Received 6 January 2022; revised 31 January 2022; accepted 18 February 2022. Date of publication 23 February 2022; date of current version 3 March 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3153226

The Frontiers of Deep Reinforcement Learning for Resource Management in Future Wireless HetNets: Techniques, Challenges, and Research Directions

ABDULMALIK ALWARAFY, MOHAMED ABDALLAH (Senior Member, IEEE), BEKIR SAIT ÇIFTLER (Member, IEEE), ALA AL-FUQAHA (Senior Member, IEEE), AND MOUNIR HAMDI (Fellow, IEEE)

Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

CORRESPONDING AUTHOR: A. ALWARAFY (e-mail: aalwarafy@hbku.edu.qa)

This publication was made possible by NPRP-Standard (NPRP-S) Thirteen (13th) Cycle grant # NPRP13S-0201-200219 from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work, and are solely the responsibility, of the authors.

ABSTRACT Next generation wireless networks are expected to be extremely complex due to their massive heterogeneity in terms of the types of network architectures they incorporate, the types and numbers of smart IoT devices they serve, and the types of emerging applications they support. In such large-scale and heterogeneous networks (HetNets), radio resource allocation and management (RRAM) becomes one of the major challenges encountered during system design and deployment. In this context, emerging Deep Reinforcement Learning (DRL) techniques are expected to be one of the main enabling technologies to address the RRAM in future wireless HetNets. In this paper, we conduct a systematic in-depth, and comprehensive survey of the applications of DRL techniques in RRAM for next generation wireless networks. Towards this, we first overview the existing traditional RRAM methods and identify their limitations that motivate the use of DRL techniques in RRAM. Then, we provide a comprehensive review of the most widely used DRL algorithms to address RRAM problems, including the value- and policy-based algorithms. The advantages, limitations, and use-cases for each algorithm are provided. We then conduct a comprehensive and in-depth literature review and classify existing related works based on both the radio resources they are addressing and the type of wireless networks they are investigating. To this end, we carefully identify the types of DRL algorithms utilized in each related work, the elements of these algorithms, and the main findings of each related work. Finally, we highlight important open challenges and provide insights into several future research directions in the context of DRL-based RRAM. This survey is intentionally designed to guide and stimulate more research endeavors towards building efficient and fine-grained DRL-based RRAM schemes for future wireless networks.

INDEX TERMS Radio resource allocation and management, deep reinforcement learning, next generation wireless networks, HetNets, power, bandwidth, rate, access control.

I. INTRODUCTION

RADIO resource allocation and management (RRAM) is regarded as one of the essential challenges encountered in modern wireless communication networks [1]. Nowadays, modern wireless networks are becoming more heterogeneous and complex in terms of the types of emerging radio access networks (RANs) they integrate, the explosive number and types of smart devices they serve, and the types of disruptive applications and services they support [2], [3]. It is envisaged that future networks will integrate land, air, space, and deep-sea wireless networks into a single network to meet the stringent requirements of a fully-connected world vision [4], [5], as shown in Fig. 1. This will ensure ubiquitous connectivity for user devices with enhanced quality of service (QoS) in terms of coverage, reliability, and throughput. In addition, future user devices will also witness

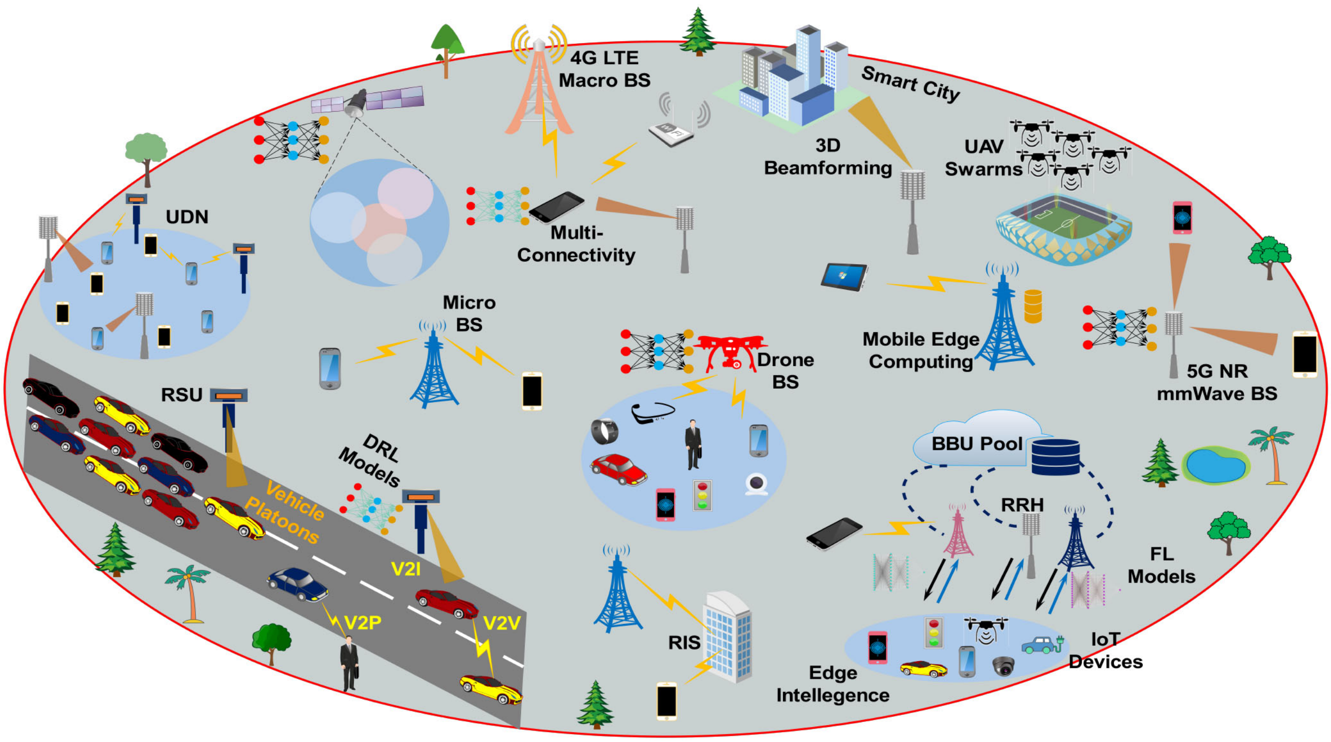

FIGURE 1. A pictorial illustration of next generation wireless networks characterized by their massive heterogeneity in terms of RANs infrastructures, types and numbers of user devices served, and types of applications and services supported.

an unprecedented increase in their numbers and types of data-hungry applications they require/support [3], [6]. It is expected that by 2023, the number of user networked devices and connections, including smart-phones, tablets, wearable devices, and sensors, will reach 29.3 billion [6], and generate a data rate exceeding 50 trillion GB [1]. All these trends will exacerbate the burdens during system design, planning, deployment, operation, and management. In particular, RRAM will become crucial in such complex and large-scale networks in order to guarantee an enhanced communications experience.

RRAM plays a pivotal role during infrastructure planning, implementation, and resource optimization of modern wireless networks. Efficient RRAM solutions will guarantee enhanced network connectivity, increased system efficiency, and reduced energy consumption. The performance of wireless networks heavily relies on two aspects. First, how network radio resources are being utilized, managed, and orchestrated, including transmit power control, spectrum channel allocations, and user access control. Second, how efficiently the system can react to the rapid changes of network dynamics, including wireless channel statistics, users mobility patterns, instantaneous radio resources availability, and variability in traffic loads. Efficient RRAM techniques must efficiently and dynamically account for such design aspects in order to ensure high network QoS and enhanced users' Quality of Experience (QoE).

Deep reinforcement learning (DRL) is a branch of artificial intelligence (AI) that enables network entities, such as base stations, user devices, edge servers, gateways, and access points, to continuously interact with the environment to make autonomous control decisions [7]–[14]. DRL Techniques have attracted considerable research recently and demonstrated efficient performance in addressing complex wireless optimization problems, including RRAM problems. Therefore, experts expect DRL methods to be one of the main enabling technologies for future wireless networks due to their ability to overcome the limitations of traditional RRAM techniques [2], [15].

A. MOTIVATIONS OF THE PAPER

The main motivations of this work stem from three aspects. First, the paramount importance of allocating radio resources in future wireless networks. Second, the limitations and shortcomings of existing state-of-the-art RRAM techniques. Third, the robustness of Deep reinforcement techniques in alleviating these limitations and providing efficient performance in the context of RRAM. Here we elaborate more on each aspect.

1) IMPORTANCE OF RRAM IN MODERN WIRELESS NETWORKS

The explosive growth in the number and types of modern smart devices, such as smartphones/tablets and wearable devices, has led to the emergence of disruptive wireless communications and networking technologies, such as 5G NR cellular networks, IoT networks, personal (or wireless body area networks), device-to-device (D2D) communications, holographic imaging and haptic communications, and vehicular networks [3], [4], [16]–[23]. Such networks are envisaged to meet the stringent requirements of the emerging applications and services via supporting high data rates, coverage, and connectivity with significant enhancements in reliability, reduction in latency, and mitigation of energy consumption.

However, achieving this goal in such large-scale, versatile, and complex wireless networks is quite challenging, as it requires a judicious allocation and management of the networks' limited radio resources [24], [25]. In particular, efficient and more advanced RRAM solutions must be developed to balance the tradeoff between enhancing network performance while guaranteeing an efficient utilization of radio resources. Furthermore, efficient RRAM solutions must also strike and intelligent tradeoff between optimizing network radio resources and satisfying users' QoE. For example, RRAM techniques must jointly enhance network spectral efficiency (SE), energy efficiency (EE), and throughput while mitigating interference, reducing latency, and enhancing rate for user devices.

Efficient and advanced RRAM schemes can considerably enhance the system's SE compared to the traditional techniques by relying on the advanced channel and/or source coding methods. RRAM is essential in broadcast wireless networks covering wide geographical areas as well as in modern cellular communication networks comprised of several adjacent and dense access points (APs) that typically share and reuse the same radio frequencies.

From a cost point of view, the deployment of wireless APs and sites, e.g., base stations (BSs), including the real estate costs, planning, maintenance, and energy, is the most critical aspect alongside with the frequency license fees. Hence, the goal of RRAM is maximizing the network's SE in terms of bits/sec/Hz/area unit or Erlang/MHz/site, under some constraints related to user fairness. For instance, the service grade must meet a minimum acceptable level of QoS, including the coverage of certain geographical areas while mitigating network outages caused by interference, noise, large-scale fading (due to path losses and shadowing), and small-scale fading (due to multi-path). The service grade also depends on blocking caused by admission control, scheduling errors, or inability to meet certain QoS demands of edge devices (EDs).

2) WHERE DO TRADITIONAL RRAM TECHNIQUES FAIL?

Future wireless communication networks are complex due to their large-scale, versatile, and heterogeneous nature. To optimally allocate and manage radio resources in such networks, we typically formulate RRAM as complex optimization problems. The objective of such problems is to achieve a particular goal, such as maximizing network sum-rate, SE, and EE, given the available radio resources and QoS requirements of user devices. Unfortunately, the massive heterogeneity nature of modern networks poses tremendous challenges during the process of formulating optimization problems as well as applying conventional techniques to solve them, such as optimization, heuristic, and game theory algorithms.

The large-scale nature of next generation networks makes it quite difficult to formulate RRAM optimization problems that are often intractable non-convex. Also, conventional techniques used to solve the RRAM problems require complete or quasi-complete knowledge of the wireless environment, including accurate channel models and real-time channel state information (CSI). However, obtaining such information in a real-time fashion in these large-scale networks is quite difficult or even impossible. Furthermore, conventional techniques are often computationally-expensive and incur considerable timing overhead. This renders them inefficient for most emerging time-sensitive applications, such as autonomous vehicles and robotics.

Moreover, game theory-based techniques are unsuitable for future heterogeneous networks (HetNets) as such techniques are devised for homogeneous players. Also, the explosive number of network APs and user devices will create extra burdens on game theory-based techniques. In particular, network players, such as BSs, APs, and user devices, need to exchange a tremendous amount of data and signaling. This will induce unmanageable overhead that largely increases delay, computation, and energy/memory consumption of network elements.

3) HOW CAN DRL OVERCOME THESE CHALLENGES AND PROVIDE EFFICIENT RRAM SOLUTIONS?

Emerging artificial intelligence (AI) techniques, such as deep reinforcement learning (DRL), have shown efficient performance in addressing various issues in modern wireless communication networks, including solving complex RRAM optimization problems [7]–[15]. In the context of RRAM, DRL methods are mainly used as an alternative to overcome the shortcomings and limitations of the conventional RRAM techniques discussed above. In particular, DRL techniques can solve complex network RRAM optimization problems and take judicious control decisions with only limited information about the network statistics. They achieve this by enabling network entities, such as BSs, RAN's APs, edge servers (ESs), gateways nodes, and user devices, to make intelligent and autonomous control decisions, such as RRAM, user association, and RAN's selection, in order to achieve various network goals such as sum-rate maximization, reliability enhancement, delay reduction, and SE/EE maximization. In addition, DRL techniques are model-free that enable different network entities to learn optimal policies about the network, such as RRAM and user association, based on their continuous interactions with the wireless environment, without knowing the exact channel models or other network statistics *a-priori*. These appealing features make DRL methods one of the main key enabling technologies to address the RRAM issue in modern wireless communication networks [2], [3].

B. RELATED WORK

There is a limited number of surveys that focus on the role of DRL in RRAM. Existing related surveys are listed in

TABLE 1. Relationship between this survey and existing surveys on DRL-based RRAM For wireless networks.

Paper	Summary of the survey's contributions	Related contents in this paper	Value added in this paper
Luong *et al.* [15]	Applications of DRL in communications and networking	Section III/IV	Particularly focus on DRL usage for RRAM and enhanced list of papers
Hussain *et al.* [1]	ML- and DL-based resource management mechanisms in cellular wireless and IoT networks	Sections II/IV	In-depth and holistic coverage of DRL algorithms used for RRAM, intensive review of existing papers related to DRL for RRAM, and the coverage of more types of wireless networks
Lin *et al.* [34]	Applications of AI approaches in resource management, such as spectrum, computing, and caching.	Section V	Particularly focus on DRL methods, including more radio resources, and intensive literature review
Liang *et al.* [24]	DL-Based resource allocation with application to vehicular networks	Sections II/IV	Focus on DRL techniques for RRAM, in-depth literature review, and including various types of modern wireless networks
Chen *et al.* [26]	Applications of ML algorithms in solving wireless networking problems	NA	Focus on applications of DRL in solving RRAM wireless problems, and coverage of more wireless networks
Gupta *et al.* [10]	General research and simulation tools used for DRL	Section III	Specifically focus on DRL algorithms along with the related research conducted in the context of RRAM
Du *et al.* [11]	Investigates how to achieve green DRL for radio resource management via energy allocation based on architecture and algorithm innovations	Section IV	Further extend to more radio resources and more modern wireless networks
Pham *et al.* [35]	A layered-based classification of resource management techniques in Wireless Access Networks	Section II	A holistic study of conventional and emerging ML-based techniques for RRAM applied to modern wireless networks and including more radio resources
Arulkumaran *et al.* [36], Zhang *et al.* [37]	Overview of DRL approaches in general, including applications and models	Section III	Focus on DRL approaches utilized in RRAM for wireless networks, and also provide detailed literature review
Zappone *et al.* [27]	Motivations, applications, visions, and case studies for the usage of DL techniques in wireless communication networks	NA	Particularly focusing on DRL techniques for wireless communication networks in the context of RRAM
Lee *et al.* [8]	DRL-based resource management schemes for 5G HetNets in energy harvesting, network slicing, cognitive HetNets, coordinated multi-point transmission, and big data	Section III	In-depth analysis of DRL methods used for RRAM including; DRL algorithms, types of wireless networks, types of radio resources investigated, and extensive literature review
Qian *et al.* [12]	Applications of RL and DRL in three technologies: mobile edge computing, software defined network, and network virtualization in 5G	NA	Focus on DRL applications for RRAM in cellular and other emerging wireless networks
Khorasgani *et al.* [28]	Key limitations and challenges in using DRL to address the problem of dynamic dispatching in the mining industry	Section IV	Extend the investigation to include various wireless networks with an extensive focus on radio resources
Xu *et al.* [38]	A comprehensive survey on resource allocation for 5G HetNets, including current research, future trends, and research challenges	Section II	Particularly focus on DRL algorithms, focus only on radio resources, and the converge of more types of wireless networks

Table 1. The table also summarizes the topics covered in these surveys along with a mapping to the relevant sections of this paper and a categorical discussion of the improvements and value-added in this paper relative to these surveys. In general, as reported in Table 1, these published surveys still have several research gaps that are addressed in this survey. We summarize them as follows.

- Some of the existing surveys focus on DRL applications in wireless communications and networking in general, without paying much attention to RRAM [10], [15]. For example, existing surveys cover topics related to DRL enabling technologies, use-cases, architectures, security, scheduling, clustering and data aggregation, traffic management, etc.
- Some of the published surveys focus on RRAM for wireless networks using ML and/or DL techniques without paying much attention to DRL techniques [1], [24], [26], [27]. For example, they consider ML techniques such as convolutional neural networks (CNN), recurrent neural networks (RNN), supervised learning, Bayesian learning, K-means clustering, Principal Component Analysis (PCA), etc.
- Even the surveys that address DRL for RRAM in wireless networks focus on specific wireless network types or applications [8], [9], [11], [12], [28], missing some of the recent research, not providing an adequate overview of the most widely used DRL algorithms for RRAM [12], or not covering the RRAM in-depth, but, rather, just covering a limited number of radio resources.

Hence, the role of this paper to fill these research gaps and overcome these shortcomings. In particular, we provide a comprehensive survey on the application of DRL techniques in RRAM for next generation wireless communication networks. We have carefully cited up-to-date surveys and related research works. We should emphasize here that the scope of this paper is focused only on radio (or communication) resources, and no computation resources are included during the study and analysis. Fig. 2 shows the radio resources or issues addressed in this survey. However, computation resource aspects such as offloading, storage, task scheduling, caching, etc., can be found in other studies such as [29]–[33] and the references therein.

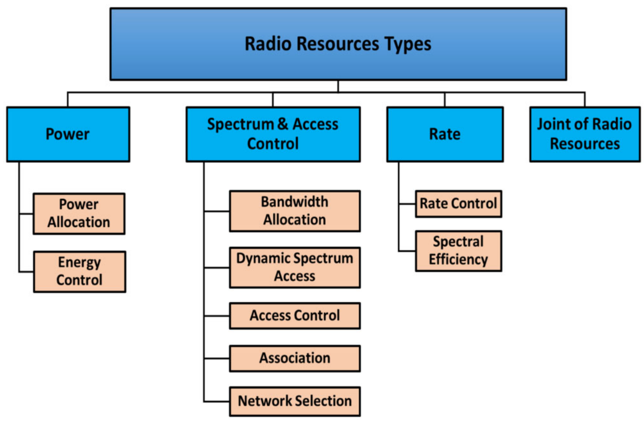

FIGURE 2. Classification based on radio resources (or issues) addressed in the papers.

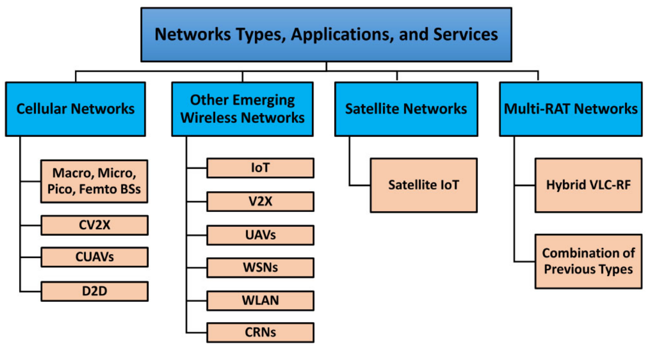

FIGURE 3. Classification based on networks types covered in the papers.

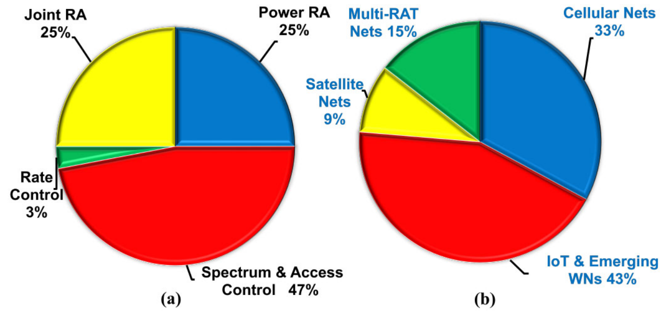

FIGURE 4. Percentages of related work based on (a) types of radio resources covered and (b) types of networks and application investigated. RA: resource allocation, WNs: wireless networks.

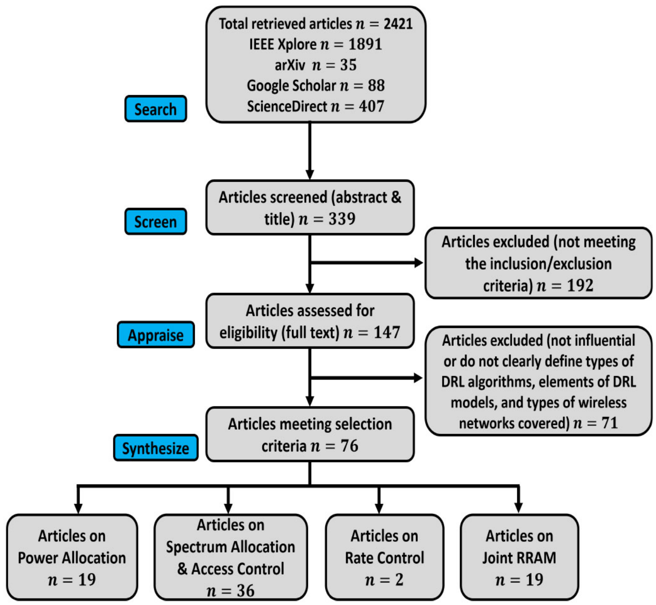

FIGURE 5. The review protocol followed in this survey.

C. PAPER CONTRIBUTIONS

The main contributions of this paper are summarized as follows.

1) We provide a detailed discussion on the state-of-the-art techniques used for RRAM in wireless networks, including their types, shortcomings, and limitations that led to the adoption of DRL solutions.
2) We identify the most widely used DRL techniques utilized in RRAM of wireless networks and provide a comprehensive overview of them. The advantages, features, and limitations of each technique are discussed. Hence, the reader is provided with an in-depth knowledge of which DRL techniques should be leveraged for each RRAM problem under investigation.
3) We conduct an extensive and up-to-date literature review and classify the papers as reported in the literature based on the type of radio resources they address (as shown in Fig. 2) and the types of wireless networks, applications, and services they consider (as shown in Fig. 3). Specifically, for each paper reviewed, we identify the problem it addresses, type of wireless network it investigates, type of DRL model(s) it implements, main elements of the DRL models (i.e., agent, state space, action space, and reward function), and its main findings. This provides the reader with in-depth technical knowledge of how to efficiently engineer DRL models for RRAM problems in wireless communications.
4) Based on the papers reviewed in this survey, we outline and identify some of the existing challenges and provide deep insights into some promising future research directions in the context of using DRL for RRAM in wireless networks.

Fig. 4 shows the percentage of the related works, classified based on the types of radio resources discussed in each paper, Fig. 4 (a), and based on the types of wireless networks studied in each paper, Fig. 4 (b). This survey is designed by carefully following the review protocol illustrated in Fig. 5. Since this survey mainly focuses on deep reinforcement learning for RRAM in wireless networks, we included the following terms during the search stage along with "AND/OR" combinations of them; "deep reinforcement learning," "DRL," "resource allocation," "resource management," "power," "spectrum," "bandwidth," "access control," "user association," "network selection," "cell selection,"

TABLE 2. List of acronyms used and their definitions.

Acronym	Definition	Acronym	Definition	Acronym	Definition
4G	Fourth Generation mobile system	GAN	Generative Adversarial Network	RAT	Radio Access Technology
5G	Fifth Generation mobile system	HetNets	Heterogeneous Networks	RB	Resource Block
6G	Sixth Generation mobile system	HSR	High-Speed Railway	RF	Radio Frequency
A2C	advantage actor-critic	IAB	Integrated Access and Backhaul	RIS	Reconfigurable Intelligent Surface
A3C	Asynchronous Actor Critic Algorithm	IIoT	Industrial Internet of Things	RL	Reinforcement Learning
ADMM	Alternating Direction Method of Multipliers	IoT	Internet of Things	RNC	Radio Network Controller
AI	Artificial Intelligence	KPI	Key Performance Indicator	RRA	Radio Resource Allocation
AP	Access Point	LEO	Low Earth Satellite	RRAM	Radio Resource Allocation and Management
BBU	Base-Band Unit	LTE	Long-Term Evolution	RRH	Remote Radio Head
BS	Base Station	M2M	Machine-to-Machine	RSU	Road Side Unit
C-RAN	Cloud Radio Access Network	MADRL	Multi-Agent Deep Reinforcement Learning	SE	Spectral Efficiency
CRN	Cognitive Radio Network	MCA	Multi-Channel Access	SINR	Signal to Interference plus Noise Ratio
CSI	Channel State Information	MCC	Mission-critical communication	SIoT	Satellite Internet of Things
CUAV	Cognitive Unmanned Aerial Vehicle	MDP	Markov Decision Process	SNR	Signal to Noise Ratio
CV2X	Cellular Vehicular Communication	MeNB	Macro eNodeB	SU	Secondary User
D2D	Device-to-Device	ML	Machine Learning	TD	Time Difference
D3QN	Dueling Double Deep Q-Network	mmWave	Millimeter Wave	TDD	Time Division Duplex
DDPG	Deep Deterministic Policy Gradient	NE	Nash Equilibrium	UAV	Unmanned Aerial Vehicles
DDQN	Double Deep Q-Network	NOMA	Non-Orthogonal Multiple Access	UDN	Ultra-Dense Network
DL	Deep Learning	NTNs	Non-Terrestrial Networks	UE	User End
DNN	Deep Neural Network	OFDM	Orthogonal Frequency Division Multiplexing	V2I	Vehicle to Infrastructure
DPG	Deterministic Policy Gradient	OMA	Orthogonal Multiple Access	V2V	Vehicle to Vehicle
DQN	Deep Q-Network	OU	Ornstein–Uhlenbeck	V2X	Vehicle to Everything
DRL	Deep Reinforcement Learning	PED	Patient Edge Device	VANETs	Vehicular Ad Hoc Networks
DSA	Dynamic Spectrum Access	PPO	Proximal Policy Optimization	VLC	Visible Light Communication
DT	Digital Twin	PU	Primary User	WLAN	Wireless Local Area Network
EE	Energy Efficiency	QoE	Quality of Experience	WMMSE	Weighted Minimum Mean Square Error
FL	Federated Learning	QoS	Quality of Service	WSN	Wireless Sensor Network
FP	Fractional Programming	RAN	Radio Access Network	XAI	Explainable AI

"rate control," "joint resources," "wireless networks," "satellite networks," "cellular networks," and "Heterogeneous networks." The number of papers found and the databases searched are detailed in Fig. 5. The inclusion criteria are papers that address the use of DRL techniques to manage and allocate the radio resources shown in Fig. 2 for the wireless networks shown in Fig. 3. The exclusion criteria are papers that: 1) address computation resources, e.g., task offloading, storage, scheduling, etc., 2) use conventional RRAM approaches, i.e., not using DRL techniques, 3) use ML/DL techniques, or 4) address non-wireless networks, e.g., wired networks, optical networks, etc. In Fig. 5, the number of papers excluded after a detailed check of the body is 71, which are directly related to our survey but not influential or do not clearly identify the types of DRL algorithms used, elements of DRL models (i.e., agents, state space, action space, and reward function), type of wireless networks covered, and/or not well written.

In general, the research questions that this survey aims to address are stated as follows. How can DRL techniques be implemented to address the RRAM problems in modern wireless networks? What are the performance advantages achieved when using DRL tools compared to the state-of-the-art RRAM approaches? What are the most effective and widely used DRL algorithms to address the RRAM problems, and how can they be implemented? What are the most important and influential papers that present DRL-based solutions for RRAM in next generation wireless networks? What are the challenges and possible research directions that stem from the reviewed papers in the context of using DRL for RRAM in wireless networks? The retrieved papers shown in Fig. 5, i.e., the 76 papers, are selected carefully to help with answering these questions, as we will elaborate in the next sections.

