Reinforcement Learning for Intelligent Healthcare Systems
A Review of Challenges, Applications, and Open Research Issues

Abdellatif, Alaa Awad; Mhaisen, Naram; Mohamed, Amr; Erbad, Aiman; Guizani, Mohsen

DOI
10.1109/JIOT.2023.3288050

Publication date
2023

Document Version
Final published version

Published in
IEEE Internet of Things Journal

Citation (APA)
Abdellatif, A. A., Mhaisen, N., Mohamed, A., Erbad, A., & Guizani, M. (2023). Reinforcement Learning for Intelligent Healthcare Systems: A Review of Challenges, Applications, and Open Research Issues. *IEEE Internet of Things Journal*, *10*(24), 21982-22007. Article 3288050. https://doi.org/10.1109/JIOT.2023.3288050

Reinforcement Learning for Intelligent Healthcare Systems: A Review of Challenges, Applications, and Open Research Issues

Alaa Awad Abdellatif, *Member, IEEE*, Naram Mhaisen, Amr Mohamed, *Senior Member, IEEE*, Aiman Erbad, *Senior Member, IEEE*, and Mohsen Guizani, *Fellow, IEEE*

***Abstract*—The rise of chronic disease patients and the pandemic pose immediate threats to healthcare expenditure and mortality rates. This calls for transforming healthcare systems away from one-on-one patient treatment into intelligent health systems, leveraging the recent advances of Internet of Things and smart sensors. Meanwhile, reinforcement learning (RL) has witnessed an intrinsic breakthrough in solving a variety of complex problems for distinct applications and services. Thus, this article presents a comprehensive survey of the recent models and techniques of RL that have been developed/used for supporting Intelligent-healthcare (I-health) systems. It can guide the readers to deeply understand the state-of-the-art regarding the use of RL in the context of I-health. Specifically, we first present an overview of the I-health systems' challenges, architecture, and how RL can benefit these systems. We then review the background and mathematical modeling of different RL, deep RL (DRL), and multiagent RL models. We highlight important guidelines on how to select the appropriate RL model for a given problem, and provide quantitative comparisons, showing the results of deploying key RL models in two scenarios that can be followed in monitoring applications. After that, we conduct an in-depth literature review on RL's applications in I-health systems, covering edge intelligence, smart core network, and dynamic treatment regimes. Finally, we highlight emerging challenges and future research directions to enhance RL's success in I-health systems, which opens the door for exploring some interesting and unsolved problems.**

***Index Terms*—Deep learning, distributed machine learning, edge computing (EC), Internet of Things (IoT), remote monitoring.**

I. Introduction

RAPID evolution of artificial intelligence (AI), Internet of Mobile Things (IoMT), software-defined networks (SDNs), and big data is paving the way for the emergence of Intelligent-Health (I-health) systems. Indeed, Internet of Things (IoT) instilled an important dimension to I-health systems through providing ease of data collection from the patients and the environment. As healthcare is a top priority worldwide, more frameworks integrating these technologies are foreseen to be realized soon [1]. Leveraging ubiquitous sensing, heterogeneous 5G network, intelligent processing, and control systems, I-health systems can in real-time monitor people's daily life and provide intelligent healthcare services to both citizens and travelers without limiting their activities. I-health can offer various applications, including remote monitoring, pandemic management, home care, and remote surgery (see Fig. 1).

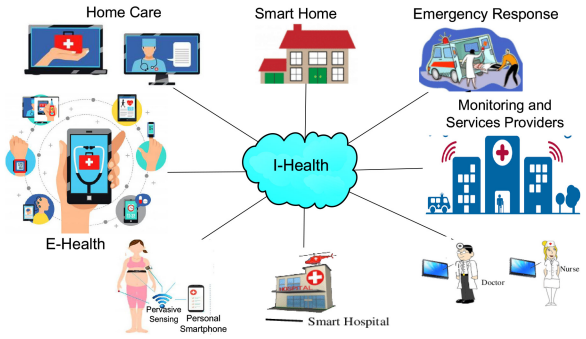


Fig. 1. I-health system components.

However, enabling high-quality healthcare services to the citizens imposes major challenges due to the rapid increase in the number of elderly and chronic disease patients. The average age of populations around the world is rising fast, so demands for healthcare are becoming even more significant. An aging population means an increase in the need for healthcare and treatment of chronic and age-related health issues, such as heart disease, cancer, chronic lower respiratory diseases, and diabetes. All these health issues related to a booming population segment will invariably place a high demand on the world's healthcare systems that offer a good quality of services, including the low cost, risk, and high safety of patients.

Typically, I-health systems comprise a diverse number of IoMT devices that generate enormous amount of data to intensively monitor the patients' state and automate emergency and intervention measures. These data should be processed, stored,

Manuscript received 25 March 2023; accepted 14 June 2023. Date of publication 26 June 2023; date of current version 7 December 2023. This work was supported by NPRP from the Qatar National Research Fund (a member of Qatar Foundation) under Grant NPRP13S-0205-200265. The work of Amr Mohamed and Aiman Erbad was supported in part by NPRP under Grant NPRP12S-0305-190231. The findings achieved herein are solely the responsibility of the authors. *(Corresponding author: Mohsen Guizani.)*

Alaa Awad Abdellatif and Amr Mohamed are with the College of Engineering, Qatar University, Doha, Qatar.

Naram Mhaisen is with the College of Electrical Engineering, Mathematics, and Computer Science, TU Delft, 2600 AA Delft, The Netherlands.

Aiman Erbad is with the College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar.

Mohsen Guizani is with the Machine Learning Department, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE (e-mail: mguizani@ieee.org).

