IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 June 2023, accepted 15 June 2023, date of publication 22 June 2023, date of current version 3 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3288698

RESEARCH ARTICLE

Intelligent Resource Management for eMBB and URLLC in 5G and Beyond Wireless Networks

RANA M. SOHAIB[1], (Member, IEEE), OLUWAKAYODE ONIRETI[1], (Senior Member, IEEE), YUSUF SAMBO[1], (Senior Member, IEEE), RAFIQ SWASH[2], (Senior Member, IEEE), SHUJA ANSARI[1], (Senior Member, IEEE), AND MUHAMMAD A. IMRAN[1], (Fellow, IEEE)

[1]James Watt School of Engineering, University of Glasgow, G12 8QQ Glasgow, U.K.
[2]Aidrivers Ltd., UB10 0NE London, U.K.

Corresponding author: Rana M. Sohaib (RanaMuhammad.Sohaib@glasgow.ac.uk)

The authors wish to acknowledge receipt of the funding we received for this research under the Engineering and Physical Sciences Research Council (EPSRC) U.K.-India Future Networks Initiative (Grant ref-EP/W016524/1) led by the University of East Anglia in U.K.

ABSTRACT In the era of 5G and beyond wireless networks, the simultaneous support of enhanced Mobile Broadband (eMBB) and Ultra-Reliable Low Latency Communications (URLLC) poses significant challenges in managing radio resources efficiently. By leveraging the puncturing technique, we propose an intelligent resource management framework for meeting the strict latency and reliability requirement of URLLC services and the high data rate for eMBB services. In particular, a semi-supervised learning and deep reinforcement learning (DRL) based architecture is proposed to manage the resources intelligently. We decompose the optimization problem into two subproblems: 1) resource block allocation (RBA) strategy for eMBB slice, and 2) URLLC scheduling. Through extensive simulations and performance evaluations, we demonstrate the effectiveness of the proposed technique in optimizing resource utilization, minimizing latency for URLLC users, and maximizing the throughput for eMBB services. Simulation findings demonstrate that the proposed methodology can ensure the URLLC reliability requirements while maintaining higher average sum rate for eMBB and higher convergence rate. The proposed framework paves the way for the efficient coexistence of diverse services, enabling wireless network operators to optimize resource allocation, improve user experience, and meet the specific requirements of eMBB and URLLC applications.

INDEX TERMS 5G, DNN, DRL, RAN slicing, eMBB, URLLC.

I. INTRODUCTION

The 5th Generation (5G) network is revolutionizing human lives by stretching the performance bounds of mobile networks to support a variety of use cases. The industrial network has become more interesting due to the increasing demand for digitalization, and there is a huge prospect of its growth in years to come. However, due to the heterogeneity of the network, it is challenging to implement different types of services on existing networks. There are different types of requirements for different types of services, which include control for precision manufacturing and automation as it needs to meet the criteria of reliability and latency [1]. This makes resource allocation more challenging in meeting the requirements of these services due to the limitations of traditional networks [2]. Next-generation wireless networks (NGWN) can overcome these issues and provides a flexible environment where resources can be managed intelligently at a low cost. According to ITU Radio communication Sector (ITU-R), 5G is categorized into three services termed enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC) [3], [4]. The purpose of eMBB services is to serve high data rates applications such as augmented reality (AR), virtual reality (VR), and ultra high definition (UHD) video with tolerable

The associate editor coordinating the review of this manuscript and approving it for publication was Miguel López-Benítez.

reliability [5]. On the other hand, URLLC services focus on higher reliability and low-latency communications by transmitting shorter packets in length with a packet error rate (PER) in the range of 10^{-6}. It covers mission-critical applications such as vehicular communications, remote health services, and industrial automation. Packets are transmitted at shorter transmission time intervals (TTI) to meet the requirements of low latency. In Long Term Evolution (LTE) systems, latency is higher because control messages occupy a large part of the transmission. So, At the physical layer of 5G new radio (NR) systems, certain modifications are proposed to fulfill the requirements of URLLC [6]. Whereas mMTC's objective is to accommodate a massive number of Internet of Things (IoT) devices where each device can communicate with each other and the base station (BS) at a low data rate.

Generally, eMBB and URLLC services are mostly discussed in 5G networks [7]. We have considered these two services in our paper and proposed a novel approach to optimize their performance. Whereas, the pre-5G architecture does not support these two services [8]. The latency and reliability issues can be overcome by leveraging the concept of network slicing (NS) [9]. In NS, a single common physical channel is partitioned into multiple logical sub-networks, where each logical sub-network has its dedicated channel [10]. Better utilization of the resources can be achieved through logical separation and virtualization, which makes NS more flexible. Each logical sub-network has its radio access approach, and network virtualization functions (NVF). Radio access network (RAN) slicing is a type of NS that focuses on the RAN portion of the network. It allows operators to provide dedicated logical networks with customer-specific functionality without losing the economies of scale of a common infrastructure. The 5G RAN slicing helps operators manage the RAN resources needed for NS to operate. RAN slicing enables radio resources to be sliced in different ways, such as allocating different physical resource blocks (PRB) to different network slices.

In RAN slicing, resource management can be challenging due to the heterogeneity of the network. There are limited available radio resources to meet the requirements of URLLC and eMBB slices. In URLLC, packets need to have a short symbol length to meet the low-latency requirements. To meet the low-latency requirement, one option is to transmit the packet by reducing its symbol period. However, this approach is only suitable for mm-Wave bands because of the less delay spread due to smaller cell size [11]. Another possible method involves utilizing mini-slots with shorter TTIs by reducing the number of symbols to a minimum [11]. Slicing can be implemented on the RAN and core network (CN) parts. In this paper, we focus on the RAN part by intelligently allocating the resources to each user equipment (UE) according to the user demand. Note that the dynamic assignment of resources combined with the rapid user demands poses a formidable challenge. We utilize the above-mentioned features to design an intelligent and effective approach to overcome the resource allocation problem of URLLC and eMBB services.

The development of a dynamic framework of radio resource allocation in RAN slicing has become a primary focus for researchers [12], [13]. The heterogeneous traffic in the network requires an optimal resource management approach to meet the quality of service (QoS) requirements. The URLLC service cannot be held back due to its stringent low-latency requirement. The arriving URLLC traffic needs to be given priority over any ongoing eMBB transmission. To achieve this aim, two schemes have been proposed in the 3GPP standard [14]: *1) puncturing*, and *2) orthogonal scheduling*. In puncturing, to meet the latency requirement, BS ends the ongoing eMBB transmission and URLLC packets are scheduled in mini-slots over the already scheduled eMBB transmission. Through puncturing, low-latency requirements can be achieved, but it can affect the capacity and reliability of the system. Whereas in orthogonal scheduling, frequency channels are withheld before the URLLC transmission. The drawback of this scheme is that in case of no URLLC traffic the frequency channels reserved for URLLC transmission will not be utilized, which results in wastage of resources.

In this paper, we evaluate the puncturing technique to manage the radio resources in NS. As stated earlier, the instantaneous scheduling of URLLC transmission, which interrupts the eMBB traffic, has a significant impact on both system capacity and reliability. Furthermore, it leads to a degradation in the performance of the eMBB service. Hence, we develop an optimization-based approach to address the resource allocation problem, where it is important to not just focus on maximizing the capacity of the system, but also consider the reliability of the URLLC and eMBB services.

A. CONTRIBUTIONS AND CHALLENGES

It is challenging to handle the co-existence of eMBB and URLLC services over a common physical resource. In this paper, we get the better of this issue by using the puncturing technique. We proposed an efficient framework to ensure the capacity and reliability of the system while meeting the low latency requirement. This study addresses not only the problem of maximizing eMBB rates while meeting latency requirements but also investigates the influence of URLLC traffic on system capacity and reliability. There are some constraints associated with URLLC services due to the fact that URLLC-based services are optimized to operate independently of other services. URLLC systems are often optimized in a standalone manner, meaning that they are designed and implemented without considering other systems that may be operating in the same environment. It is challenging for this approach to manage the URLLC transmission characteristic in a dynamic network environment. Furthermore, in the worst case, it can break the URLLC reliability constraints in order to get the optimal

solution of the optimization problem, which can affect the QoS requirements. Due to the heterogeneous environment, randomness, and stringent requirements of URLLC traffic, the radio resources need to be allocated intelligently. ML-based algorithms such as semi-supervised and DRL can solve complex optimization problems in real-time in order to allocate the resources intelligently [15]. We apply the co-training method of semi-supervised learning in the strategy. In the existing co-training approach, a predefined policy fails to consider the sampling bias of the chosen samples between the labeled and unlabeled data samples [16]. Existing works related to resource management based on ML are largely dependent on labeled data samples. Though unlabeled data samples can be generated easily it requires a complex computational process to obtain the output of each data sample. Thus, we propose a novel framework to improve learning ability. In certain scenarios, acquiring labeled data for training DRL models can be challenging due to data sparsity. The Co-training DRL (CDRL) approach can mitigate this issue by leveraging a combination of labeled and unlabeled data. The labeled data can provide valuable information for training the model, while the unlabeled data can aid in discovering hidden patterns and improving generalization. This can be particularly beneficial in the context of eMBB RB allocation, where obtaining a large amount of labeled data might be difficult. In this work, we propose a CDRL approach to address the eMBB resource allocation problem. We present a novel CDRL approach based on q-learning, to improve the policy by choosing the unlabeled samples after taking the action at each TTI. We generate the labeled data through a two-sided matching technique, and use DRL with a semi-supervised based co-training method to predict the resource block for each user associated with eMBB slice. The implementation of DRL in URLLC poses challenges due to the stringent latency, and reliability requirements. Slow convergence of DRL can also be an issue in the implementation. We have considered all these challenges in our work and proposed a novel framework incorporating optimization-based techniques with semi-supervised learning and DRL to enhance the resource allocation capabilities for eMBB and URLLC traffic in 5G and beyond wireless networks. In this work, our key contributions are:

- Firstly, the resource allocation problem is expressed as an optimization-based problem, where we aim to maximize the sum rate of the eMBB service while fulfilling the URLLC constraints.
- Secondly, we decompose the problem into two sub-problems, consisting of eMBB resource block allocation strategy, and URLLC scheduling. Each sub-problem is treated separately depending on its framework in order to obtain the optimal solution.
- In the eMBB resource block allocation strategy, we propose the CDRL approach for resource block allocation, where we use DRL with co-training to predict the resource block for each user. To learn the best sample selection policy in co-training, we propose a q-learning approach, which utilizes the policy to train the model.
- In the URLLC scheduling sub-problem, we present a DRL-based DDQN approach to meet the latency and reliability requirements and to intelligently manage the URLLC traffic over the punctured eMBB slots. We propose the DDQN approach based on Thompson sampling to overcome the problem of slow convergence.
- Finally, we evaluate the performance of the proposed schemes. Simulation results demonstrate that the proposed methodology can ensure the URLLC reliability requirements while maintaining higher average sum rate for eMBB users.