It is observed from Fig. 4 (a) that the majority of related works are on the Spectrum and Access Control radio resources, followed by both the Power radio resource and Joint radio resources. Also, as shown in Fig. 4 (b), the related works on the IoT and Other Emerging Wireless Networks have received more attention than the other wireless network types, followed by the Cellular Networks.

The rest of this paper is organized as follows. Table 2 lists the acronyms used in this paper and their definitions. Section II discusses existing RRAM techniques, including conventional methods and DRL-based methods. The definitions, types, and limitations of existing techniques are discussed. Also, the advantages of employing DRL techniques for RRAM are explained. Section III provides an overview of the DRL techniques widely employed for RRAM, including their types and architectures. In-depth classifications of the existing research works is provided in Section IV. Existing papers are classified based on the radio resources and the network types they cover. Section V provides key open challenges, lessons learned, and some insights for future research directions. Finally, Section VI concludes the paper. The organization of the paper is pictorially illustrated in Fig. 6.

II. RADIO RESOURCE ALLOCATION AND MANAGEMENT TECHNIQUES

In this section, we define the main radio resources of interest and provide a summary of the conventional techniques and tools used for RRAM in wireless networks. Also, the limitations of these conventional techniques that motivate the use of DRL solutions will be highlighted. Then we discuss how DRL techniques can be efficient alternatives to these traditional approaches.

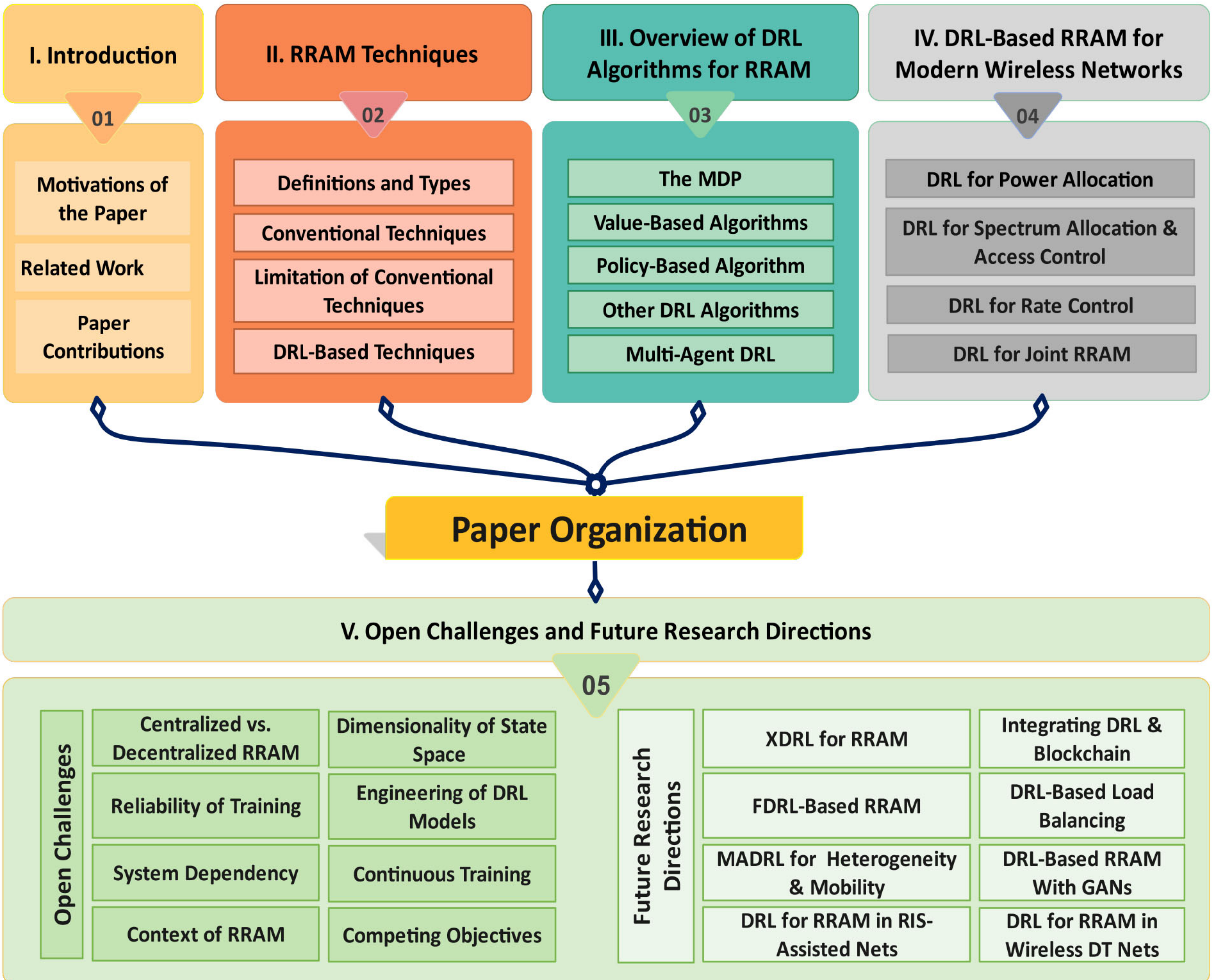

FIGURE 6. Organization of the paper.

A. RADIO RESOURCES: DEFINITIONS AND TYPES (OR ISSUES)

In general, allocation and management of wireless network resources include radio (i.e., communication) and computation resources. This paper focuses only on the RRAM issue. This involves strategies and algorithms used to control and manage wireless network parameters and resources, such as transmit power, spectrum allocation, user association/assignment, rate control, access control, etc. The main goal of wireless networks, in general, is to utilize and manage these available radio resources as efficiently as possible to provide enhanced network QoS, such as enhanced data rate, SE, EE, reliability, connectivity, and coverage while meeting users' QoS demands.

Efficient RRAM schemes can considerably enhance the system's SE compared to the traditional techniques relying on advanced channel and/or source coding methods. For example, future wireless networks are expected to cover broad geographical areas with ultra-dense network (UDN) deployments. In these UDNs, a massive number of adjacent APs typically require sharing communication resources, such as radio frequencies and channels, to utilize resources and enhance network QoS. RRAM would be essential in such UDN-based network deployments [38].

The most crucial radio resources or issues that play a fundamental role in controlling wireless networks' performance are summarized below.

- *Power resource:* Which is one of the most critical issues in the RRAM of modern HetNets. Transmit power allocation in the downlink/uplink from/to network APs, such as BSs and edge servers (ESs), is essential to guarantee a satisfactory QoS for communication links. Power control is essential from two perspectives; physical limitations and communication links. Practically, the maximum power is limited by the capability of APs' power amplifiers or government regulation. Hence, it is common to incorporate the limited power resource as a constraint during the design and implementation of HetNets. On the other hand, power control is also needed to guarantee enhanced networks' QoS and user devices' QoE. For example, in large-scale and UDNs such as the mmWave and THz band systems [2], [3], [39], signal attenuation due to path losses must be accounted for during power budget analysis. Also, the coverage of BSs' cells and the inter-and intra-cell interference issues become crucial, which are mainly determined by the transmit power level. Hence, developing adaptive and fine-grained power allocation and interference management strategies is essential to address such challenges.

- *Spectrum resource and access control:* This is also another main issue in the RRAM of modern HetNets. User devices must be allocated frequency channels to start transmitting/receiving data with acceptable SNR. Existing wireless networks, such as the sub 6 GHz, suffer from a severe bandwidth shortage which is even exacerbated with the explosive increase in the number of user devices [6]. Fortunately, the mmWave and emerging THz bands can considerably overcome this shortcoming by providing an extra 3.25 GHz and 10-100 GHz bandwidth, respectively [40]. It is also expected that future user devices will be equipped with advanced capabilities that enable them to aggregate all these three frequency bands, i.e., the sub 6GHz, mmWave, and THz, to support future technologies and services [41]. However, allocating and managing the radio channels of these frequency bands across multi-RAN to a massive number of user devices mandate developing advanced signal processing techniques. Unfortunately, such techniques require perfect knowledge of network statistics and CSI, which is quite difficult or even impossible due to the large-scale and massive heterogeneity of modern HetNets. Hence, it is expected that future HetNets will integrate DRL methods with signal processing techniques to overcome this issue.
- *User association:* With the ever-increase in the number of IoT smart devices and the varying QoS demands of emerging applications, it becomes necessary to ensure reliable network hyper-connectivity to these devices [2], [39]. User association defines which BS(s), RAN's AP(s), or edge server(s) that each user device must connect/associate to/with to guarantee its QoS demands. Taking into consideration the multi-RAN and multi-connectivity nature of modern HetNets [3], it is expected that future devices will be equipped with SDR capabilities that enable them to support multi-association/assignment to multiple RANs simultaneously [41]. Based on users' QoS demands, devices can operate in a multi-mode or multi-homing fashion. In the multi-mode fashion, each device will be associated with a single RAN AP at a time [41], [42] in a traditional fashion. Whereas in the multi-homing fashion, devices can be associated with multiple RAN APs simultaneously to aggregate RANs' radio resources. Achieving such a goal, however, is also another challenging issue. Obtaining real-time information on the network statistics, such as CSI, traffic load, RANs occupancy, and user devices' QoS demands, requires unmanageable and intolerable overhead. Hence, DRL techniques can be adopted in such a scenario to dynamically learn the channel and perform autonomous user association/assignment decisions.
- *Rate control:* Often, the main objective of RRAM is to maximize the QoS of HetNets in terms of network sum-rate or SE. This is typically achieved by formulating complex wireless network optimization problems and deriving their solutions subjected to available network radio resources while respecting the data rate demands of user devices. However, accurate solutions for such problems require full knowledge of wireless channel gain, including the large-scale and small-scale fading [43]. Obtaining such knowledge in real-time is quite difficult, especially in modern HetNets, due to their rapid increase in the underlying RANs/user devices and the type of applications. Moreover, multi-RANs data rate aggregation has also been proposed recently [41], [44] to support the multi-Gbps data rate requirements of the emerging applications. Hence, it becomes imperative to develop efficient schemes that enable rate aggregation while having limited knowledge of channels. DRL methods can be employed to achieve this goal [41], [42], [44].

B. CONVENTIONAL RRAM TECHNIQUES

In this subsection, we overview the state-of-the-art approaches and tools used for RRAM in modern HetNets. RRAM techniques can be classified into two broad categories based on their adaptivity to the wireless environment: static and dynamic approaches. Each of which can be further classified based on various criteria, such as centralized or distributed, instantaneous or ergodic, optimal or sub-optimal, single-cell or multi-cell, cooperative or non-cooperative, in addition to different combinations of these variants. In this paper, we discuss the general features of the static and dynamic techniques along with their types.

RRAM has been one of the major research interests in wireless networks using conventional approaches. It has been extensively surveyed for various wireless networks and systems. Table 3 lists some of the existing surveys for resource allocation and management using conventional methods along with the types of wireless networks and systems they study.

1) STATIC TECHNIQUES

Static approaches are designed based on a priori statistical information and cannot adapt to wireless network parameters, such as traffic load, users' mobility pattern, channel conditions/quality, network spectrum occupancy, and users' QoS demands. These techniques are simple; however, they suffer from several shortcomings, such as severe underutilization of radio resources, increased network outage, reduced network throughput, and poor network QoS.

Static RRAM techniques are employed in several traditional networks, such as cellular networks and WLANs.

TABLE 3. Existing surveys on resource allocation and management for wireless networks and systems using conventional approaches.

Paper	Types of wireless networks and systems studied
[45]–[47]	Cognitive radio networks (CRNs)
[38], [48]–[51]	Wireless HetNets
[52]	M2M communication networks
[53]–[56]	OFDM systems
[57]	MIMO-OFDM systems
[58], [59]	D2D communication networks
[60]	UAV communications
[61], [62]	Vehicular communications (V2X)
[63]	Railway communications
Value added in this paper	Focus on the applications of DRL techniques for RRAM in next generation wireless networks, such as cellular HomNets, IoT networks, satellite networks, multi-RATs networks, HetNet, etc.

Examples of static RRAM techniques include circuit-mode communication using frequency division multiple access (FDMA) and time division multiple access (TDMA) schemes and fixed radio resource allocation, such as fixed power and channel allocation.

2) DYNAMIC TECHNIQUES

On the contrary, dynamic or adaptive RRAM approaches are more efficient as they can dynamically adjust the network radio resources to accurately track variations in propagation conditions and user QoS requirements.

Dynamic RRAM schemes are widely utilized in designing modern HetNets. They have shown efficient results in reducing the expensive manual network planning and achieving tighter radio resource utilization, which will lead to enhanced network efficiency. Some RRAM schemes are centralized, where several BSs, ESs, APs, and network gateways are controlled by a central Radio Network Controller (RNC). Others are distributed, either autonomous algorithms implemented in user devices, BSs, ESs, or coordinated by exchanging information among these network entities. Examples of dynamic RRAM schemes include power control algorithms, spectrum/channel allocation algorithms, multi-access control schemes, traffic/link adaptation algorithms, channel-dependent scheduling schemes, and cognitive radio approaches.

In dynamic RRAM, we typically formulate the RRAM as complex optimization problems. The main objective of such problems is maximizing/minimizing some utility/cost functions, e.g., network sum-rate, EE, and SE, while constraining the available network's radio resources. The state-of-the-art approaches to solve these RRAM optimization problems are heuristic-based, optimization-based, and game theory-based approaches. Such approaches employ advanced algorithms to solve the RRAM problem either optimally or sub-optimally.

a) Heuristic-based techniques: These techniques allocate radio resources sub-optimally and without any performance guarantee. They are typically used to provide approximate and sub-optimal solutions in cases the solution of the formulated optimization problem is quite complex or intractable. Modern wireless systems such as 4G LTE implement some types of greedy heuristics [64]. Examples of heuristic algorithms include the recursive branch-and-bound state-space search algorithm [65] and alpha-beta search algorithm [66].

b) Optimization-based techniques: Typically, most of the RRAM optimization problems in modern HetNets are non-convex (e.g., continuous power allocation) [67], combinatorial (e.g., user association and channel access) [24], or mixed-integer nonlinear programming (MINP) (e.g., combined of continuous- and discrete-type problems) [41]. Many algorithms have been developed to systematically solve such problems and find either the global optimum solution or sub-optimal solution, e.g., [24], [68]–[71] . Such algorithms include, fractional programming (FP) [67], [72], Weighted Minimum Mean Square Error (WMMSE) [67], [72], evolutionary algorithms (e.g., particle swarm optimization (PSO) [73], [74], genetic algorithm [75], [76], ant/bee colony optimization algorithm [77], [78]), among others. These algorithms are extremely computationally-extensive and typically executed in a central RNC with full and real-time information about network statistics and CSI.

c) Game theory-based methods: Game theory techniques are used for distributed RRAM in modern HetNets when network entities (i.e., players) cooperate or compete on radio resources. Such techniques have shown efficient results, and they are widely used as tools to model complex wireless optimization problems in a decentralized fashion [1]. In particular, the RRAM problem is formulated as a cooperative or non-cooperative game/optimization problem between network entities (e.g., BSs, RANs' APs, and user devices). In cooperative game techniques, players collaboratively solve the underlying RRAM game using heuristic- or optimization-based techniques to achieve a specific network goal (e.g., sum-rate or SE/EE maximization). However, in non-cooperative game techniques, players try to solve the RRAM game in a greedy and non-collaborative fashion in order to achieve their own goal (e.g., to satisfy their own QoS demands). The main goal of most game theory algorithms is to find the Nash Equilibrium (NE) solution for the underlying RRAM problem.

C. LIMITATION OF CONVENTIONAL RRAM TECHNIQUES

Unfortunately, all these state-of-the-art approaches will encounter severe limitations in future HetNets, which mainly motivate the usage of DRL in RRAM. Here we summarize the main limitations, and the interested reader can also refer to [1].

- Most of these approaches require complete or quasi-complete knowledge of the wireless environment, including accurate channel models and real-time CSI. However, obtaining such accurate information in future HetNets is quite difficult or even impossible due to the large-scale, ultra-dense, and massive heterogeneity of the system.
- These approaches are generally not scalable, as they encounter several challenges when the number of user devices becomes very large or when used in UDNs.

The main reason is that the optimization space becomes prohibitively large to cater to the whole network, which will lead to a significant increase in computational complexity when finding optimal solutions. With the large-scale and massive heterogeneity of future networks, it becomes essential to engineer and devise more efficient and practical implementations from a computation performance perspective. Also, it becomes challenging in many scenarios to mathematically formulate RRAM optimization problems, or we may end up with non-well-defined or even intractable optimization problems. These cases are encountered for many reasons, including the uncertain nature of wireless channels, network traffic load, and users' mobility patterns. Hence, new innovative RRAM solutions must be developed to address such challenges. In this context, the data-driven AI-based RRAM techniques are feasible alternatives, and they have shown efficient adaptivity when applied on dynamic HetNets.

- Such approaches are heavily system-dependent and will not be accurate for rapidly varying environments. They need, however, reconfiguration to reflect the new system settings. Unfortunately, modern HetNets need to support highly dynamic systems characterized by massive rapidity, such as vehicular and railway networks. This renders conventional methods impractical for such scenarios.
- Most of these methods are computationally expensive and incur considerable timing overhead. This renders them inefficient for most emerging time-sensitive applications, e.g., autonomous vehicles/drones applications. Also, the computational complexity of these methods proportionally increases with the increase in network size, making them unscalable and unsuitable for modern large-scale networks. Furthermore, since most conventional algorithms are computationally expensive, they can be implemented only in sophisticated infrastructures with high computational capabilities, such as supercomputers and servers. Hence, tiny and self-powered user devices will not support them.
- RRAM optimization problems in HetNets are generally complex and non-convex [41]. Hence, leveraging conventional optimization algorithms to solve them will likely result in local optimal solutions rather than global ones. This case is regularly encountered in wireless optimization problems, which have too many local optima.
- Game theory-based techniques are unsuitable for networks characterized by massive heterogeneity in system architecture and user devices. In particular, NE solutions are obtained by assuming that all players are homogeneous, have statistically equal capabilities, and have complete network information. Unfortunately, this is not the case in modern HetNets, in which network entities are massively heterogeneous in terms of physical, communication, and computational capabilities.
- Finally, the complexity of game theory-based techniques and the amount of information exchanged between cooperating/competing players is proportional to the number of playing nodes. Unfortunately, future HetNets will be prohibitively large-scale in terms of the number of network APs and user devices [2], [6]. Hence, such techniques will fail. In particular, exchanging and updating the tremendous amount of data and signaling among the massive number of players will create extra and unmanageable overhead as well as a drastic increase in delay, computation, and energy/memory consumption of network players.

D. ADVANTAGES OF USING DRL-BASED TECHNIQUES FOR RRAM

Emerging AI tools, such as ML, DL, and DRL methods, have been recently used to effectively address various problems and challenges in different areas of wireless communications and networking, including RRAM [1], [8], [13], [15], [24], [26], [27], [79], [80]. Next generation wireless networks will generate a tremendous amount of data related to network statistics, such as user traffic, channel occupancy, channel quality, etc. AI algorithms can leverage this data to develop automated and fine-grained schemes to optimize network radio resources. This paper is solely dedicated to providing a comprehensive survey on DRL applications for RRAM in modern wireless networks. However, the applications of ML and DL techniques in various wireless networks fields can be found in [1], [24], [26], [27], [81] and the references therein.

DRL is an advanced data-driven AI technique that combines neural networks (NNs) with traditional reinforcement learning (RL). It is mainly utilized to enhance the learning rate of RL algorithms and address wireless communication and networking problems having high dimensionality [8], [9], [36], [37]. DRL techniques have gained considerable fame lately to their superiority in making judicious control decisions in uncertain environments like the wireless channels. They enable various network components such as BSs, RAT APs, edge servers (ESs), gateways nodes, and user devices to make autonomous and local decisions, such as RRAM, RATs selection, caching, and offloading, that achieve the objectives of various wireless networks, including sum-rate maximization and SE/EE maximization. Since traditional approaches will not be able to address the RRAM issue of future wireless networks, DRL methods have been proposed lately to be alternative solutions. In particular, DRL techniques are appealing for next generation communication networks due to the following distinct features.

First, they enable network controllers to solve complex network optimization problems, including RRAM and other wireless control problems, with only limited information about the wireless networks. Second, DRL methods enable network entities (e.g., BSs, RAT APs, ESs, gateways nodes, and user devices) to act as agents (i.e., decision-makers) to learn and build knowledge about the wireless environment. This is achieved by learning optimal policies, such as radio resource allocation, RATs selection, and scheduling

TABLE 4. List of the model-free DRL algorithms that are widely used in RRAM for modern wireless networks.

Family	Algorithm	Action Space	Policy Type
Value-Based	Q-Learning	Discrete (also discrete state space)	Off
	DQN	Discrete	
	Double DQN	Discrete	
	Dueling DQN	Discrete	
Policy-Based	REINFORCE	Discrete & Continuous	On
	A2C-A3C	Discrete & Continuous	On
	DDPG	Continuous	Off

decisions, based on continuous interaction between agents and the wireless environment, without knowing the accurate channel models or statistics of the underlying systems *a-priori*. DRL algorithms employ the data collected during the continuous interaction with the environment as a training data-set to train their models. Once DRL agents learned the optimal policies, they can be deployed in an online fashion to make intelligent and autonomous decisions based on local observations made on the wireless environment.

DRL techniques provide efficient solutions from both the network and user devices' points of view to overcome the problems of the conventional RRAM approaches. By employing DRL techniques, various network entities are enabled to learn wireless environments in order to optimize system configuration. Networks entities will be able to optimally and autonomously allocate the optimal transmitting power to mitigate signals interference and reduce energy consumption. For this purpose, advanced DRL techniques such as the deep deterministic policy gradient (DDPG) method and its variants can be utilized. On the other hand, DRL can also enable smart devices to autonomously access the radio channels. For this purpose, deep Q-network (DQN) and its variants can be leveraged. The wireless channels are extremely stochastic due to, e.g., the rapid mobility of user devices and channel objects. Hence, accurate and real-time knowledge of channel state information (CSI) becomes quite difficult, and DRL techniques can be efficiently used to learn wireless channel statistics.

Finally, spectrum prediction and forecasting is also another promising field enabled by DRL techniques. Emerging DL models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), can be integrated with DRL to add the "prediction" capability to the DRL algorithms. Also, conventional optimization techniques do not incorporate the context, and hence they cannot adapt and react according to the sudden variations and changes in the wireless environments. Therefore, such conventional approaches will result in unreliable and poor resource management and utilization. DRL techniques can, however, dynamically adapt and learn the context of wireless environments, which makes their RRAM solutions more accurate and reliable.

To sum up, DRL techniques are required in RRAM problems in four main scenarios; when there is insufficient knowledge about the statistics of the wireless networks, accurate mathematical models do not exist, inference information is required to be incorporated into the decision process,

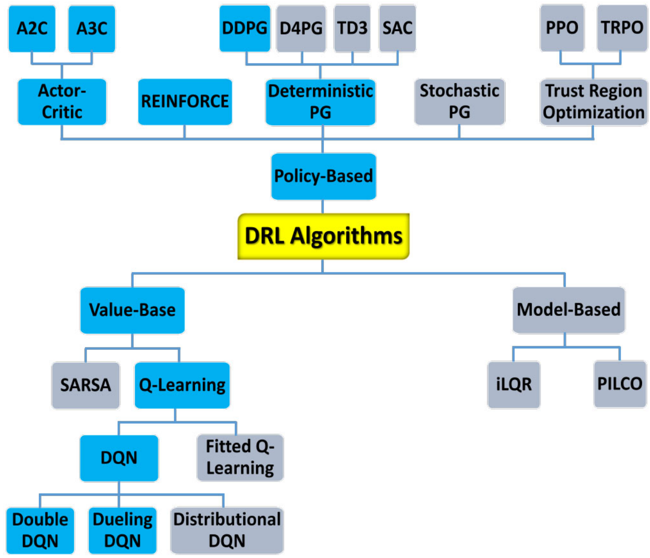

FIGURE 7. Taxonomy of all DRL algorithms [37]. Algorithms colored in blue are covered in Section III.

or a mathematical model exists, but applying conventional algorithms is not possible. In general, most of the RRAM problems in modern wireless networks fall under the above scenarios. The main reason is the large-scale and massive heterogeneity nature of networks in terms of types and numbers of underlying infrastructures, user devices, and QoS demands of applications.

All the aforementioned unique features of DRL techniques make them one of the leading AI-based enabling technologies that can be leveraged to address the RRAM in future wireless communication networks [2], [3].

III. OVERVIEW OF DRL TECHNIQUES USED FOR RRAM

In this section, we briefly review the foundations of DRL, such as the Markov Decision Process (MDP), and show how RRAM problems can be modeled as MDPs. Fig. 7 shows a detailed taxonomy of existing DRL techniques/algorithms. Reviewing all these techniques is beyond the scope of this paper, and we rather focus on the most widely used ones in the literature to address RRAM problems. Interested readers, however, can refer to [7], [15] for a thorough review of the remaining algorithms. Furthermore, we briefly review other emerging technologies used for RRAM problems, such as multi-agent DRL models. Hence, this section is deliberately designed to provide the reader with adequate knowledge of the basics, advantages, limitations, and use-cases of the most widely used DRL techniques employed in the RRAM field.

Table 4 lists the most widely used DRL techniques/algorithms in RRAM of modern wireless networks. Note that all of them are model-free learning algorithms, which means that the agent does not build a model of the wireless environment or reward; instead, it directly maps states to the corresponding actions.

Depending on the dimensionality of the RRAM problem, we can select the most appropriate DRL algorithm that fits the problem settings. For example, RRAM problems could have discrete action space, such as channel access, user association, RAN assignment, etc., or could have continuous action space, such as power allocation and continuous spectrum allocation.

A. THE MARKOV DECISION PROCESS (MDP)

Under the uncertain and stochastic environments of modern HetNets, the problem of RRAM, or any decision-making problem including control problems, are typically modeled by the so-called Markov Decision Process (MDP). It provides a mathematical framework for modeling decision-making problems whose outcome is random and controlled by a decision-maker, aka agent. The MDP also has another variant, called partially observable MDP (POMDP), which models decision-making problems in partially observable wireless environments.

The general practice in RRAM is to formulate the radio resource allocation (RRA) as an optimization problem whose objective is to maximize/minimize some network utility/cost function while constraining on the available network radio resources and optional QoS demands of user devices. However, as we discussed in Section II, tremendous challenges are encountered during formulating such problems or/and even during solving them, which renders conventional approaches inapplicable. Hence, RL/DRL techniques are utilized instead.

In order to apply DRL to solve RRA problems, we need first to convert the formulated optimization problem into the MDP framework. The resultant MDP-based model must contain seven elements: the agent(s), environment, action space $\mathcal{A}$, state space $\mathcal{S}$, instantaneous reward function r, a transition probability p, and policy π, as shown in Fig. 8. The MDP is represented mathematically by the tuple ($\mathcal{S}$, $\mathcal{A}$, p, r).

In RRAM problems, the dynamicity of the agent's learning process according to the MDP framework is shown in Fig. 8. At time t, the agent observes a state s_t from the state space $\mathcal{S}$. The state space should contain useful and effective information about the wireless environment, such as available radio resources, SNR, the number of user devices, and required QoS. Then, the agent takes action a_t from the action space $\mathcal{A}$ such as the RRA and RAN assignment. The taken action must achieve network utility goal, such as sum-rate/SE/EE maximization. Then the state moves to a new state s_{t+1} with a transition probability p, and the agent receives a feedback numerical instantaneous reward r_t, which quantifies the quality of the taken action. This interaction, i.e., (s_t, a_t, r_t, s_{t+1}), between the agent and wireless environment repeatedly continues, and the agent will utilize the received reward to adjust its strategy until it learns the optimal policy π^*. The agent's policy π defines the mapping from states to the corresponding actions $\mathcal{S} \rightarrow \mathcal{A}$, i.e., $a_t = \pi(s_t)$. Typically, we define the long-term reward as the expected accumulated discounted instantaneous reward over the time horizon T, which is given by $\mathcal{R} = \mathbb{E}[\sum_{t=1}^{T} \gamma r_t(s_t, \pi(s_t))]$. The parameter $0 \leq \gamma \leq 1$ is the discounted factor, which trades-off between instantaneous and future rewards. The main goal of the agent in MDP is to obtain π^* (i.e., allocating optimal radio resources) that maximizes the long-term reward, i.e., $\pi^* = \max_{\pi} \mathcal{R}$.