Digital Object Identifier 10.1109/JIOT.2023.3288050

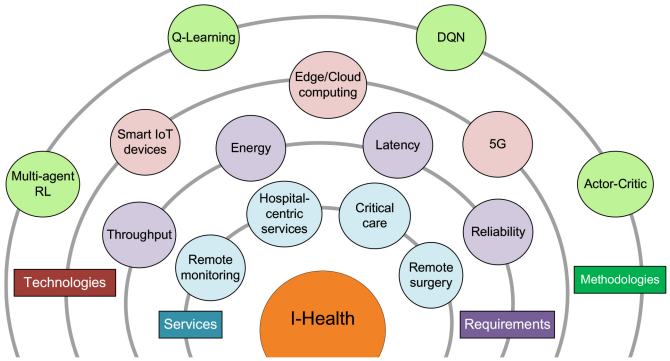


Fig. 2. I-health system overview. The layers, from the center out, represent the services that can be supported, requirements of these services, technologies that enable the implementation of these services, and various RL methodologies that allow I-health system to satisfy these requirements.

and accessed anytime anywhere, which increases the importance of cloud computing and edge computing (EC) to reduce response time and lower bandwidth cost [2], [3]. However, this imposes a significant load on the system design to handle millions of devices and process such massive amounts of data. Hence, several techniques have been proposed to analyze and manage these data efficiently, such as reinforcement learning (RL) techniques, which have been proposed as a promising approach for building such complex I-health systems.

As a subfield in AI, RL has emerged to study the optimal sequential decision-making process for an agent in nondeterministic environments. RL has obtained major theoretical and technical evolution in recent years, given its rising applicability to real-life problems. Indeed, different RL models aim at obtaining the optimal policies leveraging prior experiences obtained from interaction with a (learned) model (i.e., guided trial and error). This turns RL to be more appealing than diverse control-based schemes for healthcare systems, since it is typically hard to define a precise model that simulates the complex human body interactions to different administered treatments, due to nonlinear, varying, and delayed interaction between the treatments and human bodies. Fig. 2 summarizes the different tiers, diverse services, technical requirements, technologies, and RL methodologies that will be discussed in this survey. We envision that supporting different types of services and applications in I-health systems requires efficient integration between different technologies. For instance, to support remote monitoring applications, EC is a key component to realize ultralow latency, while cloud computing (or smart core network) enables the optimization of resources usage and interactions across distributed nodes. At the heart of this architecture, RL can play a crucial role to optimize the I-health system performance, decision making, and data flows throughout the overall system.

Recently, RL techniques are gaining much interest in healthcare systems, as they exhibit fast processing capabilities with real-time predictions. An RL agent learns optimal policies by interacting with the environment, taking actions based on the current state and receiving immediate rewards. This iterative process enables the agent to learn and improve its actions to maximize long-term rewards, leading to near-optimal performance. Conventional RL schemes utilized a lookup table to store the obtained rewards [4], however, this was sufficient for straightforward learning models with a small state space [5]. With increasing the action and state spaces (in complex learning problems), it would be hard to obtain the optimal policy in a reasonable time leveraging such lookup tables. Thus, a combination of RL with deep learning has been proposed to cope up with these limitations. In this context, deep neural network (DNN) [6] has been integrated with RL models to generate new models, namely, deep RL (DRL), that aims at improving the conventional RL models' performance, while handling complex control problems [7]. These DRL models include different variants, such as deep policy gradient RL [8], deep Q-networks (DQNs) [9], distributed proximal policy optimization (DPPO) [10], and asynchronous advantage actor–critic [11].

This survey reviews different types of RL and DRL techniques and their applications in I-health systems. To the best of our knowledge, the existing surveys mainly focus on the applications of DRL in IoT systems. Also, most of the presented works are restricted to model-free single-agent DRL methods. Additionally, the concept of I-health as a future healthcare system is relatively new and not adequately tackled in the existing literature. Thus, this survey will focus on other less explored RL techniques, in addition to the model-free DRL, and their applications in I-health systems, while highlighting I-health system's challenges, architecture, and future research directions.

A. Paper Organization

The remainder of this article is organized as follows. We first start in Section II by discussing the related surveys that tackled RL solutions, while clearly positioning our work among others in the literature. Then, Section III presents the architecture, challenges, and advantages of the considered I-health system. Section IV discusses the fundamentals and background of different RL, DRL, and multiagent RL models, while highlighting some quantitative results and tips about how to select the appropriate RL model based on the considered problem. Section V reviews the state-of-the-art approaches that adopt different RL models for optimizing the performance of healthcare systems. In particular, we categorized the presented RL applications for healthcare systems under three main areas: 1) edge intelligence; 2) smart core network; and 3) dynamic treatment regimes. Section VI presents our vision for the open challenges and future research directions that worth further investigation in the future. Finally, Section VII concludes our paper. To sum up, Fig. 3 illustrates how this survey is organized.

II. Related Work

In this section, we review the main related surveys that discussed different AI schemes, including RL, in IoT and I-health systems. Then, we highlight our contributions with respect to these related surveys. Table I recaps the important surveys that targeted RL applications, challenges, and related issues.

RL schemes have been widely used in different domains and applications, such as wireless network management, finance, healthcare, and autonomous vehicles. However, their