Given the differing requirements of eMBB and URLLC, it is challenging to optimize resource allocation simultaneously to satisfy both types of services. The high data rate demands of eMBB may lead to increased latency and reduced reliability for URLLC services if resources are not allocated efficiently. Conversely, prioritizing URLLC requirements may result in under-utilization of resources and lower data rates for eMBB users. Optimizing each layer individually allows for fine-tuning and maximizing the performance metrics specific to that layer, enabling better performance for both eMBB and URLLC users. By customizing the optimization process, it becomes possible to enhance the performance metrics relevant to each layer, leading to improved outcomes for both types of users. By employing different DRL algorithms customized for specific traffic types, resource allocation efficiency can be enhanced. In this work, our aim is to propose an approach that can effectively converge to near-global optimal solutions or provide satisfactory performance in practical settings. Because a resource allocation in network slicing is a complex and multi-dimensional optimization problem. It involves numerous variables, constraints, and objectives. The solution space can be vast and non-linear, making it difficult to analytically derive global optimal solutions. The problem complexity and the presence of local optima can limit the assurance of reaching the global optimum. However, combining CDRL with DDQN and Thompson sampling to solve the coexistence problem of eMBB and URLLC users provide a way to leverage the strengths of each algorithm to achieve near-global optimal solutions.

We have organized this paper in the following manner. In section II, we review the related work before introducing our system model in section III including the URLLC data rate and eMBB data rate after puncturing. Further, the problem formulation is presented in section IV. Then, we present the proposed resource block allocation (RBA) strategy in section V, and an intelligent URLLC scheduling framework based on deep reinforcement learning (DRL) is presented in section VI. In section VII, the simulation results of the proposed algorithms have been presented. Section VIII presents the conclusion of the paper.

II. RELATED WORK

A. URLLC AND eMBB REQUIREMENTS

Extensive research work on the RAN resource management approach is being carried out in both industry and academia. Mainly, it focuses on how to develop an effective RAN resource management approach, and how to address the issues related to it. In [17], the authors presented a slicing approach for the LTE network to manage the resources efficiently, so the services can be provided to different mobile network operators (MNOs). A slicing and scheduling approach has been proposed in [18] to ensure services by allocating resource blocks (RBs) to each virtual network. For a single-cell orthogonal frequency-division multiple access (OFDMA) network, the authors proposed an effective sub-carrier and power allocation approach in [19]. Due to the limitations of the aforementioned works, it cannot meet the QoS requirements of the NGWN. With the advancement of applications, 5G systems needs to support a massive number of devices by meeting the strict low-latency requirements. In [6], the authors mentioned the main URLLC requirements and also highlighted its issues at the physical layer. In [20], the authors showed that overlapping URLLC traffic over eMBB transmission after every mini-slot can significantly improve the performance of a system in terms of resource efficiency. For the design of URLLC, the authors have discussed theoretical aspects such as massive MIMO, and medium access control (MAC) protocols in [21]. In [22], the constraints of URLLC have been discussed and future research direction of URLLC was given for the NGWN and termed eXtreme URLLC (xURLLC). To avoid transmission delay, the blocklength in URLLC should be finite. Whereas, Shannon's capacity theorem is applicable when blocklength is infinite. In [23], authors have analyzed the resource management problem for URLLC service given the achievable data rate in the context of finite blocklength. The optimization problem focuses on optimizing the power allocation and bandwidth allocation subject to the reliability and latency constraints.

In [24], the authors have proposed an approach for Vehicle-to-Vehicle (V2V) networks based on an optimization problem that aims to reduce the power subject to latency and reliability limitations. Here, they applied the extreme value theory and defined the reliability measure with regard to the maximum queue length paired between vehicles. The work in [25] evaluated the joint optimization of the V2V communications, where it aims to optimize the radio resources, modulation schemes, power control, and increase the capacity of cellular users while ensuring the stringent requirements of vehicle users in terms of latency and reliability. To solve the joint optimization problem and to reach the optimal solution, the authors have used binary search methods and Lagrange dual decomposition. The study conducted in [26] proposed an approach based on concurrent scheduling of URLLC and eMBB traffic, with the objective of maximizing the capacity available to eMBB users while simultaneously ensuring compliance with stringent latency and reliability requirements. The authors discussed the effect on eMBB service due to the incoming URLLC traffic.

The authors in [27] presented a proportional fairness scheme, where radio resources are allocated to incoming URLLC transmission while guaranteeing the reliability requirements of eMBB and URLLC services. In [28], the authors discussed eMBB and URLLC transmission services in terms of cloud RAN, where multi-cast and unicast transmissions are marked for eMBB and URLLC services, respectively. A general revenue-based maximization problem was presented as mixed-integer nonlinear programming for RAN slicing. The work in [29] proposed an approach for eMBB and URLLC services to find the optimal policy to the resource scheduling problem. For multiplexing of eMBB/URLLC traffic, authors in [30] studied the orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) schemes and discussed the trade-offs between them. The results are simulated with different decoding schemes such as puncturing and successive interference cancellation (SIC), and it shows that the OMA minimizes the interference among eMBB and URLLC traffic, but degrades the performance of the URLLC service. Whereas, NOMA with SIC scheme improves the URLLC performance while enhancing the capacity of eMBB service. In [27], a risk-sensitive framework was presented in order to manage the radio resources in NS. The resource allocation problem was specified as an optimization problem that aims to increase the capacity of the eMBB slice while considering the risk measure function. The aforementioned works do not discuss the impact of data transmission with URLLC requirements over eMBB slots, so we present an in-depth analysis and look to develop a dynamic resource allocation approach to schedule the URLLC and eMBB users effectively.

B. MACHINE LEARNING (ML) FOR RESOURCE MANAGEMENT

The allocation of radio resources in NS can be an issue and it can be resolved by implementing the machine learning (ML) algorithms [9], [31]. In wireless communication, supervised learning-based models such as deep neural network (DNN) have been used widely by researchers. DNN can solve complex problems and it helps in finding the optimal solution to an optimization problem. In [32], the authors proposed DNN based algorithm to manage the radio resources, where DNN was used to predict the transmit power policy. The work in [33] proposed a deep learning (DL) based algorithm to optimize the energy efficiency and spectrum efficiency in cognitive radio. The authors presented a convolutional neural network (CNN) based optimization problem in [34], which aims to determine the transmit power while maximizing the energy efficiency and spectrum efficiency with less computation time. Simulation results show that after training the model, the presented approach helps to predict the transmit power taking less computation time compared to other schemes. As it has been presented in the above-mentioned studies [32], [33], [34], the DNN strategy can be utilized

without distinctly finding the solution of the complex optimal control approach of the wireless network. DL strategy can be utilized as an intelligent tool to solve complex optimization problems in resource management, such as resource block allocation, power control, and scheduling. In a real-time environment, DL based resource management approach can determine the network and user state in the wireless network, which helps to manage the radio resources accordingly. This type of intelligent approach is very crucial to meet the URLLC requirements in 5G and beyond wireless systems [35], [36]. The label samples can affect the performance of the DL model. It is not quite difficult to get a large number of unlabeled samples in the ML-based approach towards resource management. However, in this scenario, there is a requirement to use more computation to obtain the result of each sample [37]. So, there is a need for an algorithm to enhance the learning performance and minimize dependency. In this paper, we present a novel approach based on semi-supervised DRL with a co-training method to solve the resource allocation problem. Labeled samples are taken from the predicted approach and trained, and incorporated with a large number of unlabeled data. Co-training is a semi-supervised learning approach, where two learners are initialized by the learner. It utilizes the estimated labels on the unlabeled data and samples are chosen based on the highest confidence. The wastage of resources can be evaded and the issue of poor generalization can be solved by applying the semi-supervised based approach.

In recent times, many studies are conducted in order to manage the radio resources by using the DRL [15]. In [38], the authors presented a framework based on actor-critic reinforcement learning (RL) to optimize power, resource allocation, and joint selection of transmission mode in V2V based device-to-device (D2D) enabled Internet of Vehicle (IoV) networks. It aims to increase the capacity of vehicle-to-infrastructure (V2I) nodes. The authors in [39] proposed a DRL-based framework to meet the URLLC requirements subject to power control and rate in the downlink of an OFDMA system. ML has been applied in various RRM tasks, such as spectrum sensing, channel prediction, interference management, and resource allocation. Another promising application of ML in RRM is to predict channel conditions and optimize transmission parameters. In a recent study, authors in [40] proposes a novel approach for resource allocation in RAN using Hierarchical Deep Learning (HDL) to meet the diverse QoS requirements of eMBB and URLLC services. The paper presents a novel approach for resource allocation in RANs using HDL, but it also has some limitations. The HDL model uses a two-level approach, with the first level performing resource allocation for eMBB and the second level optimizing resource allocation for URLLC. The proposed approach aims to maximize network utilization, minimize resource wastage, and ensure that the QoS requirements of both eMBB and URLLC services are met. However, the authors in [40] have not considered the impact of interference from neighboring cells, which may affect the QoS. Further, the proposed approach assumes a centralized resource allocation scheme, which is not be suitable for large-scale networks with distributed and dynamic traffic patterns.

The work in [41] presented a DRL-based deep q-learning framework for the co-existence problem of eMBB and URLLC. The existing works do not highlight the larger action space problem (i.e., increased number of possible actions at each time slot) while taking the decision about the allocation of RB. An agent starts exploring meaningless actions (e.g., actions that cannot fulfill the URLLC constraints), which results in a slow convergence rate that affects the performance of the DRL method. We proposed a novel framework incorporating a DRL-based double deep q-learning network (DDQN) with Thompson sampling to enhance the resource allocation capabilities for URLLC traffic in 5G and beyond wireless networks. Authors in [42] present a dynamic RL approach for resource provisioning in virtualized networks, specifically targeting D2D-based communications. The proposed framework adopts a three-stage layered structure, wherein the initial stage introduces a dynamic virtual resource allocation scheme based on DRL. In [43], authors propose an approach for autonomously provisioning and customizing resources in virtualized RAN to accommodate mixed traffics. The proposed scheme leverages a DRL algorithm to dynamically allocate resources based on the specific requirements of different traffic types. Existing studies have used the e-greedy approach as the exploration-exploitation strategy in DRL-based approaches to address the resource allocation problem in network slicing. Using Thompson Sampling in the DRL approach can potentially provide advantages over the greedy method [44]. Greedy methods typically focus on exploitation by always selecting the action with the highest estimated reward. While this can be effective in some cases, it may lead to sub-optimal solutions or being stuck in local optima. Whereas Thompson sampling uses a probabilistic approach where each action's selection is influenced by a distribution. This distribution allows for exploration by occasionally selecting sub-optimal actions to gather more information about their potential rewards. By exploring different options, Thompson Sampling can potentially discover better resource allocation strategies that may not be immediately apparent through a purely greedy approach. The network conditions, user demands, and traffic patterns may vary over time. Using Thompson Sampling enables the DRL agent to continually learn and adapt to these changing conditions. It can adjust its resource allocation decisions based on the most recent information, leading to improved performance and responsiveness. Efficient exploration can lead to quicker identification of optimal or near-optimal allocation strategies.

