedy algorithm.

First, with fewer tenants and associated users arriving, the acquired user satisfaction will be greater, because each tenant's user has access to their desired demand. However, with an increase in the number of tenants and users, congestion occurs. This situation can cause the network to become inefficiently saturated, which leads to an increasing number of user backing off from the service or being rejected, as seen at $U = 300$ in the case of ORAN and

greedy approaches from [48]. However, the proposed framework reduces rejection by provisioning services from the optimal network within the guaranteed bounds close to the tenant's effective demand due to its QoS/QoE monitoring feature. The monitoring feature helps the efficient distribution of traffic flow among the services of set $\mathcal{S}$ to ensure efficient admission control and resource allocation. In contrast, existing schemes have higher rejection rates due to competition among users for limited desired resources and adoption of the greedy approach. This deficiency in the existing scheme results in the degradation of user satisfaction and network resource utilization.

6.1.3 Traffic/load distribution across heterogeneous services

A detailed analysis of the traffic distribution across heterogeneous services belonging to set $\mathcal{S}$ is presented here. Figure 7 illustrates the average network utilization with and without the proposed demand forecasting and admission control framework. The results are obtained on a fully loaded network, for instance, if 300 tenants arrive on the network. It can be observed that on admission without forecasting, 4G and 5G networks are inefficiently saturated with 100% network utilization. This network saturation results in the tenant's QoE dropping due to congestion and more tenants being rejected or backing off from the service network, whereas, in the 2G and 3G network, resources are underutilized with 62%, and 66% network utilization, respectively. In O-RAN enabled networks, these circumstances become costly for the network operator due to under/over-utilized resources in the respective networks. This increases not only the operational cost but also reduces overall network performance and tenant acquired QoE.

Network utilization achieved by the proposed framework is superior to the legacy approach in heterogeneous services provisioning from set $\mathcal{S}$ at more than 95%. The proposed FAC framework minimizes the drawbacks of the legacy approach by dynamically forecasting traffic demand for tenants' optimal admission through the fuzzy-logic-based network selection. The fuzzy-logic approach helps the efficient distribution of the traffic load based on the demanded services and available capacity of various heterogeneous networks. The proposed forecasting framework's self-organization feature ensures that the networks do not saturate with 100% load. In congestion, the proposed framework permits the tenant to accept resources over guaranteed bounds close to network capacity. In this way, each service will accept only relevant demand to accommodate more tenants in the agreed QoE. In contrast, without forecasting, traffic is randomly admitted by the network on their desired demand and sensitivity, which leads to congestion and over/under network utilization.

6.2 Impact of optimization and service monitoring

Information about the tenants' demand is crucial for fuzzy-logic optimization. However, uncertainty in the tenant forecasted demand can lead to inefficient admission. When a tenant wishes to gain access to the network, the proposed framework learns and leverages the information from the tenants' history to improve the efficiency of the demand forecasting and admission control mechanism. These improvements reduce network saturation and increase network utilization, as shown in Figs. 8 and 9.

6.2.1 Priority-based supervised admission with optimization

Figure 8 illustrates the performance of the proposed framework concerning bandwidth utilization with varying traffic demand and compared with schemes documented in the existing literature. The gain is computed by the average bandwidth utilization by the tenants on forecasted and legacy demand as in [27]. The relative gains in bandwidth utilization by the proposed framework are 81.43% and 72.22% on RL-NSB [27] and 94.69% and 92% on MTF [26] at $U = 5$. This utilization gain is by the proposed forecasting modifier ($\mathcal{P}\epsilon$) and admission priority factor (φ). $\mathcal{P}\epsilon$ optimizes the forecasted demand through fuzzy-logic-based optimal network selection and service monitoring to maximize resource utilization. φ prioritizes tenant admission to the network after reviewing the tenant history to earn more revenue through efficient resource utilization. In addition, existing schemes admit users on their arrival on the network according to their resource demand forecasting.