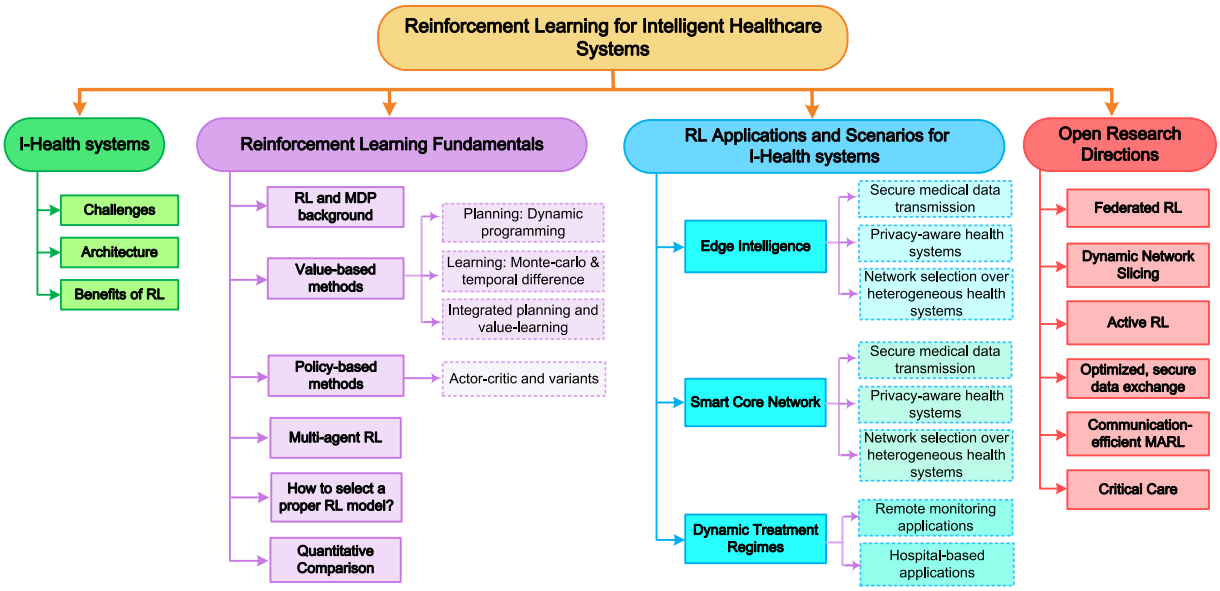


Fig. 3. Taxonomy of RL for I-health systems: approaches, applications, challenges, and open research issues.

TABLE I
TAXONOMY OF THE RELATED SURVEYS BASED ON THE DISCUSSED AI SCHEMES

Ref	Objective	SL	DL	RL	DRL	I-health
[12]	It focuses on the RL interpretation methods that are used to analyze the internal design of different RL models.			✓	✓	
[13]	It reviews the state-of-the-art of single and multi-agent RL shemes with an emphasis on deep MARL for AI-enabled wireless networks.			✓	✓	
[14]	It review the state-of-the-art of DRL algorithms for IoT applications.				✓	
[15]	It reviews the state-of-the-art of DRL schemes for AIoT systems.				✓	
[16]	It presents a comprehensive survey on the application of RL for IoT security.			✓	✓	
[21]	It focuses on the applications of DRL for communications and networking problems, such as network access and rate control, caching and offloading, security and connectivity preservation, as well as traffic routing and resource scheduling.				✓	
[22]	It presents a survey on diverse ML algorithms used for supporting intelligent end-to-end communication services, without focusing on I-health systems.	✓	✓	✓	✓	
[23]	It presents a survey on the applications of RL in autonomous driving.			✓	✓	
[24]	This survey focuses on a general description for IoT system, and wireless sensor networks. It reviews the learning and big data studies related to IoT.	✓				
[25]	It provides an overview on smart transportation systems including ML techniques and IoT applications in Intelligent Transportation Systems (ITS).	✓				
[26]	This survey focuses on the IoT for cloud/fog/edge computing and the use of ML in MEC systems.	✓				
[27]	It surveys the existing techniques of deep learning in the fields of translational bioinformatics, medical imaging, pervasive sensing, medical informatics and public health.	✓	✓			
[28]	It surveys the existing DRL techniques developed for cyber security and various vital aspects, including DRL-based security models for cyber-physical systems.				✓	
[29]	It reviews the AI applications for stroke, specifically it focuses on three major areas, i.e., early detection and diagnosis, dynamic treatment, as well as the outcome prediction and prognosis evaluation.	✓				
[30]	It reviews the related work for the implementation of AI into existing clinical workflows, including data sharing and privacy, transparency of algorithms, data standardization, and interoperability across multiple platforms, while considering the major concerns for patient safety.	✓				
[31]	It presents a general overview for the neural networks and deep neural networks challenges, training algorithms, architecture, and implementations, without focusing on I-health systems.	✓		✓	✓	
[32]	It investigates some of the RL and DRL solutions that targeted WBAN.			✓	✓	
[33]	It surveys the adoption of RL, DRL, and inverse IRL in dynamic treatment regime and some healthcare applications.			✓	✓	✓
This survey	It presents a comprehensive survey about different RL and DRL applications in different system levels of I-health systems.			✓	✓	✓

black-box nature renders their analysis and implementation complex, especially in critical systems such as I-health systems. Thus, Alharin et al. [12] reviewed the main RL interpretation methods that are used to understand the internal design of different RL models. Moreover, several papers surveyed the potentials of DRL in different domains,

such as autonomous IoT-based systems, communications and networking, 6G networks, IoT Security, and autonomous driving [13], [14], [15], [16], [17], [18], [19], [20]. In these tutorials, the fundamentals and applications of different RL schemes have been investigated for various aforementioned systems, including rewards maximization, policy convergence, multiagents connection, and performance optimization. However, the challenges, solutions, and open research issues of RL in I-health systems have not been reviewed before. For instance, Feriani and Hossain [13] presented a survey on the fundamentals and potentials of single and multiagent DRL frameworks in future 6G networks. Advantages and challenges of using DRL algorithms for IoT applications, such as smart grid and intelligent transportation systems, are discussed in [14]. Uprety and Rawat [16] reviewed various types of cyber-attacks in IoT systems, while summarizing diverse RL-based security solutions presented in the literature for securing IoT devices.

Applications of DRL in communications and networking aspects are reviewed in [21] and [22]. With the rapid growth of pervasive IoT applications, different nodes and devices would have to make decisions locally to support decentralized and autonomous network functions, while maximizing the network performance. Given the high dynamics and uncertainty of diverse environments, RL has shown its efficiency in obtaining the optimal policies for different network agents (i.e., nodes) while dealing with highly dynamics networks. In [21], a tutorial of DRL concepts and models is presented, while focusing on DRL schemes proposed to tackle the emerging problems in communications and networking, such as dynamic network association, throughput maximization, wireless caching, data/task offloading, network security, data aggregation and dynamic resource sharing, as well as connectivity preservation. Tang et al. [22] reviewed the recent ML-based network optimization methods, implemented from the data-link layer to the application layer, to deal with the high complexity and dynamic environments in 6G networks. In [23], a review for the potentials, challenges, and advantages of employing DRL in real-world autonomous driving systems is presented.