III. SYSTEM MODEL

We consider the downlink transmission scenario of heterogeneous network. In the considered scenario, the coverage area of a macro cell is populated with a random distribution of multiple small cells, and set of all BS is denoted by

TABLE 1. List of Abbreviations & Notations.

Notation	Definition
DRL	Deep reinforcement learning
eMBB	Enhanced mobile broadband
URLLC	Ultra-reliable low latency communication
TTI	Transmission time intervals
NS	Network slicing
RAN	Radio access network
RRM	Radio resource management
HDL	Hierarchical deep learning
CDRL	Co-training deep reinforcement learning
SINR	Signal-to-Interference-Noise-Ratio
DDQN	Double deep Q-learning
RBA	Resource block allocation
RB	Resource block
PGACL	Policy gradient actor-critic learning
SSL	Semi-supervised learning
DNN	Deep neural network
$\mathcal{B}$	Set of BS
$\mathcal{W}_b^e, \mathcal{W}_b^u$	Set of eMBB and URLLC users, respectively
K	Total number of mini-slots
N	Total number of RBs
$\xi_{n,k}^{b,w}(t)$	Puncturing decision variable
B	Initial assigned bandwidth to eMBB users
$a_{w,n}^b$	RB allocation strategy
$\zeta_{b,n}^{e,w}(t)$	SINR of eMBB users
$\zeta_{b,n}^{u,w}(t)$	SINR of URLLC users
$p_{b,n}^{e,w}$	eMBB Transmitted power
$p_{b,n}^{u,w}$	URLLC Transmitted power
$g_{b,n}^{e,w}$	eMBB Channel gain
$g_{b,n}^{u,w}$	URLLC Channel gain
$r_{b,n}^e(t)$	Sum rate of eMBB users
$r_{b,n}^{u,w}(t)$	Sum rate of URLLC users
$Y_{b,n}^{u,w}$	Dispersion of the channel
$\psi(t)$	Total number of URLLC data that arrived at TTI t
$Q^\pi(s_t, a_t)$	Cumulative discounted reward of policy π
Cov_{a_t}	Co-variance for every action a_t

$\mathcal{B} = \{1, 2, \ldots, b, \ldots, |\mathcal{B}|\}$. We focus on two kinds of downlink requests, eMBB and URLLC. As shown in Fig.1, there are different kinds of UEs such as AR/VR, smart transportation, and smartphones scattered randomly and connected to each BS. In this model, several edge servers are placed at the edge of a network, and these edge servers are linked to a larger centralized cloud server. The set of eMBB and URLLC users present in the network can be denoted as $\mathcal{W}_b^e = \{1, \ldots, W_b^e\}$, and $\mathcal{W}_b^u = \{1, \ldots, W_b^u\}$, respectively. The available radio resources in 5G-NR can be presented in frequency and time domain, whereas frequency and time domain is divided into a number of N radio resources or RB, where each RB has a bandwidth defined as B in the frequency domain. The set of RBs can be defined as $\mathcal{N} = \{1, \ldots, n, .., N\}$. In the time domain, every TTI has a duration of 1 ms, so in one time slot, there are total N number of available RBs. Each RB has 7 symbols and consists of 12 sub-carriers, so there are 84 resource elements (RE) in a single RB. The available time slot is further split into K smaller units called short TTI or mini-slots. Generally, to enhance the SE, eMBB service spans multiple TTIs. Due to stringent latency requirements, the incoming URLLC traffic cannot be held back during the ongoing eMBB communication service.

So, we puncture the eMBB slots and transmits the URLLC traffic. In this regard, we schedule URLLC service at short TTI (duration of 0.5 ms), and eMBB service with the duration of 1 ms. Because of this, the instantaneous scheduling of URLLC transmission, which involves interrupting eMBB traffic, can have a significant impact on both the system's capacity and reliability, leading to a degradation in the performance of the eMBB service. So, a proper framework is required to meet the QoS requirements.

A. eMBB THROUGHPUT

Transmitting the URLLC traffic over the punctured eMBB slots can affect the bit rate of eMBB services. We introduce a decision variable for the purpose of puncturing as stated below.

$$\xi_{n,k}^{b,w}(t) = \begin{cases} 1, & \text{if the } k^{th} \text{ mini-slot is punctured by the } w^{th} \\ & \text{URLLC user, } \forall\, n \in \mathcal{N}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

There's a problem in allocating the total number of radio resources to the users, because each RB needs to be assigned to active user. We assume that one RB of each BS is occupied by a single user. Mathematically RBA strategy can be represented as:

$$a_{w,n}^b(t) = \begin{cases} 1, & \text{If the RB } n \text{ of BS } b \text{ is assigned to the} \\ & \text{eMBB user } w, \forall\, b \in \mathcal{B}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The signal-to-noise-and-interference-ratio (SINR) of the eMBB user w can be computed as:

$$\zeta_{b,n}^{e,w}(t) = \frac{p_{b,n}^{e,w}(t) g_{b,n}^{e,w}(t)}{\underbrace{\sum_{\substack{b' \in \mathcal{B} \\ b' \neq b}} p_{b',n}^{e,w}(t) g_{b',n}^{e,w}(t)}_{\text{eMBB interference}} + \underbrace{\sum_{\substack{b' \in \mathcal{B} \\ b' \neq b}} p_{b',n}^{u,w}(t) g_{b',n}^{u,w}(t)}_{\text{URLLC interference}} + \sigma^2}, \tag{3}$$

where $p_{b,n}^{e,w}(t)$, and $g_{b,n}^{e,w}(t)$ represents the transmitted power and channel gain, respectively, of eMBB user w of BS b over RB n, and σ^2 is the noise power. The throughput of an eMBB user w of BS b on RB n at time slot t can be approximated as:

$$r_{b,n}^{e,w}(t) = B\left(1 - \frac{\sum_{k=1}^{K} \xi_{n,k}^{b,w}(t)}{K}\right) \log_2\left(1 + \zeta_{b,n}^{e,w}(t)\right), \tag{4}$$

where the term $\frac{\sum_{k=1}^{K} \xi_{n,k}^{b,w}(t)}{K}$ represents the loss of eMBB rate due to puncturing. Thus, the total sum rate achieved by the eMBB user w can be expressed as:

$$r_{b,w}^e(t) = \sum_{n \in \mathcal{N}} a_{w,n}^b(t) r_{b,n}^{e,w}(t). \tag{5}$$

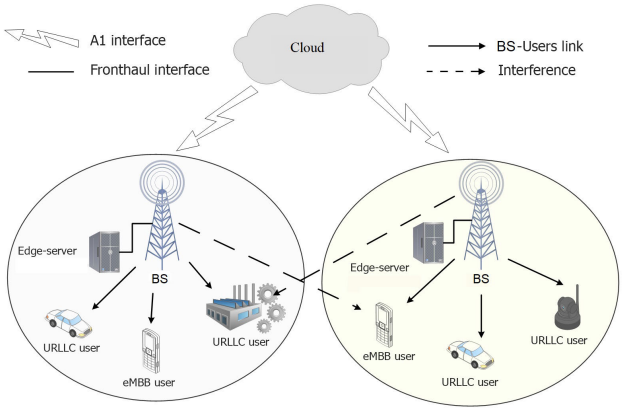


FIGURE 1. System model.

B. URLLC THROUGHPUT

To avoid transmission delay, the blocklength in URLLC should be finite. Whereas, Shannon's capacity theorem is applicable when blocklength is infinite. In [23], the authors have analyzed the resource management problem for URLLC service given achievable data rate in the finite blocklength regime. Thus, the work in [45] describes the achievable data rate in URLLC for finite blocklength as follows,

$$r_{b,n}^{u,w}(t) = \sum_{n\in\mathcal{N}} B_n\Big(\frac{\sum_{k=1}^{K}\xi_{n,k}^{b,w}(t)}{K}\Big)\Big[\log_2\left(1+\zeta_{b,n}^{u,w}(t)\right) - \sqrt{\frac{Y_{b,n}^{u,w}}{v_{b,n}^{u,w}(t)}}.Q^{-1}(x)\Big], \tag{6}$$

where $\zeta_{b,n}^{u,w}(t)$ refers to the SINR of URLLC user, expressed as

$$\zeta_{b,n}^{u,w}(t) = \frac{p_{b,n}^{u,w}(t)g_{b,n}^{u,w}(t)}{\underbrace{\sum_{\substack{b'\in\mathcal{B}\\ b'\neq b}} p_{b',n}^{u,w}(t)g_{b',n}^{u,w}(t)}_{\text{URLLC interference}} + \underbrace{\sum_{\substack{b'\in\mathcal{B}\\ b'\neq b}} p_{b',n}^{e,w}(t)g_{b',n}^{e,w}(t)}_{\text{eMBB interference}} + \sigma^2}. \tag{7}$$

Here, $Y_{b,n}^{u,w}$ indicates the dispersion of the channel, and determines the channel randomness of user, and can be represented as:

$$Y_{b,n}^{u,w} = 1 - \frac{1}{\left(1+\zeta_{b,n}^{u,w}(t)\right)^2} \tag{8}$$

The number of symbols in each mini-slot is represented by $v_{b,n}^{u,w}(t)$, and $Q^{-1}(x)$ represents the Gaussian inverse CDF Q-function, where x indicates the error rate.