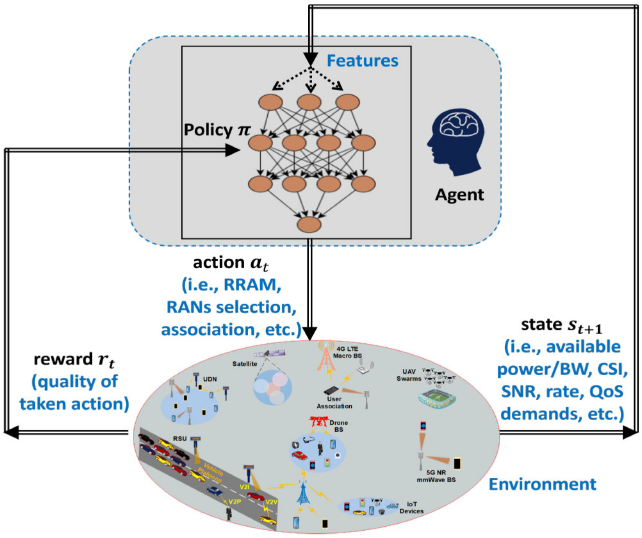

FIGURE 8. Framework of DRL models [15].

Next, we discuss the most widely used DRL algorithms to handle MDP problems, i.e., RRAM problems. As shown in Fig. 7, these algorithms belong to two main families of methods; the value-based and the policy-based methods.

B. VALUE-BASED ALGORITHMS

This family of methods is used to estimate the value function of the agent. This value function is then utilized to implicitly and greedily obtain the optimal policy. Two value functions exists; the value function $V^{\pi}(s)$ and the state-action function $Q(s_t, a_t)$. Both represent the expected accumulated discounted rewards received when taking action a_t (in state s_t for $V^{\pi}(s)$) (or at pair (s_t, a_t) for $Q(s_t, a_t)$) and then following the policy π thereafter. These functions are important as they represent the link between the MDP mathematical formulation and the DRL formulation, and they are given by [7]:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t, s_{t+1}) | a_t \sim \pi(.|s_t), s_0 = s\right],$$

$$Q^{\pi}(s,a) = \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t, s_{t+1})| \\ a_t \sim \pi(.|s_t), s_0 = s, a_0 = a\Bigg].$$

The optimal value function $V^*(s)$ and state-action function $Q^*(s,a)$ are obtained by solving the following Bellman equations [7], [15]:

$$V^*(s) = \max_{a_t}\big[r_t(s_t, a_t) + \gamma \mathbb{E}_{\pi} V^*(s_{t+1})\big],$$

$$Q^*(s,a) = r_t(s_t, a_t) + \gamma \mathbb{E}_{\pi}\Big[\max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})\Big].$$

Recall that the main goal of MDP is to obtain the optimal policy π^* (i.e., mapping states to optimum actions), which is given by $\pi^* = \underset{\pi}{\text{argmax}}\ \mathcal{R} = \underset{\pi}{\text{argmax}} = \mathbb{E}[\sum_{t=1}^{T} \gamma r_t(s_t, \pi(s_t))]$. Hence, the optimal actions can be obtained to be the ones that maximize the above value functions, and the optimal policy will be the one that maximizes these values functions [7]. In particular, the Q-function $Q^{\pi}(s,a)$ is commonly used, and the problem of obtaining the optimal policy becomes $\pi^*(s) = \underset{a}{\text{argmax}}\ Q^{\pi^*}(s_t, a_t)$. The ultimate goal of all the value-based DRL algorithm is to approximate this function as discussed next.

1) Q-LEARNING TECHNIQUE

In RL, Q-learning is one of the most widely used algorithms to address MDPs. It obtains the optimal values of the Q-function iteratively using the following Bellman equation:

$$Q(s_t, a_t) = Q(s_t, a_t) \\ + \alpha_t\Big[r_t(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\Big]$$

where α_t is the learning rate that defines how much the new information contributes to the existing Q-value. The main idea of this Bellman rule relies on finding the Temporal Difference (TD) between the current Q-value $(Q(s_t, a_t))$ and the predicted Q-value $(r_t(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$. The Q-learning algorithm uses this rule to construct a table of all possible Q values for each stat-action pair. The algorithm terminates when we reach a certain number of iterations or when all Q-values have converged. In such a case, the optimal policy will determine the optimal action to take at each state such that $Q^{\pi^*}(s_t, a_t)$ is maximized for all states in the state space, i.e., $\pi^* = \underset{a_{t+1}}{\text{argmax}}\ Q^{\pi^*}(s_t, a_t)$.

However, the Q-learning algorithm has many limitations when applied for RRAM in modern HetNets. First, it is applicable only to problems with low dimensionality of both state and action spaces, making it unscalable. Second, it is applicable only on RRAM with discrete state and action spaces, such as channel access and RANs assignment. If, however, they are applied to problems with continuous action spaces, e.g., power allocation, the action space must be digitized. This renders them inaccurate due to quantization error.

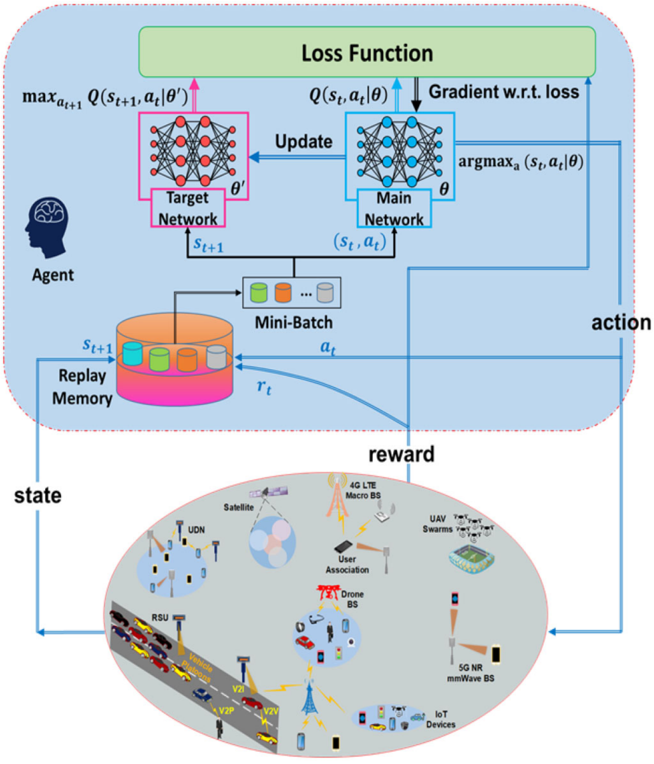

FIGURE 9. Illustration of the DQN architecture.

2) DEEP Q NETWORK (DQN) TECHNIQUE

Since the Q-learning algorithm relies on building a table for the Q values, it will fail to obtain the optimal policy when the state and action spaces become prohibitively large. This case is commonly encountered in the RRAM problems of modern HetNets. To overcome this issue, the DQN algorithm has been developed, which inherits the advantages of Q-learning and DL techniques. The main idea is to replace the table in the Q-learning algorithms with a DNN that approximates the Q values, i.e., $Q(s_t, a_t|\theta)$, where θ represents the training parameters (i.e., weights) of the DNN. Fig. 9 shows the DQN architecture. The replay memory is denoted by $\mathcal{D}$, and it is mainly used to break the correlation between the training samples, i.e., (s_t, a_t, r_t, s_{t+1}), by making them independently and identically distributed i.i.d. During the learning process of the policy, we store the training transitions generated during the interaction with wireless environment in $\mathcal{D}$. The DQN's agent will then randomly select minibatch transition samples from $\mathcal{D}$ to train its DNN. To enhance the DQN model's stability, the target Q network is used, whose weights θ' will be periodically updated to track those of the main Q network.

Since the DQN algorithm is mainly used to learn the optimal policy, i.e., $\pi^* = \underset{a}{\text{argmax}}\ Q^{\pi^*}(s_t, a_t)$, the optimal Q-function is derived from the following iterative Bellman equation:

$$Q(s_t, a_t) = r_t(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_t),$$

and the DQN algorithm is then optimized by iteratively updating θ to minimize the following Bellman loss function;

$$L(\theta_t) = \mathbb{E}_{s_t,a_t,r_t,s_{t+1}\in\mathcal{D}}\Big[r_t(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_t|\theta') - Q(s_t, a_t|\theta)\Big]^2.$$

The DQN algorithm is applicable to a wide variety of RRAM problems, specifically for problems characterized by their discrete action space. As we will elaborate in-depth in Section IV, the DQN technique can be used efficiently for channel allocation, access control, spectrum access, user association, and RANs assignment. The DQN algorithm can also be used for RRAM problems with continuous action space, such as power control, by discretizing the action space. However, such a methodology makes DQN vulnerable to serious quantization error that may considerably deteriorate its accuracy. There are also other limitations in the basic DQN, and various DRL algorithms have been proposed to overcome them, as we discuss in the following sections.

3) DOUBLE DQN ALGORITHM

The Double DQN technique has been proposed in [82] to enhance the basic DQN algorithm. The DQN algorithm tends to overestimate the Q values, which can degrade the training process and lead to suboptimal policies. The overestimation results from the fact that the same training transitions are utilized in selecting and evaluating an action in the Bellman equation. As a solution, the authors in [82] propose to use two Q value functions, one for selecting the best action and the other to evaluate the best action. The action selection is still based on the online weights θ, while the second weights parameters θ' are used to evaluate the value of this policy. So, as in the conventional Q learning, the value of the policy is still estimated based on the current Q values. The weights θ' are updated via switching between θ and θ'.

The target Q values are derived from the following modified Bellman equation [82]:

$$Q(s_t, a_t) = r_t(s_t, a_t) + \gamma Q\Big(s_{t+1}, \operatorname*{argmax}_{a_{t+1}} Q(s_{t+1}, a_t|\theta_t), \theta'_t\Big),$$

and the Double DQN algorithm uses the following modified Bellman loss function to update its weights;

$$L(\theta_t) = \mathbb{E}_{s_t,a_t,r_t,s_{t+1}\in\mathcal{D}}\Big[r_t(s_t, a_t) + \gamma Q\Big(s_{t+1}, \operatorname*{argmax}_{a_{t+1}} Q(s_{t+1}, a_t|\theta_t), \theta'_t\Big) - Q(s_t, a_t|\theta_t)\Big]^2.$$

The Double DQN algorithm is also widely used in RRAM problems, as we will discuss in the next section. Although this algorithm has advantages over the basic DQN algorithm, they both share the same shortcomings.

4) DUELING DQN ALGORITHM

This algorithm is another enhancement to the basic DQN algorithm [83]. Recall that the goal of the network is to estimate the Q values, i.e., $Q(s_t, a_t)$. This function can be divided into two terms; the state-value function $V(s)$, which tells the importance of being in a particular state, and the action-value function (or the advantage function) $A(s, a)$, which tells the importance of selecting a particular action among all available actions. Hence, the Q value function can be written as $Q(s, a) = V(s) + A(s, a)$. The authors in [83] utilized this concept and suggested having two independent paths of fully-connected layers instead of having only a single path as the case in the basic DQN. One path will estimate $V(s)$, and the other will estimate $A(s, a)$. The two paths will eventually be combined to produce a single output, which is $Q(s, a)$. Here, the loss function is obtained similar to the DQN and Double DQN algorithms.

C. POLICY-BASED ALGORITHM

The policy-based techniques are part of the policy gradient family of methods. They provide an alternative way to solve MDP problems having high dimensionality and continuous action spaces. Recall that the main idea of the value-based methods discussed before is to find the state-action value function $Q(s, a)$. This function is defined as the expected total discounted reward received by taking a particular action from the state. If these Q values are known, the optimal policy is obtained by selecting actions that maximize the Q values in each state. However, in environments with continuous action spaces, such as power control in wireless systems, the Q function cannot be obtained as it is impossible to conduct a full search in a continuous action space to obtain the optimal action. Hence, value-based methods are inaccurate for such problems, and the policy-based methods are applied instead.

In policy-based approaches [7], [84], we avoid calculating Q values and directly obtain the optimal policy $\pi_\theta(a|s)$ that maximizes the agent's expected accumulated reward J, i.e., $J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t)]$. The policy gradient approaches learn the optimal weights θ^* via performing gradient ascent on the function J. In particular, the policy gradients are derived from trajectories obtained via the current policy, such that in each gradient update the agent interacts with the environment to collect new and fresh trajectories, and this is why policy-gradient methods are called on-policy algorithms.

1) REINFORCE ALGORITHM

The main idea of this algorithm is to increase the probabilities of good actions and reduce the probabilities of bad ones. The REINFORCE algorithm differs from the Q learning methods in three aspects. First, REINFORCE algorithm does not need a replay buffer $\mathcal{D}$ during training as it belongs to the on-policy family, which requires only fresh training transitions. Although this enhances its convergence speed, it needs more interaction with the environment. Second, the REINFORCE algorithm implicitly performs the exploration process, as it depends on the probabilities returned by the network, which incorporate uniform random agent behavior. Third, no target network is required in the REINFORCE method as the Q values are obtained from the experiences in the environment.

The disadvantage of the REINFORCE algorithm is that it suffers from high variance, meaning that any small shift in the return leads to a different policy. This limitation motivated the actor-critic algorithms.

2) ACTOR-CRITIC ALGORITHM

The actor-critic methods are mainly developed to enhance the convergence speed and stability (i.e., reducing the variance) of the policy-gradient method. Like the policy-based methods, it utilizes the accumulated discounted reward J to obtain the gradient of policy ∇J, which provides the direction that enhances the policy. This algorithm learns a critic to reduce the variance of gradient estimates since it utilizes various samples, whereas the REINFORCE algorithm utilizes only a single sample trajectory.

To select the best action in any state, the total discount reward of the action is used, i.e., $Q(s, a)$. The total reward can be decomposed into state-value function $V(s)$ and advantage function $A(s, a)$, i.e., as $Q(s, a) = V(s) + A(s, a)$. So, another DNN is utilized to estimate $V(s)$, which is trained based on the Bellman equation. The estimated $V(s)$ is then leveraged to obtain the policy gradient and update the policy network such that the probabilities of actions with good advantage values are increased. Hence, the *actor* is the policy network $\pi(a|s)$ that takes actions by returning the probability distribution of actions, while the *critic* network evaluates the quality of the taken actions, $V(s)$. This algorithm is also called the advantage actor-critic method (A2C).

In the A2C algorithm, the weights of actor network θ_π and critic network θ_v are updated using the accumulated policy gradients $\partial\theta_\pi$ and value gradients $\partial\theta_\mathrm{v}$, respectively, to move in the direction of the policy gradients and the opposite direction of the value gradients.

3) A3C ALGORITHM

The asynchronous advantage actor-critic (A3C) algorithm is an extension of the basic A2C [85]. This algorithm is used to solve the high variance issue in gradients that results in non-optimal policies. A3C algorithm conducts a parallel implementation of the actor-critic algorithm, where the actor and critic share the network layers. A global NN is trained to output action probabilities and an estimate of the advantage function $A(s_t, a_t|\theta_\pi, \theta_\mathrm{v})$ given by $\sum_{i=0}^{k-1} \gamma^i r_{t+1} + \gamma^k V(s_{t+k}|\theta_\mathrm{v}) - V(s_t|\theta_\mathrm{v})$, where k depends on the state and upper-bounded by the maximum number of time steps.

Several parallel actor learners are instantiated with copies of both the environment and global NN weights. Each learner independently interacts with its environment and gathers training transitions to derive the gradients with respect to its NN weights. Learners will then propagate their gradients to the global NN to update its weights. This mechanism ensures a periodic update of the global model with diverse transitions from each learner.

4) DEEP DETERMINISTIC POLICY GRADIENT (DDPG) ALGORITHM

DDPG is one of the most widely used DRL techniques in addressing RRAM problems for wireless networks characterized by their high dimensionality and continuous action space [86]. DDPG algorithm belongs to the actor-critic family, and it combines both Q-learning and policy gradients algorithms. It consists of actor and critic networks. The actor network takes the state as its input, and it outputs the exact "deterministic" action, not probability distribution over actions as in the actor-critic algorithm. Whereas the critic is a Q-value network that takes both the state and action as inputs, and it outputs the Q-value as a single output.

The deterministic policy gradient (DPG) algorithm is proposed in [87] to overcome the limitation caused by the *max* operator in the Q-learning algorithm, i.e., $\max_{a_{t+1}} Q(s_{t+1}, a_t)$. It simultaneously learns both the Q-function and the policy. In particular, the DPG algorithm has a parameterized actor function $\mu(s|\theta^\mu)$ with weights θ, which learns the deterministic policy that gives the optimal action corresponding to $\max_{a_{t+1}} Q(s_{t+1}, a_t)$. The critic $Q(s, a)$ is leaned via minimizing the Bellman loss function as in the Q-learning algorithm.

The learning process of the actor policy is updated using gradient ascent with respect to θ^μ in order to solve the objective given by the following chain rule [87]:

$$J(\theta) = \mathbb{E}_{s\in\mathcal{D}}\big[Q(s, \mu(s|\theta^\mu))\big],$$
$$\nabla_{\theta^\mu} J = \mathbb{E}_{s\in\mathcal{D}}\Big[\nabla_a Q(s, a|\theta^Q)|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s=s_t}\Big].$$

The DDPG algorithm proposed in [86] is built based on the DPG algorithm, where both the policy and critic are DNNs, as shown in Fig. 10. The DDPG algorithm creates a copy of both the actor and critic networks, $Q'(s, a|\theta^{Q'})$ and $\mu'(s|\theta^{\mu'})$, respectively, to compute the target values. The weights of these target networks, $\theta^{Q'}$ and $\theta^{\mu'}$, are then updated to slowly track the weight of the learned network to provide more stable training using $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$ with $\tau \ll 1$. The critic network is updated to minimize the following Bellman loss;

$$L(\theta^Q) = \mathbb{E}_{s_t, a_t, r_t, s_{t+1}\in\mathcal{D}}\Bigg[(r_t(s_t, a_t) + \gamma \max_{a_{t+1}} Q\Big(s_{t+1}, \mu\Big(s_{t+1}|\theta^{\pi'}\Big)|\theta^{Q'}\Big) - Q\Big(s_t, a_t|\theta^Q\Big)\Big)^2\Bigg].$$

Note that the DDPG algorithm is off-policy, which means that we use a replay buffer $\mathcal{D}$ to store training transitions.

The exploration-exploitation issue is addressed by adding the Ornstein–Uhlenbeck (OU) process or some Gaussian noise $\mathcal{N}$ to the action selected by the policy, i.e., $\mu(s_t|\theta_t^\mu) + \varepsilon\mathcal{N}$ [86].

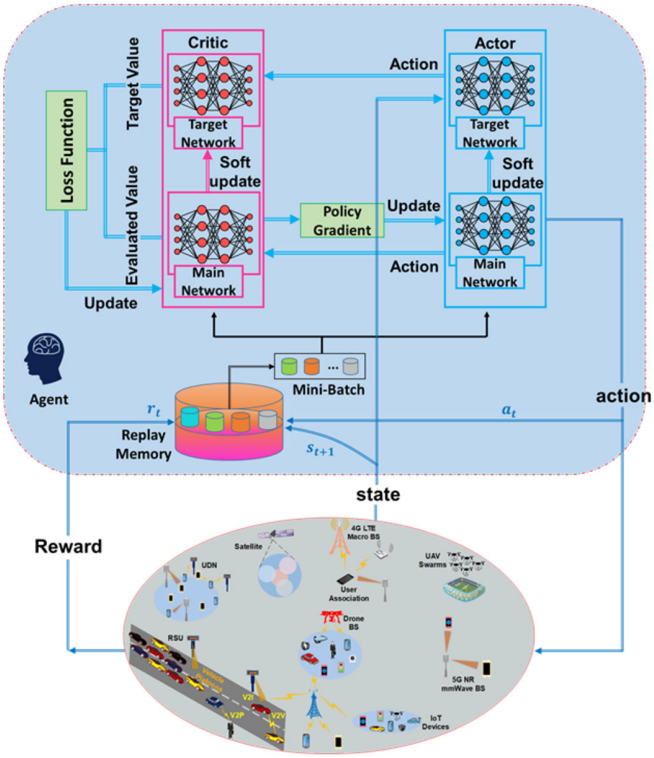

FIGURE 10. Illustration of the DDPG actor-critic architecture [88].

D. OTHER DRL ALGORITHMS

The DRL algorithms discussed above are the commonly used approaches to address the problem of RRAM in wireless networks, as we will discuss in the next section. Although there are several other algorithms, they are rarely utilized for such types of problems. Therefore, they are not included in this article. However, generally speaking, all the other variants are mainly developed to enhance the performance of the basic algorithms discussed above. For completeness, this section highlights some of these variants for the interested reader.

Other variants of the value-based algorithms are developed to enhance the performance of vanilla DQN algorithm in terms of stability, convergence speed, implementation complexity, sample/learning efficiency, etc. Such variants include prioritized experience replay DQN [89], distributed prioritized experience replay DQN [90], distributional DQN [91], Rainbow DQN [92], and recurrent DQN [93].

For the policy-based algorithms, several algorithms are envisioned to enhance the overestimation issue, such as the Twin Delayed DDPG (TD3) [94], enhance stability and robustness, such as the Soft Actor-Critic (SAC) [95], and to enhance stability, convergence, and sample efficiency, such as the distributed distributional DDPG (D4PG) [96].

E. MULTI-AGENT DRL ALGORITHMS

Multi-agent DRL (MADRL) is a natural generalization of the single-agent DRL that allows multiple agents to concurrently learn optimal RRAM policies based on their interactions with the environment and with each other. These agents can either be deployed cooperatively, in which all agents interact with each other to learn the same global policy, or non-cooperatively, in which each agent learns its own policy. MADRL provides several performance advantages over the single-agent case regarding the quality of the learned policies, convergence speed, etc. However, it encounters several challenges such as scalability, partial observability, and agents' non-stationarity. Nguyen *et al.* [97] provide a survey on MADRL systems and their applications. Different methods are reviewed along with their advantages and disadvantages. In [98], the authors provide a selective overview of the theories and algorithms for MARL.

MADRL is widely employed in addressing various RRAM problems in modern wireless networks. The authors in [14] provide an overview of the MADRL algorithms and highlight their applications in future wireless networks. The learning frameworks in MADRL are also investigated. The application of MARL in solving problems for vehicular networks is studied in [99]. In [100], an overview of the evolution of cooperative MARL algorithms is presented with an emphasis on distributed optimization.

Most of the RRAM problems in modern HetNets are of a multi-agent nature [14]. Network entities such as user devices, BSs, and APs can act as cooperative/non-cooperative multi-agents to learn optimal RRA policies and solve complex network optimization problems. For example, channel access control may be formulated as a MADRL problem in which each user device represents a learning agent that senses the radio channels and coordinates with other agents to avoid collisions. Next, we discuss how RRAM problems in HetNets are formulated and solved using these algorithms.

IV. DRL-BASED RESOURCE ALLOCATION AND MANAGEMENT FOR FUTURE HETEROGENEOUS NETWORKS

This section provides an extensive and in-depth review of the related works for RRAM using DRL techniques. We classify them based on the radio resources (or issues) they investigate as well as based on the wireless network types they cover, as shown in Figs. 2 and 3, respectively. It must be noted that this survey is dedicated to only the application of DRL algorithms for radio resources, i.e., no computation resources are covered, which can be found in [15].

DRL algorithms enable various network entities to efficiently learn the wireless networks, which allows them to make optimal control decisions that achieve some network utility function. For example, DRL methods can be deployed to maximize network sum-rate, minimize network energy consumption, or enhance spectral efficiency. In this section, we review the applications of DRL methods in the following RRAM issues: power allocation, spectrum allocation and access control, rate control, and the joint use of these radio resources.

A. DRL FOR POWER ALLOCATION

Energy-efficient communication is one of the main objectives of modern wireless networks. It is achieved via efficient

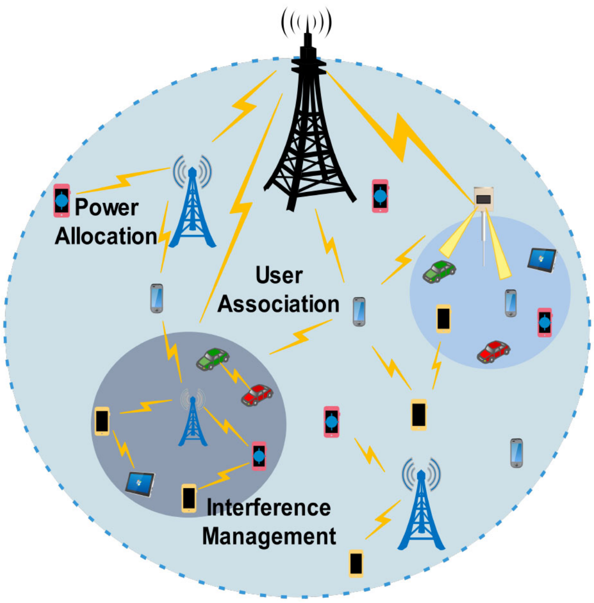

FIGURE 11. Importance of power allocation in modern wireless communication networks.

power allocation to ensure high QoS, better coverage, and enhanced data rate, as shown in Fig. 11. Power allocation is mainly involved in vital network operations such as modulation and coding schemes, path loss compensation, interference management, etc. On the other hand, almost all modern user devices and IoT sensors are battery-powered with very limited battery capacity and charging capabilities. Hence, designing energy-efficient resource allocation schemes, protocols, and algorithms becomes fundamental in dynamic wireless network environments.

Several conventional approaches have been applied for power allocation and management. Most of them rely on solving power-constrained optimization problems, such as FP algorithm [72] and WMMSE algorithm [101]. These approaches are iterative and model-driven, which means that they need a mathematically tractable and accurate model. They are typically executed in a centralized fashion in which a network controller has full CSI. In such a mechanism, BSs, wireless APs, and/or user devices require to wait until the centralized controller's iterations converge and send the outcome back over backhaul links. However, as discussed in Section II, such approaches become impractical due to the large-scale nature of modern wireless networks and the difficulty in obtaining accurate and instantaneous CSI. Hence, DRL techniques are used instead due to their superiority in obtaining optimal power allocation policies based on limited CSI.

1) IN CELLULAR NETWORKS

In the following paragraphs, we review works that employ DRL algorithms to address the power allocation problem in cellular, cellular IoT, and wireless homogeneous networks (HomNets) depicted in Fig. 3.

Power allocation in small-cell multi-user cellular systems is fundamental to increase system performance while reducing inter-cell interference. In [102], the authors propose single- and multi-agent actor-critic DRL methods to tackle the problem of downlink sum-rate maximization through power allocation in multi-cell, multi-user cellular networks. In their model, the agents are the base stations (BSs), whose state space is continuous and comprises network CSI and the transmit power allocation by previous BSs. The action space is continuous, representing the power allocation, while the reward function is the cellular network sum SE. Experimental results demonstrate that their proposed DRL-based method can both achieve higher SE than conventional optimization algorithms, such as fractional programming (FP) and weighted minimum mean-squared error (WMMSE), while performing two times faster than these conventional methods.

On the same context, the authors in [103] address the power allocation issue by building on their initial investigation in [104]. A multi-agent DQN-based DRL algorithm is proposed in which each BS-user link is considered as an agent. The state space is continuous, comprised of a logarithmic normalized interferer, the link's corresponding downlink rate, and the transmitting power. The action space is discrete, corresponding to the downlink power allocation, while the reward is continuous, which is a function of the downlink data rate of the communication link. Experimental results indicate that their proposed DQN outperforms benchmark algorithms such as FP, WMMSE, random power allocation, and maximum power allocation in terms of achievable averaged sum-rate and the convergence time when considering different user densities.