Although there are other surveys related to RL and DRL, most of them ignore healthcare system architecture, applications' characteristics, and related network challenges. For instance, [34] and [35] review the usage of DRL algorithms for computer vision and natural language processing. The survey in [36] focuses on the applications of deep learning schemes in wireless networks, while the survey in [37] focusing on the problems of continuous control and large discrete action space. To the best of our knowledge, only few surveys that partially address some applications of RL in healthcare systems, i.e., [32], [33], and [38]. The survey in [32] discusses the potential of RL-based solutions in wireless body area network (WBAN). In particular, this survey focuses on the energy management issues in WBAN, which include internetwork interference management over the multiagent environment, energy consumption minimization, power control for in-body sensors, tradeoff between energy efficiency and transmission delay, as well as power control to mitigate jamming. Then, it discusses the main solutions that address these issues, such as dynamic power control, sensor access control, and energy harvesting. The survey in [38] discusses some of the RL models that are used in dynamic treatment regimes, as well as chronic and critical diseases treatments. Coronato et al. [33] focused on RL applications in healthcare systems, such as precision medicine, dynamic treatment regime, personalized rehabilitation, and medical imaging, while offering recommendations on selecting the most proper RL approach to apply. The previous two surveys have concentrated on the utilization of RL in healthcare applications, however, they ignored the proposed RL initiatives for edge intelligence, smart core network, as well as networking and communication aspects of healthcare systems. This motivates us to develop a survey that presents a comprehensive literature review on the architecture, requirements, applications, and challenges of I-health systems, while focusing on different RL models to tackle these challenges and requirements in different system levels (i.e., end-user, edge, and core network).

A. Our Contribution

To the best of our knowledge, this is the first survey that addresses the potentials of RL in all layers of healthcare systems. Specifically, the main contributions of this survey lie in the following aspects.

1) We first review the major challenges of I-health systems and propose a generic I-health system architecture that integrates diverse components of any I-health system. Then, we discuss why RL is needed to cope with the increasing demand of I-health systems by addressing these challenges.
2) We present a comprehensive tutorial on single-agent and multiagent DRL frameworks. In particular, we briefly explain the main concepts of diverse RL schemes and their fundamental building blocks, including the value-based and policy gradient models, while discussing the pros and cons of each category and how to select the appropriate RL model to use. Moreover, we empirically assess the performance of different value-based methods (i.e., DQN and actor–critic methods).
3) We focus on the potentials and applications of RL models in I-health systems, where the relevant studies are categorized into three main sectors, i.e., edge intelligence, smart core network, and dynamic treatment regimes. In each sector, we discuss different RL models that are used to fulfil various smart healthcare services' requirements, following the I-health system architecture. The proposed I-health architecture not only helps in building a taxonomy to recap and classify existing studies, but also provides a general framework to investigate the potentials of different AI techniques in healthcare systems.
4) Finally, we discuss several future research directions related to the design of efficient, scalable, and distributed RL schemes in I-health systems to enhance the available healthcare services while enabling new, smart services for future healthcare systems.

III. I-Health Systems: Challenges, Architecture, and Benefits of RL

In this section, we first present the main challenges for implementing I-health systems. Then, the proposed I-health system architecture is introduced to address these challenges. After that, we highlight the major benefits that can be obtained by incorporating RL schemes within the proposed I-health architecture.

A. Challenges of I-Health Systems

Despite the promising evolution toward enabling remote health systems, several challenges still have to be addressed toward providing intelligent healthcare services.

Highly Dynamic Environment: The evolution of e-health allows for collecting, processing, and analyzing piles of information from several devices/locations (e.g., hospitals, clinics, etc.) in order to provide efficient healthcare services. The advances of edge nodes (e.g., smartphones), equipped with built-in sensors, cameras, and high-performance computing and storage resources enable each patient/user to generate massive amount of data anytime and anywhere. The patients can generate, collect, and communicate irregularly large volumes of real-time data about its own operation and surrounding environment. Thus, given the anomaly in generating and collecting the medical data in addition to the networks' dynamics, I-health systems turn to be heterogeneous and highly dynamic environment. Moreover, such dynamics are too complex to be captured with either simple stochastic assumptions or static conventional optimization. Such highly, nonstationary environment typically calls for adopting experience-based learning.

Large Number of Potential Users: Given the rapid increase in the number of the chronically ill and elderly people, most of the healthcare facilities are required to serve thousands of patients daily [39]. This number can be easily duplicated in case of pandemic, such as the recent COVID-19 outbreak [40], which puts a serious load on different healthcare facilities. Given such enormous number of patients/users in the I-health system, providing accurate, sequential personalized clinical decisions is a very challenging task.

Distributed and Imbalanced Data: The heterogeneity of the I-health system, and the distribution and size of the collected data from different patients significantly vary. The locally generated data in I-health systems is distributed across multiple devices/nodes in the network, e.g., smartphones, sensors, and hospitals, which results in imbalanced data distribution. Most of the ML algorithms may suffer from biased and inaccurate prediction due to the imbalanced data. Thus, heterogeneous and imbalanced data along with the high-dynamic environment put major challenges in designing an efficient I-health system that supports reliable and real-time healthcare services.

Limited Computational and Communication Resources: The large number of patients/users participating in the I-health system, the availability of the needed computational and communication resources at different network levels can be challenging. The generated traffic from different locations of the I-health system grows linearly with the number of participating users. Furthermore, the heterogeneity of the edge nodes or patients' equipment, in terms of computational capabilities and energy availability (i.e., battery level) introduces an extra constraint to optimize the performance in I-health system.

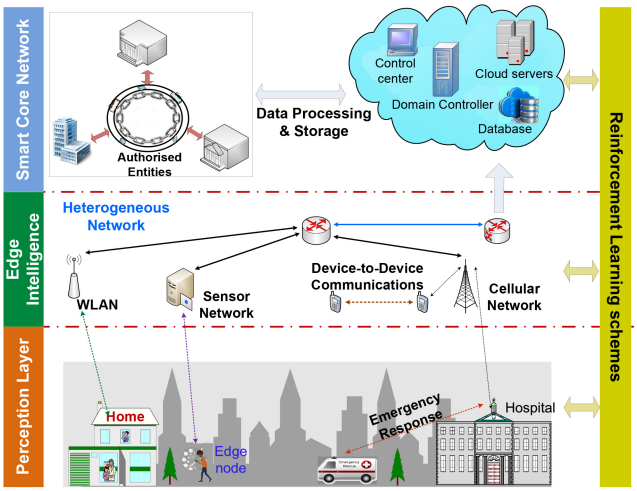


Fig. 4. Proposed I-health system architecture.