IV. PROBLEM FORMULATION

In this paper, we aim to maximise the sum rate of the eMBB user, while fulfilling the URLLC constraints. The resource allocation problem is formulated as an optimization-based problem. In the beginning, we assign transmission power and RBs to eMBB UEs at each TTI. We assume that total power is equal over all sub-carriers. Then, we puncture the eMBB slots and transmit the URLLC traffic over them. Puncturing can affect the capacity and reliability of the system. So, we propose a new approach to maximize the eMBB rate subject to URLLC constraints while minimizing the effect on eMBB reliability. For URLLC users, we suppose that the users create small packets fragments, and the packet arrival rate at mini-slot $k \in \mathcal{K} = \{1, \ldots, k, .., K\}$ at TTI t follows a Poisson point process (PPP) distribution. We denote the number of arrived small packets with a random variable $\psi_k(t)$ such that

$$\psi(t) = \sum_{k\in\mathcal{K}} \psi_k(t) \tag{9}$$

$\psi(t)$ indicates the total number of URLLC packets that arrived at TTI t. Thus, the reliability of URLLC service can be obtained by the following equation:

$$P\left[\sum_{w\in\mathcal{W}_b^u} r_{b,n}^{u,w}(t) \leq \kappa\psi(t)\right] \leq \eta_u, \quad \forall b \in \mathcal{B} \tag{10}$$

where κ refers to the packet size of URLLC service. The above equation indicates the outage probability should not exceed the threshold value η. So, the optimization problem of joint resource allocation of eMBB and URLLC can be mathematically formulated as follows:

$$P: \max_{a,\xi} \left\{\sum_{w\in\mathcal{W}_b^e} r_{b,w}^{e}\right\} \tag{11a}$$

$$\text{subject to} \sum_{w\in\mathcal{W}_b^e} a_{w,n}^{b}(t) \leq 1, \quad \forall n \in \mathcal{N}, b \in \mathcal{B} \tag{11b}$$

$$\sum_{w\in\mathcal{W}_b^u} \xi_{n,k}^{b,w}(t) \leq 1, \quad \forall n \in \mathcal{N}, b \in \mathcal{B} \tag{11c}$$

$$P\left[\sum_{w\in\mathcal{W}_b^u} r_{b,n}^{u,w}(t) \leq \kappa\psi(t)\right] \leq \eta_u, \ \forall b \in \mathcal{B} \tag{11d}$$

$$a_{w,n}^{b}(t) \in \{0,1\}, \quad \forall w \in \mathcal{W}^e, n \in \mathcal{N} \tag{11e}$$

$$\xi_{n,k}^{b,w}(t) \in \{0,1\}, \quad \forall w \in \mathcal{W}^u, n \in \mathcal{N} \tag{11f}$$

where (11a) aims to maximize the eMBB rate. Constraint (11b) indicates the RB allocation limitations, and it ensures that only a single user should be associated with a RB. Whereas, (11c) represents the puncturing constraint. Constraint (11d) guarantees the URLLC reliability. The key objective is to execute dynamic allocation of the resources in order to increase the capacity (in terms of sum rate) of the eMBB users subject to different constraints. It can be observed in (11) that optimization problem **P** is a NP-hard non-convex problem, and it is challenging to find the optimal solution in general. There is a requirement for an intelligent approach for solving this optimization problem. The resource allocation approaches for URLLC and eMBB services are different. URLLC services need to meet the low-latency requirements and also prioritized access to the network, while

eMBB services require high data-rate and optimized network utilization. These differences in resource allocation strategies make it difficult to optimize both services simultaneously, and decomposing the optimization problem can help to optimize each service's resource allocation separately. To find the optimal solution to the resource allocation optimization problem, we break the problem **P** into two sub-problems, *P1: RB allocation for eMBB slice*, and *P2: URLLC scheduling*.

V. RB ALLOCATION STRATEGY FOR eMBB SLICE

The RB allocation problem can be expressed as:

$$P1: \max_{a} \left\{ \sum_{w \in \mathcal{W}_b^e} r_{b,w}^e \right\} \tag{12a}$$

$$\text{subject to} \sum_{w \in \mathcal{W}_b^e} a_{w,n}^b(t) \leq 1, \quad \forall n \in \mathcal{N}, b \in \mathcal{B} \tag{12b}$$

$$a_{w,n}^b(t) \in \{0,1\}, \quad \forall w \in \mathcal{W}^e, n \in \mathcal{N} \tag{12c}$$

We propose a novel CDRL approach, where we use DRL with a semi-supervised based co-training method to predict the resource block for each user associated with eMBB slice. First, we modify the **P1** in (12) into a loss function and then achieve the optimal solution of the RB allocation by minimizing the loss function such that:

$$\min_{\hat{A}} \|\hat{A} - \arg\max r_{b,w}^e\|^2$$

$$\text{subject to} \sum_{w \in \mathcal{W}_b^e} a_{w,n}^b(t) \leq 1, \quad \forall n \in \mathcal{N},\ b \in \mathcal{B}$$

$$a_{w,n}^b(t) \in \{0,1\}, \quad \forall w \in \mathcal{W}^e,\ n \in \mathcal{N} \tag{13}$$

where $\hat{A}$ indicates the forecasted RB allocation strategy. In Algorithm 1, we have presented the two-sided matching approach in order to produce the initial RB allocation strategy. RBs and different users associated with different slices are considered as two contestants seeking the maximization of their specific objective function. Co-training is a semi-supervised learning method where two models are trained by utilizing a large number of unlabeled data. From Algorithm 1, we have generated the labeled data which consists of gain values $g_{b,n}^{u,w}(t)$ and RB allocation strategy $a_{w,n}^b(t)$. Algorithm 1 based on two-sided matching technique serve as an initial RB allocation mechanism for the CDRL approach for eMBB RB allocation. The initial RB allocation based on the two-sided matching technique provides a foundation for further optimization by providing an initial allocation of RBs based on user preferences, which can then be refined and optimized using the CDRL approach. The CDRL algorithm can learn from the initial RB allocation and user feedback to improve the RB allocation policy over time. By continuously interacting with the environment and optimizing the allocation based on the learned policy, the CDRL approach can improve the RB allocation efficiency and adapt to changing network conditions. The parameters $\Omega_{w,n}^b$ and Ω_w^b indicates the users assigned to RB (n) and the users of unallocated RBs, respectively.

Algorithm 1 Initial RB Allocation Strategy Based on Two-Sided Matching Method

RB allocation A is initialized
for a BS b from the set of BS $\mathcal{B}$ **do**
  Initialize $\Omega_{w,n}^b$ as users assigned to RB(n)
  Specify Ω_w^b for users of unallocated RBs
  while $\Omega_w^b \neq \{\}$ **do**
    for users **do**
      Select the RB (n) with the highest signal-to-interference-noise-ratio (SINR) based on channel quality indicator (CQI)
      if $\Omega_{w,n}^b = 1$ **then**
        $a_{w,n}^b(t) = 1$
        Update $\Omega_{w,n}^b$ and Ω_w^b
      end if
      if $\Omega_{w,n}^b = 2$ **then**
        The utility function of the two users assigned to RB(n) needs to be calculated
        Choose the users that increases the sum rate of the RB
        Update $\Omega_{w,n}^b$ and Ω_w^b
      end if
    end for
  end while
end for

It can be presented in matrix form as follows:

$$d_l = \{(G_1, A_1), (G_2, A_2), \ldots, (G_l, A_l)\}, \tag{14}$$

whereas, unlabeled data can be presented as:

$$d_u = \{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_l\}, \tag{15}$$

where l is the number of data samples. Our main objective is to predict the label value A from unlabeled data G. The existing co-training method is based on a policy of choosing the samples which have high-confidence values. In this paper, we have used DRL based q-learning approach to improve the policy by choosing the unlabeled samples after taking the action a_t at each TTI. First, we decompose the unlabeled samples into various sub-samples according to their similar traffic behavior. The DRL agent employs a policy to chose one sub-sample at each TTI t instead of selecting one sample, which can enhance the computational efficiency and reduce the latency, and then the two learners are updated. The decomposition of unlabeled samples can be presented as follows:

$$\mho_u = \{\mho_1, \mho_2, \ldots, \mho_j\}, \tag{16}$$

where j is the number of data samples. First, the two learners are trained with a small amount of labeled data d_l at the start of training. At each TTI, the DRL agent takes a decision (action), and then the unlabeled sub-samples are chosen to train the learners. The backbone of our proposed model is the q-learning approach, where best quality unlabeled

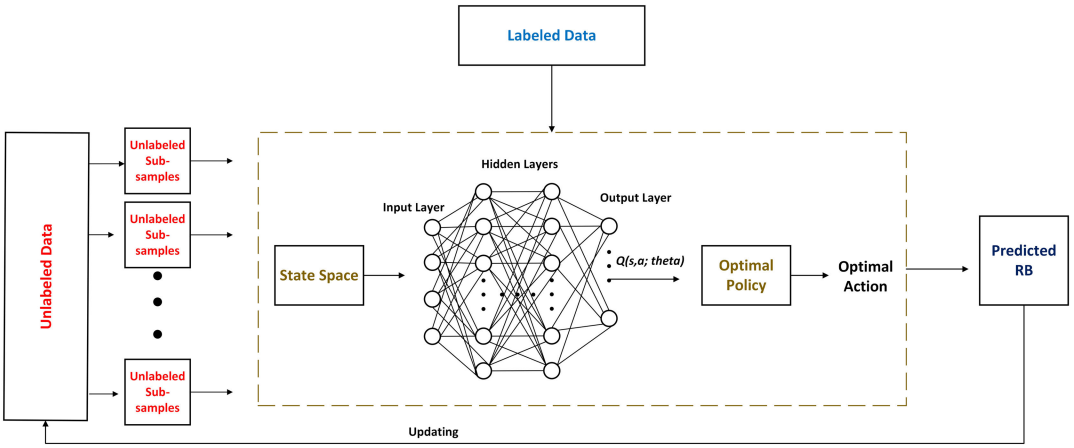


FIGURE 2. CDRL framework.

sub-samples are chosen for co-training by the agent after understanding the optimal policy through training. The state-space s_t is observed by the agent at each TTI t and takes the best possible action a_t, and then the two learners Z_1 and Z_2 are updated with $\mho_u$. Our objective is to train the learner Z, which can accurately predict the RB allocation such that

$$Z : G \rightarrow A$$

Let's assume that $Z(g/\theta)$ indicates the Gaussian distribution, then the distribution can be presented as [37]:

$$Z(g/\theta) = \sum_{i=1}^{d_l} \delta_i Z(g/\theta_i) \tag{17}$$

There is a parallel vector latent variable against each data sample. This variable is determined by the mixture coefficient δ_i, and corresponding component h can be obtained by using this coefficient. The probability of h and g can be given as $P(v_i)|g_i, h_i$. The optimal classification can be formulated as:

$$Z(g) = \max_k \sum_j P(v_i = j|g_i, h_{i=j}) P(h_i = j|gi), \tag{18}$$

where;

$$P(h_i = j|g_i) = \frac{\delta_j Z(g_i|\theta_j)}{\sum_{i=1}^{d_l} \delta_i Z(g_i/\theta_{d_l})}$$

From the above equation, it can be observed that training samples can be used to predict $Z(g)$.

A. STATE SPACE

The agent should be well familiar with the distribution of the unlabeled sample in order to choose the best sub-samples. We examine the probability distribution of two learners and it can be formulated as:

$$s_t = Cat(\beta|\gamma) \tag{19}$$

$$p_\theta(g|s_t, h) = f(g; s_t, h, \theta) \tag{20}$$

where β and γ represents the probability distribution of two learners Z_1 and Z_2 respectively, and Cat indicates the concatenation operation. Whereas, $f(g; s_t, h, \theta)$ represents a nonlinear likelihood function.