A pioneer work is presented in [105], in which the authors design a multi-agent DQN and DDPG-based DRL framework to address the problem of power allocation in HetNets. A centralized-training-distributed-execution algorithm is designed in which the APs are the agents, each of which implements a local DNN. The state space of each local DNN is continuous, representing the local state information, while the local action space is continuous, representing the transmit power. Then, multiple-actor-shared-critic method (MASC) is proposed to separately train each of these local DNN in an online fashion. The main idea is that the MASC training method is composed of multiple actor DNNs and a shared critic DNN. An actor DNN is first established in the core network for each local DNN, and the structure of each actor DNN is the same as the corresponding local DNN. Then, a shared critic DNN is established in the core network for these actor DNNs. Historical global information is provided into the critic DNN, and the output of the critic DNN will evaluate whether the output power of each actor DNN is optimal or not from a global view. The reward function is continuous, representing the data rate between each AP and its associated user. Simulation results show that their proposed algorithm outperforms the WMMSE and FP algorithms in terms of both convergence rate and computational complexity.

Similar to the work in [105], the authors in [106] address the problem of sum-rate maximization via continuous power allocation in wireless mobile networks based on a distributive multi-agent DDPG algorithm. Unlike authors' previous work in [107], which was based on the DQN technique, the authors extended their work to leverage the unique advantages of the DDPG algorithm when addressing problems with continuous state space nature. Particularly, in [105], the agents are each transmitter (e.g., mobile devices, links, etc.) whose state is a combination of three feature groups; the local information, interfering neighbors, and interfered neighbors feature groups. Each agent's action is to choose the transmit power level, while the reward is a function of the sum-rate maximization problem. Simulation results show that their proposed method gives better performance results than the conventional FP methods and comparable results with the WMMSE methods.

D2D underlying cellular communication has emerged as one of the main enabling technologies for modern wireless networks. Establishing communication links in such highly dynamic environments is an essential issue. In this context, the authors in [108] present a centralized multi-agent DQN-based DRL algorithm to address the problem of power allocation of D2D cellular communication in a time-varying environment. The agents are the D2D transmitters, whose state space is continuous, comprised of the SINR and channel gain of users. The action space is discrete, representing the transmit power of each D2D user, while the reward is a function of system throughput. Simulation results show that their proposed algorithm outperforms the traditional RL methods in terms of network capacity and user's achieved QoS.

5G UDNs are characterized by their high vulnerability to inter-cell interference, which can be greatly reduced via judicious power management. Towards this, Saeidian *et al.* [109] propose a data-driven approach based on a multi-agent DQN algorithm to tackle the downlink power control in dense 5G cellular networks. The agents are the BSs, whose state space is continuous, comprised of path-gain, SINR, downlink rate, and downlink power. The action space is discrete, representing the downlink power, while the reward is a function of the network-wide harmonic-mean of throughput. Simulation results indicate that their approach can improve data rates at the cell edge while ensuring a reduced transmitted power compared to the baseline fixed power allocation approaches.

Non-orthogonal multiple access (NOMA) technology has recently emerged as an efficient tool to enhance the QoS and EE of millimeter-wave (mmWave) communication systems by enhancing the power level of received signals. The authors in [110] propose a multi-agent DQN-based DRL framework to optimize the EE in downlink full-duplex cooperative NOMA of mmWave UDNs. The agents are the relay near users, whose state space is continuous, consisting of information related to wireless environment and channel, the user's battery capacity, energy power transfer coefficient, self-interference cancellation residue coefficient, and the buffer size of nearby relay users. The action space is to specify the required user pairing between the near relay user group and edge user group, along with the pre-processing of EE power allocation. The reward is a function of the EE of the mmWave network. Experimental results are compared with a conventional centralized iteration algorithm, which demonstrate both the superiority of their proposed algorithm in terms of the convergence speed and the efficiency to provide near-optimal results.

DRL methods have also been investigated for beamforming design in cellular networks. The authors in [111] propose a single-agent DDPG-based model to address the problem of SE maximization via hybrid beamforming design in mmWave MIMO cellular systems. The action space is continuous, comprised of the digital beamformer and analog combiner. The state space is also continuous representing the digital beamformer and analog combiner at the previous time step. The reward is a continuous function defined in terms of network SE. Simulation results show the efficiency of their proposed model in terms of SE, bit error rate, and computation time.

2) IN IOT AND OTHER EMERGING WIRELESS NETWORKS

In the following paragraphs, we review works that utilize DRL algorithms to address the power allocation issue in IoT and other emerging wireless networks shown in Fig. 2.

Developing efficient spectrum sharing schemes is regarded as one of the main persistent objectives and challenges in CRNs. In [112], the authors propose a non-cooperative single-agent DQN-based DRL scheme to address the problem of spectrum sharing via power control in CRNs. In their model, the agent is the SU, whose action space is discrete, corresponding to selecting the transmit power from a pre-defined power set. The state space is discrete, defined by four parts; the transmit power of PU and SU, the path loss between PU and a sensor that measures the RSS, the path loss between the SU and a sensor that measures the RSS, and some Gaussian random variable. The reward is a discrete function defined by the achieved SINR level and the minimum SINR requirements of both PU and SU. Simulation results show that their proposed algorithm is robust against random variation in state observations, and the SU interacts with PU efficiently until they reach a state in which both users successfully transmit their own data.

In another interesting work in [25], the authors present a non-cooperative multi-agent algorithm to address the problem of power allocation in D2D underlying communication networks based on three DQNs, namely, DQN, Double DQN, and Dueling DQN. The agents are the D2D transmitters in each D2D pair, whose state space is discrete, comprised of the level of the interference indicator function. The action space is discrete, representing the set of transmitting power levels, while the reward is a function of the system EE. Simulation results show the ability of their DQN-based models to provide energy-efficient power allocation for the underlying D2D network.

UAV IoT networks are attracting considerable attention recently due to their ability to provide enhanced QoS communication in harsh and vital environments. However, power management is one of the key challenges in such networks. In this context, the authors in [113] address the problem of downlink power control in ultra-dense UAV networks with the aim of maximizing the network's EE. A multi-agent DQN-based DRL model is proposed in which the agents are the UAVs in the network. The state space is continuous, representing the remaining energy of the UAV and the interference caused by neighboring UAVs. The action space is discrete, representing the set of possible discrete transmit power values, while the reward function is the EE of the UAV network. Simulation results are compared with Q-learning and random algorithms, which show the superiority of their proposed scheme in terms of both the convergence speed and EE.

In the same context for multi-UAV wireless networks, the authors in [114] propose a multi-agent DDPG-based DRL to address the problem of joint trajectory design and power allocation. In their scheme, the agents are the UAVs, whose state space is a discrete binary indicator function representing whether the QoS of the user ends (UEs) are satisfied or not. The action space is also discrete, corresponding to selecting UAVs' trajectory and transmission power. The reward is a continuous function defined by the joint trajectory design and power allocation as well as the number of UEs covered by the UAVs. Simulation results show that the proposed algorithm achieves higher network utility and capacity than the other optimization methods in wireless UAV networks with reduced computational complexity.

Another interesting work [107] proposes a multi-agent DQN-based DRL method to study the problem of transmit power control in wireless networks. The agents are the transmitters whose state space is continuous, consisting of three main feature groups; local information, interfering neighbors, and interfered neighbors. The action space is discrete corresponds to discrete power levels, while the reward is a function of the weighted sum-rate of the whole network. Experimental results demonstrate that the proposed distributed algorithm provides comparable and even better performance results to the state-of-the-art optimization-based algorithms available in the literature.

High-speed railway (HSR) systems are one of the emerging IoT applications for next-generation wireless networks. Such systems are characterized by their rapid variations in the wireless environment, which mandate the development of light-weighted RRAM solutions. As a response to this, Xu and Ai [88] propose a multi-agent DDPG-based DRL model to address the problem of sum-rate maximization via power allocation in hybrid beamforming-based mmWave HSR systems. In their approach, each mobile relay (MR) acts as an agent. The action space is continuous, corresponding to the transmit power level of each MR agent. Also, the state space is continuous, defined by; each MR own signal channel, local observation information of each MR, i.e., beamforming design, each MR achievable rate, and each MR transmit power in the last time step. The reward function is continuous, defined by the achievable sum-rate of the network. Simulation results demonstrate that the SE of their proposed algorithm is comparable to the full digital beamforming scheme, and it outperforms conventional approaches such as maximum power allocation, random power allocation, DQN, and FP.

Federated deep reinforcement learning (FDRL) is an emerging AI paradigm that integrates FD and DRL methods. FDRL can be utilized as an efficient technique to enhance the RRAM solutions in large-scale distributed systems. As an example, an interesting approach is proposed in [115], in which the authors propose a cooperative multi-agent actor-critic-based FDRL framework for distributed wireless networks. The authors particularly address the problem of energy/spectrum efficiency maximization via distributed power allocation for network edge users. In their proposed model, the agents are the edge users, whose action space is continuous, defined as the power allocation strategies. The state space is continuous, defined by the allocated transmit power, the SINR on the assigned RBs, and the reward of the previous training time step. The system is defined in terms of a local continuous cost function expressed in terms of SINR, power, path loss, and environmental noise. Using simulation results, the authors demonstrate that their proposed framework achieves better performance accuracy in terms of power allocation than other approaches such as greedy, non-cooperation power allocation, and traditional FL.

3) IN SATELLITE NETWORKS

In the following paragraphs, we review works that employ DRL techniques to address the power allocation issue in satellite networks as well as emerging satellite IoT systems.

Managing downlink transmit power in satellite networks is also one of the major persistent challenges. To this end, the authors in [116] extended their work in [117] and present a single-agent Proximal Policy Optimization (PPO)-based DRL model to solve the problem of power allocation in multi-beam satellite systems. In their model, the agent is the processing engine that allocates power within the satellite, whose state space is continuous, comprises the set of demand requirements per beam, and the optimal power allocations for the two previous time steps. The action space is continuous, representing the allocation of the power for each beam, while the reward is a function of both the link data rate achieved by the beam and the power set of the agent. Experimental results demonstrate the robustness of their proposed DRL algorithm in dynamic power resource allocation for multi-beam satellite systems.

NOMA technique has shown efficient results in improving the performance of terrestrial mmWave cellular systems [118]. This has motivated the use of NOMA for satellite communication systems. However, managing the radio resources in such a system becomes an imperative issue. In this context, Yan *et al.* [119] conducted a pioneer work to study the problem of power allocation for NOMA-enabled SIoT using a single-agent DQN-based DRL scheme. In their system, the agent is the satellite, whose action space is discrete, corresponding to selecting the power allocation coefficient for each NOMA user. The state space is continuous, consisting of the average SNR, link budget, and delay-QoS requirements of NOMA users, while the reward is discrete, which is a function of the effective capacity of each NOMA user. Experimental results demonstrate that their proposed DRL-based power allocation scheme can produce optimal/near-optimal actions, and it provides superior performance to both the fixed power allocation strategies and OMA scheme.

4) IN MULTI-RAT NETWORKS

Multi-RAT wireless HetNets is one of the main enabling technologies for modern wireless systems, including 6G networks [3]. In HetNets, several RATs with different operating characteristics coexist to enhance network coverage and reliability while providing enhanced QoE to users. The underlying RATs have non-overlapping radio resources; therefore, there would not be typically interference in the network.

Since a stand-alone network with a single RAT would not be able to support the stringent QoS requirements of emerging disruptive applications, modern user devices are equipped with advanced capabilities that enable them to aggregate various radio resources to boost their QoE. Modern user devices can operate in a multi-mode scenario, in which each user device can be connected to a single RAT at any time. Alternatively, user devices can operate in a multi-homing scenario such that they can be connected simultaneously to various RATs to aggregate their radio resources, such as bandwidth and data rate. Multi-RAT networks include the coexistence of RATs, such as the licensed band networks, unlicensed bands networks, hybrid systems, and any combination of the wireless networks that are shown in Fig. 3.

Visible Light Communication (VLC) is a promising RAT that can support multi-Gbps of data rates over wireless links [120]. It is mainly developed for indoor applications; however, it is gaining considerable attention lately for outdoor applications as well [121]. This has motivated researchers to propose solutions that integrate VLC with conventional radio systems to boost data rates. Managing radio resources in these integrated systems, however, becomes a challenge. In this context, in [42], the authors propose a multi-agent Q-learning-based two-time scale scheme to address the power allocation issue for multi-Homing hybrid RF/VLC networks. In their technique, the agents are the RF and VLC APs, whose action space is discrete, corresponding to selecting the downlink power level that ensures the QoS's satisfaction of the multi-homing users. The state space is discrete, which is a function of users' achievable and target rates from the RF and VLC APs. The reward is also discrete, which is a function of the achieved and target rates from all RF and VLC APs. Experimental results demonstrate that not only the users' target rates are satisfied, but also the ability of their algorithm to adapt to the network's dynamics.

For the same network settings as in [42], Ciftler *et al.* [44] propose a DRL-based scheme to enhance the results and overcome the shortcomings. While the work in [42] was based on the vanilla Q-learning algorithm, the work in [44] has shown the advantages of utilizing the DQN algorithm to improve the convergence rate and accuracy. In particular, the authors in [44] propose a non-cooperative multi-agent DQN-based algorithm to address the problem of power allocation in hybrid RF/VLC networks. The agents are the RF and VLC APs whose action space is discrete, representing the transmit power. The state space is continuous, comprised of the actual and target rates, while the reward function is continuous and is a function of target rate band, target rate, and actual rate. Using simulation results, the authors demonstrate that the DQN-based algorithm converges with a rate of 96.1% compared with the Q-learning-based algorithm's convergence rate of 72.3%.

Findings and lessons learned: In this section, we review the applications of DRL techniques for power allocation and management in modern wireless networks. The reviewed papers are summarized in Table 5. We observe that various DRL techniques can efficiently solve the power allocation optimization problems in diversified wireless network scenarios, and their performance outperforms the state-of-the-art heuristic approaches. Besides, as we discussed in the previous paragraphs, DRL methods can provide comparable results to the conventional centralized optimization-based approaches that have full knowledge of the wireless environments as reported in [106], or even better results as reported in [105]. Moreover, note that the main motivations of using DRL techniques in all the papers presented in this subsection are the complexity of the formulated power allocation problems, the limited information about network dynamics and CSI, and the difficulty in applying conventional methods to solve the formulated power allocation problems.

We also observe that most of the papers implement multi-agent DRL interactions, and the value-based DRL algorithms, such as DQN and Q-learning, are utilized more than the policy-based counterparts. However, since the power allocation problem falls in the continuous action space, the use of value-based algorithms to address these types of problems makes the learned policies vulnerable to discretization errors that degrade the accuracy and reliability of the learned models. Hence, the emerging policy-based algorithms, such as DDPG and actor-critic, have received more attention lately, and they have shown more accurate and reliable results compared to the value-based counterparts with additional complexity, as discussed in [88], [102], [105], [106]. In addition, we observe that the definition of the state space and the reward function for the RRAM problems must be

TABLE 5. A summary list of papers related to DRL for power allocation.

Network Type		Ref.	Radio Resource (or Issues Addressed)	Learning Algorithm	
				Mode	Algorithm
Cellular Networks	Multi-cell cellular	Khan *et al.* [102]	Downlink power allocation	Single- & multi-agent	Actor-critic
	Cellular networks	Meng *et al.* [103]	Downlink power allocation	multi-agent	DQN
	HetNets	Zhang *et al.* [105]	Power control	multi-agent	DQN & DDPG
	Wireless mobile networks	Nasir *et al.* [106]	Continuous power control	Multi-agent	DDPG
	D2D cellular	Bi *et al.* [108]	Power allocation	multi-agent	DQN
	Dense 5G cellular	Saeidian *et al.* [109]	Downlink power control	Multi-agent	DQN
	NOMA mmWave UDNs	Zhang *et al.* [110]	EE power allocation	Multi-agent	DQN
	mmWave MIMO Cellular	Wang *et al.* [111]	Hybrid beamforming design	Single-agent	DDPG
Emerging IoT Nets	CRNs	Li *et al.* [112]	Power control	Single-agent	DQN
	D2D networks	Nguyen *et al.* [25]	Power allocation	multi-agent	DQN, DDQN, & Dueling DQN
	Ultra-dense UAV	Li *et al.* [113]	Downlink power control	Multi-agent	DQN
	Multi-UAV	Zhao *et al.* [114]	Power allocation	Multi-agent	DDPG
	Wireless Networks	Nasir *et al.* [107]	Transmit power control	Multi-agent	DQN
	mmWave HSR systems	Xu *et al.* [88]	Power allocation	Multi-agent	DDPG
	Distributed networks	Yan *et al.* [115]	Distributed power allocation	Multi-agent	Actor-critic
Satellite	Multi-beam satellites	Luis *et al.* [116]	Power allocation	Single-agent	PPO
	NOMA-enabled SIoT	Yan *et al.* [119]	Power allocation	Single-agent	DQN
Multi-RAT	Hybrid RF/VLC networks	Kong *et al.* [42]	Power allocation	Multi-agent	Q-learning
	Hybrid RF/VLC networks	Ciftler *et al.* [44]	Power allocation	multi-agent	DQN

deliberately engineered as they play a crucial role in the convergence and accuracy of the learned policies. For policy-based power allocation algorithms, it is more convenient to define the reward as a continuous function since the learning process depends on taking its derivative, which is not necessarily the case with the value-based algorithms.

It is also observed that DRL-based power allocation algorithms can be deployed in a centralized and distributed fashion, depending on the deployment scenario. Distributed scenarios provide more accurate and reliable policies than centralized ones at the expense of added complexity and signaling overhead, especially as the number of agents increases. Therefore, the tradeoff between the centralized and distributed policies heavily depends on the scenario under investigation. For example, it is preferable to deploy DRL models in a distributed fashion to address the power allocation problem for time-sensitive applications. However, for ultra-reliable applications, it is preferable to adopt centralized DRL deployment. Moreover, most of the papers consider the rate maximization, SE, and EE as key performance metrics (e.g., [88], [102], [110]). However, other KPI metrics must be considered as well during the design of DRL frameworks, such as latency, reliability, and coverage, especially for emerging real-time and time-sensitive IoT applications.

We also observe from Table 5 that both the cellular HomNets and emerging IoT wireless networks gain more attention than satellite and multi-RAT networks that still in their early stages and require more in-depth investigation.

B. DRL FOR SPECTRUM ALLOCATION AND ACCESS CONTROL

One of the significant challenges in modern wireless communication networks that still needs more investigation is spectrum management and access control. In this context, DRL techniques have attracted considerable research interest recently due to their robustness in making optimal decisions in dynamic and stochastic environments. This section presents the related works to the applications of DRL algorithms for radio spectrum allocation in modern wireless networks. This includes issues, such as dynamic network access, user association or cell selection, spectrum access or channels selection/assignment, and the joint of any of these issues, as shown in Fig. 3.

In modern wireless networks, a massive number of user devices may request to access the wireless channel simultaneously. This may drastically overload and congest the channel, causing communication failure and unreliable QoS. Hence, efficient communication schemes and protocols must be developed to address this issue in channel access via employing various access scheduling and prioritization techniques. RRAM for modern wireless networks requires considering dynamic load balancing and access management methods to support the massive capacity and connectivity requirements of the future wireless networks while utilizing their radio resources efficiently. DRL methods have been used recently to address these issues, and they have demonstrated efficient results in the context of massive channel access.

On the other hand, user devices in cellular networks are required to associate or be assigned to BS(s) or network AP(s) to get a service. The association process could be symmetric, i.e., both uplink and downlink are from the same BS/AP, or it may be asymmetric in which the uplink and downlink may associate to different BSs/APs. This association or cell selection process must be carefully addressed as it strongly affects the allocation of network radio resources. Unfortunately, such types of problems are typically non-convex and combinatorial [41] and need accurate network information to obtain the optimal solution. In this context, DRL techniques have also shown efficient results in addressing user association and cell selection issues for modern wireless networks.

1) IN CELLULAR NETWORKS

In the following paragraphs, we review works that employ DRL algorithms to address the spectrum and access control problem in cellular networks depicted in Fig. 2.

Users-BSs association and bandwidth allocation in UAV-assisted cellular networks are also among the main emerging challenges. Towards this end, interesting work is proposed in [122] based on the multi-agent DQN model to address the joint user association, spectrum allocation, and content caching in an LTE network consisting of UAVs serving ground users. In their model, the agents are the UAVs, which have storage units and have the ability to cached contents in LTE-BSs. These UAV agents can access the licensed as well as the unlicensed spectrum bands, and a remote cloud-based server is used to control them. The licensed cellular spectrum band is used in the transmissions from the cloud to the UAVs. Each UAV agent has to obtain 1) its user association, 2) bandwidth assignment indicators in the licensed spectrum band, 3) time slot indicators in the unlicensed spectrum band, and 4) content that the users request. The input of the DQL is the other agents' actions (the UAV-user association schemes), and the output is the set of users that the UAV can handle. Simulation results demonstrate that their proposed DQL strategy enhances the number of users up to 50% compared to the standard *Q*-learning strategy.

Based on their initial work in [123], the authors in [124] propose a multi-agent Dueling Double DQN (D3QN)-based DRL model to handle the joint BS and channel selections in macro and femto BS networks sharing a set of radio channels. In their scheme, the agents are the UEs, whose state space is a discrete binary vector that shows whether UEs' SINR higher than the minimum QoS requirement or not. The action space is discrete, corresponding to the BS and channel association. The reward function is discrete in which the UE agent will receive a utility as a reward if the QoS is met; otherwise, it will receive a negative value for the reward. Simulation results demonstrate that their proposed strategy outperforms the standard *Q*-learning strategy in terms of generalization, system capacity, and convergence speed.

The problem of user association in cellular IoT networks is studied in an interesting work in [125]. The goal is to assign IoT devices to particular cellular users to maximize the sum-rate of the IoT network. Two single-agent DQN DRL algorithms are proposed; the first one utilizes global information to make decisions for all IoT devices at one time, while the other algorithm uses local information to make a distributed decision for only a single IoT device at one time. In their model, the BS acts as the agent whose state space is continuous, consisting of both historical CSI and interference information. The action space is discrete, representing both all possible association schemes of the network and the individual association for only a single IoT device. The reward function of the first DQN algorithm is the sum-rate of all IoT devices, while for the second DQN includes both the current transmission rate of IoT devices and the interference with other IoT devices. Experimental results demonstrate that their proposed DRL framework both scalable and achieves performance comparable to the optimal user association policy.

Emerging integrated access and backhaul (IAB) cellular networks are characterized by their dynamic environment and large-scale deployment. In another interesting work in [126], the authors study the problem of spectrum allocation in the IAB networks. The problem is first formulated as a non-convex mix-integer and non-linear programming, and then a DRL framework based on single-agent Double DQN and actor-critic algorithms is proposed to solve it. In their model, the agent is a center-located controller or distributed UE. The state space is discrete, indicating the status of UEs' QoS, and the action space is discrete, corresponding to the allocation matrix for the donor BS and IAB nodes. The reward function is modeled to optimize the proportional fairness allocation of the network. Experimental results demonstrate that their framework has promising results compared to other conventional spectrum allocation policies.

The problem of load balancing in large-scale and dynamic wireless networks is also another important issue. In this context, the authors in [127] present a multi-agent *Q*-learning-based algorithm to address the problem of user association for load balancing in cellular vehicular networks. In their scheme, the agents are the BSs, whose action space is discrete, representing the associations with the network's vehicles. The state space is a hybrid (continuous and discrete), consisting of the service resources and its service demands, SINR matrix, and association matrix. The reward is a continuous function defined through the association and SINR matrices. The main advantage of this paper is that the performance of their proposed algorithm is evaluated using experiments on real-field taxi movements. The authors show that their approach provides higher quality load balancing compared to conventional association methods.

Most recently, Zheng *et al.* [128] propose a single agent actor-critic-based DRL algorithm to address the problem of channel assignment for the emerging hybrid NOMA-based 5G cellular networks. The agent is the BS, whose action space is discrete, corresponding to assigning channels for users. The state space is a hybrid (continuous and discrete) comprised of three elements; the CSI matrix, achieved users' data rate in the previous time slot, and the assigned channels in the previous time slot. The reward is a discrete function defined in terms of users' SE, the number of channels that use NOMA for transmission, and the number of users whose data rate is zero. Simulation results demonstrate that their proposed method outperforms some conventional approaches, such as greedy, random, match theory-based, and Genetic Algorithms, in terms of both network SE and sum-rate.

The problem of spectrum management in wireless DSA is addressed in [129] based on distributed multi-agent DQN. In their approach, the agents are each DSA user, whose action space is discrete, corresponding to the transmit power change for each channel. The state space is discrete, defined as the transmit power on wireless channels. The reward is a continuous function defined by the SE and the penalty caused by the interference to PUs. Experimental results show that their proposed model with echo state network-based DQN achieves a higher reward with both achievable data rate and PU protections.

Antenna selection is widely used for physical layer security in multi-antenna-based cellular networks. In this context, the authors in [130] propose a single-agent DQN algorithm to predict the optimal transmit antenna in the MIMO wiretap channel. The state space is discrete, defined in terms of the security capacity and maximum SNR of the MIMO wiretap channel. The action space is discrete, corresponds to selecting the transmit antenna. The reward function is discrete, defined in terms of the achieved SNR at the antenna. Experimental results demonstrate that their proposed algorithm proactively predicts the optimal antenna while reducing the secrecy outage probability of MIMO wiretap system compared to the support vector machine and conventional approaches.

2) IN IOT AND OTHER EMERGING WIRELESS NETWORKS

In the following paragraphs, we review works that employ DRL algorithms to address the spectrum and access control problem in IoT and emerging wireless networks illustrated in Fig. 2.

IoT sensor networks are characterized by their high dynamicity, which necessitates efficient channel access for the connecting nodes. In [131], the authors build on their initial work in [132] and propose a single-agent DQN-based DRL scheme to tackle the problem of dynamic channel access for IoT sensor networks. In their scheme, the agent is the sensor, and its action is discrete, corresponding to selecting one channel to transmit its packets at each time slot. The state space is discrete, which is a combination of rewards and actions in the previous time slots. The reward function is also discrete, which is "+1" if the selected channel is in low interference in such case a successful transmission occurs; otherwise, the reward is "−1" in such case the selected channel is in high interference, and a transmission failure occurs. Simulation results show that their proposed scheme achieves an average reward of 4.4 compared to 4.5 obtained using the conventional myopic policy [133], which needs a compact knowledge of the transition matrix of the system.

Energy consumption is considered one of the persistent challenges for emerging wireless sensor networks. In this context, an interesting work is proposed in [134] in which the authors develop a single-agent DQN-based DRL model to address the channel selection in energy harvesting-based IoT sensor networks. In that work, the agent is one BS, which controls the channel assignments for energy harvesting-enabled sensors. The problem of the agent is to predict the battery level of the sensors and to assign channels to sensors such that the total rate is maximized. The DQL model used to solve this problem has two long-short-term-memory (LSTM) neural network (NN) layers, one for predicting the sensor's battery state and one for obtaining channel access policy based on the predicted states obtained from the first layer. The agent's action is all the available sensors that require to access the channels. The state contains the history of channel access scheduling, true and predicted battery information history and the current sensor's CSI. Simulation results show that the total rates obtained using the DQL scheme are 6.8 kbps compared to 7 kbps obtained from the optimal scheme rate.

Managing spectrum allocation is one of the main objectives in cognitive radio networks (CRNs). The main idea is to efficiently utilize the available spectrum via enabling SUs to use the spectrum resources when the PUs are inactive. The authors in [135] propose a multi-agent DQN-based model to address the cooperative spectrum sensing issue in CRNs. In their scheme, the agents are the SUs whose action is discrete, corresponding to sensing the spectrum for possible transmission without interfering with the PUs. The state space is discrete, and it is comprised of four elements representing cases when the spectrum is sensed as occupied, the spectrum is not sensed in a particular time slot, the spectrum is sensed as idle, and one of the other SUs broadcast the sensing result first. The reward function is the binary indicator, which is "+1" if the spectrum is sensed as idle and "0" otherwise. Simulation results show that their proposed algorithm has a faster convergence speed and better reward performance than the conventional Q-learning algorithm.

For the same network in [135], the authors in [136] extend the work and propose a multi-agent DQN-based DRL scheme to address the problem of dynamic joint spectrum access and mode selection (SAMS) in CRNs. The agents are the secondary users (SUs) whose action space is discrete, corresponding to selecting the access channel and access mode. The state space of each SU agent is discrete, comprised of the action taken by the mth SU agent, the ACKs of all SUs agents, and the ACK of the mth SU agent. The reward function is discrete, which is "1" if the action selection process is successful; otherwise, there is a collision, and the agent will receive a "0" reward. Simulation results demonstrate that their proposed DQN algorithm provides comparable results to the Max benchmark after the model's convergence.