B. I-Health System Architecture

To tackle the above challenges, we propose the I-health architecture in Fig. 4. The proposed architecture is comprised of three main layers: 1) perception layer; 2) edge intelligence layer; and 3) smart core network layer. This architecture stretches from the physical layer, where data is generated/collected, to the control and management layer. It includes the following main components.

Perception Layer: A combination of sensing and IoMT devices represents the set of data sources, which provide real-time monitoring and ubiquitous sensing for diverse E-health applications.

Edge Intelligence Layer: This layer focuses on processing the acquired data from different sources (e.g., transferring it from raw sensory data into actionable insights), in addition to the association with the heterogeneous network (HetNet) infrastructure. Thus, it represents the intermediate processing, communication, and storage stage between heterogeneous data sources and core network. In particular, an intelligent edge node can gather the medical and nonmedical data from different sources, to analyze, classify, and extract information of interest, then it forwards the processed data or extracted information to the core network through HetNet [41]. The latter incorporates cellular networks, wireless local area networks (WLANs), device-to-device (D2D) communications, and sensor networks. Such a HetNet enables seamless switching among different types of technologies in order to optimize medical data delivery. Interestingly, different health-related applications (apps) can be also implemented in such intelligent edge nodes. These apps can play a crucial role in long-term chronic diseases management, while assisting patients to actively involve in their treatment regime by ubiquitously interacting with their doctors anytime and anywhere.

Smart Core Network: It takes comparative advantages of powerful computing sources, i.e., from the edge to the cloud,

to process, analyze, and store the collected data. Moreover, authorized entities, such as the government and big organizations, can have certain privilege and authorization at this layer to access the collected information and define the requirements or policies for decision making and control. The proposed architecture enables the two-way communication, i.e., sensing and control. This two-way communication enables acquiring the knowledge about the physical world, while monitoring and managing every device/component in the system to make it operate properly and smartly. Hence, leveraging the acquired data from the physical layer, different RL schemes can be implemented at each layer of the proposed architecture to facilitate various decision-making processes.

C. Lessons Learned: Why RL Is Needed for I-Health Systems?

The gathered data at an I-health system can be processed at different system levels (i.e., edge devices, fog nodes, and the cloud) from perception layer to core network layer. Thus, different RL schemes can be represented by a transversal layer, i.e., a vertical layer crossing over all layers (see Fig. 4), where they can be implemented to efficiently process the gathered data and configure optimally the different parameters at each layer. Thus, the main motivations of applying RL in I-health systems can be summarized as follows.

1) *Optimizing Data Processing and Transmission in I-Health Systems:* Leveraging RL allows for optimizing the integration of distributed data processing and communication in the resource-constrained environment, such as I-health systems. The ability of RL to derive optimal decision-making strategies from observational data without having a model for the underlying process makes it a promising approach in optimizing I-health systems' behavior.
2) *Defining Optimal Dynamic Treatment Regimes:* One of the main targets in I-health system is to provide effective treatment regimes that: a) dynamically adapt to the varying patients' state; b) enable personalized medication through responding to diverse patients' responses to various treatment regimes; and c) enhance the long-term benefits of patients. RL framework perfectly fits in designing optimal dynamic treatment regimes that can be viewed as sequential decision-making problems [42]. Each agent in the RL framework aims to learn the optimal policy from the interactions with the environment in order to maximize its reward. This is precisely mapping to what happens in dynamic treatment regimes, where the set of rules followed in dynamic treatment regimes are equivalent to the policies in RL, while the treatment outcomes are equivalent to the rewards in RL. However, the learned treatment policies through RL, using observational data, may be vulnerable to slight variations in the design of the learning model, which calls for integrating clinicians' input into the RL learning process in order to avoid different sources of bias and safety issues of RL [43].
3) *Supporting Automated Medical Diagnosis Processes:* Driven by the recent advances in big data analysis and ML schemes, there is much interest to promote the diagnostic process toward enabling automated medical diagnosis [44], [45]. In order to provide error-resistant process in diagnosis while assisting the clinicians for an effective decision making, several studies work on formulating the diagnostic classification problem as a sequential decision-making process. This allows for leveraging the RL in solving such dynamic problem with a small amount of labeled data generated from relevant resources [46].

IV. Reinforcement Learning: Overview and Theoretical Background

This section presents a brief overview to the theoretical background, fundamental concepts, as well as different models and advanced techniques in RL. Section IV-A first describes the general model of Markov decision processes (MDPs) and how RL can be used to deduced optimal polices for that model. Then, Section IV-B discusses the main value-based methods that are used to solve the MDP problem, while Section IV-C investigates different policy gradients-based methods that search directly in the space of policies to find the optimal policy. Section IV-D discusses the motivation of leveraging multiagent reinforcement learning (MARL) framework to model and solve diverse control problems, while highlighting the main proposed algorithms in this area. In Section IV-E, we propose a generic way to identify how to select the appropriate RL model to use based on the targeted application's requirements. Finally, we present some quantitative comparisons in Section IV-F in order to empirically depict the difference between various techniques of RL.

A. RL and MDP Formulation: Overview

In RL, the problem of interest is often framed as an MDP. An MDP provides a formal framework for modeling decision-making problems where an agent interacts with an environment over time. RL algorithms are designed to solve MDPs and optimize the agent's decision-making policy [47]. Indeed, if a problem is well described as a standard MDP with stationary transition probabilities amongst the environment states [48], then RL will probably be effective to find an efficient solution for such a problem, not only a good solution, but also a long-term policy that maintains maximal gains over time. However, if the problem cannot be mapped onto an MDP, then the theory behind RL cannot provide any guarantees of useful results. This is because RL assumes that the problem can be represented as an MDP, and the guarantees of convergence and optimality are based on this assumption. Thus, it is not always necessary or optimal to use RL for every problem. In some cases, conventional greedy techniques, such as rule-based systems or simple heuristics can perform as well as, or even better than, RL methods [48]. For instance, in the domain of autonomous helicopter control, Bagnell and Schneider [49] found that a hand-coded greedy controller performed better than an RL-based controller. Similarly, Konidaris and Barto [50] showed that skill chaining technique can be used to solve complex continuous

control problems more efficiently than RL methods that rely on trial-and-error learning.