Algorithm 2 CDRL Approach for eMBB RB Allocation

1: **Input**: Labeled samples of RB allocation d_l, labeled validation samples d_l', unlabeled sub-samples $\mho_u$;
2: **for** $t = 1$ to T **do**
3: **for** $j = 1$ to 2 **do**
4: Train $Z_{t,j}$ with labeled samples d_l;
5: Take action $a_t = \max_a Q(s_t, a)$;
6: Use Z_1 to label the sub-samples $\mho_u'$;
7: Upgrade Z_2 with pseudo-labeled sub-samples $\mho_u'$, and labeled samples d_l;
8: Z_2 is used to label the sub-samples $\mho_u'$;
9: Upgrade Z_1 with pseudo-labeled sub-samples $\mho_u'$, and labeled samples d_l;
10: Determine the reward r_t based on validation labeled samples d_l';
11: Determine s_{t+1};
12: Update parameter θ;
13: Compute loss function using (32);
14: **end for**
15: **end for**

B. ACTION SPACE

The q-learning agent chooses the best possible action by choosing the best quality unlabeled sub-samples at TTI t after learning the optimal policy such that

$$a_t = \max_a Q(s_t, a) \tag{21}$$

C. REWARD

The reward of each learner can be formulated as:

$$r_t = \begin{cases} r_1 \times r_2, & \text{if } r_1 \ \& \ r_2 > 0, \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

where r_1 and r_2 represent the model accuracy of learners Z_1 and Z_2, respectively determined on the labeled testing data

samples at TTI t. The agent aims to take a decision or action a_t at each TTI t which can increase the future discount reward.

$$R_t = \sum_{t=1}^{\tau} \Lambda^t r_t, \tag{23}$$

where Λ refers to the discount factor. The main focus is to maximize the reward R_t by finding an optimal policy. The q-agent in the q-learning network will learn the optimal policy by interacting with the two learners which act as the environment. The loss function can be presented as:

$$Loss(\theta_i) = \mathbb{E}_{s,a}\Big[(Y(\theta_{i-1}) - Q(s, a; \theta_i))^2\Big] \tag{24}$$

where,

$$Y(\theta_{i-1}) = \mathbb{E}_{s'}\Big[(r + \Lambda \max_{a'} Q(s', a'; \theta_{i-1}|(s, a))\Big] \tag{25}$$

The above equation indicates that θ learn the optimal policy by using the gradient descent method. During testing, the two learners Z_1 and Z_2, and the q-learning agent were simulated together without the labeled validation samples. The agent learns the optimal policy and takes action a_t and chooses the unlabeled sub-samples. Finally, learner Z can be defined as:

$$Z = \varphi Z_1 + (1 - \varphi)Z_2 \tag{26}$$

where φ indicates the weight factor based on learning policy. The detail is provided in Algorithm 2. A depiction of CDRL framework is presented in Fig. 2.

VI. URLLC SCHEDULING

Due to the heterogeneity of URLLC traffic, it has become essential to intelligently and dynamically assign the radio resources to the incoming URLLC traffic. Thus, we present a DRL-based URLLC scheduling approach to manage the radio resources for the incoming URLLC traffic. We can state the URLLC scheduling problem as follows:

$$P2: \max_{\xi} \left\{ \sum_{w \in \mathcal{W}_b^e} r_{b,w}^e \right\} \tag{27a}$$

$$\text{subject to} \sum_{w \in \mathcal{W}_b^u} \xi_{n,k}^{b,w}(t) \leq 1, \quad \forall n \in \mathcal{N}, b \in \mathcal{B} \tag{27b}$$

$$P\Big[\sum_{w \in \mathcal{W}_b^u} r_{b,n}^{u,w}(t) \leq \kappa \psi(t)\Big] \leq \eta_u, \quad \forall b \in \mathcal{B} \tag{27c}$$

$$\xi_{n,k}^{b,w}(t) \in \{0, 1\}, \quad \forall w \in \mathcal{W}^u, n \in \mathcal{N} \tag{27d}$$

The URLLC scheduling obtained by the CDRL algorithm can not fulfill the low latency and reliability reliability constraint due to the slow convergence of DRL-based q- network. So, we use CDRL approach for eMBB resource allocation and propose a novel approach based on Thompson sampling for URLLC scheduling. We ensure that the constraints (27b-27d) meets the requirements while actively engaging with the environment. In this algorithm, we present a DDQN based approach to meet the latency and reliability requirements and to intelligently manage the URLLC traffic over the punctured eMBB slots. A RL model is described by action, state, and reward.

State-Space: We define the state space s_t by describing the throughput of each user of BS b associated with eMBB service without puncturing depending on the channel gain, allocated RBs, incoming URLLC traffic, and transmission power. So, the throughput of each user associated with eMBB service without puncturing can be presented as:

$$\hat{r}_{b,w}^e(t) = \sum_{n \in \mathcal{N}} a_{w,n}^b(t) B \log_2\big(1 + \zeta_{b,n}^{e,w}(t)\big).$$

Thus, the state space s_t can be defined as:

$$s_t = \big[\hat{r}_{b,w}^e(t), g(t), \psi(t)\big] \tag{28}$$

where $\psi(t)$ is defined in (9) and $g(t)$ is channel gain.

Action Space: The action space a_t can be described as the $N \times K$ puncturing matrix which indicates the K number of mini-slots within each RB that have been punctured.

$$a_t = \{\xi_{n,k}^{b,w}(t), \quad \forall n \in \mathcal{N}, w \in \mathcal{W}_b^e\} \tag{29}$$

Reward: Considering the QoS requirements of different slices and associated applications, we present a reward function which can be given as:

$$r_t = \overbrace{\big(\max_{\xi} \sum_{w \in \mathcal{W}_b^e} r_{b,w}^e\big)}^{I} - \vartheta(t)\overbrace{\big(\sum_{w \in \mathcal{W}_b^u} r_{b,n}^{u,w}(t) - \kappa\psi(t)\big)}^{II} \tag{30}$$

where $\vartheta(t)$ indicates the time varying weight coefficients of part II. We introduce this coefficient to ensure the URLLC reliability constraint. The following equation can be used to describe it:

$$\vartheta(t+1) = \max\{\vartheta(t) + \eta(t) - \eta_u, 0\}, \tag{31}$$

where $\eta(t)$ represents the achieved outage probability as stated in (10). Part I represents the eMBB rate we want to maximize, whereas part II indicates the URLLC constraint. The agent aims to select an optimal policy π in order to increase the reward, which means with the lowest outage probability and the highest sum rate are achieved. The policy $\pi = \pi_a^K$ can be defined as the given network state s_t observed by the agent and the agent takes action a_t on the number of punctured mini-slots K from each allocated RB a. Then by using (30), the reward is calculated by the agent based on decisions taken, and new state information of the network is given to the agent. Let us assume that $Q^\pi(s_t, a_t)$ indicates the q-function, the cumulative discounted reward for the given network state with a policy π can be formulated as:

$$Q^\pi(s_t, a_t) = \mathbb{E}\Big[\sum_{t=1}^{\infty} \Lambda(t) r_t(s_t, a_t)|s_0 = s_t, \pi\Big] \tag{32}$$

where $\Lambda(t)$ and s_0 represents the discount factor and initial state, respectively. The above function only takes the current

reward into account. According to [46], it can be rewritten as:

$$Q^{\pi}(s_t, a_t) = r_t(s_t, a_t) + \sum_{t=1}^{\infty} \Lambda Q^{\pi}(s_{t+1}, a_{t+1}) \tag{33}$$

A DNN is used for the approximation of the above function. The main objective of the earlier mentioned approach is to find the optimal policy π which can increase the reward. The optimal policy π can be expressed as follows:

$$\pi = \max Q^{\pi}(s_t, a_t) \tag{34}$$

To optimize the policy π in (34), different RL techniques can be employed such as policy gradient and q-learning. Therefore, the work in [47] shows that the q-learning technique converges slowly and it is hard for it to solve the optimal policy. Whereas, policy gradient method results in high variance and converges to a local optimum. Thus, we propose the DDQN method with Thompson sampling to learn the policy which results in a faster convergence rate.

A. DDQN WITH THOMPSON SAMPLING

We present the Thompson sampling method with DDQN in order to improve the convergence rate and balance the exploitation and exploration. The Thompson sampling is based on probability-based exploration, where the agent takes an action randomly depending on the best probability. Thompson sampling is a very effective and efficient method in the context of exploitation and exploration, because the agent never selects the actions with less probability, and avoids consuming time on meaningless explorations which result in a faster convergence rate [48]. Therefore, combining the DDQN with Thompson sampling results in reliable and effective resource management for URLLC traffic. It helps in handling large state spaces as it avoids exhaustive exploration of the entire space. Only actions with higher probabilities of being optimal are more likely to be selected. By repeatedly sampling and selecting actions based on the estimated probabilities, the algorithm gradually learns which actions are more likely to yield better results. In our previous work [49], we employed Thompson sampling to enhance network efficiency and fulfill the stringent URLLC requirements within a resource-constrained and highly dynamic V2X (Vehicle-to-Everything) environment. DDQN method was proposed by Hasselt [50] to solve the overestimation problem in q-learning. There are two different DNN utilized by the DDQN: 1) deep q network (DQN), and 2) target network. It can be mathematically expressed as follows:

$$y \leftarrow r_{t+1} + \Lambda \hat{Q}^{\pi}(s_{t+1}, \hat{a}) \tag{35}$$

where,

$$\hat{a} = \max_{a} Q^{\pi}_{DQN}(s_{t+1}, a) \tag{36}$$

Further, $\hat{Q}^{\pi}(s_{t+1}, \hat{a})$ refers to the target network where DQN chooses the maximum Q-value by taking the best action a of the next state. Then the target network $\hat{Q}$ calculates the approximated Q-value by taking action $\hat{a}$. The Q-value of DQN is updated based on the approximation from the target network $\hat{Q}$. Then the parameters of the $\hat{Q}$ are updated based on the DQN parameters. The architecture of DDQN comprises a DNN where the Q-value indicates a linear function. Thus, for any network state s_t and action a_t, it can be expressed as follows:

$$Q^{\pi}(s_t, a_t) = \phi_{\theta}(s_t)^T \omega_{a_t} \tag{37}$$

where ω_{a_t} and $\phi_{\theta}(s_t)$ denote the weight of the last layer and linearity of the output layer parameterized by θ, respectively. Similarly, the output layer and weight of the target network can also be represented by the $\phi_{\hat{\theta}}(.)$ and $\hat{\omega}_{a_t}$, respectively. Further, (35) and (36) can be rewritten as:

$$Q^{\pi}(s_t, a_t) = \phi_{\theta}(s_t)^T \omega_{a_t} \rightarrow y = r_{t+1} + \Lambda \phi_{\hat{\theta}}(s_{t+1})^T \hat{\omega}_{a_t} \tag{38}$$

where,

$$\hat{a_t} = argmax_a \phi_{\theta}^T \omega_{a_t} \tag{39}$$

The loss function can be computed as:

$$\nabla(Q^{\pi}, \hat{Q}^{\pi}) = \mathbb{E}\Big[\big(Q^{\pi}(s_t, a_t) - \hat{Q}^{\pi}(s_t, a_t)\big)^2\Big] \tag{40}$$

We employed Gaussian Bayesian linear regression in order to approximate the posterior on the weight of the last layer and the Q-network function. In this paper, we estimate the distribution by using Gaussian Bayesian linear regression over the Q-values and formulate an effective and balance exploration-exploitation scheme by utilizing Thompson sampling. The posterior distribution is estimated as:

$$\overline{w}_{a_t} = \frac{1}{\wp^2} Cov_{a_t} \Psi^{\theta}_{a_t} \hat{Q}^{\pi}(s_t, a_t) \tag{41}$$

$$w_{a_t} \sim \mathcal{M}(\overline{w}_{a_t}, Cov_{a_t}) \tag{42}$$

where $\overline{w}_{a_t}$ and $\wp$ indicate the mean and variance of likelihood, respectively. Through (42) the agent employs Thompson sampling to sample w_{a_t} around mean $\overline{w}_{a_t}$ and co-variance Cov for every decision a_t. DDQN agent keeps the prior and at the beginning of each TTI updates the posterior and extracts weight of the last layer and follows the optimal policy π. The training details are given in Algorithm 3. Fig. 3 shows the block diagram of the proposed framework.