Xu *et al.* [137] propose a single-agent DQN and DDQN-based DRL approaches to address the problem of dynamic spectrum access in wireless networks. In their model, the agent is a wireless node (e.g., a user) whose action is discrete, corresponding to sensing the discrete frequency channel for possible data transmission. The state space is discrete, defining if the channel is occupied or idle at time slot t. The reward function is discrete, which is ranging from 0 to 100 for successful transmission; otherwise, the reward is "−10" if the channel state is occupied and the user transmission fails. It is shown using simulation results that both DQN and DDQN can learn different nodes' communication patterns and achieve near-optimal performance without prior knowledge of system dynamics.

Allocating spectrum resources is also a major challenge in vehicular IoT networks. Based on their initial work in [138], the authors in [139] propose a distributed single- and multi-agent DQN-based DRL schemes to address the spectrum sharing problem in V2X networks. In their proposed system, multiple V2V links reuse the frequency spectrum preoccupied with V2I links. The agents are the V2V links whose action space is discrete, corresponding to spectrum sub-band and power selection. Each agent's local observation space includes local channel information (such as its own signal channel gain, interference channels from other V2V transmitters, interference channel from its own transmitter to the BS, and the interference channel from all V2I transmitters), the remaining V2V payload, and the remaining time budget. The reward is continuous, which is a function of both the instantaneous sum capacity of all V2I links and the effective V2V transmission rate until the payload is delivered. Experimental results show that the agents cooperatively learn a policy that enables them to simultaneously improve the sum capacity of V2I links and payload delivery rate of V2V links. The authors also show that their proposed models for the single-agent and multi-agent settings provide very close performance to the conventional exhaustive search.

Multi-sensor network is an emerging technology that is expected to play a key role in future wireless networks. In this context, the authors in [140] propose a single-agent DQN model to address the joint channel access and packet forwarding in a multi-sensor scenario. In the proposed scheme, one sensor is the agent, which acts as a relay to forward packets arriving from its surrounding sensors to the sink. The agent has a buffer to store arriving packets. The agent's action is to choose channels for the packet forwarding, the packets transmitted on these channels, and a modulation scheme at each time slot to maximize its utility (defined as the ratio of the transmitted packets number to the transmit power). The state is the combination of the buffer and channel states. The input of the DQL model is the state, while the output is the corresponding action selection. Simulation results demonstrate that the proposed DQL scheme enhances system utility (i.e., 0.63) compared to the conventional random action selection scheme (i.e., 0.37).

One of the major challenges in mmWave wireless networks is establishing radio links and coping with the high vulnerability of intermittent communication. This issue is even exacerbated in mmWave V2X due to the high mobility of vehicles. Towards this end, Khan *et al.* [141] propose a multi-agent A3C-based DRL to address the problem of vehicle-cell association in mmWave V2X networks. The agents are the RSUs whose action is discrete, corresponding to determining the optimal vehicle-RSU association for RSU. The state space is a hybrid (discrete and continuous) defined in terms of the last channel observations, rate threshold violation indicator, and experienced data rate of vehicles. The reward function is continuous, defined in terms of the average rate per vehicle and threshold rate. Using experimental results, it is shown that their proposed algorithm achieves around 15% sum-rate gains and a 20% reduction in vehicular user outages compared to baseline approaches.

The problem of dynamic spectrum access in CRNs is investigated in [142] through combining DRL and evolutionary game theory. In particular, uncooperative multi-agent DQN is considered in which the agents are the SUs whose action is discrete, corresponding to selecting the access channel. The state space is discrete, which includes two main parts; the channel selected by the agent and the utility obtained after transmission on the selected channel. The reward function is defined in terms of evolutionary game theory. Simulation results indicate the performance enhancement of their proposed algorithm over the case without learning in terms of average system capacity.

Another interesting work is presented in [143] in which the authors propose a multi-agent DQN-based DRL algorithm to address the problem of optimum multi-user access control in Non-Terrestrial Networks (NTNs). In their model, UEs are the independent agents that report their experiences and local observations to a centralized trainer controller located at the backhaul network. The latter will then utilize the collected experiences to update the global DQN parameter. The agent's state space is continuous, comprised of the connected NT-BS of UEs at the previous time slots, the RSS of UEs, the number of connected UEs of each NT-BS, and the transmission rate of UEs. The action space is discrete, representing the binary indicator functions of UEs, while the reward is a function of the transmission rate of UEs. Experimental performance results show that their proposed scheme is efficient in addressing the fundamental issues in the deployment of NTNs infrastructure, and it outperforms the traditional algorithms in terms of both the data rate and the number of handovers.

The integration of various DRL algorithms to improve the efficiency and accuracy of the learned RRAM policies has shown promising results lately. In this context, Tomovic and Radusinovic [144] propose an interesting single-agent DRL model based on the integration of Double deep Q-learning architecture and RNN to address the problem of DSA in multi-channel wireless networks. In particular, the agent is the SU node, whose action space is discrete, representing the selection of a channel for sensing. The state space is also discrete, comprised of a history of the binary observations and history of taken actions. The reward function is a discrete binary function, which is "1" if the observation is "1" and "0" otherwise. Simulation results show that their proposed method is able to find a near-optimal policy in a smaller number of iterations, and it can support a wide range of communication environment conditions.

In other work in [145], the authors propose both a single-agent and multi-agent deep actor-critic DRL-based framework for dynamic multi-channel access in wireless networks. In their system, the agents are the users whose action space is discrete, corresponding to selecting a channel. The observation space is also discrete, which is defined based on the status of the channel and collision status. The reward function is discrete, which is "$+1$" if the selected channel is good; otherwise, it is "-1". Using simulation results, the authors show that their proposed actor-critic framework outperforms the DQN-based algorithm, random access, and the optimal policy when there is full knowledge of the channel dynamics.

The problem of DSA for the CRN is studied in [146] based on an uncoordinated and distributed multi-agent DQN model. The agents are CRs, whose action is discrete, representing the possible transmit powers. The state space is discrete, reflecting whether the limits for DSA are being met or not, depending on the relative throughput change at all the primary network links. The reward is also discrete, which is a function of the throughput of the links and the environment's state. Experimental results reveal that their proposed scheme finds policies that yield performance within 3% of an exhaustive search solution, and it finds the optimal policy in nearly 70% of cases.

Industrial IoT (IIoT) has emerged recently as an innovative networking ecosystem that facilitates data collection and exchange in order to improve network efficiency, productivity, and other economic benefits [147]. RRAM in such a sophisticated paradigm is also a challenge that needs more investigation. The work in [148] can be considered to be a pioneer in which the authors propose a solution for spectrum resource management for IIoT networks, with the goal of enabling spectrum sharing between different kinds of UEs. In particular, a single-agent DQN algorithm is proposed in which the agent is the BS. The action space is discrete, which corresponds to the allocation of spectrum resources for all UEs. The observation space is a hybrid (continuous and discrete) consisting of four elements; the current action (i.e., the allocation of spectrum resources), the data priority of type I UEs, the buffer length of type II UE, and the communication status of the first type of UEs. The reward function is continuous, defined to address their optimization problem. It is divided into four objectives; 1) maximizing the spectrum resource utilization; 2) quickly transmitting the high-priority data; 3) meeting the threshold rate requirement of the first type of UEs; 4) ensuring that the second type of UEs completes the transmission in time. Using simulation results, it is demonstrated that their proposed algorithm achieves better network performance with a faster convergence rate compared with other algorithms.

Most recently in [149], the authors propose a multi-agent Double DQN-based DRL model to address the problem of DSA in distributed wireless networks. In particular, they design a channel access scheme to maximize channel throughput regarding fair channel access. The agents in their scheme are the users. The action space is discrete, which is "0" if the user does not attempt to transmit packets during the current time slot, and it is "1" if it has attempted to transmit. The state space is discrete, consisting of four main elements; each user action taken on the current time slot, channel capacity (which could be negative, positive, or zero), a binary acknowledgment signal showing if the user transmits successfully or not, and a parameter that enables each user to estimate other users' situations. The reward is a discrete binary function that takes the value of "1" if the user transmits successfully; otherwise, it is "0" meaning that the user transmitted with collision. It is shown using simulation results that their scheme can maximize the total throughput while trying to make fair resource allocation among users. Also, it is shown that their proposed scheme outperforms the conventional Slotted-Aloha scheme in terms of sum-throughput.

Vehicular ad hoc networks (VANETs) are one of the promising networks for next generation wireless networks, where networks are formed and information is relayed among vehicles. Wang *et al.* [150] address the problem of DSA in VANETs, by proposing an interesting scheme that combines multi-hop forwarding via vehicles and DSA. The optimal DSA policy is defined to be the joint maximization of channel utilization and minimization of the packet loss rate. A multi-agent DRL network structure is presented that combines RNN and DQN for learning the time-varying process of the system. In their scheme, each user acts as an agent whose action space is discrete, corresponding to choosing a channel for transmission at time slot t. The state space is discrete, composed of three components; a binary transmission condition η, which is "1" if the transmission is successful and "0" otherwise, the channel selection action, and the channel status indicator after each dynamic access process. The reward is a discrete binary function, which takes a positive value if $\eta = 1$, otherwise it takes the value of "0". Simulation results show that their proposed scheme: 1) reduces the packet loss rate from 0.8 to around 0.1, 2) outperforms Slotted-Aloha and DQN in terms of reducing collision probability and channel idle probability by about 60%, and 3) enhances the transmission success rate by around 20%.

Due to their ability to improve communication in harsh environments, UAV networks have gained considerable research lately [151]. For example, most recently in [152], the authors propose efficient multi-agent DRL-based schemes to address the problem of joint cooperative spectrum sensing and channel access in clustered cognitive UAV (CUAV) communication networks. Three algorithms are proposed: 1) a time slot multi-round revisit exhaustive search based on virtual controller (VC-EXH), 2) a Q-learning based on independent learner (IL-Q), and 3) a DQN based on independent learner (IL-DQN). The agents are the CUAVs in the network. The action space of any CUAV agent is a discrete function defined by the steps that this agent moves clockwise in time slot t relative to the channel selected in time slot $t-1$ on the PU channel ring. The state space is a discrete set consisting of two main elements: 1) the number of CUAVs agents that have selected a particular channel to sense and access in the previous time slot and 2) a binary indicator function that shows the occupancy status of a particular channel in the previous time slot. The reward is a discrete function defined in terms of spectrum sensing, channel access, utility, and cost. Experimental results show that all the three algorithms proposed show efficient results in terms of convergence speed and the enhancement of utilization of idle spectrum resources.

An interesting work is conducted in [153] in which the authors propose a multi-agent deep recurrent Q-network-based model to address the problem of DSA in dynamic heterogeneous environments with partial observations. In their work, the authors consider a case-study with multiple heterogeneous PUs sharing multiple independent radio channels. The agents are the SUs, whose action space is discrete, corresponding to deciding whether to transmit in a particular band or wait during the next time slot. The state space is discrete, representing whether the channels are occupied, idle, or unknown. The reward function is discrete, represented by two values; 100 per channel for successful transmission and -50 per channel for collision. Using simulation results, the authors show that their proposed algorithm handles various dynamic communication environments, and its performance outperforms the myopic conventional methods and is very close to the optimization-based approaches that have a full observation of the environment.

3) IN SATELLITE NETWORKS

In the following paragraphs, we review works that employ DRL algorithms to address the spectrum and access control problem in satellite networks and emerging satellite IoT systems.

The work in [154] proposes a single-agent DQN-based DRL algorithm that considers the problem of channel assignment in multi-beam satellite systems. In their scheme, the agent is the satellite, whose action is discrete, including an index that indicates the channel that the newly arrived user has occupied. The agent's reward is discrete, which contains a positive value if the service is satisfied, and a negative value if the service is not satisfied or blocked. The state space is also discrete, which comprises three elements; the current users, the current channel assignment matrix, and a list of the new user arrivals. Experimental results demonstrate that their proposed scheme decreases the blocking probability and improves the carried traffic up to 24.4% as well as enhances the spectrum efficiency compared to the conventional fixed channel assignment approach.

In the same context, the authors in [155] propose a single-agent DQN-based DRL algorithm to address the problem of dynamic channel allocation in multi-beam satellite systems. In particular, an image-like tensor formulation on the system environments is considered in order to extract traffic spatial and temporal features. The agent in their model is the satellite, whose action space is discrete, corresponding to determining the resource allocation schemes. The state space is continuous, consisting of two elements; the system resource allocation state and the users' service request state. The reward function is discrete, which is defined in terms of the optimization objective function.

SIoT has emerged lately as a promising wireless system that provides global satellite IoT services with reliable and ubiquitous coverage. Recently, the work in [156] can be considered to be a pioneer in which the authors propose a single-agent DQN-based approach for energy-efficient channel allocation in SIoT. The agent in their model is the LEO satellite, whose action space is discrete, corresponding to mapping from newly coming node tasks to channels to be allocated. The state space is discrete, including information about user tasks, such as the size and location of tasks. The reward is continuous, which is divided into two normalized reward function components; the power efficiency reward and the normalized value of the service blocking rate. Both of these reward components are functions of power set up by the agent, the optimal power decided by the location of the beam, and the number of served nodes. Experimental results demonstrate that their proposed algorithm saves energy consumption of around 67.86% compared to some conventional approaches.

In the same context, the authors in [157] propose a centralized single-agent DQN-based scheme to address the problem of dynamic channel allocation in SIoT. The agent in their model is the satellite, whose action is discrete, corresponding to selecting which sensors to allocate channels to. The state space is discrete, comprised of three parts; the number of tasks in each time step, the bandwidth that a sensor node requires, and the duration of a new task. The reward is continuous, which is a function of the duration of data transmission for the sensor. Using simulation results, it is shown that their proposed algorithm both provides higher transmission success rates and reduces data transmission latency by at least 87.4% compared to the conventional channel allocation algorithms.

An interesting work is reported by Zheng *et al.* [158] in which the authors propose a single-agent Q-learning-based RL model to address the problem of combination allocation of fixed channel pre-allocation and dynamic channel scheduling in a network architecture of LEO satellites that utilizes a centralized resource pool. In their model, the satellite serves as an agent whose action is discrete, corresponding to assigning channels to users. The state space is discrete, defined by the channel assignment of users in each beam. The reward is continuous, which is a function of the user's supply-demand ratio. Experimental results demonstrate that their proposed approach enhances the system supply-demand ratio by 14% and 18% compared to the static channel allocation and the Lagrange algorithm channel allocation methods, respectively.

4) IN MULTI-RAT NETWORKS

In the following paragraphs, we review works that employ DRL algorithms to address the spectrum and access control problem in multi-RAT HetNets. This includes the coexistence of various variants of the wireless networks shown in Fig. 2.

Managing the spectrum bands in unlicenced cellular networks is also another persistent challenge. In this context, the authors in [159] present a multi-agent DQN-based model that jointly tackles the dynamic channel selection and interference management in Small Base Stations (SBSs) cellular networks that share a set of unlicensed channels in Long Term Evolution (LTE) networks. In the proposed scheme, the SBSs are the agents who choose one of the available channels for transmitting packets in each time slot. The agent's action is channel access and channel selection probability. The DQL input includes the channels' traffic history of both the SBSs and Wireless Local Area Networks (WLAN), while the output is the agent's predicted action vectors. Simulation results reveal that their proposed DQL strategy enhances the average data rate by up to 28% compared to the conventional Q-learning scheme.

For the same network settings in [159], the authors in [160] propose a single-agent DQN-based model to tackle the dynamic spectrum allocation for multiple users that share a set of K channels. In their scheme, the agent is the user whose action is either choosing a channel with a particular attempt probability or selecting not to transmit. The agent's state includes the history of the actions of the agent and its current observations. The DQL model input is the previous actions along with their observations, while the output is the Q-values corresponding to the actions. Simulation results demonstrate that their proposed DQL strategy achieves a double data rate compared to the state-of-the-art Slotted-Aloha scheme.

The integration of cellular networks and indoor networks has also shown efficient results in enhancing the QoS of wireless communication in terms of coverage and data rate. Towards this end, Wang and Lv [161] propose an efficient single-agent prediction-DDPG-based DRL algorithm to study the problem of the dynamic multichannel access (MCA) for the hybrid long-term evolution and wireless local area network (LTE-WLAN) aggregation in dynamic HetNets. The agent is the central BS controller, whose state space is continuous, consisting of both the channels' service rates and the users' requirement rates. The action space, on the other hand, is discrete, representing the users' index. Two reward functions are provided; online traffic real reward and online traffic prediction reward, each of which are functions of users' requirements, channels' supplies, degree of system fluctuation, the relative resource utilization, and the quality of user experience. Using simulation results, the authors demonstrate the efficiency of the proposed prediction-DDPG model in solving the dynamic MCA problem compared to conventional methods.

Another interesting work in [162], the authors investigate the joint allocation of the spectrum, computing, and storing resources in multi-access edge computing (MEC)-based vehicular networks. In particular, the authors propose multi-agent DDPG-based DRL algorithms to address the problem in a hierarchical fashion considering a network comprised of macro eNodeB (MeNB) and Wi-Fi APs. The agents are the controller installed at MEC servers. The agents' action space is discrete including the spectrum slicing ratio set, spectrum allocation fraction sets for the MeNB and for each Wi-Fi AP, computing resource allocation fraction, and storing resource allocation fraction. The state space is discrete representing information of the vehicles within the coverage area of the MEC server, including vehicles' number, x-y coordinates, moving state, position, and task information. The reward function is discrete, defined in terms of the delay requirement, and requested storing resources required to guarantee the QoS demands of an offloaded task. Provided experimental results reveal that their proposed schemes achieve high QoS satisfaction ratios compared with the random assignment techniques.

The integration of various cellular wireless networks is also one of the main enabling technologies for the next generation wireless networks. Recently in [163], the authors propose an efficient single-agent DQN algorithm based on Monte Carlo Tree Search (MCTS) to address the problem of dynamic spectrum sharing between 4G LTE and 5G NR systems. In particular, the authors used the MuZero algorithm to enable a proactive BW split between 4G LTE and 5G NR. The agent is a controller located at the network core, whose action space is discrete, corresponding to a horizontal line splitting the BW to both 4G LTE and 5G NR. The state space is discrete, defined by five elements: 1) an indicator if the user is an NR user or not, 2) the number of bits in the user's buffer, 3) an indicator of whether the user is configured with multimedia broadcast single frequency network (MBSFN) or not, 4) the number of bits that can be transmitted for the user in a given subframe, and 5) the number of bits that will arrive for each user in the future subframes. The reward function is a continuous function defined as a summation of the exponential of the delayed packet per user. Experimental results show that their proposed scheme provides comparable performance to the state-of-the-art optimal solutions.

Findings and lessons learned: This section reviews the applications of DRL for dynamic spectrum allocation and access control in modern wireless networks. These types of radio resources are inherently coupled with user association, network/RAT selection, dynamic multi-channel access, and DSA. Table 6 summarizes the reviewed papers in this section. In general, the application of DRL for spectrum allocation and access control problems has received considerable attention lately. We observe that most DRL algorithms, when deployed for non-IoT networks, are implemented in centralized fashions at network controllers, such as BSs, RSUs, and satellites [125], [141], [154]. This is done to utilize the controllers' powerful and advanced hardware capabilities in collecting network information and designing cross-layer policies. Hence, we observe that DRL models are deployed as a single-agent at the network controllers [148]. On the contrary, DRL provides a flexible tool in diversified IoT networks and systems, conventionally involving dynamic

TABLE 6. A summary list of papers related to DRL for spectrum allocation and access control.

Network Type		Ref.	Radio Resource (or Issues Addressed)	Learning Algorithm	
				Mode	Algorithm
Cellular Networks	UAV-assisted LTE	Chen *et al.* [122]	Joint user association, spectrum allocation, & content caching	Multi-agent	DQN
	Macro & femto BS	Zhao *et al.* [123], [124]	Joint BS & channel selections	Multi-agent	Dueling DDQN
	Cellular IoT	Zhang *et al.* [125]	User association	Single-agent	DQN
	IAB cellular	Lei *et al.* [126]	Dynamic spectrum allocation	Single-agent	DDQN & actor-critic
	CV2X	Li *et al.* [127]	User association	Multi-agent	Q-learning
	Hybrid NOMA-based 5G	Zheng *et al.* [128]	Dynamic spectrum allocation	Single-agent	actor-critic
	Wireless DSA	Song *et al.* [129]	Dynamic spectrum allocation	Multi-agent	DQN
	MIMO systems	Hu *et al.* [130]	Transmit antenna selection	Single-agent	DQN
IoT & Other Emerging Wireless Networks	IoT sensor networks	Wang *et al.* [131], [132], [164]	Dynamic multi-channel access	Single-agent	DQN
	Energy harvesting-based IoT sensors	Chu *et al.* [134]	Dynamic spectrum allocation	Single-agent	DQN
	CRNs	Zhang *et al.* [135]	Dynamic multi-channel access	Multi-agent	DQN
	CRNs	Yang *et al.* [136]	Joint spectrum access & mode selection	Multi-agent	DQN
	Wireless networks	Xu *et al.* [137]	DSA	Single-agent	DQN & DDQN
	V2X	Liang *et al.* [138], [139]	Dynamic spectrum sharing	Single-& multi-agent	DQN
	Multi-sensor scenario	Zhu *et al.* [140]	Joint channel access & packet forwarding	Single-agent	DQN
	mmWave V2X	Khan *et al.* [141]	User association	Multi-agent	A3C
	CRNs	Yang *et al.* [142]	DSA	Multi-agent	DQN
	NTNs	Cao *et al.* [143]	Dynamic multi-channel access	Multi-agent	DQN
	Multi-channel wireless networks	Tomovic *et al.* [144]	DSA	Single-agent	RNN-base DDQN
	Wireless networks	Zhong *et al.* [145]	Dynamic multi-channel access	Single-& multi-agent	actor-critic
	CRNs	Tondwalkar *et al.* [146]	DSA	Multi-agent	DQN
	IIoT networks	Shi *et al.* [148]	Dynamic spectrum sharing	Single-agent	DQN
	Distributed wireless networks	Janiar *et al.* [149]	DSA	Multi-agent	DDQN
	VANETs	Wang *et al.* [150]	DSA	Multi-agent	RNN-based DQN
	CUAV	Jiang *et al.* [152]	Joint spectrum sensing & channel access	Multi-agent	DQN
	Heterogeneous environments	Xu *et al.* [153]	DSA	Multi-agent	RNN-based DQN
Satellite Nets	Multi-beam satellite systems	Liu *et al.* [154]	Dynamic spectrum allocation	Single-agent	DQN
	Multi-beam satellite systems	Hu *et al.* [155]	Dynamic spectrum allocation	Single-agent	DQN
	SIoT	Zhao *et al.* [156]	Dynamic spectrum allocation	Single-agent	DQN
	SIoT	Liu *et al.* [157]	Dynamic spectrum allocation	Single-agent	DQN
	LEO satellites	Zheng *et al.* [158]	Joint channel pre-allocation & dynamic channel scheduling	Single-agent	Q-learning
Multi-RAT	Small BSs cellular	Challita *et al.* [159]	Joint dynamic channel selection & interference management	Multi-agent	DQN
	Small BSs cellular	Naparstek *et al.* [160]	Dynamic spectrum allocation	Single-agent	DQN
	LTE-WLAN HetNets	Wang *et al.* [161]	Dynamic multi-channel access	Single-agent	prediction-DDPG
	MEC-based V2X	Peng *et al.* [162]	Joint allocation of spectrum, computing, & storing	Multi-agent	DDPG
	4G LTE and 5G NR systems	Challita *et al.* [163]	Dynamic spectrum sharing	Single-agent	DQN

system modeling and multi-agent interactions, such as CRNs and distributed systems. Also, note that the main motivations of using DRL techniques in almost all the papers presented in this subsection are the complexity of the formulated spectrum allocation and access control problems, the inability to obtain accurate CSI, and the inadequacy of conventional methods to solve the formulated problems.

In addition, the management of such types of radio resources falls in general in the discrete action space. Therefore, the value-based algorithms are utilized more than the policy-based ones, and they have shown efficient results, as we discussed in the surveyed papers. We also observe that embedding prediction-based DRL algorithms, such as RNN, with the conventional DNN models has shown efficient results in enabling DRL to perform a proactive spectrum prediction. Such integrated models have been seen in [144], [150], [153] and we expect that they will attract more attention in the future. In addition, it is always preferable to utilize the DQN-based algorithms to the Q-learning algorithm as they provide better performance in terms of convergence speed and accuracy of the learned policies. Moreover, as is the case with the other DRL models, the definitions of the state space and reward function are crucial, and they must provide representative and rich information about the system and environment to the agent in order to learn efficient and reliable RRAM policies.

We also observe from Table 6 that the use of DRL techniques for IoT and emerging wireless networks receives more attention than other wireless networks, especially for the cognitive radio-based systems as in [152].

The exponential increase in smart IoT devices mandates making autonomous decisions locally, especially for delay-sensitive IoT applications and services. In this context, we anticipate that the research on spectrum allocation and access control using distributed multi-agent DRL algorithms for future IoT networks will attract more attention as in [139], [141], [150], [152].

C. DRL FOR RATE CONTROL

This refers to the adaptive rate control in the uplink and downlink of wireless networks. With the explosive increase in the number of user devices and the emergence of massive types of data-hungry applications [3], it becomes essential to keep high network KPIs in terms of data rates and users' QoE. Adaptive rate control schemes must ensure satisfactory QoS in highly dynamic and unpredictable wireless environments. In this context, DRL techniques can be efficiently deployed to solve adaptive rate control problems instead of conventional approaches that possess high complexity and heavily rely on accurate network information and instantaneous CSI.

In the following paragraphs, we review works that employ DRL algorithms to address the rate control issue in cellular networks.

5G network slicing is a technique based on the network virtualization concept that enables dividing the single network connections into multiple unique virtual connections to provide various radio resources to various types of traffic. Liu *et al.* [165] conduct a pioneer DRL-based work to address the problem of network resource allocation, in terms of rate, for 5G network slices. The problem is decomposed into a master-slave, and a multi-agent DDPG-based DRL scheme is then proposed to solve it. The agents are located in every network slice, whose action space is continuous, defining the resource allocation to users in the network slice. The state space is continuous and has two main parts; the first one shows how much utility the user obtained compared to its minimum utility requirement, while the second part shows the auxiliary and dual variables from the master problem. The reward is a continuous function defined in terms of utility, utility requirements, and auxiliary and dual variables. Simulation results demonstrate that their proposed algorithm outperforms the baseline approaches and gives a near-optimal solution.

High mobility networks are characterized by their rapid variations that render link establishment a major issue. In this context, the authors in [166] propose an interesting work using a single-agent DQN-based DRL algorithm to address the problem of dynamic uplink/downlink radio resources allocation in terms of network capacity in high mobility 5G HetNets. Their proposed algorithm is based on the Time Division Duplex (TDD) configuration in which the agent is the BS, whose action space is discrete, corresponding to the configurations of TDD sub-frame allocation at the BS. The state space is discrete, comprised of different kinds of features of the BS, including uplink/downlink occupancy, buffer occupancy, and channel condition of all uplinks/downlinks to the BS. The reward is discrete, defined as a function of the uplink and downlink channel utility, which mainly depends on channel occupancy with chosen TDD configuration. Using experimental results, the authors show that their proposed algorithm achieves performance improvement in terms of both network throughput and packet loss rate, compared to some conventional TDD resource allocation algorithms.

Findings and lessons learned: This section reviews the use of DRL techniques for adaptive rate control in next generation wireless networks. In general, there is limited research that is solely dedicated to addressing the rate radio resource issue. We consider [165] and [166] as pioneer works in this type of RRAM. Most of the research in the literature is dedicated to video streaming applications, and the paper [15] highlighted some of them. However, as we discussed in the previous sections, the data rate control issue is typically addressed via controlling other radio resources such as power, user association, and spectrum. In addition, the adaptive rate control is typically addressed as a joint optimization with other radio resources, as we will elaborate in the next section, e.g., as in [167], [168].

We also observe that DRL-based solutions for cellular networks receive more attention than other wireless networks, and there is a lack of research on adaptive rate control for IoT and satellite networks. This also deserves more in-depth investigation and analysis.