The MDP is the primary choice of modeling the problems that have a temporal structure, such as monitoring and control, which frequently arises in health monitoring applications. In these problems, the aim is to deduce a policy to be executed in real time (at every time step). This is in contrast to optimization formulation, whose aim is to achieve an optimal design point [51]. Thus, MDP formulation should be considered whenever there is a decision-making problem under uncertainty. Indeed, MDP and their extensions provide a general framework of acting optimally under uncertainty. On the contrary, if the state of the system to be optimized can be captured as a vector which is constant or varies rarely, then regular optimization formulations are preferred [52]. The MDP formulation is an abstraction of the problem of data-driven control from interaction. It models any learning problem with three signals passing back and forth between an agent and its environment: 1) the first signal refers to the choices made by the agent (the actions); 2) the second signal represents the basis on which the choices are made (the states); and 3) the third one defines the agent's reward. Despite being this framework not sufficient to model all decision-learning problems adequately, it has proved to be widely valuable and applicable.

An MDP is defined as the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p, \gamma, \rangle$. At every time step t, the agent receives a representation of the environment state $S_t \in \mathcal{S}$. The agent executes an action $A_t \in \mathcal{A}$ using a policy $\pi(a|s)$ in order to receive a reward $R_t \in \mathcal{R}$. Then, it moves forward to the next state S_{t+1}. p is referred to as the dynamics of the MDP. It is a probability distribution $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{R} \to [0, 1]$ defined as $p(s', r|s, a) \doteq \Pr S_t = s', R_t = r|S_{t=1} = s, A_{t-1} = a$. Note that useful information can be extracted from p. Those are the state transitions $\Pr S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a = \sum_{r\in\mathcal{R}} p(s', r|s, a)$, as well as the reward value, i.e.,

$$\mathbb{E}R_t|S_{t-1} = s, A_{t-1} = a = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p\big(s', r|s, a\big). \tag{1}$$

We refer to those terminologies repeatedly throughout this article.

The total feature discounted sum of rewards until some horizon H is denoted as $R_t = \sum_{t'=t}^{H} \gamma^{t'-t} r_t$, with the discount factor $\gamma \in [0, 1]$. The state-action value function of a specific policy π is defined as follows:

$$Q^{\pi}(s, a) = \mathbb{E}_{a\sim\pi, s'\sim\mathcal{T}}[R_t|s_t = s, a_t = a] \tag{2}$$

which summarizes the sum of the rewards resulting from taking the action a in state s by following the policy π. The state value function $V^{\pi}(s) = \mathbb{E}_{a\sim\pi}[Q^{\pi}(s, a)]$ assesses the quality of the state when following the policy π.

In what follows, we explore general approaches to solve a given MDP. Such solutions aim to find an optimal policy that results in the maximum reward.

B. Value-Based Methods

The main solution of MDPs is derived from the Bellman optimality equations, which is based on dynamic programming (DP) and describes the optimal action-value of a state-action pair in terms of the next ones (recursive definition)

$$q_*(s, a) = \sum_{s',r} p\big(s', r|s, a\big)\big[r + \gamma \max_{a'} q_*\big(s', a'\big)\big]. \tag{3}$$

The equation states that the optimal action-value function for a state-action pair is expressed as the expected value of the reward from taking action a in state s, and then taking the best possible action in the next state. This two-stage view summarizes the future expected rewards through $q_*(s', a')$, and enables the optimal policy to be easily recovered by acting greedily with respect to it

$$\pi_* = \arg\max_a q_*(s, a). \tag{4}$$

There are two main specific instances of value-based methods that can be utilized based on the knowledge of the environment's model. In the following section, we explore these two instances and discuss the model knowledge.

1) Planning (Dynamic Programming): To solve the MDP framework, exact DP is used when the environment model is known. The model is a concept that helps the agent predict the environment's transition and reward. The state transition function is used to describe this. If the transition model is known, generalized policy iteration (GPI) algorithms can be used to find an optimal policy [48] (as in Algorithm 1).

Algorithm 1: GPI

```
V(s) ∈ ℝ and π(s) ∈ A(s) arbitrarily;
while π-stable is false do
    while Δ is above an accuracy threshold do
        v ← V(s);
        V(s) ← Σ_{s',r} p(s', r|s, π(s))[r + γV(s')];
        Δ ← max(Δ, |v − V(s)|);
    end
    π-stable ← true;
    for each state do
        a_old ← π(s);
        π(s) ← arg max_a Σ_{s',r} p(s', r|s, π(s))[r + γV(s')];
    end
    if a_old ≠ π(s) then
        π stable ← false;
    end
    return π ≈ π_*
end
```

The GPI algorithm mainly consists of two alternating steps, a *policy evaluation* step and *policy improvement* step. The evaluation aims to calculate the value function (or the action-value function) of the policy being followed by the agent. In contrast, the improvement step modifies the policy in a way that is guaranteed to increase the value function. The evaluation step might occur across multiple environment's interaction steps before an improvement step is performed. If the evaluation is done at every interaction step, GPI reduces to value iteration (VI) algorithm; if the evaluation is allowed to continue until convergence, GPI reduces to regular policy iteration (PI). In practice, VI is more widely used as it allows faster improvement of the policy.

2) Learning (Monte Carlo and Temporal Difference): RL can estimate the state transition model through samples, which abstracts the need to identify and design models that precisely describe the environment dynamics. This feature also allows the learning-based approach to adapt to changes or inaccuracies in the model since the approximation from samples is a continuous process. Hence, learning the state transition model through interaction with the environment has two main benefits: 1) it eliminates the need for designing precise environment dynamics models and 2) enables adaptation to inaccuracies or changes in the model. These benefits make RL a popular optimization technique, although other approaches, such as online convex optimization also offer learning and adaptability features [53]. However, speed of adaptability is still a current research focus.