It can be noticed that resource utilization is continuously increasing as the number of tenants increases. Such as at $U = 20$ resource utilization by the proposed framework is above 95%. Similarly, the achieved relative gains at $U = 20$ with $\mathcal{P}\epsilon$ and ϵ are 3% and 2% with RL-NSB, and 55% and 54% with MTF, respectively. As the existing schemes admit the users on their demand forecasting only, it takes time to converge to an optimal solution and improve the admission process. Therefore, existing schemes show less utilization at the beginning and converge to higher utilization as the number of tenants and processing time increases. However, in the proposed approach, with fewer tenants arriving, the resources are allocated over the tenants' expected QoE bounds, and the remaining resources will be placed in the pool to be used by other operators for providing services to their associated tenants. This

provides the operator with an incentive to lease as many resources as possible to earn more revenue corresponding to resource utilization. In the case of congestion, through negotiation, tenants are accommodated at their guaranteed demand to minimize rejection or backing-off from the network. The proposed admission control and resource allocation mechanism are efficient for forecasted demand, due to the priority factor and self-organized forecasting mechanism. This yields better performance in terms of resource utilization and acquired QoE.

6.2.2 Resource allocation with optimization

For optimal network selection, the fitness function and its relationship with the data are the keys to optimization. This determines the appropriate network for the tenant and fair resource distribution among tenants in the network. Figure 9 illustrates the fairness of resource allocation on $U = [315, 330]$ tenants, with 100 users each, across various approaches. The results achieved are compared with the work in [49]. In the proposed admission scheme, fairness is computed as the tenants' acquired average throughput on the given load, whereby each user shares the same proportion in terms of resource utilization and acquired QoE. The result indicates that the proposed QoE-based admission scheme acquires efficient allocation with a fairness index of approximately 1 compared to the bankruptcy game allocation scheme with its fairness index floating around 0.99. This is because in bankruptcy game scheme, users randomly form coalitions for network admission and resource allocation. This leads to more users rejections due to competition on limited resources and congestion generated by inefficient admission control. The relative gain in fairness by the proposed framework is 0.6% and 0.65% at $U = 315$ and $U = 330$. The rise in gain is low but noticeable on the entire range of U. This gain is achieved on the availability of optimal network solutions and multi-variate priority feature for tenant's admission and resource allocation, which is obtained via fuzzy-logic-based network selection in the proposed framework. It helps to accommodate as many tenants as possible along with guaranteed resource allocation, to claim a lower number of tenants rejection or back-off from the network. Thus, efficient admission control and resource allocation lead to network utilization being maximized, as well as fairness among tenants, as summarized in Table 3.

7 Conclusion

In this paper, we proposed a dynamic traffic forecasting and admission control framework for a federated open radio access network (O-RAN). In this framework, a three-stage approach, namely demand and capacity analyzer, network selection and configuration, and QoS/QoE and traffic flow management have been proposed. This framework predicts the future traffic demand for the optimal network selection among multiple heterogeneous access networks and resource management to ensure better tenant QoE and O-RAN network utilization. The considered demand characteristics are bandwidth, latency sensitivity, quality of service demand, service priority type, and packet loss ratio. In this work, a fuzzy-logic-based network selection scheme with a multi-variate admission priority feature is introduced for optimal admission control and service allocation to the tenants. Moreover, a QoS/QoE-based service monitoring approach is also presented to update the demand via a forecasting modifier to allocate resources approximately closer to the tenant's actual demand, which improves overall network QoS and tenant QoE. The proposed framework outperforms the existing legacy approaches in terms of enhanced tenant QoE and fairness of resource allocation, efficient network utilization, and better user satisfaction levels for various heterogeneous services of future wireless networks. For future work, we aim to enhance the framework with signaling optimization on homogeneous user-application-specific characteristics for service provisioning of complex networks of the future.

Funding Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interests.

References

1. Tseliou G, Samdanis K, Adelantado F, Pérez XC, Verikoukis C (2016) A capacity broker architecture and framework for multi-tenant support in LTE-A networks. In: 2016 IEEE international conference on communications (ICC), IEEE, pp 1–6