D. DRL FOR JOINT RRAM

Due to the massive complexity and large-scale nature of modern wireless networks, it becomes necessary to design efficient schemes that account for the joint radio resources. In many scenarios, the design problem in wireless networks might end up with competing objectives. For example, in UDNs, increasing the power level is beneficial in combating path loss and enhancing the received signal quality. However, this might cause serious interference to the neighboring user devices and BSs. Hence, the joint design of power level control and interference management becomes mandatory. Conventional approaches for solving joint RRAM problems require complete and instantaneous knowledge about network statistics, such as traffic load, channel quality, and radio resources availability. However, obtaining such information is not possible in such large-scale networks. In this context, DRL techniques can be adopted to learn system dynamics and communication context to overcome the limited knowledge of wireless parameters.

This section intensively reviews the most important and influential works that implement DRL algorithms for the problem of joint RRAM in modern wireless networks. Particularly, we present related works that jointly optimize the radio resources shown in Fig. 2, such as power allocation, spectrum resources, user association, dynamic channel access, cell selection, etc.

1) IN CELLULAR NETWORKS

In the following paragraphs, we review works that employ DRL algorithms to address the joint RRAM issue in cellular networks shown in Fig. 2.

Cellular vehicular communication (CV2X) is regarded as one of the main enabling technologies for next generation wireless networks. RRAM in such networks has received significant momentum using conventional methods, and they are now gaining notable attention using DRL methods. For example, an interesting work is reported in [169], in which the authors study the problem of joint optimization of transmission mode selection and resource allocation for CV2X. They propose single-agent settings in which DQN and federated learning (FL) models are integrated to improve the model's robustness. The agent in their model is each V2V pair. The action space is discrete, representing the resource block (RB) allocation, communication mode selection, and transmit power level of the V2V transmitter. The state space is a hybrid (continuous and discrete) consisting of five main parts; the received interference power at the V2V receiver and the BS on each RB at the previous subframe, the number of selected neighbors on each RB at the previous subframe, the large-scale channel gains from the V2V transmitter to its corresponding V2V receiver and the BS, current load, and remaining time to meet the latency threshold. The reward is a continuous function defined in terms of the sum-capacity of vehicular UEs as well as the QoS requirements of both vehicular UEs and V2V pairs. Using experimental results, the authors show that their proposed two-timescale federated DRL algorithm outperforms other decentralized baselines.

RRAM in small cell networks is one of the ongoing challenges for cellular operators. Towards this end, Jang and Yang [170] propose a multi-agent DQN-based algorithm to address the problem of sum-rate maximization via a joint optimization of resource allocation and power control in small cell wireless networks. The agents in their proposed model are the small cell BSs, whose action space is discrete, corresponding to selecting the resource allocation and power control of small BS on RB. The state space is continuous, including all the CSI that the small BS collects on RB, such as local CSI, local CSI at the transmitter, etc. The reward is a continuous function expressed by the average sum-rate of its own serving users and the other small BSs. Experimental results show that their proposed approach both outperforms the conventional algorithms under the same CSI assumptions and provides a flexible tradeoff between the amount of CSI and the achievable sum-rate.

In the same context, another interesting work is presented in [171] in which the authors propose a model-driven multi-agent Double DQN-based framework for resource allocation in UDNs. In particular, They first develop a DNN-based optimization framework comprised of a series of ADMM iterative procedures that uses the CSI as the learned weights. Then, channel information absent Q-learning resource allocation algorithm is presented to train the DNN-based optimization scheme without massive labeling data, where the EE, SE, and fairness are jointly optimized. The agents are each D2D transmitter, whose action space is discrete, corresponding to selecting a subcarrier and corresponding transmission power. The state space is a hybrid (continuous and discrete) consisting of two parts; user association information and interference power. The reward function is continuous, comprised of two components; the network EE and the fairness of service quality, which is expressed by the variance of throughput between authorized users. Using experimental results, it is demonstrated that their proposed algorithm has a rapid convergence speed, well characterizes the extent of optimization objective with partial CSI, and outperforms other existing resource allocation algorithms.

D2D-enabled cellular networks are also one of the key enabling technologies for next generation cellular systems. RRAM in such networks is one major concern, especially for mmWave-based cellular networks, as the D2D links require frequent link re-establishment to combat the high blockage rate. The authors in [172] propose a multi-agent Double DQN-based scheme to address the problem of joint subcarrier assignment and power allocation in D2D underlying 5G cellular networks. The agents in their model are the D2D pairs, whose action space is discrete, corresponding to determining the transmit power allocation on the available subcarriers. The state space is a hybrid (continuous and discrete), comprised of four components: 1) local information (including the previous transmit power, previous SE achieved by transmitters, channel gain, and SINR), 2) the interference that each agent causes at the BS side, 3) the interference received from agent's interfering neighbors and the SE achieved by agent's neighbors, and 4) the interference that each agent causes to its neighbors. The reward is a continuous function comprised of three elements: 1) the SE achieved by each agent, 2) the SE degradation of the agent's interfered neighbors, and 3) the penalty due to the interference generated at the BS. Experimental results show that their proposed algorithm outperforms both the exhaustive and random subcarrier and even power (RSEP) assignment methods in terms of SE of D2D pairs.

Mission-Critical Communications (MCC) is an emerging service in the next generation wireless networks. It is envisioned to enable First Responders, such as firefighters and emergency medical personnel, to replace conventional radio with advanced communication capabilities available to next generation smartphones and IoT devices. Most recently, a pioneer work is conducted by Wang *et al.* [173] in which the authors propose a multi-agent DQN-based DRL scheme to address the problem of spectrum allocation and power control for MCC in 5G networks. In MCC, multiple D2D users reuse non-orthogonal wireless resources of cellular users without BS in order to enhance the network's reliability. The paper aims to help the D2D users autonomously select the channel and allocate power to maximize system capacity and SE while minimizing interference to cellular users. The agents are the D2D transmitters whose action space is discrete, corresponding to channel and power level selection. The state space is discrete, defined in a three-dimensional matrix, which includes information on the channel state of uses, the state of power level, and the number of the D2D pairs. The reward function is discrete, defined in terms of the total system capacity and constraints. Simulation results show that their proposed learning approach significantly improves spectrum allocation and power control compared to traditional methods.

RRAM in OFDM-based systems is also one of th main challenging issues. In this context, the authors in [174] propose a multi-agent DQN-based model to address the problem of joint user association and power control in OFDMA-based wireless HetNet. The agents are the UEs, whose action space is discrete, corresponding to jointly associate with the BS and determine the transmit power. The state space is discrete, which is defined by the situation of all UEs association with BS and power control. The reward function is continuous, which is defined in terms of the sum-EE of all UEs. Using simulation results, it is shown that their proposed method outperforms the Q-learning method in terms of convergence and EE.

Another interesting work is reported in [175] in which the authors propose a single-agent DQN-based DRL model to address the problem of joint optimization of user association, resource allocation, and power allocation in HetNets. The agent is the BS, whose action is discrete, corresponding to power allocation to users. The state space is discrete, defined by the channel gain matrix and the set of users association. The reward function is continuous, defined by the utility function of users' achieved data rate. Using simulation results, the authors show that their proposed algorithm outperforms some of the existing methods in terms of SE and convergence speed.

2) IN IOT AND OTHER EMERGING WIRELESS NETWORKS

In the following paragraphs, we review works that employ DRL algorithms to address the joint RRAM issue in IoT and emerging wireless networks depicted in Fig. 2.

For the same system settings in [106], [107], the authors in [67] extended their work and propose a multi-agent DDPG-based DRL framework to address the problem of the joint spectrum and power allocation in wireless networks. Two DRL-based algorithms are proposed, which are executed and trained simultaneously in two layers in order to jointly optimize the discrete subband selection and continuous power allocation. The agent in their approach is each transmitter. In the top layer, the action space of all agents is discrete, representing the discrete subband selection, while the bottom layer has a continuous action space corresponding to the transmit power allocation. The state space is a hybrid (continuous and discrete), containing information on achieved SE, transmit power, sub-band selection, rank, and downlink channel gain. The reward is shared by both layers, which is a continuous function defined in terms of the externality of agents to interference and the spectral efficiency. Using experimental results, the authors show that their proposed framework outperforms the conventional fractional programming algorithm.

Based on their initial work in [176], the authors in [177] extended their work and propose a distributed multi-agent DQN-based DRL scheme to address the problem of joint channel selection and power control in D2D networks. The agents in their model are the D2D pairs, whose action space is discrete, corresponding to selecting a channel and a transmit power. The state space of each agent is a hybrid (continuous and discrete) which contains three sets of information; local information, non-local information from the agent's receiver-neighbor set, and non-local information from the agent's transmitter-neighbor set. The reward function of each agent is continuous, which is decomposed into the following elements; its own received signal power, its own total received SINR, its interference caused to transmitter-neighbors, the received signal power, and the total received SINR of transmitter-neighbors. Using simulation results, it is shown that the performance of their scheme closely approaches that of the FP-based algorithm even without knowing the instantaneous global CSI.

In [178], the authors extended their previous works in [179], [180] and present a distributed multi-agent DQN-based model to address the problem of joint sub-band selection and power level control in V2V communication networks. Their proposed model is applicable to both unicast and broadcast scenarios. The agents are the V2V link or vehicles whose action space is discrete, corresponding to the selection of the frequency band and transmission power level that generate small interference to all V2V and V2I links while ensuring enough resources to meet latency constraints. The state space is continuous, containing the following information; the CSI of the V2I link, the received interference signal strength in the previous time slot, the channel indices selected by neighbors in the previous time slot, the remaining load for transmission, and the time left before violating the latency constraint. The reward function is continuous, consisting of three components; the capacity of V2I links, the capacity of V2V links, and the latency constraint. Using experimental results, it is shown that agents learn to satisfy the latency constraints on V2V links while minimizing the interference to V2I communications.

Another pioneering work is reported in [181] in which the authors propose a single-agent Double DQN-based DRL to address the problem of joint channel selection and power allocation with network slicing in CRNs. Their study aims to provide high SE and QoS for cognitive users. The agent is the overall CRN, whose action space is discrete, corresponding to the channel selection and power allocation of SUs. The state space is continuous, defined by the SINR of the PU. The reward function is continuous, which is a function of the system SE, user QoS, interference temperature, and the interference temperature threshold. Experimental results show that their proposed algorithm improves the SE and QoS and provides faster convergence and more stable performance than the Q-learning and DQN algorithms.

NOMA-based systems are characterized by their ability to provide enhanced QoS in cellular networks. However, allocating radio resources in such systems is quite challenging. In this context, the problem of joint subcarrier assignment and power allocation in an uplink multi-user 5G-based NOMA systems is addressed in [182]. A multi-agent two-step DRL algorithm is proposed; the first step employs the DQN algorithm to output the optimum subcarrier assignment policy, while the second step employs the DDPG algorithm to dynamically allocate the transmit power for the network's users. The agent is a controller located at the BS, whose action space is a hybrid (discrete and continuous), corresponding to the subcarrier assignment decisions and power allocation decisions. The state space is continuous, which is defined by the users' channel gains at each subcarrier. The reward function is defined as the sum EE of the NOMA system. Experimental results show that their proposed algorithm provides better EE than the fixed and DQN-based power allocation schemes.

Unlike the work in [182], a pioneer work is reported in [183] in which the authors propose a multi-agent DDPG-based model to address the problem of joint power and spectrum allocation in NOMA-based V2X networks. In particular, the authors are looking to maximize the sum-rate of V2I communications. The agents are the V2V communication links. The state space is discrete, defined by a set of actions performed by V2I and V2V communication links. The set includes the transmit power of both V2I and V2V links as well as the spectrum multiplexing factor of both V2V and V2V links. The state space is continuous, defined by five parts; the local channel gain information of each V2V link, interference channels from other V2V communication links, interference channel from each link's own transmitter to the BS, interference channel from all V2I transmitters, and the state of queue length in the buffer of each V2V transmitter. The reward function is continuous, defined by the achieved sum-rate of V2I communication links and the delivery probability of V2V communication links. Compared with both the DQN algorithm and random resource allocation scheme, simulation results show that their proposed algorithm outperforms both of them in terms of maximizing the sum-rate of V2I communication links while meeting the latency and reliability requirements of V2V communications.

On the other hand, another interesting work is conducted by Munaye *et al.* [168] in which they propose a multi-agent DQN-based DRL model to address the problem of joint radio resources of bandwidth, throughput, and power in UAV-assisted IoT networks. The agents are the UAVs, whose action space is discrete, corresponding to jointly selecting channel allocation of bandwidth, throughput, and power. The state space is discrete, comprising three components; the air-to-ground channel used by users, the rate of power consumption, and the interference. The reward is a discrete function, defined in terms of throughput, power allocation, bandwidth, and SINR levels. Simulation results show that their proposed algorithm outperforms other algorithms in terms of accuracy, convergence speed, and error.

Reconfigurable intelligent surface (RIS) technology has emerged recently as one of the main technologies for future wireless networks [187]. RISs employ many passive reflecting elements with controllable phase shifts and negligible power consumption, which provide a favorable wireless propagation environment for transmitted signals. In particular, RISs can be used to overcome blockage by providing virtual LoS links between transmitters and receivers, interference cancellation, and physical layer security [187], [188]. However, RISs encounter massive challenges related to environment uncertainty and real-time channel estimation issues [187], [188]. Hence, DRL approaches have attracted considerable research lately as efficient tools to assist the RIS technology. In this context, the authors in [184] provide an interesting work based on the multi-agent dueling DQN model to address the problem of power minimization in UAV-RIS-based multi-cell HetNets. In particular, they proposed to solve their problem in two stages. The first stage employs dueling DQN to solve the problem of UAVs' trajectories/velocities, RISs' phase control, and subcarrier allocations for microwave band. The second stage employs alternating methods to solve active beamforming and subcarrier allocation for mmWave. The agents are the UAV-RISs, whose state space is continuous defined by the trajectory of the UAV-RISs and all channel gains, i.e., from the BS to the RISs, RISs to users, and small-cell BSs to users. The action space is discrete, defined by the possible direction and speeds of the UAVs, RISs' phase shifts, and association indicator. The reward function is continuous defined in terms of power consumption. Simulation results show that their proposed algorithm reduces the transmit power consumption by 6 dBm compared to other baseline methods.

3) IN MULTI-RAT NETWORKS

In the following paragraphs, we review works that employ DRL algorithms to address the joint RRAM problem in multi-RAT HetNets. This includes the coexistence of various variants of the wireless networks as illustrated in Fig. 2.

Integrating RF and VLC RATs is a promising solution to enhance networks' QoS. Towards this, recently in [185], the authors present a multi-agent DQN-based algorithm to address the problem of joint optimization of bandwidth, power, and user association in hybrid RF/VLC systems. The APs are the agents whose action is discrete, representing the bandwidth, association function, and power level. The state space is discrete, which is a function of the problem constraints such as system bandwidth, association function, and power levels. The reward is discrete, which is a function of the rates delivered by the APs. Experimental results show that their algorithms improve the achievable sum-rate and number of iterations for convergence by 10% and 54% compared to that obtained using conventional optimization approaches.

Another interesting work is proposed recently by Alwarafy *et al.* [41]. The authors propose a hierarchical multi-agent DQN and DDPG-based algorithm to address the problem of sum-rate maximization in multi-RAT multi-connectivity wireless HetNets. The authors addressed the problem of multi-RATs assignment and continuous power allocation that maximize the network sum rate. In their model, single and multi-agents are proposed. The edge server acts as a single agent employed by DQN, while RATs APs behave as multi-agents employed by DDPG. For the single-agent DQN model, the action space is discrete, corresponding to the RATs-EDs assigning process. The state space of the DQN is continuous, comprised of the link gains and the required data rates of users. The reward function of the DQN agent is continuous, defined by the difference between the achieved rate and the required rate by users. For the multi-agent DDPG models, the action space is continuous, representing the power allocation of each RAT AP agent. The state space is a hybrid (continuous and discrete) consisting of four elements: the RATs-EDs assignment process performed by the DQN agent, the minimum data rate of users, the gains of the links, and the achieved data rate. Experimental results show that their algorithm's performance is approximately 98.1% and 95.6% compared to the conventional CVXPY solver that assumes full knowledge of the wireless environment.

Hybrid access networks are a special architecture for broadband access networks where different types of access networks are integrated to improve bandwidth. Huang *et al.* [186] propose a single-agent DQN model to

TABLE 7. Summary of the related works that address the joint RRAM.

Network Type		Ref.	Types of Joint Radio Resources (or Issues Addressed)	Learning Algorithm	
				Mode	Algorithm
Cellular Networks	CV2X	Zhang *et al.* [169]	Transmission mode selection & resource allocation	Single-agent	DQN
	Small cell networks	Jang *et al.* [170]	Resource allocation & power control	Multi-agent	DQN
	UDNs	Liao *et al.* [171]	Subcarrier selection & transmission power	Multi-agent	DDQN
	D2D underlying 5G cellular	Zhang *et al.* [172]	Subcarrier assignment & power allocation	Multi-agent	DDQN
	Mission-critical in 5G	Wang *et al.* [173]	Spectrum allocation & power control	Multi-agent	DQN
	OFDMA-based networks	Ding *et al.* [174]	User association & power control	Multi-agent	DQN
	HetNets Cellular	Zhang *et al.* [175]	User association, resource allocation, & power allocation	Single-agent	DQN
Emerging IoT Nets	Wireless networks	Nasir *et al.* [67]	Spectrum & power allocation	Multi-agent	DDPG
	D2D networks	Tan *et al.* [177]	Channel selection & power control	Multi-agent	DQN
	V2V networks	Ye *et al.* [178]	Sub-band selection & power level control	Multi-agent	DQN
	CRNs	Yuan *et al.* [181]	Channel selection & power allocation	Single-agent	DDQN
	5G-based NOMA systems	Zhang *et al.* [182]	Subcarrier assignment & power allocation	Multi-agent	DQN & DDPG
	NOMA-based V2X networks	Xu *et al.* [183]	Spectrum & power allocation	Multi-agent	DDPG
	UAV-assisted IoT networks	Munaye *et al.* [168]	Bandwidth, throughput, & power	Multi-agent	DQN
	UAV-RIS-based HetNets	khalili *et al.* [184]	UAVs' trajectories, RISs' phase shifts, & subcarrier allocations	Multi-agent	Dueling DQN
Multi-RAT	Hybrid RF/VLC systems	Shrivastava *et al.* [185]	Bandwidth, power, & user association	Multi-agent	DQN
	Multi-RAT HetNets	Alwarafy *et al.* [41]	RAT selection & power control	Single& multi-agent	DQN & DDPG
	mmWave mobile hybrid access	Huang *et al.* [186]	Spectrum & power resource allocation	Single-agent	DQN
	Heterogeneous health systems	Chkirbene *et al.* [167]	RAT selection, data split control, & compression ratio control	Single-agent	DDPG

address the problem of delay minimization via joint spectrum and power resource allocation in mmWave mobile hybrid access network. The agent is located in the roadside BS, whose action space is discrete, corresponding to allocating spectrum and power resources for data. The state space is discrete, consisting of information about the current power and spectrum of the resource pool, required spectrum and power, and the number of spectrum and power levels. The reward signal is a continuous function defined in terms of queueing delay and the resource length required for each data. Using simulation results, it is shown that their proposed model guarantees the URLLC delay constraint when the load does not exceed 130%, which outperforms other conventional methods such as random and greedy algorithms.

Healthcare systems are one of the main services in next generation wireless systems. Unlike the work in [41], a pioneer work is presented in [167] to address the problem of network selection with the aim of optimizing medical data delivery over heterogeneous health systems. In particular, an optimization problem is formulated in which the network selection problem is integrated with adaptive compression to minimize network energy consumption and latency while meeting applications' QoS requirements. A single-agent DDPG-based DRL model is proposed to solve it. The agent is a centralized entity that can access all radio access networks (RANs) information and Patient Edge Node (PEN) data running in the core network. The action space is discrete, corresponding to the joint selection of data split over the existing RANs and the adequate compression ratio. The state space is a hybrid (continuous and discrete) defined by two elements: the fraction of time that the PENs should use over a particular RAN and the PEN investigated in the current timestamp. The reward is a continuous function, which is defined in terms of: the fraction of data of PEN that will be transferred through RAN, the energy consumed by PEN to transfer bits over RAN, distortion, expected latency of RANs, the monetary cost of PENs to use RANs, the resource share, the fraction of time that the PENs should use over a particular, and the data rate. Simulation results demonstrate that their proposed scheme outperforms the greedy techniques in terms of minimizing energy consumption and latency while satisfying different PENs requirements.

Findings and lessons learned: This section reviews the use of DRL methods for joint radio resources of power, spectrum, access control, user association, and rate. Table 7 summarizes the reviewed papers in this section. We observe that DRL tools can be efficiently deployed to address different types of joint radio resources for diversified network scenarios. The results obtained using DRL models are better than the heuristic methods [168], [183] and comparable to the state-of-the-art optimization approaches [67], [177]. Also, note that the main motivations of using DRL techniques in addressing the joint radio resources problems presented in this subsection are the complexity of these formulated problems, the limited information about system dynamics and CSI, and the difficulty in applying traditional RRAM methods to solve such problems.

We also observe that multi-agent DRL deployment based on value-based algorithms receives more attention than policy-based algorithms. The reason is that users tend to have more control over their channel selection, data control, and transmission mode selection, and hence we find a more popular implementation of DRL agents at local IoT devices. In addition, the integration of value-based and policy-based algorithms for joint RRAM is also an interesting concept that requires more investigation, especially for multi-agent deployment scenarios. In particular, depending on the type of radio resources under investigation, resources with continuous nature such as power typically implement policy-based algorithms, while resources with discreet nature such as channel allocation and user association typically implement value-based algorithms. Simultaneous dealing with continuous and discrete types of radio resources may integrate both the policy- and value-based DRL algorithms to learn a global

TABLE 8. Advantages and disadvantages/shortcomings of DRL methods when applied for RRAM problems in next generation HetNets.

Advantages	Disadvantages/Shortcomings
- Can solve complex RRAM optimization problems relying on limited network information	- The prohibitive high dimensionality of state and action spaces
- Enable network entities to learn efficient RRAM policies for wireless environment	- The high sensitivity to DNNs' hyperparameters
- Can be deployed online to make autonomous decisions based on local observations of network	- The system dependency of learned DRL policies
- Can overcome limitations of conventional RRAM methods	- The need to continuously train and update DRL models
- Can be equipped with prediction capabilities to enable RRAM forecasting	- Hard to beat well-designed algorithms if domain knowledge exists
- Used when accurate RRAM mathematical models do not exist	

policy as in [41], [182], or even adopting the value-based algorithms as in, e.g., [172], [174], [185] with an expense of added quantization error.

We also observe that DRL methods for cellular networks as well as IoT wireless networks gain more attention than multi-RAT networks, particularly for D2D and V2V communications. In addition, there is a lack of research on applications of DRL for emerging IoT applications, such as healthcare systems as investigated recently in [167], which is also a promising field that requires more attention. Furthermore, we observe a lack of research on DRL applications for joint RRAM in satellite networks, which also deserves more in-depth investigation.

V. OPEN CHALLENGES AND FUTURE RESEARCH DIRECTIONS

Throughout the previous section, we have demonstrated the superiority of DRL algorithms over traditional methods in addressing complex RRAM problems for modern wireless networks. However, there are still several challenges and open issues that either not explored yet or need further exploration. This section provides highlights these open challenges and provides insights on future research directions in the context of DRL-based RRAM for next generation wireless networks. Table 8 summarizes the advantages and disadvantages/shortcomings of DRL methods when applied for RRAM in next generation wireless networks.

A. OPEN CHALLENGES

1) CENTRALIZED VS. DECENTRALIZED RRAM TECHNIQUES

Future wireless networks are characterized mainly by their massive heterogeneity in wireless RANs, the number of user devices, and types of applications. Centralized DRL-based RRAM schemes are efficient in guaranteeing enhanced network QoS and fairness in allocating radio resources. They also ensure that RRAM optimization problems will not get stuck in local minima due to their holistic view of the system. However, formulating and solving RRAM optimization problems become tough tasks in such large-scale HetNets. Hence, centralized DRL-based RRAM solutions are typically unscalable. This motivates distributed multi-agent DRL-based algorithms that enable edge devices to make resource allocation decisions locally. Stochastic Game-based DRL algorithms are one promising research direction in this context [14]. However, the rapid increase in the number of edge devices (players) makes information exchange in such networks unmanageable. Also, the partial observability of agents might lead to suboptimal RRAM policies. Therefore, it is an open challenge to develop DRL-assisted algorithms that optimally balance between the centralization and distribution issue in RRAM. A possible solution is to develop hybrid ecosystems that implement some DRL models at the network's edge, e.g., at the ESs or user devices, instead of deploying all DRL algorithms on a centralized network.

2) DIMENSIONALITY OF STATE SPACE IN HETNETS

In modern wireless HetNets, service requirements and network conditions are rapidly changing. Hence, single-agent DRL algorithms must be designed to capture and respond to these fast network changes. To this end, it is required to reduce the state space and action space during the learning process, which inevitably degrades the quality of the learned policies. The existence of multi-agents and their interactions will also complicate the agents' environment and prohibitively increase the dimensionality of state space, which will slow down the learning algorithms. A possible solution to this issue is to split the large state spaces into smaller ones through state-space decomposition. The idea is to use smaller DNNs to learn the dynamics of the decomposed sub-state spaces, while another DNN considers the relatively less frequent interactions between the different sub-state spaces [189]. This approach enables us to distribute computation and accelerate agents' training.

3) RELIABILITY OF TRAINING DATASET

Although the DRL-based solutions for RRAM we reviewed previously demonstrate efficient performance results, almost all the models are developed based on simulated training and testing datasets. The simulated dataset is typically produced based on some stochastic models, which provide simplified versions of practical systems and greatly ignore hidden system patterns. This methodology greatly weakens the reliability of the developed policies as their performance on practical networks would be skeptical. Hence, it is imperative to develop more effective and reliable approaches that generate precise simulation datasets and capture practical system settings as much as possible. This ensures high reliability and confidence during the training and testing modes of the developed RRAM policies. Developing such approaches is still a challenge due to the large-scale nature and rapid variations of future wireless environments.

On the other hand, the DRL models are sensitive to any change in the input data. Any minor changes in the input data will cause considerable change in the models' output. This mainly deteriorates the reliability of DRL algorithms, especially when deployed for modern IoT applications that require ultra-reliability, such as remote surgery or any other mission-critical IoT applications. Hence, ensuring high reliability for DRL models is a challenging issue. A possible solution to such issues is to exploit real-field measurement data collected from various cellular and IoT wireless scenarios to train and test the DRL-based RRAM models. This will increase the reliability of the learned policies and also enables DRL model generalization.

4) ENGINEERING OF DRL MODELS FOR RRAM

Since DRL employs DNNs as function approximators for the reward functions, DRL models will inherit some of the challenges that exist in the DNN world. For example, it is still quite challenging to optimize the DNN hyperparameters, such as the type of DNNs used (e.g., convolutional, fully connected, or RNN), the number of hidden layers, the number of neurons per hidden layer, the learning rate, batch size, etc. DRL models suffer from high sensitivity to these hyperparameters. This challenge is even exacerbated in multi-agent settings as all agents share the same radio resources and must converge simultaneously to some policies. A possible solution is to implement some optimization techniques from the deep learning field, such as grid and random search methods, to find the optimal configuration of these hyperparameters [190].

On the other hand, the engineering of DRL parameters such as state space and reward function is challenging for RRAM. The state space must be engineered to capture useful and representative information about the wireless environment, such as the available radio resources, users' QoS requirements, channel quality, etc. Such information is crucial and heavily defines the learning and convergence behaviors of DRL agents. Again, the presence of multi-agents will even make it more challenging, as discussed in [14]. Also, since DRL models are reward-driven learning algorithms, the design of the reward function is also essential to guide the agent during the policy-learning stage. Formulating reward functions that capture the network objective and account for the available radio resources is also challenging.