Initially, the BS assigns RBs to eMBB service users according to the optimal policy obtained by the CDRL approach. Then it sends the state space to Algorithm 3. The experience replay buffer of the proposed Algorithm 3 is initialized based on the results obtained by the CDRL approach. This information can serve as input to the URLLC RB allocation decision-making process. Then, Algorithm 3 which is based on Thompson sampling chooses an action based on its observed environment, and perceive the immediate reward r_t and next state $s(t+1)$, and accumulates the state space, action, reward and next state in the experience replay buffer. The transfer of states between the eMBB and URLLC components facilitates a collaborative learning process. It allows the components to leverage relevant

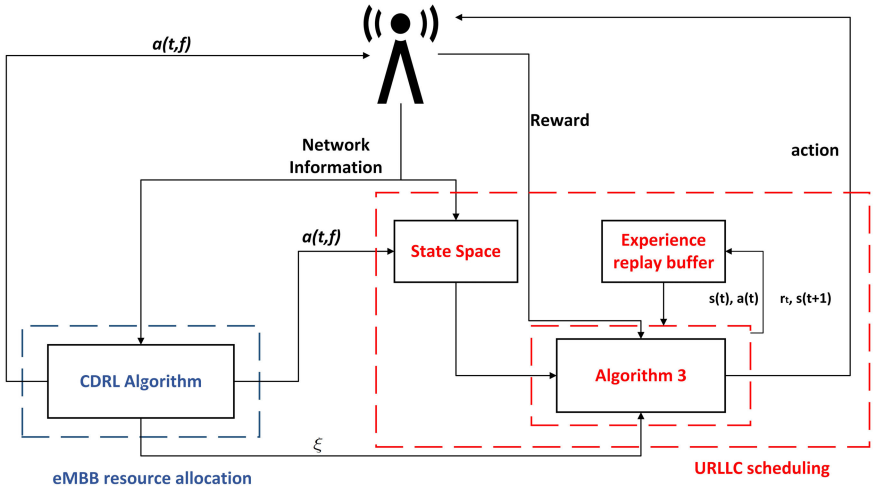


FIGURE 3. Block diagram of the proposed framework.

Algorithm 3 DDQN With Thompson Sampling for Intelligent URLLC Scheduling

Initialize Q-value function
Initialize θ, $\hat{\theta}$, Cov_a, ω_{a_t}, and $\hat{\omega}_{a_t}$
Set replay memory=$\emptyset$
for each TTI **do**
  Observe the network state $s_t = \left[\hat{r}^e_{b,w}(t), g(t), \psi(t)\right]$
  Samples a Q-function
  if t mod posterior update period=0 **then**
    Update mean $\overline{w}_{a_t}$ and co-variance Cov_a of posterior distribution
  end if
  if t mod posterior sampling period=0 **then**
    Extract samples using (42)
  end if
  Set $\hat{\theta} \leftarrow \theta$ after every target update
  Execute using (39)
  Store transition (s_t, a_t, r_t, s_{t+1}) in replay memory
  if replay memory is full **then**
    Sample a mini-batch from the replay memory
  end if
  Update $\hat{Q}^{\pi}(s_t, a_t)$ using (38)
  Update parameter θ by minimizing a loss function
end for

information from each other to improve the overall RB allocation performance and ensure the specific requirements of both eMBB and URLLC users are considered. Finally, the weight coefficient value $\vartheta(t)$ is updated.

VII. PERFORMANCE ANALYSIS

We show the performance of our proposed algorithms in this section through inclusive empirical analysis for different parameters. The network dynamics are modeled by considering factors which includes channel conditions, interference levels, traffic variations, and resource utilization.

TABLE 2. Simulation parameters.

Parameter	Value
Cell radius	4 km
Transmission power	40 dBm
No. of users	22,500
Total system bandwidth	20 MHz
TTI duration	1 ms
Length of time frame	10 ms
No. of RE	84
Bandwidth of each RB	180 kHz
No. of symbols in RB	7
No. of sub-carriers in RB	12
Channel	Rayleigh fading
Packet size of URLLC	32
Packet size of eMBB	Infinite
Traffic model of URLLC	PPP
Traffic model of eMBB	Full-buffered

The model can simulate the evolution of these factors over time, enabling the DRL agent to observe and learn from the network dynamics during the training process. The dynamics of the network model are incorporated into the DRL training by allowing the DRL agent to observe the current state of the network, take actions, and observe the resulting state transitions and rewards. By interacting with the model, the DRL agent can learn to make optimal resource allocation decisions in response to changes in the network dynamics. The Thompson sampling algorithm can adaptively explore, and exploit actions based on their estimated probabilities of being optimal. This allows the algorithm to dynamically adjust its scheduling decisions in response to changes in network conditions and requirements. In this work, we evaluate our results by comparing them with different approaches such as PGACL [11]: a risk-averse based approach to increase the reliability, Q-learning, DQN, optimal approach, and random search. PGACL achieves policy learning with a rapid convergence rate by integrating policy and value learning. The algorithm leverages the gradient method. PGACL is made up of the actor and the

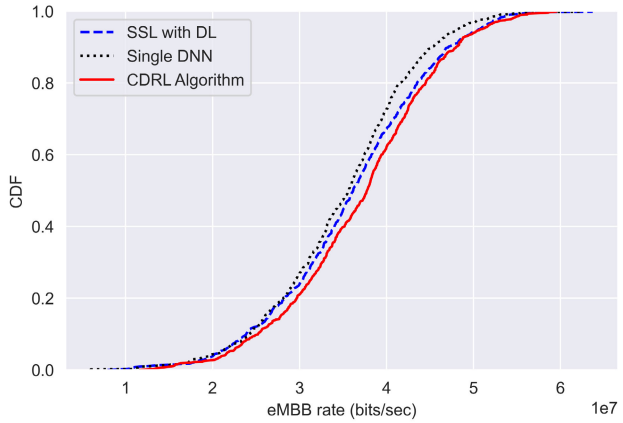


FIGURE 4. CDF of the eMBB sum rate obtained by different schemes.

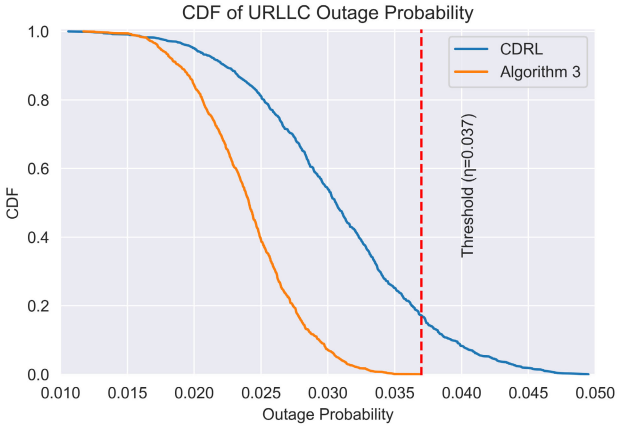


FIGURE 5. CCDF of the URLLC reliability.

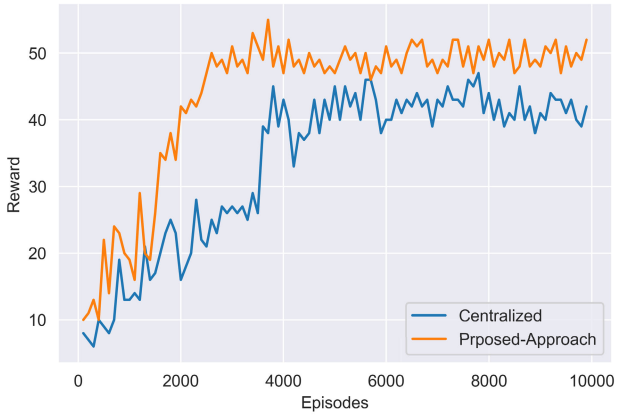


FIGURE 6. Convergence analysis.

critic. The actor component is responsible for policy control based on the current state of the network, determining the actions to be taken. On the other hand, the critic component evaluates the effectiveness of the chosen policy by utilizing the reward function, providing feedback on the quality of the selected actions.

A. SIMULATION FRAMEWORK

The eMBB and URLLC services are utilized by a diverse set of users randomly scattered across a 3-cell cluster in a 4km area and control packets are sent between network nodes and devices. The duration of 1ms is assigned to TTI, and further, each TTI is decomposed into seven orthogonal mini-slots. A RB consists of 84 RE having 12 sub-carriers and 7 OFDM symbols with a bandwidth of 180 kHz for each RB and the system bandwidth is 20 MHz. The pathloss is defined as $120.8 + 37.5 \log_{10}(d)$ dB, where d refers to the distance between user and basestation. Simulation parameters are provided in Table 2.

B. PERFORMANCE EVALUATION OF CDRL ALGORITHM

In this section, we analyze the performance of the CDRL algorithm for RB allocation and compare the results with single DNN [40] and semi-supervised learning (SSL) with DL.