Learning through samples can be done in two main ways: 1) the conventional Monte Carlo (MC) style or 2) the temporal difference (TD) methods. An MC-based method simply generates a lot of experience tuples and utilizes them to estimate the values of a state. In RL, the experience tuple refers to a state, action, next state, and reward tuple (s, a, s', r). The MC-based method averages the subsequent rewards from a given state across multiple episodes to get its value $v^{\pi}(s)$ [the same principle applies to state-action value function estimation $(q^{\pi}(s))$]. MC is known to be an un-biased estimator but can be slow [48]. On the other hand, TD methods directly implement the Bellman equations using raw experience, similar to MC methods. However, they update estimates using other learned estimates through bootstrapping, resulting in faster approximation but biased estimators. In other words, the TD methods can be realized through replacing the exact value functions in the bellman equations with samples. Bootstrapping-based methods are much faster in approximating the values. However, they are biased estimators. Nonetheless, there are several bounds on the bias of that estimator, and in practice, it usually delivers very good results and is widely adopted [54]. The best representative of TD methods, and the most adopted RL approach is referred to as Q-learning.

Q-learning [55] is a model-free algorithm that is based on the iterative execution of the Bellman equation. Specifically, the Q-learning algorithm can be realized by transferring the Bellman equation into an assignment

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\Big[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\Big]. \quad (5)$$

Extensive theoretical analysis that guarantees the convergence can be found in [56]. The iterative application of the Bellman equation will eventually converge to the optimal action-value function Q^*. Once this optimal function is generated, the optimal function can be recovered by acting greedily with respect to this function.

Neural networks (NNs) are commonly used as function approximators to evaluate the Q-function in the deep Q-learning algorithm (as in Algorithm 2). However, the introduction of function approximation poses theoretical challenges as convergence is no longer guaranteed, despite its apparent ease of implementation. Nonetheless, various techniques have been developed to stabilize learning. Those include maintaining a replay buffer, two versions of the Q-network (Q-function), an online and a previous copy, which are used as per the algorithm.

Algorithm 2: Q-Learning With Function Approximation

input : Environment simulator
output: θ, i.e., the NN parameters for the approximation Q_*.
Initialize parameters of the first network θ randomly
Initialize parameters of second (target) network $\bar{\theta} \leftarrow \theta$
for *episodes = 1:M* **do**
  Initialize state s_0
  for *time step t= 1:I* **do**
    /**Interaction with the environment**/ Select state
    update action a_t based on ϵ-greedy policy
    Execute a_t, observe s_{t+1} and r_{t+1}
    Store the tuple of experience $(s_t, a_t, s_{t+1}, r_{t+1})$ in $\mathcal{D}$
    /**Updating the estimates**/
    Randomly sample a minibatch $\mathcal{F} = \{(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)}, r_{t+1}^{(i)})\}_{i=1}^{|\mathcal{F}|}$ from $\mathcal{D}$
    Calculate Q-targets using the target network $Y^{(i)} \leftarrow r_{t+1}^{(i)} + \gamma \max_a Q(s_{t+1}^{(i)}, a; \bar{\theta})$
    Fit $Q(s^{(i)}, a^{(i)}; \theta)$ to the target $Y^{(i)}$:$\theta \leftarrow \theta - \eta \nabla_\theta L(\theta)$
    Every *target* steps, update the target network $\bar{\theta} \leftarrow \tau\theta + (1-\tau)\bar{\theta}$
  end
end

3) Integrated Planning and Value Learning: Some RL methods can learn both the model of the environment and the value function through updates from both real interaction tuples and simulated ones. The simulated ones are taken from a model that the agent learns (i.e., an approximation to the actual environment model). Indeed, model-based RL can be useful to incorporate prior knowledge or when real interaction is expensive, but in most cases where a simulator is available, model-free learning can be sufficient.

C. Policy-Based Methods

As the original goal of RL is to learn the optimal policy, a different approach is to search directly in the space of policies, rather than first to find the value function and then recover a policy out of it. This class of methods is known as the policy-gradients or policy-based methods. The policy is parametrized by a parameter vector, which is optimized through the known stochastic gradient descent (SGD) methods. In most practical instances, the direct application of the policy gradient does not yield high performance. This is due to the nature of SGD-based algorithms of being of high variance and prone to being trapped in local minima. To minimize the variance's effect, a policy gradient algorithm with a fixed baseline is proposed.

The optimization in policy gradient methods is done on the cost function: $J(\theta) \doteq v_{\pi_\theta}(s_0)$, which is the cost of starting from the initial state s_0, and following the parametrized policy π_θ. From the policy gradient algorithm [48], the gradient of this function can be written as follows:

$$\nabla J = (G_t - b(S_t)) \frac{\nabla \pi(A_t|S_t, \theta_t)}{\pi(A_t|S_t, \theta_t)} \quad (6)$$

where b is any function of the state, namely, a baseline. Its main purpose is to reduce the variance in estimating the value of the state. G_t is the sum of future discounted reward resulting from π. If it is the zero function, the equation reduces to the "reinforce" equation. However, most of the works that utilize policy gradients combine it with a value-based method. This combination is known as actor–critic and is explained in the next section. Another popular option is the value function of the state. If this state value function is updated through bootstrapping. The resulting method is called actor–critic.

In general, the main benefit of policy search methods is that they offer practical means of handling very large action spaces, and even continuous spaces (infinite number of actions). This is possible because we are controlling the parameters of the policy, and not recovering it from a value function. Thus, we can choose the policy π to be a probability distribution over the actions and learn the statistics of this distribution (e.g., the policy is a mapping from state to a normal distribution with learned mean and variance). This is different from value-based methods that learn the action-value for each of so many actions. The drawback of these policy-based methods is that they require the generation of the full episode to get a single gradient step in the policy space, which may be slow (i.e., MC style gradient approximation), although similar to what the TD methods introduced.

1) Actor–Critic and Variants: Actor–critic methods are policy gradients that use the state value function as a baseline $[b(s) = V(s)]$ and update this function through bootstrapping[1] [57]. In the literature, we notice that in most instances where policy search methods are used, it is indeed an actor–critic variant. Actor critic enables the online time-step wise learning as opposed to the full episode needed in conventional policy gradient algorithms. A variant of the actor–critic algorithm is shown in Algorithm 3. The variant is a deep deterministic policy gradient (DDPG). The main difference compared to conventional actor critic algorithms is that it maps a state directly to an action, rather that to a distribution of actions, which allows for continuous action spaces. Similar to what was presented in the previous algorithm, we adapt neural-network-based function approximation to express the value function as well as the policy. Note that the $\mathcal{L}$ denotes the loss function (squared error between predicted state/action value and the actual one as retrieved from the previous replay buffer $\mathcal{D}$.