5) SYSTEM DEPENDENCY OF DRL MODELS

DRL models are system-dependent as they are trained and tested for specific wireless environments and networks. Therefore, they provide effective results when employed to solve specific types of problems for which they are trained. However, if there would be a significant change in the characteristics of the wireless environment or the nature of the RRAM problem, such as network topology and available radio resources, the DRL model must be retrained as the old model is no longer reflecting the new training experiences. In modern wireless HetNets, such cases are frequently encountered, especially with real-time applications or in highly dynamic environments. In such a case, it becomes quite challenging for DRL agents to update and retrain their DNNs with rapidly changing input information from the HetNet environment [1]. A possible solution is to design DRL-based RRAM models in a manner that supports generalization via transfer learning and meta-learning. Multi-tasks DRL approaches [191], [192] are efficient frameworks to support these aspects.

On the other hand, if domain knowledge is available or easy to obtain, it becomes hard for the DRL algorithms to beat the well-designed algorithms based on the full domain knowledge. This fact has been observed and reported in the surveyed papers in Section IV.

6) CONTINUOUS TRAINING OF DRL MODELS

DRL algorithms require big datasets to train their models, which is typically associated with a high cost [15]. The network system pays this cost during the information collection process due to, e.g., the high delays, extra overhead, and energy consumption. The emergence of a large number of real-time applications and services has even increased this training cost. In this context, DRL models require to be continuously retrained with fresh data collected from the wireless environment to be up-to-date and ensure accurate and long-term control decisions. It is not practical to conduct manual retraining of the models in such large-scale HetNets settings. Also, manually monitoring and updating DRL models in multi-agent scenarios becomes an expensive task. Therefore, continuous retraining can solve this issue, in which a dedicated autonomous system is employed to continuously assess and retrain old DRL models.

7) CONTEXT OF RRAM

The implementation of DRL algorithms basically depends on the use-cases. The context and deployment scenarios in which RRAM is required must be considered during the development of DRL models. For example, RRAM in health-sector IoT applications is different from the environmental IoT applications counterparts. Due to the high sensitivity of data in the health-sector applications, extra data preprocessing must be performed, including data compression and encryption [167]. This will directly affect the number of radio resources to be allocated for such applications. Hence, DRL models must be aware of the context aspect of applications, which is considered another challenge. A possible solution is to develop context-aware DRL models that are able to learn context variables in an unsupervised manner and adapt the policy to the existing context, e.g., as in [193].

8) COMPETING OBJECTIVE DESIGN OF DRL MODELS

Next generation wireless networks are expected to provide enhanced system QoS in terms of high data rate, high

EE/SE, and reduced latency in order to support the emerging IoT vital applications. Depending on the deployment scenario, formulating multi-objective RRAM optimization problems usually ends with many competing objectives and/or constraints. For instance, in cellular UDNs, high resource utilization of, e.g., power allocation or channel may cause severe interference. Also, for IoT applications such as vehicular communications, we require to ensure ultra-reliable and low-latency communication links, which are usually competing objectives. Therefore, developing multi-objective DRL-based RRAM models that accommodate these competing requirements is still a persisting challenge. For example, frameworks to facilitate the development of multi-agent algorithms similar to those presented in [194] can be adopted for RRAM problems.

B. FUTURE RESEARCH DIRECTIONS

1) DRL WITH EXPLAINABLE AI (XDRL) FOR RRAM

Explainable AI (XDRL) has recently emerged as an efficient technology to improve the performance of DRL models. It is mainly envisioned to unlock the "black-box" nature of conventional ML approaches and provide interpretability and explainability for DRL models [195]. In particular, XDRL explains the reasons behind certain predictions made by DRL models (or ML models in general) by fully understanding the precise working principle of these models. Hence, ensuring trust, reliability, and transparency in the DRL algorithms' policy development and decision-making processes. The research on XDRL technologies in wireless communication is still at its initial stages, and there are still some key issues for future research in the context of RRAM for next generation wireless networks [196]. For example, DRL models can get stuck easily into local optimal solutions when utilized to solve complex RRAM problems. This issue can be significantly avoided with the help of XDRL. Fortunately, the heterogeneity of information in modern wireless HetNets will significantly help to achieve the interpretation for DRL algorithms. In this context, developing RRAM schemes for wireless HetNets, through entity recognition, Shapley value-based methods, entity-relationship extraction, and representation learning (e.g., Hindsight Experience Replay, Hierarchical DRL, and self-attention) makes the DRL models' interpretation more reliable, accurate, and intuitive, which is a promising research direction.

2) INTEGRATING DRL AND BLOCKCHAIN TECHNIQUES

Blockchain-based RRAM has emerged recently as one of the promising enabling technologies for future wireless HetNets [3]. It has gained considerable momentum lately due to its ability to provide intelligent, secure, and highly efficient distributed resource sharing and management. The integration of DRL with Blockchain is also an interesting research direction, as in [197]–[199]. For example, DRL algorithms can be distributively deployed within participants or within the centralized spectrum-access systems to facilitate spectrum auctions and transactions [199]. Also, DRL can be utilized to ensure efficiency of the consensus process, enhance energy-efficient resource allocation, and reduce computation overhead in Blockchain-enabled wireless networks [200]. In addition, many of the auction's winner-determination problems in future wireless HetNets are expected to be extremely complex and intractable due to the massive increase in the number of participants, e.g., bidders and sellers. Hence, DRL algorithms are efficient tools that can be utilized to solve such types of problems.

3) FEDERATED DRL (FDRL)-BASED RRAM

Federated learning (FL) framework is envisioned mainly to preserve data privacy in ML algorithms [201]–[203]. In FL, ML algorithms are locally distributed at the wireless network edge, and the data is processed locally and not shared globally. The local ML models are then utilized for training a centralized global model. In this context, the federated DRL learning (FDRL) scheme can be leveraged when many user devices require making autonomous decisions locally. In such a case, DRL agents do not exchange their local observations, and also, not necessarily all agents receive reward signals [204].

Developing fine-grained policies in DRL becomes challenging when the state space is small and the training dataset is very limited [205]. In FDRL, the direct exchange of data between agents is not possible as this will preach the privacy promise of FL scheme. Instead, local DRL models can be developed and trained for agents with the help of other agents while preserving users' data privacy, as in [206]. Hence, developing algorithms and schemes that guarantee data and models privacy during both information sharing and models updating is an interesting research direction.

FDRL framework can also be exploited in the RRAM of modern HetNets. For example, it can be deployed for solving complex wireless network optimization problems, such as power control in cellular UDNs. In this context, FDRL can ensure a global solution for optimization problems without sharing information between BSs; each BS solves its optimization problem locally and shares the results with neighboring BSs. Also, FDRL can be adapted in distributed optimization settings, such as user association and channel access, to ensure optimal global solutions.

4) DRL-BASED LOAD BALANCING FOR SELF-SUSTAINING NETWORKS

Load balancing in modern wireless UDNs is another promising research direction. The objective is to balance the wireless networks by moving some users from the heavily congested BSs to uncongested ones, thus improving BSs utilization and providing enhanced QoS provisioning. Although the load balancing field has been heavily investigated in the literature using conventional resource management approaches, as in [207]–[209], there still a research gap in applying DRL for such a field. In this context, DRL can be adopted to realize the self-sustaining (or self-organization) vision of next generation wireless networks [3]. Hence, developing single/multi-agent DRL models to achieve intelligent load balancing in future HetNets, is a possible research direction.

5) MADRL ALGORITHMS IN SUPPORT OF MASSIVE HETEROGENEITY AND MOBILITY

Load balancing in modern wireless UDNs is another promising research direction. The objective is to balance the wireless networks by moving some users from the heavily congested BSs to uncongested ones, thus improving BSs utilization and providing enhanced QoS provisioning. Although the load balancing field has been heavily investigated in the literature using conventional resource management approaches, as in [207]–[209], there still a research gap in applying DRL for such a field. In this context, DRL can be adopted to realize the self-sustaining (or self-organization) vision of next generation wireless networks [3]. Hence, developing single/multi-agent DRL models to achieve intelligent load balancing in future HetNets, is a possible research direction. Such models must be agile to network dynamics, including varying users' mobility patterns and network resources availability.

6) DRL-BASED RRAM WITH GENERATIVE ADVERSARIAL NETWORKS (GANS) FOR RRAM

Ensuring the reliability of DRL algorithms is one of the major challenges and objectives in DRL-based RRAM methods. In many real-life scenarios, we may require to deploy DRL models to allocate resources in vital systems requiring ultra-reliability, such as IoT healthcare applications [167]. In this context, there are proposals on Generative Adversarial Networks (GANs), which have emerged recently as an effective technique to enhance the reliability of DRL algorithms [210].

In practice, the shortage of realistic training datasets required to train DRL models and learn optimal policies is a challenging issue. To overcome this, GANs are utilized, which generate large amounts of realistic datasets synthetically by expanding the available limited amounts of real-time datasets. From a DRL perspective, GANs-generated synthetic data is more effective and reliable than traditional augmentation methods [79]. This is because DRL agents will be exposed to various extreme challenging and practical situations by merging the realistic and synthetic data, enabling DRL models to be trained on unpredicted and rare events. Another advantage of GAN over traditional data augmentation methods is that it eliminates dataset biases in the synthetic data, which greatly enhances the quality of the generated data and leads to more reliability in DRL models' training and learning processes.

In general, the research in the GANs-based DRL methods for RRAM is still in its early stages, and we believe that it will take further pace in the future. For example, developing experienced DRL-based algorithms for URLL communication using GANs in which DRL models are pre-trained based on a mix of real and synthetic data is a promising research direction as in [211].

7) DRL FOR RRAM IN RIS-ASSISTED WIRELESS NETWORKS

Reconfigurable Intelligent Surfaces (RIS) have emerged recently as an innovative technology to enhance the QoS of future wireless networks [212], [213]. RISs can be deployed in cellular networks as passive reflecting elements to provide near line-of-sight communication links to users, hence enhancing communication reliability, increasing throughput, and reducing latency [214], [215]. Deploying RIS to assist cellular communication, however, requires judicious RRAM schemes to optimize network performance. This research field is still nascent, and there is much to do for future research and investigation, especially in the context of DRL-based RRAM techniques. Towards this, it is required to develop end-to-end DRL-based algorithms that jointly optimize the configuration of the RIS system, i.e., elements' phases and amplitudes, and radio resources of BSs. For instance, designing DRL models that intelligently and optimally allocate the downlink BSs' transmit power and/or BSs' beamforming configuration from one side and the amplitude and phase shifts of the RIS elements on the other side is a promising research direction, as in [43]. We also believe that the currently ongoing research in RIS-assisted wireless networks, e.g., [43], [216]–[218] will be cornerstones.

8) DRL FOR RRAM IN WIRELESS DIGITAL TWIN NETWORKS

Digital twin (DT) has recently emerged as a promising technology for future wireless networks [219]. DT is a virtual representation of the components and dynamics of a given physical system, which is envisioned to bridge the connection gap between physical systems and digital spaces. The digital replicas of physical systems, such as user devices, BSs, and machines, are constructed at the server based on historical data and real-time running status. DT utilizes tools from ML, data analytics, and multiphysics simulation to study and analyze the dynamics of physical systems. Therefore, DT enables system monitoring, real-time interaction, and reliable communication between physical systems and digital space in order to optimize the operation of physical systems [220]. With these promising features, DT is getting considerable interest recently in enhancing the performance of wireless communication networks for applications, such as computation offloading, content caching, and RRAM. For example, a promising research direction is to develop DRL algorithms to address various problems in wireless DT networks, such as the DT placement and migration problems [221], in capturing the dynamics of UAV-based networks [222], in blockchain-based networks to enhance network security and users privacy [223].

Table 9 summarizes the open challenges and future research directions provided in this section.

TABLE 9. Summary of challenges and future research directions in the context of using DRL for RRAM in future wireless networks.

Open Challenges	Developing DRL-based algorithms that optimally balance the centralization and distribution issue of RRAM in future large-scale massive HetNets.
	Reducing the dimensionality of state space in distributed MADRL algorithms during the learning process without slowing down or degrading the quality of learned RRAM policies.
	Developing more effective and reliable training approaches that generate accurate simulation datasets and capture practical system settings.
	Optimizing DRL models' hyperparameters, especially in MADRL scenarios, and engineering the state space and reward functions to capture representative information about system dynamics.
	Designing agile DRL algorithms that can quickly update and retrain the DNNs in response to the rapid change of input information from the highly dynamic HetNet environment.
	Performing continuous retraining for the DRL models, especially MADRL, with fresh data in future large-scale and rapidly changing wireless environments.
	Developing DRL models that are aware of the context aspect and use cases of various emerging applications.
	Developing DRL algorithms that accommodate competing multi-objectives relevant to emerging applications.
Future Research Directions	Developing efficient and reliable DRL algorithms for RRAM in next-generation HetNets based on the XAI concept through, e.g., entity recognition, entity-relationship extraction, and representation learning.
	Developing DRL-based Blockchain techniques to address the problem of distributed resource sharing and management for future HetNets, e.g., to facilitate spectrum auctions and transactions, solving the problem of auction's winner-determination, etc.
	Developing FDRL algorithms that ensure global solutions for complex RRAM optimization problems while guaranteeing data and models privacy during information sharing and models updating.
	Developing DRL models to achieve intelligent load balancing in future self-sustaining (or self-organization) HetNets.
	Developing light-weighted and agile networked MADRL algorithms that enable cooperation between agents with different heterogeneous reward functions and adapt to environments with rapid mobility.
	Developing ultra-reliable RRAM schemes by integrating DRL algorithms and GANs techniques to support emerging IoT applications with high-reliability demands.
	Developing end-to-end DRL-based algorithms that jointly optimize the configuration of RIS systems, i.e., elements' phases and amplitudes, and radio resources of networks, e.g., downlink transmit power.
	Developing DRL algorithms to address various RRAM problems in wireless digital twin networks, e.g., the DT placement and migration problems.

VI. CONCLUSION

This paper presents a comprehensive survey on the applications of DRL techniques in RRAM for next generation wireless HetNets. We thoroughly review the conventional approaches for RRAM, including their types, advantages, and limitations. We then illustrate how the emerging DRL approaches can overcome these shortcomings to enable DRL-based RRAM. After that, we illustrate how the RRAM optimization problems can be formulated as an MDP before solving them using DRL techniques. Furthermore, we conduct an extensive overview of the most efficient DRL algorithms that are widely leveraged in addressing RRAM problems, including the value- and policy-based algorithms. The advantages, limitations, and use-cases for each algorithm are provided. We then conduct a comprehensive and in-depth literature review and classified the existing related works based on both the radio resources they are addressing and the type of wireless networks they are considering. To this end, the types of DRL models developed in these related works and their main elements are carefully identified. Finally, we outline important open challenges and provide insights into future research directions in the context of DRL-based RRAM.

REFERENCES

[1] F. Hussain, S. A. Hassan, R. Hussain, and E. Hossain, "Machine learning for resource management in cellular and IoT networks: Potentials, current solutions, and open challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1251–1275, 2nd Quart., 2020.

[2] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[3] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May/Jun. 2020.

[4] Z. Zhang *et al.*, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.

[5] "6G Summit Connecting the Unconnected." [Online]. Available: https://6gsummit.org (accessed Feb. 18, 2022).

[6] "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," Cisco, San Jose, CA, USA, White Paper, 2019.

[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[8] Y. L. Lee and D. Qin, "A survey on applications of deep reinforcement learning in resource management for 5G heterogeneous networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2019, pp. 1856–1862.

[9] F. Obite, A. D. Usman, and E. Okafor, "An overview of deep reinforcement learning for spectrum sensing in cognitive radio networks," *Digit. Signal Process.*, vol. 113, Jun. 2021, Art. no. 103014.

[10] S. Gupta, G. Singal, and D. Garg, "Deep reinforcement learning techniques in diversified domains: A survey," *Arch. Comput. Methods Eng.*, vol. 28, pp. 4715–4754, Feb. 2021.

[11] Z. Du, Y. Deng, W. Guo, A. Nallanathan, and Q. Wu, "Green deep reinforcement learning for radio resource management: Architecture, algorithm compression, and challenges," *IEEE Veh. Technol. Mag.*, vol. 16, no. 1, pp. 29–39, Mar. 2021.

[12] Y. Qian, J. Wu, R. Wang, F. Zhu, and W. Zhang, "Survey on reinforcement learning applications in communication networks," *J. Commun. Inf. Netw.*, vol. 4, no. 2, pp. 30–39, Jun. 2019.

[13] Z. Xiong, Y. Zhang, D. Niyato, R. Deng, P. Wang, and L.-C. Wang, "Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 44–52, Jun. 2019.

[14] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for ai-enabled wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1226–1252, 2nd Quart., 2021.

[15] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

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

[17] P. Yang, Y. Xiao, M. Xiao, and S. Li, "6G wireless communications: Vision and potential techniques," *IEEE Netw.*, vol. 33, no. 4, pp. 70–75, Jul./Aug. 2019.

[18] K. David and H. Berndt, "6G vision and requirements: Is there any need for beyond 5G?" *IEEE Veh. Technol. Mag.*, vol. 13, no. 3, pp. 72–80, Sep. 2018.

[19] S. Elmeadawy and R. M. Shubair, "6G wireless communications: Future technologies and research challenges," in *Proc. Int. Conf. Electr. Comput. Technol. Appl. (ICECTA)*, 2019, pp. 1–5.

[20] T. Huang, W. Yang, J. Wu, J. Ma, X. Zhang, and D. Zhang, "A survey on green 6G network: Architecture and technologies," *IEEE Access*, vol. 7, pp. 175758–175768, 2019.

[21] A. Alwarafy, K. A. Al-Thelaya, M. Abdallah, J. Schneider, and M. Hamdi, "A survey on security and privacy issues in edge-computing-assisted Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4004–4022, Mar. 2021.

[22] A. I. Sulyman, A. Alwarafy, G. R. MacCartney, T. S. Rappaport, and A. Alsanie, "Directional radio propagation path loss models for millimeter-wave wireless networks in the 28-, 60-, and 73-GHz bands," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6939–6947, Oct. 2016.

[23] A. Alwarafy, A. Albaseer, B. S. Ciftler, M. Abdallah, and A. Al-Fuqaha, "AI-based radio resource allocation in support of the massive heterogeneity of 6G networks," in *Proc. IEEE 4th 5G World Forum (5GWF)*, Oct. 2021, pp. 464–469.

[24] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep-learning-based wireless resource allocation with application to vehicular networks," *Proc. IEEE*, vol. 108, no. 2, pp. 341–356, Feb. 2020.

[25] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and M.-N. Nguyen, "Non-cooperative energy efficient power allocation game in D2D communication: A multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 100480–100490, 2019.

[26] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 4th Quart., 2019.

[27] A. Zappone, M. D. Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019.

[28] H. Khorasgani, H. Wang, and C. Gupta, "Challenges of applying deep reinforcement learning in dynamic dispatching," 2020, *arXiv:2011.05570*.

[29] G. S. Rahman, T. Dang, and M. Ahmed, "Deep reinforcement learning based computation offloading and resource allocation for low-latency fog radio access networks," *Intell. Converged Netw.*, vol. 1, no. 3, pp. 243–257, 2020.

[30] A. Mohammed, H. Nahom, A. Tewodros, Y. Habtamu, and G. Hayelom, "Deep reinforcement learning for computation offloading and resource allocation in blockchain-based multi-UAV-enabled mobile edge computing," in *Proc. 17th Int. Comput. Conf. Wavelet Active Media Technol. Inf. Process. (ICCWAMTIP)*, 2020, pp. 295–299.

[31] S. Sheng, P. Chen, Z. Chen, L. Wu, and Y. Yao, "Deep reinforcement learning-based task scheduling in IoT edge computing," *Sensors*, vol. 21, no. 5, p. 1666, 2021.

[32] X. Chen and G. Liu, "Energy-efficient task offloading and resource allocation via deep reinforcement learning for augmented reality in mobile edge networks," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10843–10856, Jul. 2021.

[33] Q. Liu, T. Han, and E. Moges, "Edgeslice: Slicing wireless edge computing network with decentralized deep reinforcement learning," 2020, *arXiv:2003.12911*.

[34] M. Lin and Y. Zhao, "Artificial intelligence-empowered resource management for future wireless communications: A survey," *China Commun.*, vol. 17, no. 3, pp. 58–77, Mar. 2020.

[35] Q. T. A. Pham, K. Piamrat, and C. Viho, "Resource management in wireless access networks: A layer-based classification-version 1.0," Rep. PI-2017, 2014, p. 23.

[36] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

[37] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, Mar. 2019.

[38] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, "A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 668–695, 2nd Quart., 2021.

[39] T. S. Rappaport *et al.*, "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78729–78757, 2019.

[40] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.

[41] A. Alwarafy, B. S. Ciftler, M. Abdallah, and M. Hamdi, "DeepRAT: A DRL-based framework for multi-RAT assignment and power allocation in hetnets," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2021, pp. 1–6.

[42] J. Kong, Z.-Y. Wu, M. Ismail, E. Serpedin, and K. A. Qaraqe, "Q-learning based two-timescale power allocation for multi-homing hybrid RF/VLC networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 443–447, Apr. 2020.

[43] G. Lee, M. Jung, A. T. Z. Kasgari, W. Saad, and M. Bennis, "Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[44] B. S. Ciftler, M. Abdallah, A. Alwarafy, and M. Hamdi, "DQN-based multi-user power allocation for hybrid RF/VLC networks," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[45] A. Ahmad, S. Ahmad, M. H. Rehmani, and N. U. Hassan, "A survey on radio resource allocation in cognitive radio sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 888–917, 2nd Quart., 2015.

[46] M. El Tanab and W. Hamouda, "Resource allocation for underlay cognitive radio networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1249–1276, 2nd Quart., 2017.

[47] M. Naeem, A. Anpalagan, M. Jaseemuddin, and D. C. Lee, "Resource allocation techniques in cooperative cognitive radio networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 729–744, 2nd Quart., 2014.

[48] S. Manap, K. Dimyati, M. N. Hindia, M. S. A. Talip, and R. Tafazolli, "Survey of radio resource management in 5G heterogeneous networks," *IEEE Access*, vol. 8, pp. 131202–131223, 2020.

[49] M. Peng, C. Wang, J. Li, H. Xiang, and V. Lau, "Recent advances in underlay heterogeneous networks: Interference control, resource allocation, and self-organization," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 700–729, 2nd Quart., 2015.

[50] Y. Teng, M. Liu, F. R. Yu, V. C. M. Leung, M. Song, and Y. Zhang, "Resource allocation for ultra-dense networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2134–2168, 3rd Quart., 2019.

[51] K. Piamrat, A. Ksentini, J.-M. Bonnin, and C. Viho, "Radio resource management in emerging heterogeneous wireless networks," *Comput. Commun.*, vol. 34, no. 9, pp. 1066–1076, 2011.

[52] N. Xia, H. Chen, and C. Yang, "Radio resource management in machine-to-machine communications—A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 791–828, 1st Quart., 2018.

[53] S. Sadr, A. Anpalagan, and K. Raahemifar, "Radio resource allocation algorithms for the downlink of multiuser OFDM communication systems," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 3, pp. 92–106, 3rd Quart., 2009.

[54] E. Yaacoub and Z. Dawy, "A survey on uplink resource allocation in OFDMA wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, pp. 322–337, 2nd Quart., 2012.

[55] R. O. Afolabi, A. Dadlani, and K. Kim, "Multicast scheduling and resource allocation algorithms for OFDMA-based systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 240–254, 1std Quart., 2013.

[56] S. Chieochan and E. Hossain, "Adaptive radio resource allocation in OFDMA systems: A survey of the state-of-the-art approaches," *Wireless Commun. Mobile Comput.*, vol. 9, no. 4, pp. 513–527, 2009.

[57] D. Niyato and E. Hossain, "Radio resource management in MIMO-OFDM- mesh networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 45, no. 11, pp. 100–107, Nov. 2007.

[58] W. Zhao and S. Wang, "Resource sharing scheme for device-to-device communication underlaying cellular networks," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4838–4848, Dec. 2015.

[59] L. Song, D. Niyato, Z. Han, and E. Hossain, "Game-theoretic resource allocation methods for device-to-device communication," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 136–144, Jun. 2014.

[60] Y. Kawamoto, H. Nishiyama, N. Kato, F. Ono, and R. Miura, "Toward future unmanned aerial vehicle networks: Architecture, resource allocation and field experiments," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 94–99, Feb. 2019.

[61] A. Masmoudi, K. Mnif, and F. Zarai, "A survey on radio resource allocation for V2X communication," *Wireless Commun. Mobile Comput.*, vol. 2019, Oct. 2019, Art. no. 2430656.

[62] M. Allouch, S. Kallel, A. Soua, O. Shagdar, and S. Tohme, "Survey on radio resource allocation in long-term evolution-vehicle," *Concurrency Comput. Pract. Exp.*, vol. 34, no. 7, 2021, Art. no. e6228.

[63] S. Xu, G. Zhu, B. Ai, and Z. Zhong, "A survey on high-speed railway communications: A radio resource management perspective," *Comput. Commun.*, vol. 86, pp. 12–28, Jul. 2016.

[64] S. Dhilipkumar, C. Arunachalaperumal, and K. Thanigaivelu, "A comparative study of resource allocation schemes for D2D networks underlay cellular networks," *Wireless Personal Commun.*, vol. 106, no. 3, pp. 1075–1087, 2019.

[65] J. Clausen, *Branch and Bound Algorithms-Principles and Examples*, Univ. Copenhagen, Copenhagen, Denmark, 1999, pp. 1–30.

[66] D. S. Nau, V. Kumar, and L. Kanal, "General branch and bound, and its relation to A^* and AO^*," *Artif. Intell.*, vol. 23, no. 1, pp. 29–58, 1984.

[67] Y. S. Nasir and D. Guo, "Deep reinforcement learning for joint spectrum and power allocation in cellular networks," 2020, *arXiv:2012.10682*.

[68] Y. Xu *et al.*, "Robust resource allocation for two-tier HetNets: An interference-efficiency perspective," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 3, pp. 1514–1528, Sep. 2021.

[69] Y. Xu, G. Li, Y. Yang, M. Liu, and G. Gui, "Robust resource allocation and power splitting in SWIPT enabled heterogeneous networks: A robust minimax approach," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10799–10811, Dec. 2019.

[70] Y. Xu, H. Xie, and R. Q. Hu, "Max-min beamforming design for heterogeneous networks with hardware impairments," *IEEE Commun. Lett.*, vol. 25, no. 4, pp. 1328–1332, Apr. 2020.

[71] Y. Xu, H. Xie, C. Liang, and F. R. Yu, "Robust secure energy-efficiency optimization in SWIPT-aided heterogeneous networks with a nonlinear energy-harvesting model," *IEEE Internet Things J.*, vol. 8, no. 19, pp. 14908–14919, Oct. 2021.

[72] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[73] X. Zhang, X. Zhang, and Z. Wu, "Utility- and fairness-based spectrum allocation of cellular networks by an adaptive particle swarm optimization algorithm," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 1, pp. 42–50, Feb. 2020.

[74] X. He, X. Li, H. Ji, and H. Zhang, "Resource allocation for secrecy rate optimization in UAV-assisted cognitive radio network," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2021, pp. 1–6.

[75] M. Kim and I.-Y. Ko, "An efficient resource allocation approach based on a genetic algorithm for composite services in IoT environments," in *Proc. IEEE Int. Conf. Web Services*, 2015, pp. 543–550.

[76] Y. El Morabit, F. Mrabti, and E. H. Abarkan, "Spectrum allocation using genetic algorithm in cognitive radio networks," in *Proc. 3rd Int. Workshop RFID Adapt. Wireless Sens. Netw. (RAWSN)*, 2015, pp. 90–93.

[77] M. B. Satria, I. W. Mustika, and Widyawan, "Resource allocation in cognitive radio networks based on modified ant colony optimization," in *Proc. 4th Int. Conf. Sci. Technol. (ICST)*, 2018, pp. 1–5.

[78] M. Tian, H. Deng, and M. Xu, "Immune parallel artificial bee colony algorithm for spectrum allocation in cognitive radio sensor networks," in *Proc. Int. Conf. Comput. Inf. Telecommun. Syst. (CITS)*, 2020, pp. 1–4.

[79] M. Naeem, S. T. H. Rizvi, and A. Coronato, "A gentle introduction to reinforcement learning and its application in different fields," *IEEE Access*, vol. 8, pp. 209320–209344, 2020.

[80] 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.

[81] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao, and K. Wu, "Artificial-intelligence-enabled intelligent 6G networks," *IEEE Netw.*, vol. 34, no. 6, pp. 272–280, Nov./Dec. 2020.