2. Rappaport TS, Sun S, Mayzus R, Zhao H, Azar Y, Wang K, Wong GN, Schulz JK, Samimi M, Gutierrez F (2013) Millimeter wave mobile communications for 5G cellular: it will work! IEEE Access 1:335–349
3. Roh W, Seol JY, Park J, Lee B, Lee J, Kim Y, Cho J, Cheun K, Aryanfar F (2014) Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results. IEEE Commun Mag 52:106–113
4. Qiao J, Shen XS, Mark JW, Shen Q, He Y, Lei L (2015) Enabling device-to-device communications in millimeter-wave 5g cellular networks. IEEE Commun Mag 53:209–215
5. Wang D, Song B, Chen D, Du X (2019) Intelligent cognitive radio in 5G: AI-based hierarchical cognitive cellular networks. IEEE Wirel Commun 26:54–61
6. Yu QY, Lin HC, Chen HH (2019) Intelligent radio for next generation wireless communications: an overview. IEEE Wirel Commun 26:94–101
7. Richart M, Baliosian J, Serrat J, Gorricho J (2016) Resource slicing in virtual wireless networks: a survey. IEEE Trans Netw Serv Manage 13:462–476
8. Andrews JG, Buzzi S, Choi W, Hanly SV, Lozano A, Soong AC, Zhang JC (2014) What will 5G be? IEEE J Sel Areas Commun 32:1065–1082
9. 3GPP (2017) Technical specification group radio access network; study on new radio access technology; radio interface protocol aspects. 3GPP TR 38.804 V14.0.0
10. Lee YL, Loo J, Chuah TC, Wang LC (2018) Dynamic network slicing for multitenant heterogeneous cloud radio access networks. IEEE Trans Wirel Commun 17:2146–2161
11. Banchs A, Breitbach M, Costa X, Doetsch U, Redana S, Sartori C, Schotten H (2015) A novel radio multiservice adaptive network architecture for 5G networks. In: Vehicular technology conference (VTC Spring), 2015 IEEE 81st, IEEE 1–5
12. Alliance O (2018) O-RAN: towards an open and smart RAN. White paper
13. Gavrilovska L, Rakovic V, Denkovski D (2020) From cloud ran to open ran. Wirel Pers Commun 1–17
14. Aryafar E, Keshavarz-Haddad A, Wang M, Chiang M (2013) RAT selection games in HetNets. In: 2013 Proceedings IEEE INFOCOM. IEEE, pp 998–1006
15. 3GPP (2013) GPRS enhancements for E-UTRAN access. 3GPP TS 23.401 Release 12
16. 3GPP (2013) Architecture enhancements for non-3GPP accesses. 3GPP TS 23.402 Release 12
17. 3GPP (2012) GPP system to wireless local area network (WLAN) interworking; system description. 3GPP TS 23.234, V11.0.0, 3
18. Bouali F, Moessner K, Fitch M (2016) A context-aware user-driven framework for network selection in 5G multi-RAT environments. In: 2016 IEEE 84th vehicular technology conference (VTC-Fall). IEEE, pp 1–7
19. 3GPP (2018) Radio frequency (RF) requirements for multicarrier and multiple radio access technology (Multi-RAT) base station (BS). 3GPP TR 37.900 V15.0.0
20. Wang J, Roy H, Kelly C (2019) OpenRAN: the next generation of radio access networks. Telecom Infra Project
21. xRAN Forum (2016) The mobile access network, beyond connectivity. xRAN Forum
22. Nokia (2020) What is open ran and why is it important? Nokia
23. Niknam S, Roy A, Dhillon HS, Singh S, Banerji R, Reed JH, Saxena N, Yoon S (2020) Intelligent O-RAN for beyond 5G and 6G wireless networks. arXiv preprint arXiv:2005.08374
24. Techplayon (2019) Open RAN: (O-RAN) reference architecture. Techplayon O-RAN Alliance
25. Wireless P (2020) 5G 4G 3G 2G WI-FI OPENRAN controller. Parallel Wireless
26. Sciancalepore V, Samdanis K, Costa-Perez X, Bega D, Gramaglia M, Banchs A (2017) Mobile traffic forecasting for maximizing 5G network slicing resource utilization. In: IEEE INFOCOM 2017-IEEE conference on computer communications. IEEE, pp 1–9
27. Sciancalepore V, Costa-Perez X, Banchs A (2019) RL-NSB: reinforcement learning-based 5G network slice broker. IEEE/ACM Trans Netw 27:1543–1557
28. Raikwar AR, Sadawarte RR, More RG, Gunjal RS, Mahalle PN, Railkar PN (2017) Long-term and short-term traffic forecasting using holt-winters method: a comparability approach with comparable data in multiple seasons. Int J Syn Emot (IJSE) 8:38–50
29. Dudek G (2016) Neural networks for pattern-based short-term load forecasting: a comparative study. Neurocomputing 205:64–74