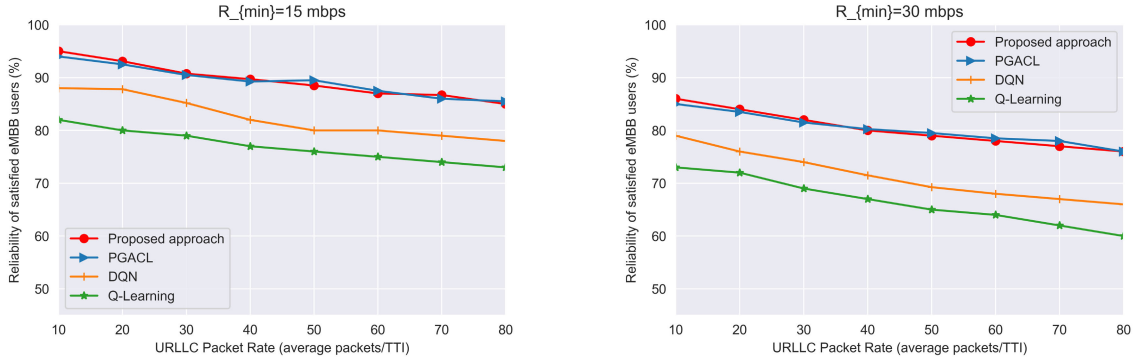
In Fig. 4, the eMBB sum rate obtained by semi-supervised learning with DL and single DNN for RBA has been shown. It can be seen that system performed differently for different schemes. CDRL result performs better than the other schemes, value ranging from 20 Mbit/sec to 60 Mbit/sec. It is evident that the proposed CDRL algorithm performs better than DL schemes. This shows that co-training with DRL can solve the problem of RBA for eMBB service users in NS.

C. RELIABILITY EVALUATION OF URLLC

First, we analyze the worst URLLC reliability scenario achieved by Algorithm 3 based on DDQN with Thompson sampling and compare the performance with Q-learning and PGACL. URLLC reliability analysis is shown in Fig. 5 by plotting the CCDF. It can be observed from the CCDF plot that DDQN with Thompson sampling reduces the tail-risk of URLLC outage probability. The proposed Algorithm 3 guarantees that its values do not violate the threshold η, whereas Algorithm 2 violate the reliability threshold. Our proposed method adjusts the weight parameters according to the behavior of URLLC traffic. This helps to achieve reliable URLLC transmission. Thus, Algorithm 2 fail to guarantee strict URLLC reliability requirements due to their inability to adjust according to channel variations. The Q-learning based method converges slowly and it is hard for it to solve the optimal policy for stringent URLLC service, which results in poor performance. It can be observed from Fig. 5, that the outage probability achieved by the Algorithm 2 performs poorly when the threshold value is 0.037 with a violation probability value around 0.13.

D. CONVERGENCE ANALYSIS

Next, we analyze the convergence behavior of the proposed approach and compare it with the centralized method, where every user has complete awareness of the environment. In this case, the agent takes the decision selection of all agents,



(a) $R_{min} = 15Mbps$ (b) $R_{min} = 30Mbps$

FIGURE 7. eMBB reliability for different URLLC packet rate $\psi(t)$.

increasing the dimension of the action space which effects the convergence rate.

In Fig. 6, we plot the convergence reward value over a number of episodes. It can be seen that the centralized method experiences a poor convergence performance initially and then converges after some episodes. However, the proposed approach based on DDQN coupled with Thompson sampling performs better than centralized approach, at the beginning it converges fast and achieves a better reward value. So, our proposed method performs better in a heterogeneous environment and finds an optimal policy with a fast convergence rate.

E. eMBB RELIABILITY ANALYSIS

Due to incoming URLLC traffic, it is necessary to analyze the reliability of the eMBB service. The reliability of eMBB is determined by calculating the number of eMBB users who achieve a data rate higher than a specific target rate (R_{min}) and dividing it by the total number of eMBB users. This helps us determine the percentage of eMBB users who experience satisfactory service levels in a particular scenario characterized by specific channel conditions and URLLC traffic.

It can be seen in Fig. 7 that the PGACL based risk-averse formulation and proposed method achieves higher reliability. The PGACL based risk-averse formulation performs better than other schemes because the variance of eMBB users punctures only those users with higher SNR, which results in better reliability of eMBB service. However, our proposed algorithm achieves comparable eMBB reliability with PGACL and a much higher sum-rate, because the URLLC service is scheduled over eMBB time slots given the cost function to increase the sum rate of the system while ensuring the QoS requirements of users associated with the eMBB service. The proposed approach ensures the eMBB's reliability by efficiently finding the optimal policy of radio resource management.

Furthermore, it can also be noticed that as the target data rate R_{min} increases, the eMBB reliability decreases with it. The proposed algorithm and PGACL based risk-averse

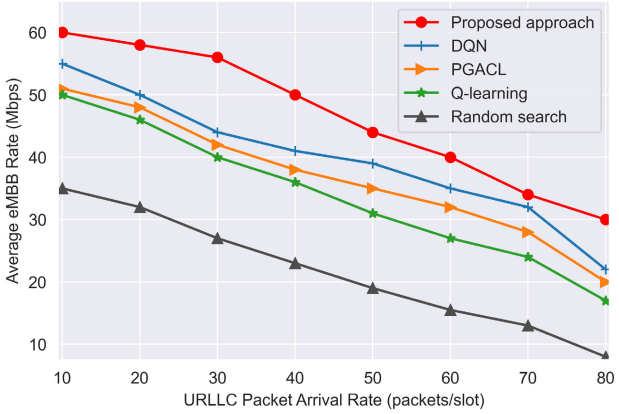


FIGURE 8. Average eMBB users rate for different ψ.

formulation keep the higher reliability at almost 90% when the target data rate is 15 Mbps, while Q-learning fails to keep a tolerable eMBB reliability. Furthermore, when the target data rate is increased to $R_{min} = 30$ Mbps the reliability achieved by the proposed algorithm and PGACL based risk-averse approach is near 80%, while the other schemes fail to achieve the tolerable reliability. It is because the agent in our proposed approach allocate the RBs to the users which has higher SINR and meet the objective function. It can also be seen that as the number of URLLC users increases, the reliability of the eMBB service decreases, because more eMBB slots needs to be punctured which effects the eMBB reliability.

F. eMBB RATE PERFORMANCE

We study the effect of puncturing on eMBB data rate, and compare the results with other methods for various loads of incoming URLLC traffic by plotting the average data rate of the eMBB service. In Fig. 8, it can be seen that the incoming URLLC traffic affects the data rate of the eMBB service because the users associated with the URLLC service are given priority and more radio resources are assigned to URLLC users in order to meet the stringent latency requirements of URLLC service. Furthermore, as compared

TABLE 3. Comparative analysis of the proposed approach and several ML-assisted approaches.

ML Methods	Convergence	High Variance	QoS Guarantees	Overhead	Scalability	Data Requirements
DNN [32] [33] [34]	Slow	Yes	No	Low	Moderate	High
Q-learning [47]	Slow	Yes	No	Low	Low	High
DQN [41] [42] [43] [49]	Moderate	Yes	No	Moderate	Moderate	Moderate
PGAC [38] [47]	High	Yes	Yes	Moderate	High	High
Proposed approach	Fast	No	Yes	Moderate	High	Moderate

to other methods the proposed algorithm achieves a higher average data rate for eMBB users up to 48 Mbps when URLLC load is 45. The random search policy performs poorly because it randomly finds the optimal policy, and is based on a simple architecture. The PGACL based risk-averse approach achieves less average data rate than DQN and the proposed algorithm. The proposed approach achieves a higher average data rate at the beginning of the arrival of URLLC traffic and starts decreasing when the arrival rate of URLLC traffic is increased, hence keeping the higher average data rate than other methods.

G. COMPARATIVE ANALYSIS

Table 3. provides a summary of the ML-based methods employed in to address the resource management problem. The table highlights the convergence behaviour, variance, QoS, communication overhead and data requirements of each approach. It also acknowledges the scalability in large networks. The other approaches, such as Q-learning, DQN, PGAC, and DNN, are compared based on the above mentioned features to the field of reinforcement learning and resource management.

VIII. CONCLUSION AND FUTURE WORK

In this work, we have analyzed the issues related to the coexistence of eMBB and URLLC services in 5G and beyond networks. Using the puncturing technique, we proposed an efficient framework to ensure the capacity and reliability of the system while meeting the low-latency requirements. Moreover, we have employed ML-based algorithms such as semi-supervised and DRL methods to solve the complex optimization problems in real-time in order to allocate the resources intelligently. A co-training method of semi-supervised learning is used in the RB allocation strategy phase. We have addressed the URLLC scheduling sub-problem by proposing a DRL-based DDQN approach with Thompson sampling to meet the latency and reliability requirements and to intelligently manage the URLLC traffic over the punctured eMBB slots. The simulation results verified that the algorithms proposed in this study aim to fulfill the reliability requirements of URLLC users while simultaneously ensuring the reliability and achieving a higher average sum rate for eMBB users.

Training the CDRL model can be computationally intensive and time-consuming. In particular, the convergence time of the algorithm may be lengthy, especially in complex network scenarios. Hence, balancing the need for accurate optimization and real-time decision-making can pose a challenge. Furthermore, the integration of intelligent resource management algorithms may introduce additional communication overhead to the network. This could be due to the exchange of information between network elements, coordination mechanisms, or feedback loops, potentially affecting overall network performance and efficiency. In the future, we look to explore the applications of advanced ML to address these challenges and limitations.

ACKNOWLEDGMENT

The authors would like to thank the studentship support from AiDrivers Ltd.

REFERENCES

[1] W. Ning, Y. Wang, M. Liu, Y. Chen, and X. Wang, "Mission-critical resource allocation with puncturing in industrial wireless networks under mixed services," *IEEE Access*, vol. 9, pp. 21870–21880, 2021, doi: 10.1109/ACCESS.2021.3056202.

[2] T. P. Raptis, A. Passarella, and M. Conti, "A survey on industrial internet with ISA100 wireless," *IEEE Access*, vol. 8, pp. 157177–157196, 2020.

[3] *Study on New Radio (NR) Access Technology Physical Layer Aspects*, document TR38.802, version 14.0.0, 3GPP, Mar. 2017.

[4] *ITU-R M.[IMT-2020.TECH PERF REQ]-Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)*, document ITUR M 2410-0, Nov. 2017.

[5] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[6] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, Jun. 2018.

[7] *Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, document ITU-R M.2083-0, International Telecommunication Union, Geneva, Switzerland, Feb. 2015.

[8] J. Van De Belt, H. Ahmadi, and L. E. Doyle, "Defining and surveying wireless link virtualization and wireless network virtualization," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1603–1627, 3rd Quart., 2017.

[9] R. M. Sohaib, O. Onireti, Y. Sambo, and M. A. Imran, "Network slicing for beyond 5G systems: An overview of the smart port use case," *Electronics*, vol. 10, no. 9, p. 1090, May 2021.