[82] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, 2016, pp. 2094–2100.

[83] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1995–2003.

[84] R. Sutton, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems*, vol. 12. Cambridge, MA, USA: MIT Press, 2000, pp. 1057–1063.

[85] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.

[86] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*.

[87] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, 2014, pp. 387–395. [Online]. Available: http://proceedings.mlr.press/v32/silver14.html

[88] J. Xu and B. Ai, "Experience-driven power allocation using multi-agent deep reinforcement learning for millimeter-wave high-speed railway systems," *IEEE Trans. Intell. Transp. Syst.*, early access, Feb. 3, 2021, doi: 10.1109/TITS.2021.3054511.

[89] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," 2015, *arXiv:1511.05952*.

[90] D. Horgan *et al.*, "Distributed prioritized experience replay," 2018, *arXiv:1803.00933*.

[91] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 449–458.

[92] M. Hessel *et al.*, "Rainbow: Combining improvements in deep reinforcement learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 3215–3222.

[93] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," in *Proc. AAAI Fall Symp. Series*, 2015, pp. 1–7.

[94] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.

[95] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[96] G. Barth-Maron *et al.*, "Distributed distributional deterministic policy gradients," 2018, *arXiv:1804.08617*.

[97] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3826–3839, Sep. 2020.

[98] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," 2019, *arXiv:1911.10635*.

[99] I. Althamary, C.-W. Huang, and P. Lin, "A survey on multi-agent reinforcement learning methods for vehicular networks," in *Proc. 15th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, 2019, pp. 1154–1159.

[100] D. Lee, N. He, P. Kamalaruban, and V. Cevher, "Optimization for reinforcement learning: From a single agent to cooperative agents," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 123–135, May 2020.

[101] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.

[102] A. A. Khan and R. S. Adve, "Centralized and distributed deep reinforcement learning methods for downlink sum-rate optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8410–8426, Dec. 2020.

[103] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multiuser cellular networks: Deep reinforcement learning approaches," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6255–6267, Oct. 2020.

[104] F. Meng, P. Chen, and L. Wu, "Power allocation in multi-user cellular networks with deep Q learning approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–6.

[105] L. Zhang and Y.-C. Liang, "Deep reinforcement learning for multiagent power control in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2551–2564, Apr. 2021.

[106] Y. S. Nasir and D. Guo, "Deep actor-critic learning for distributed power control in wireless mobile networks," in *Proc. 54th Asilomar Conf. Signals Syst. Comput.*, 2020, pp. 398–402.

[107] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[108] Z. Bi and W. Zhou, "Deep reinforcement learning based power allocation for D2D network," in *Proc. IEEE 91st Veh. Technol. Conf. (VTC-Spring)*, 2020, pp. 1–5.

[109] S. Saeidian, S. Tayamon, and E. Ghadimi, "Downlink power control in dense 5G radio access networks through deep reinforcement learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[110] Z. Zhang, H. Qu, J. Zhao, and W. Wang, "Deep reinforcement learning method for energy efficient resource allocation in next generation wireless networks," in *Proc. Int. Conf. Comput. Netw. Internet Things*, 2020, pp. 18–24.

[111] Q. Wang, K. Feng, X. Li, and S. Jin, "PrecoderNet: Hybrid beamforming for millimeter wave systems with deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1677–1681, Oct. 2020.

[112] X. Li, J. Fang, W. Cheng, H. Duan, Z. Chen, and H. Li, "Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach," *IEEE Access*, vol. 6, pp. 25463–25473, 2018.

[113] L. Li, Q. Cheng, K. Xue, C. Yang, and Z. Han, "Downlink transmit power control in ultra-dense UAV network based on mean field game and deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15594–15605, Dec. 2020.

[114] N. Zhao, Z. Liu, and Y. Cheng, "Multi-agent deep reinforcement learning for trajectory design and power allocation in multi-UAV networks," *IEEE Access*, vol. 8, pp. 139670–139679, 2020.

[115] M. Yan, B. Chen, G. Feng, and S. Qin, "Federated cooperation and augmentation for power allocation in decentralized wireless networks," *IEEE Access*, vol. 8, pp. 48088–48100, 2020.

[116] J. G. Luis, M. Guerster, I. del Portillo, E. Crawley, and B. Cameron, "Deep reinforcement learning architecture for continuous power allocation in high throughput satellites," 2019, *arXiv:1906.00571*.

[117] J. J. G. Luis, M. Guerster, I. del Portillo, E. Crawley, and B. Cameron, "Deep reinforcement learning for continuous power allocation in flexible high throughput satellites," in *Proc. IEEE Cogn. Commun. Aerosp. Appl. Workshop (CCAAW)*, 2019, pp. 1–4.

[118] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu, and S. M. Sait, "A survey of rate-optimal power domain NOMA with enabling technologies of future wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2192–2235, 4th Quart., 2020.

[119] X. Yan, K. An, Q. Zhang, G. Zheng, S. Chatzinotas, and J. Han, "Delay constrained resource allocation for NOMA enabled satellite Internet of Things with deep reinforcement learning," *IEEE Internet Things J.*, early access, 2020. [Online]. Available: https://orbilu.uni.lu/bitstream/10993/45468/1/Delay%20Constrained%20Resource%20Allocation%20for%20NOMA.pdf

[120] A. Alwarafy, M. Alresheedi, A. F. Abas, and A. Alsanie, "Performance evaluation of space time coding techniques for indoor visible light communication systems," in *Proc. Int. Conf. Opt. Netw. Design Model. (ONDM)*, 2018, pp. 88–93.

[121] A. Memedi and F. Dressler, "Vehicular visible light communications: A survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 161–181, 1st Quart., 2021.

[122] M. Chen, W. Saad, and C. Yin, "Liquid state machine learning for resource allocation in a network of cache-enabled LTE-U UAVs," in *Proc. IEEE Global Commun. Conf.*, 2017, pp. 1–6.

[123] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–6.

[124] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141–5152, Nov. 2019.

[125] Q. Zhang, Y.-C. Liang, and H. V. Poor, "Intelligent user association for symbiotic radio networks using deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4535–4548, Jul. 2020.

[126] W. Lei, Y. Ye, and M. Xiao, "Deep reinforcement learning-based spectrum allocation in integrated access and backhaul networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 3, pp. 970–979, Sep. 2020.

[127] Z. Li, C. Wang, and C.-J. Jiang, "User association for load balancing in vehicular networks: An online reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 2217–2228, Aug. 2017.

[128] J. Zheng, X. Tang, X. Wei, H. Shen, and L. Zhao, "Channel assignment for hybrid NOMA systems with deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 10, no. 7, pp. 1370–1374, Jul. 2021.

[129] H. Song, L. Liu, J. Ashdown, and Y. Yi, "A deep reinforcement learning framework for spectrum management in dynamic spectrum access," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11208–11218, Jul. 2021.

[130] Y. Hu *et al.*, "Optimal transmit antenna selection strategy for MIMO wiretap channel based on deep reinforcement learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2018, pp. 803–807.

[131] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 2, pp. 257–265, Jun. 2018.

[132] S. Wang, H. Liu, P. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access," in *Proc. Int. Conf. Comput. Netw. Commun. (ICNC)*, 2017, pp. 257–265.

[133] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, Sep. 2009.

[134] M. Chu, H. Li, X. Liao, and S. Cui, "Reinforcement learning-based multiaccess control and battery prediction with energy harvesting in IoT systems," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2009–2020, Apr. 2019.

[135] Y. Zhang, P. Cai, C. Pan, and S. Zhang, "Multi-agent deep reinforcement learning-based cooperative spectrum sensing with upper confidence bound exploration," *IEEE Access*, vol. 7, pp. 118898–118906, 2019.

[136] N. Yang, H. Zhang, and R. Berry, "Partially observable multi-agent deep reinforcement learning for cognitive resource management," in *Proc. IEEE Global Commun. Conf.*, 2020, pp. 1–6.

[137] Y. Xu, J. Yu, W. C. Headley, and R. M. Buehrer, "Deep reinforcement learning for dynamic spectrum access in wireless networks," in *Proc. IEEE Military Commun. Conf. (MILCOM)*, 2018, pp. 207–212.

[138] L. Liang, H. Ye, and G. Y. Li, "Multi-agent reinforcement learning for spectrum sharing in vehicular networks," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2019, pp. 1–5.

[139] L. Liang, H. Yee, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Oct. 2019.

[140] J. Zhu, Y. Song, D. Jiang, and H. Song, "A new deep-Q-learning-based transmission scheduling mechanism for the cognitive Internet of Things," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2375–2385, Aug. 2018.

[141] H. Khan, A. Elgabli, S. Samarakoon, M. Bennis, and C. S. Hong, "Reinforcement learning-based vehicle-cell association algorithm for highly mobile millimeter wave communication," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 1073–1085, Dec. 2019.

[142] P. Yang *et al.*, "Dynamic spectrum access in cognitive radio networks using deep reinforcement learning and evolutionary game," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2018, pp. 405–409.

[143] Y. Cao, S.-Y. Lien, and Y.-C. Liang, "Deep reinforcement learning for multi-user access control in non-terrestrial networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1605–1619, Mar. 2021.
[144] S. Tomovic and I. Radusinovic, "A novel deep Q-learning method for dynamic spectrum access," in *Proc. 28th Telecommun. Forum (TELFOR)*, 2020, pp. 1–4.
[145] C. Zhong, Z. Lu, M. C. Gursoy, and S. Velipasalar, "A deep actor-critic reinforcement learning framework for dynamic multichannel access," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 1125–1139, Dec. 2019.
[146] A. Tondwalkar and D. A. Kwasinski, "Deep reinforcement learning for distributed uncoordinated cognitive radios resource allocation," 2019, *arXiv:1911.03366*.
[147] K. Al-Gumaei *et al.*, "A survey of Internet of Things and big data integrated solutions for industrie 4.0," in *Proc. IEEE 23rd Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, vol. 1, 2018, pp. 1417–1424.
[148] Z. Shi, X. Xie, H. Lu, H. Yang, M. Kadoch, and M. Cheriet, "Deep-reinforcement-learning-based spectrum resource management for Industrial Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3476–3489, Mar. 2021.
[149] S. B. Janiar and V. Pourahmadi, "Deep-reinforcement learning for fair distributed dynamic spectrum access in wireless networks," in *Proc. IEEE 18th Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2021, pp. 1–4.
[150] Y. Wang, X. Li, P. Wan, and R. Shao, "Intelligent dynamic spectrum access using deep reinforcement learning for VANETs," *IEEE Sensors J.*, vol. 21, no. 14, pp. 15554–15563, Jul. 2021.
[151] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Distributed multi-agent meta learning for trajectory design in wireless drone networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3177–3192, Oct. 2021.
[152] W. Jiang and W. Yu, "Multi-agent reinforcement learning based joint cooperative spectrum sensing and channel access for cognitive UAV networks," 2021, *arXiv:2103.08181*.
[153] Y. Xu, J. Yu, and R. M. Buehrer, "The application of deep reinforcement learning to distributed spectrum access in dynamic heterogeneous environments with partial observations," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4494–4506, Jul. 2020.
[154] S. Liu, X. Hu, and W. Wang, "Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems," *IEEE Access*, vol. 6, pp. 15733–15742, 2018.
[155] X. Hu, S. Liu, R. Chen, W. Wang, and C. Wang, "A deep reinforcement learning-based framework for dynamic resource allocation in multibeam satellite systems," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1612–1615, Aug. 2018.
[156] B. Zhao, J. Liu, Z. Wei, and I. You, "A deep reinforcement learning based approach for energy-efficient channel allocation in satellite Internet of Things," *IEEE Access*, vol. 8, pp. 62197–62206, 2020.
[157] J. Liu, B. Zhao, Q. Xin, and H. Liu, "Dynamic channel allocation for satellite Internet of Things via deep reinforcement learning," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, 2020, pp. 465–470.
[158] F. Zheng, Z. Pi, Z. Zhou, and K. Wang, "Leo satellite channel allocation scheme based on reinforcement learning," *Mobile Inf. Syst.*, vol. 2020, Dec. 2020, Art. no. 8868888.
[159] U. Challita, L. Dong, and W. Saad, "Proactive resource management in LTE-U systems: A deep learning perspective," 2017, *arXiv:1702.07031*.
[160] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.
[161] S. Wang and T. Lv, "Deep reinforcement learning based dynamic multichannel access in hetnets," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2019, pp. 1–6.
[162] H. Peng and X. Shen, "Deep reinforcement learning based resource management for multi-access edge computing in vehicular networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2416–2428, Oct.–Dec. 2020.
[163] U. Challita and D. Sandberg, "Deep reinforcement learning for dynamic spectrum sharing of LTE and NR," 2021, *arXiv:2102.11176*.
[164] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," 2018, *arXiv:1802.06958*.
[165] Q. Liu, T. Han, N. Zhang, and Y. Wang, "Deepslicing: Deep reinforcement learning assisted resource allocation for network slicing," in *Proc. IEEE Global Commun. Conf.*, 2020, pp. 1–6.
[166] F. Tang, Y. Zhou, and N. Kato, "Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5G HetNet," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2773–2782, Dec. 2020.
[167] Z. Chkirbene, A. A. Abdellatif, A. Mohamed, A. Erbad, and M. Guizani, "Deep reinforcement learning for network selection over heterogeneous health systems," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 1, pp. 258–270, Jan./Feb. 2022.
[168] Y. Y. Munaye, R.-T. Juang, H.-P. Lin, G. B. Tarekegn, and D.-B. Lin, "Deep reinforcement learning based resource management in UAV-assisted IoT networks," *Appl. Sci.*, vol. 11, no. 5, p. 2163, 2021.
[169] X. Zhang, M. Peng, S. Yan, and Y. Sun, "Deep-reinforcement-learning-based mode selection and resource allocation for cellular V2X communications," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6380–6391, Jul. 2019.
[170] J. Jang and H. J. Yang, "Deep reinforcement learning-based resource allocation and power control in small cells with limited information exchange," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13768–13783, Nov. 2020.
[171] X. Liao, J. Shi, Z. Li, L. Zhang, and B. Xia, "A model-driven deep reinforcement learning heuristic algorithm for resource allocation in ultra-dense cellular networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 983–997, Jan. 2020.
[172] X. Zhang, Z. Lin, B. Ding, B. Gu, and Y. Han, "Deep multi-agent reinforcement learning for resource allocation in D2D communication underlaying cellular networks," in *Proc. 21st Asia-Pacific Netw. Oper. Manage. Sympo. (APNOMS)*, 2020, pp. 55–60.
[173] D. Wang, H. Qin, B. Song, K. Xu, X. Du, and M. Guizani, "Joint resource allocation and power control for D2D communication with deep reinforcement learning in MCC," *Phys. Commun.*, vol. 45, Apr. 2021, Art. no. 101262.
[174] H. Ding, F. Zhao, J. Tian, D. Li, and H. Zhang, "A deep reinforcement learning for user association and power control in heterogeneous networks," *Ad Hoc Netw.*, vol. 102, 2020, Art. no. 102069.
[175] Y. Zhang, C. Kang, Y. Teng, S. Li, W. Zheng, and J. Fang, "Deep reinforcement learning framework for joint resource allocation in heterogeneous networks," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC–Fall)*, 2019, pp. 1–6.
[176] J. Tan, L. Zhang, and Y.-C. Liang, "Deep reinforcement learning for channel selection and power control in D2D networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2019, pp. 1–6.
[177] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, "Deep reinforcement learning for joint channel selection and power control in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1363–1378, Feb. 2021.
[178] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.
[179] H. Ye and G. Y. Li, "Deep reinforcement learning for resource allocation in V2V communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2018, pp. 1–6.
[180] H. Ye and G. Li, "Deep reinforcement learning based distributed resource allocation for V2V broadcasting," in *Proc. 14th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, 2018, pp. 440–445.
[181] S. Yuan, Y. Zhang, W. Qie, T. Ma, and S. Li, "Deep reinforcement learning for resource allocation with network slicing in cognitive radio network," *Comput. Sci. Inf. Syst.*, vol. 18, no. 3, pp. 979–999, 2020.
[182] Y. Zhang, X. Wang, and Y. Xu, "Energy-efficient resource allocation in uplink NOMA systems with deep reinforcement learning," in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)* 2019, pp. 1–6.
[183] Y.-H. Xu, C.-C. Yang, M. Hua, and W. Zhou, "Deep deterministic policy gradient (DDPG)-based resource allocation scheme for NOMA vehicular communications," *IEEE Access*, vol. 8, pp. 18797–18807, 2020.
[184] A. Khalili, E. M. Monfared, S. Zargari, M. R. Javan, N. Mokari, and E. A. Jorswieck, "Resource management for transmit power minimization in UAV-assisted RIS HetNets supported by dual connectivity," *IEEE Trans. Wireless Commun.*, early access, Aug. 31, 2021, doi: 10.1109/TWC.2021.3107306.

[185] S. Shrivastava, B. Chen, C. Chen, H. Wang, and M. Dai, "Deep Q-network learning based downlink resource allocation for hybrid RF/VLC systems," *IEEE Access*, vol. 8, pp. 149412–149434, 2020.
[186] Q. Huang, X. Xie, and M. Cheriet, "Reinforcement learning-based hybrid spectrum resource allocation scheme for the high load of URLLC services," *EURASIP J. Wireless Commun. Netw.*, vol. 2020, no. 1, pp. 1–21, 2020.
[187] H. Yang *et al.*, "Intelligent reflecting surface assisted anti-jamming communications based on reinforcement learning," in *Proc. IEEE Global Commun. Conf.*, 2020, pp. 1–6.
[188] H. Yang *et al.*, "Intelligent reflecting surface assisted anti-jamming communications: A fast reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1963–1974, Mar. 2021.
[189] E. Wong, K. Leung, and T. Field, "State-space decomposition for reinforcement learning," Dept. Comput., Imperial College London, London, U.K., Rep., 2021.
[190] L. Zahedi, F. G. Mohammadi, S. Rezapour, M. W. Ohland, and M. H. Amini, "Search algorithms for automated hyper-parameter tuning," 2021, *arXiv:2104.14677*.
[191] N. V. Varghese and Q. H. Mahmoud, "A survey of multi-task deep reinforcement learning," *Electronics*, vol. 9, no. 9, p. 1363, 2020.
[192] K. Lei, Y. Liang, and W. Li, "Congestion control in SDN-based networks via multi-task deep reinforcement learning," *IEEE Netw.*, vol. 34, no. 4, pp. 28–34, Jul./Aug. 2020.
[193] H. Eghbal-Zadeh, F. Henkel, and G. Widmer, "Context-adaptive reinforcement learning using unsupervised learning of context variables," in *Proc. Workshop Pre-Registration Mach. Learn.*, 2021, pp. 236–254.
[194] T. T. Nguyen, N. D. Nguyen, P. Vamplew, S. Nahavandi, R. Dazeley, and C. P. Lim, "A multi-objective deep reinforcement learning framework," *Eng. Appl. Artif. Intell.*, vol. 96, Nov. 2020, Art. no. 103915.
[195] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[196] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, "Explainability in deep reinforcement learning," *Knowl. Based Syst.*, vol. 214, Feb. 2021, Art. no. 106685.
[197] Y. He, Y. Wang, C. Qiu, Q. Lin, J. Li, and Z. Ming, "Blockchain-based edge computing resource allocation in IoT: A deep reinforcement learning approach," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2226–2237, Feb. 2021.
[198] F. Guo, F. R. Yu, H. Zhang, H. Ji, M. Liu, and V. C. M. Leung, "Adaptive resource allocation in future wireless networks with blockchain and mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1689–1703, Mar. 2020.
[199] S. Hu, Y.-C. Liang, Z. Xiong, and D. Niyato, "Blockchain and artificial intelligence for dynamic resource sharing in 6G and beyond," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 145–151, Aug. 2021.
[200] L. Yang, M. Li, P. Si, R. Yang, E. Sun, and Y. Zhang, "Energy-efficient resource allocation for blockchain-enabled Industrial Internet of Things with deep reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2318–2329, Feb. 2020.
[201] O. A. Wahab, A. Mourad, H. Otrok, and T. Taleb, "Federated machine learning: Survey, multi-level classification, desirable criteria and future directions in communication and networking systems," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1342–1397, 2nd Quart., 2021.
[202] N. H. Tran, W. Bao, A. Zomaya, M. N. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 1387–1395.
[203] A. M. Albaseer, M. Abdallah, A. Al-Fuqaha, and A. Erbad, "Fine-grained data selection for improved energy efficiency of federated edge learning," *IEEE Trans. Netw. Sci. Eng.*, early access, Jul. 29, 2021, doi: 10.1109/TNSE.2021.3100805.
[204] S. Wang *et al.*, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.
[205] X. Wang, C. Wang, X. Li, V. C. Leung, and T. Taleb, "Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9441–9455, Oct. 2020.
[206] H. H. Zhuo, W. Feng, Q. Xu, Q. Yang, and Y. Lin, "Federated reinforcement learning," 2019, *arXiv:1901.08277*.
[207] B. Das and S. Roy, "Load balancing techniques for wireless mesh networks: A survey," in *Proc. Int. Symp. Comput. Bus. Intell.*, 2013, pp. 247–253.
[208] L. Zhu, W. Shen, S. Pan, R. Li, and Z. Li, "A dynamic load balancing method for spatial data network service," in *Proc. 5th Int. Conf. Wireless Commun. Netw. Mobile Comput.*, 2009, pp. 1–3.
[209] H. Desai and R. Oza, "A study of dynamic load balancing in grid environment," in *Proc. Int. Conf. Wireless Commun. Signal Process. Netw. (WiSPNET)*, 2016, pp. 128–132.
[210] F. Khayatian, Z. Nagy, and A. Bollinger, "Using generative adversarial networks to evaluate robustness of reinforcement learning agents against uncertainties," *Energy Build.*, vol. 251, Nov. 2021, Art. no. 111334.
[211] A. T. Z. Kasgari, W. Saad, M. Mozaffari, and H. V. Poor, "Experienced deep reinforcement learning with generative adversarial networks (GANs) for model-free ultra reliable low latency communication," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 884–899, Feb. 2021.
[212] R. Alghamdi *et al.*, "Intelligent surfaces for 6G wireless networks: A survey of optimization and performance analysis techniques," *IEEE Access*, vol. 8, pp. 202795–202818, 2020.
[213] C. Huang *et al.*, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.
[214] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.
[215] C. Pan *et al.*, "Reconfigurable intelligent surface for 6G and beyond: Motivations, principles, applications, and research directions," 2020, *arXiv:2011.04300*.
[216] A. Taha, Y. Zhang, F. B. Mismar, and A. Alkhateeb, "Deep reinforcement learning for intelligent reflecting surfaces: Towards standalone operation," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2020, pp. 1–5.
[217] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.
[218] Z. Yang, Y. Liu, Y. Chen, and J. T. Zhou, "Deep reinforcement learning for ris-aided non-orthogonal multiple access downlink networks," in *Proc. IEEE Global Commun. Conf.*, 2020, pp. 1–6.
[219] L. U. Khan, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Digital-twin-enabled 6G: Vision, architectural trends, and future directions," 2021, *arXiv:2102.12169*.
[220] W. Kritzinger, M. Karner, G. Traar, J. Henjes, and W. Sihn, "Digital twin in manufacturing: A categorical literature review and classification," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1016–1022, 2018.
[221] Y. Lu, S. Maharjan, and Y. Zhang, "Adaptive edge association for wireless digital twin networks in 6G," *IEEE Internet Things J.*, vol. 8, no. 22, pp. 16219–16230, Nov. 2021.
[222] W. Sun, N. Xu, L. Wang, H. Zhang, and Y. Zhang, "Dynamic digital twin and federated learning with incentives for air-ground networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 1, pp. 321–333, Jan./Feb. 2022.
[223] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, "Communication-efficient federated learning and permissioned blockchain for digital twin edge networks," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2276–2288, Feb. 2021.

ABDULMALIK ALWARAFY received the B.S. degree in electrical engineering with a minor in communication from IBB University, Yemen, and the M.Sc. degree in electrical engineering with a minor in communication from King Saud University, Saudi Arabia. He is currently pursuing the Ph.D. degree with the College of Science and Engineering, Hamad Bin Khalifa University, Qatar. His current research interests include radio resource allocation and management for mobile networks, and deep reinforcement learning techniques for 6G and beyond wireless communication networks.

MOHAMED ABDALLAH (Senior Member, IEEE) received the B.Sc. degree from Cairo University, Giza, Egypt, in 1996, and the M.Sc. and Ph.D. degrees from the University of Maryland at College Park, College Park, MD, USA, in 2001 and 2006, respectively.

From 2006 to 2016, he held academic and research positions with Cairo University and Texas A&M University at Qatar, Doha, Qatar. He is currently a Founding Faculty Member with the rank of Associate Professor with the College of Science and Engineering, Hamad Bin Khalifa University, Doha. He has published more than 150 journals and conferences and four book chapters, and co-invented four patents. His current research interests include wireless networks, wireless security, smart grids, optical wireless communication, and blockchain applications for emerging networks. He is a recipient of the Research Fellow Excellence Award at Texas A&M University at Qatar in 2016, the Best Paper Award in multiple IEEE conferences, including IEEE BlackSeaCom 2019 and the IEEE First Workshop on Smart Grid and Renewable Energy in 2015, and the Nortel Networks Industrial Fellowship for five consecutive years, 1999–2003. His professional activities include an Associate Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and the IEEE OPEN ACCESS JOURNAL OF COMMUNICATIONS, the Track Co-Chair of the IEEE VTC Fall 2019 Conference, the Technical Program Chair of the 10th International Conference on Cognitive Radio-Oriented Wireless Networks, and a technical program committee member of several major IEEE conferences.

BEKIR SAIT ÇIFTLER (Member, IEEE) received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 2011, the M.S. degree from the TOBB University of Economics and Technology, Ankara, in 2013, and the Ph.D. degree in electrical and computer engineering from Florida International University, Miami, FL, USA, in 2017. His current research interests include wireless localization and tracking for Internet of Things, software-defined radios, 5G and mmWave networks, and MIMO systems.

ALA AL-FUQAHA (Senior Member, IEEE) received the Ph.D. degree in computer engineering and networking from the University of Missouri–Kansas City, Kansas City, MO, USA, in 2004. He is currently a Professor with Hamad Bin Khalifa University. His research interests include the use of machine learning in general and deep learning in particular in support of the data-driven and self-driven management of large-scale deployments of the IoT and smart city infrastructure and services, wireless vehicular networks, cooperation and spectrum access etiquette in cognitive radio networks, and management and planning of software-defined networks. He is an ABET Program Evaluator. He has also served as the Chair, the Co-Chair, and a Technical Program Committee Member of multiple international conferences, including the IEEE VTC, IEEE Globecom, IEEE ICC, and IWCMC. He serves on editorial boards of multiple journals, including the IEEE COMMUNICATIONS LETTER and the *IEEE Network Magazine*.

MOUNIR HAMDI (Fellow, IEEE) received the B.S. degree (Hons.) in electrical engineering (computer engineering) from the University of Louisiana at Lafayette, Lafayette, LA, USA, in 1985, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, USA, in 1987 and 1991, respectively.

He was a Chair Professor and a Founding Member of the Hong Kong University of Science and Technology (HKUST), Hong Kong, where he was the Head of the Department of Computer Science and Engineering. From 1999 to 2000, he was a Visiting Professor with Stanford University, Stanford, CA, USA, and the Swiss Federal Institute of Technology, Zürich, Switzerland. He is currently the Founding Dean of the College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar. His area of research is in high-speed wired/wireless networking, in which he has published more than 360 publications, graduated more 50 M.S./Ph.D. students, and awarded numerous research grants. In addition, he has frequently consulted for companies and governmental organizations in the USA, Europe, and Asia. He received the Best 10 Lecturer Award and the Distinguished Engineering Teaching Appreciation Award from HKUST. He is frequently involved in higher education quality assurance activities as well as engineering programs accreditation all over the world. He is a Fellow of the IEEE for his contributions to design and analysis of high-speed packet switching, which is the highest research distinction bestowed by IEEE. He is also a frequent keynote speaker in international conferences and forums. He is/was on the editorial board of more than ten prestigious journals and magazines. He has chaired more than 20 international conferences and workshops. In addition to his commitment to research and academic/professional service, he is also a dedicated teacher and a quality assurance educator.