30. Dudek G (2019) Multilayer perceptron for short-term load forecasting: from global to local approach. Neu Comput Appl 1–13
31. Hippert HS, Pedreira CE, Souza RC (2001) Neural networks for short-term load forecasting: a review and evaluation. IEEE Trans Power Syst 16:44–55
32. Narmanlioglu O, Zeydan E, Kandemir M, Kranda T (2017) Prediction of active UE number with bayesian neural networks for self-organizing LTE networks. In: 2017 8th International conference on the network of the future (NOF). IEEE, pp 73–78
33. Miao D, Sun W, Qin X, Wang W (2016) Msfs: multiple spatio-temporal scales traffic forecasting in mobile cellular network. In: 2016 IEEE 14th international conference on dependable, autonomic and secure computing, 14th international conference on pervasive intelligence and computing, 2nd international conference on big data intelligence and computing and cyber science and technology congress (DASC/PiCom/DataCom/CyberSciTech). IEEE, pp 787–794
34. Zhang Z, Liu F, Zeng Z, Zhao W (2017) A traffic prediction algorithm based on bayesian spatio-temporal model in cellular network. In: 2017 International symposium on wireless communication systems (ISWCS). IEEE, pp 43–48
35. Inaba T, Elmazi D, Sakamoto S, Oda T, Ikeda M, Barolli L (2015) A secure-aware call admission control scheme for wireless cellular networks using fuzzy logic and its performance evaluation. J Mobile Multim 213–222
36. Goudarzi S, Anisi MH, Abdullah AH, Lloret J, Soleymani SA, Hassan WH (2019) A hybrid intelligent model for network selection in the industrial internet of things. Appl Soft Comput 74:529–546
37. Kaloxylos A, Barmpounakis S, Spapis P, Alonistioti N (2014) An efficient RAT selection mechanism for 5G cellular networks. In: 2014 International wireless communications and mobile computing conference (IWCMC). IEEE, pp 942–947
38. Khan AA, Abolhasan M, Ni W, Lipman J, Jamalipour A (2019) A hybrid-fuzzy logic guided genetic algorithm (H-FLGA) approach for resource optimization in 5G VANETs. IEEE Trans Veh Technol 68:6964–6974
39. Zeng H, Zhu X, Jiang Y, Wei Z, Wang T (2019) A green coordinated multi-cell NOMA system with fuzzy logic based multi-criterion user mode selection and resource allocation. IEEE J Sel Topics Sig Process 13:480–495
40. Silva KC, Becvar Z, Cardoso EH, Francês CR (2018) Self-tuning handover algorithm based on fuzzy logic in mobile networks with dense small cells. In: 2018 IEEE wireless communications and networking conference (WCNC). IEEE, pp 1–6
41. Shrimali B, Bhadka H, Patel H (2018) A fuzzy-based approach to evaluate multi-objective optimization for resource allocation in cloud. Int J Adv Technol Eng Exp 5:140–150
42. Raza MQ, Khosravi A (2015) A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. Renew Sust Energy Rev 50:1352–1372

43. Ghosh S, Razouqi Q, Schumacher HJ, Celmins A (1998) A survey of recent advances in fuzzy logic in telecommunications networks and new challenges. IEEE Trans Fuzzy Syst 6:443–447
44. Gupta A, Jha RK (2015) A survey of 5G network: architecture and emerging technologies. IEEE Access 3:1206–1232
45. Choi Yi, Park N (2017) Slice architecture for 5G core network. In: 2017 ninth international conference on ubiquitous and future networks (ICUFN). IEEE, pp 571–575
46. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6:182–197
47. Jiang M, Condoluci M, Mahmoodi T (2016) Network slicing management & prioritization in 5G mobile systems. In: Proceedings of European wireless 2016; 22th European wireless conference. VDE, pp 1–6
48. Liang L, Wu Y, Feng G, Jian X, Jia Y (2019) Online auction-based resource allocation for service-oriented network slicing. IEEE Trans Veh Technol 68:8063–8074
49. Jia Y, Tian H, Fan S, Zhao P, Zhao K (2018) Bankruptcy game based resource allocation algorithm for 5G cloud-RAN slicing. In: 2018 IEEE wireless communications and networking conference (WCNC). IEEE, pp 1–6