[10] (2018). *Network Slicing Use Case Requirements*. [Online]. Available: https://www.gsma.com/futurenetworks/wp-content/uploads/2018/07/Network-Slicing-Use-Case-Requirements-fixed.pdf

[11] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, Jul. 2021, doi: 10.1109/TWC.2021.3060514.

[12] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, May 2017.

[13] A. Kaloxylos, "A survey and an analysis of network slicing in 5G networks," *IEEE Commun. Standards Mag.*, vol. 2, no. 1, pp. 60–65, Apr. 2018.

[14] *Technical Specification Group Services and System Aspects; Release 15 Description*, document TR 21.915, version 1.1.0, 3GPP, Mar. 2019.

[15] Y. Fu, S. Wang, C. Wang, X. Hong, and S. McLaughlin, "Artificial intelligence to manage network traffic of 5G wireless networks," *IEEE Netw.*, vol. 32, no. 6, pp. 58–64, Nov. 2018.

[16] R. Zhang and A. I. Rudnicky, "A new data selection principle for semi-supervised incremental learning," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 780–783.

[17] M. I. Kamel, L. B. Le, and A. Girard, "LTE wireless network virtualization: Dynamic slicing via flexible scheduling," in *Proc. IEEE 80th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2014, pp. 1–5.

[18] M. Hu, Y. Chang, Y. Sun, and H. Li, "Dynamic slicing and scheduling for wireless network virtualization in downlink LTE system," in *Proc. 19th Int. Symp. Wireless Pers. Multimedia Commun. (WPMC)*, Shenzhen, China, Nov. 2016, pp. 153–158.

[19] Y. Zhang, L. Zhao, D. Lopez-Perez, and K. Chen, "Energy-efficient virtual resource allocation in OFDMA systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[20] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Oulu, Finland, Jun. 2017, pp. 1–5.

[21] P. Popovski, C. Stefanovic, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.

[22] J. Park, S. Samarakoon, H. Shiri, M. K. Abdel-Aziz, T. Nishio, A. Elgabli, and M. Bennis, "Extreme URLLC: Vision, challenges, and key enablers," 2020, *arXiv:2001.09683*.

[23] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 402–415, Jan. 2019.

[24] C. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1292–1295, Jun. 2018.

[25] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X. Wang, "A latency and reliability guaranteed resource allocation scheme for LTE V2V communication systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3850–3860, Jun. 2018.

[26] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 1970–1978.

[27] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi, and C. S. Hong, "EMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 740–743, Apr. 2019.

[28] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881–895, Apr. 2019.

[29] A. K. Bairagi, Md. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, and C. S. Hong, "Coexistence mechanism between eMBB and uRLLC in 5G wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1736–1749, Mar. 2021.

[30] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[31] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[32] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.

[33] H. He, C. Wen, S. Jin, and G. Y. Li, "Deep learning-based channel estimation for beamspace mmWave massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 852–855, Oct. 2018.

[34] W. Lee, M. Kim, and D. Cho, "Deep power control: Transmit power control scheme based on convolutional neural network," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1276–1279, Jun. 2018.

[35] Y. Liu, H. Zhang, K. Long, A. Nallanathan, and V. C. M. Leung, "Energy-efficient subchannel matching and power allocation in NOMA autonomous driving vehicular networks," *IEEE Wireless Commun.*, vol. 26, no. 4, pp. 88–93, Aug. 2019.

[36] Y. Li, H. Zhang, K. Long, S. Choi, and A. Nallanathan, "Resource allocation for optimizing energy efficiency in NOMA-based fog UAV wireless networks," *IEEE Netw.*, vol. 34, no. 2, pp. 158–163, Mar. 2020.

[37] H. Zhang, H. Zhang, K. Long, and G. K. Karagiannidis, "Deep learning based radio resource management in NOMA networks: User association, subchannel and power allocation," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2406–2415, Oct. 2020.

[38] H. Yang, X. Xie, and M. Kadoch, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157–4169, May 2019.

[39] A. T. Z. Kasgari and W. Saad, "Model-free ultra reliable low latency communication (URLLC): A deep reinforcement learning framework," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[40] M. Setayesh, S. Bahrami, and V. W. S. Wong, "Resource slicing for eMBB and URLLC services in radio access network using hierarchical deep learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 8950–8966, Nov. 2022.

[41] Y. Li, C. Hu, J. Wang, and M. Xu, "Optimization of URLLC and eMBB multiplexing via deep reinforcement learning," in *Proc. IEEE/CIC Int. Conf. Commun. Workshops China (ICCC Workshops)*, Aug. 2019, pp. 245–250.

[42] G. Sun, G. O. Boateng, D. Ayepah-Mensah, G. Liu, and J. Wei, "Autonomous resource slicing for virtualized vehicular networks with D2D communications based on deep reinforcement learning," *IEEE Syst. J.*, vol. 14, no. 4, pp. 4694–4705, Dec. 2020, doi: 10.1109/JSYST.2020.2982857.

[43] G. Sun, K. Xiong, G. O. Boateng, D. Ayepah-Mensah, G. Liu, and W. Jiang, "Autonomous resource provisioning and resource customization for mixed traffics in virtualized radio access network," *IEEE Syst. J.*, vol. 13, no. 3, pp. 2454–2465, Sep. 2019, doi: 10.1109/JSYST.2019.2918005.

[44] K. Azizzadenesheli, E. Brunskill, and A. Anandkumar, "Efficient exploration through Bayesian deep Q-networks," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2018, pp. 1–9, doi: 10.1109/ITA.2018.8503252.

[45] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[46] E. Ghadimi, F. D. Calabrese, G. Peters, and P. Soldati, "A reinforcement learning approach to power control and rate adaptation in cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[47] A. M. Kaushik, F. Hu, and S. Kumar, "Intelligent spectrum management based on transfer actor-critic learning for rateless transmissions in cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 5, pp. 1204–1215, May 2018.

[48] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 2249–2257.

[49] R. M. Sohaib, O. Onireti, Y. Sambo, R. Swash, and M. Imran, "Intelligent energy efficient resource allocation for URLLC services in IoV networks," in *Proc. IEEE 33rd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Kyoto, Japan, Sep. 2022, pp. 1–6, doi: 10.1109/PIMRC54779.2022.9978038.

[50] V. Hasselt, H. A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*. vol. 30, no. 1, 2016, pp. 1–11.

RANA M. SOHAIB (Member, IEEE) received the M.S. degree in electrical engineering from the Institute of Space Technology, Islamabad, Pakistan, in 2016. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Glasgow, U.K. His research interests include radio resource management, network performance optimization, network automation, and machine learning to solve networks problems.

OLUWAKAYODE ONIRETI (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electrical engineering from the University of Ilorin, Ilorin, Nigeria, in 2005, and the M.Sc. degree (Hons.) in mobile and satellite communications and the Ph.D. degree in electronics engineering from the University of Surrey, Guildford, U.K., in 2009 and 2012, respectively. He is currently a Lecturer with the University of Glasgow, U.K. He has been actively involved in projects, such as satellites for digitalization of railways (SODOR), coordinated multipoint open radio access networks, and the electromagnetic environment hub. He has authored or coauthored two books, six book chapters, and more than 60 technical papers in leading journals and peer-reviewed conferences. His research interests include self-organizing cellular networks, energy-efficient networks, wireless blockchain networks, millimeter wave communications, and cooperative communications.

YUSUF SAMBO (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from Ahmadu Bello University, Zaria, in 2010, the M.Sc. degree (Hons.) in mobile and satellite communications, in 2011, and the Ph.D. degree in electronic engineering from the Institute for Communication Systems (ICS, formally known as CCSR), University of Surrey, in 2016. He was a Lecturer in telecommunications engineering with Baze University, Abuja, from June 2016 to September 2017. In October 2017, he joined the Communications, Sensing and Imaging (CSI) Research Group, University of Glasgow, as a Postdoctoral Research Associate working on multiple projects worth over £3 million. Since December 2017, he has been the 5G Testbed Lead and coordinated the delivery of the Scotland 5G Centre Testbed, University of Glasgow, where he is currently a Lecturer with the James Watt School of Engineering. He is an active reviewer for several IEEE TRANSACTIONS and other top journals and has served as a technical program committee member for several IEEE conferences. He has also contributed to organizing IEEE conferences and workshops.

RAFIQ SWASH (Senior Member, IEEE) received the Ph.D. degree in Holoscopic 3D Imaging Systems: Camera/Processing/Display from Brunel University London, U.K. He is the founder of AiDrivers Ltd., a Lecturer at Brunel University London, and a Visiting Professor at Changchun Institute of Optics, China. He has been actively involved in a multitude of research projects namely, the lead research scientist and contributed hugely in multi-million funded innovation research projects.

SHUJA ANSARI (Senior Member, IEEE) received the M.Sc. (Hons.) and Ph.D. degrees in engineering from Glasgow Caledonian University, in 2015 and 2019, respectively. He is currently a Lecturer in autonomous systems and connectivity with the James Watt School of Engineering, University of Glasgow (UofG). He is also a Chartered Engineering who has a strong background with over a decade of experience in telecommunications. He is also the 5G and IoT use case lead for the wave-1 urban 5G project funded by the Scotland 5G Centre and the Project Manager of Glasgow COMPORAN funded by DSIT. His research interests include wireless communications, systems integration, terrestrial/airborne mobile networks, security and privacy in communications, and the Internet of Things (IoT).

MUHAMMAD A. IMRAN (Fellow, IEEE) received the M.Sc. (Hons.) and Ph.D. degrees from the Imperial College London, U.K., in 2002 and 2007, respectively. He is currently the Dean of the Glasgow College, UESTC, and a Professor in communication systems with the James Watt School of Engineering, University of Glasgow (UofG). He is also an affiliate Professor with the University of Oklahoma, USA, an Adjunct Research Professor with the Artificial Intelligence Research Centre (AIRC), Ajman University, Ajman, United Arab Emirates, and a Visiting Professor with the 5G Innovation Centre, University of Surrey, U.K. He is also leading research with the Scotland 5G Center, UofG. He has over 18 years of combined academic and industry experience, working primarily in the research areas of cellular communication systems. He has been awarded 15 patents, has authored/coauthored over 400 journals and conference publications and has been principal/co-principal investigator on over £6 million in sponsored research grants and contracts. He has supervised more than 40 successful Ph.D. graduates. He has an award of excellence in recognition of his academic achievements, conferred by the President of Pakistan. He was also awarded the IEEE Comsoc's Fred Ellersick Award, in 2014, the FEPS Learning and Teaching Award, in 2014, and the Sentinel of Science Award, in 2016. He was twice nominated for the Tony Jean's Inspirational Teaching Award. He is a shortlisted finalist for The Wharton-QS Stars Awards, in 2014, the QS Stars Reimagine Education Award 2016 for Innovative Teaching, and VC's Learning and Teaching Award from the University of Surrey. He is a Senior Fellow of the Higher Education Academy, U.K.