Algorithm 3: DDPG Template

Randomly initialize the parameters of the Q-network ϕ and policy network θ;
Initialize target networks parameters with $\bar{\phi} \leftarrow \phi$, $\bar{\theta} \leftarrow \theta$;
for *episodes* $i = 1{:}M$ **do**
  Initialize state s_0;
  for *time step* $t = 1{:}H$ **do**
    Select action $a_t = \mu_\theta(s) + e^{\frac{\varphi i}{M}}\varepsilon$, $\varepsilon \sim \mathcal{N}(0, 0.1)$;
    Execute action a_t and observe next state s_{t+1}, reward r_{t+1};
    Store (s, a, s_{t+1}, r_{t+1}) in replay buffer $\mathcal{D}$;
    Sample minibatch $\mathcal{F} = \{(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)}, r_{t+1}^{(i)})\}_{i=1}^{|\mathcal{F}|}$ from $\mathcal{D}$;
    Compute targets $Y^{(i)} \leftarrow r_{t+1}^{(i)} + \gamma Q\big(s_{t+1}^{(i)}, \mu(s'); \bar{\theta}\big)$;
    Compute loss function $\mathcal{L}$:
    $\mathcal{L}(\phi, \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s',d)\in\mathcal{B}} \Big(Q_\phi(s, a) - y(s', r, d)\Big)^2$;
    Update Q network parameters by gradient descent:$\phi \leftarrow \phi - \eta_\phi \nabla_\phi \mathcal{L}(\phi, \mathcal{D})$;
    Update policy network parameters by gradient ascent:
    $\theta \leftarrow \theta + \eta_\theta \nabla_\theta \frac{1}{|\mathcal{B}|} \sum_{s\in\mathcal{B}} Q_\phi\big(s, \mu_\theta(s)\big)$;
    Update target networks paremters
    $\phi_{targ} \leftarrow (1-\rho)\phi_{targ} + \rho\phi$
    $\theta_{targ} \leftarrow (1-\rho)\theta_{targ} + \rho\theta$;
  end
end

D. Multiagent Reinforcement Learning

In this section, we discuss the main reasons one might consider the MARL framework to model the system at hand and point readers to the prominent algorithms in this area. In monitoring and actuating applications (vital health signs and others), a central controller usually controls and coordinates between several modalities, each of which is an element in the controller's action vector. For example, consider the controller action $\mathbf{a} = a_1, a_2, \ldots, a_n$. While algorithms discussed in the previous section can theoretically learn the best control policy, it can be seen that the size of the action space $|\mathcal{A}|$ increases exponentially with each added modality, which deems exploration among these actions infeasible. Therefore, enhancing the scalability in such scenarios is the main motive behind utilizing MARL in monitoring (health) applications. Specifically, the MARL framework does not attempt to learn one joint policy by a central controller, but rather models each of the controlled modalities as an independent agent who optimizes its policy and explores *only* among its actions, while possibly coordinating with other agents in attempting to maximize the global reward.

Note that in some cases, the distribution is done by design (i.e., there exists no central controller due to security or communication concerns). Once agents are decentralized, there are several design options according to which MARL is categorized. If the reward signal is the same for all agents, the setting is known as cooperative MARL. If the rewards have to sum to a constant value, the setting is known as competitive; all other settings are referred to as mixed. We note that the study of mixed and especially competitive settings is mainly conducted in game-theoretic and economic scenarios [58]. In IoT applications, the cooperative scenario is the dominant one [59], [60].

In both of these settings, another design decision arises, which is whether communication between agents is possible. If so, there exist algorithms that allow agents' to utilize each others' experience through shared or communicated experience tuples [58]. In addition, learned communication algorithms enable agents to learn a communication protocol through

[1]Note that by boot strapping, we replace the full pay-off G_t (sum of future rewards) by the online estimation of it $r_{t+1} + \gamma V(s_{t+1})$ (the momentary rewards plus estimated value of the next state). For more details readers are refereed to [54, Sec. 13.5].

TABLE II
PROS AND CONS OF THE MAIN RL METHODS

	Value function-based	Policy-based methods	Value + Policy based methods
Primary algorithms	Solving BOE and recovering the policy	SGD on parametrized policy	SGD on parametrized policy + parametrized value function
Main strength	Fast learning through TD	Direct optimization on the policy	Variations can handle continuous actions
Main weakness	Biased function value	High variance	more parameters to tune

exchanging messages among themselves [61]. However, communication between agents is often not assumed in practical monitoring scenarios due to communication and privacy constraints. This leaves two main options for designing cooperative MARL without communication, namely: 1) independent learning (IL) and 2) centralized training with decentralized execution (CTDE). IL is the most popular option, where each agent considers all others as part of its environment. Although this approach is prevalent due to its simplicity [58], it violates the stationarity requirement assumed for MDPs. Specifically, the distribution of next states and rewards for each agent might change considerably, depending on what other agents are doing [62]. Another option is the CTDE framework, which is more common for situations where a simulator is generally available. The CTDE algorithm leverages the fact that during training, agents can efficiently communicate with each other and freely exchange information (e.g., global state, action, rewards, or network parameters) since the training is mostly done in simulators. Thus, the critic can utilize the global state information (e.g., the concatenation of all agents' observations) to guide the training of the actor, which in turn takes only the local observation as an input. Then, at execution, agents only utilize policies that are based on their local observation [63], [64]. Note that during training, the centralized critic can guide the training of the policies. However, at any time, the policy can be used independently of the critic.

There exist hybrid approaches that learn a centralized but factored Q-value function, such as QMIX [65] and QTRAN [66]. These works aim to decompose a global value function into an additive, or even nonlinear, decomposition of the individual value functions. However, not all coordination scenarios are easily factorizable (i.e., we cannot always describe the global utility function by individual utility ones).

E. Lessons Learned: How to Select the Appropriate RL Model?

We describe a generic way to identify the application requirements and the suitable RL model to be used. First, we summarize in Tables II and III the main pros and cons of diverse RL methods, as well as the categorization tool that can be used to specify the application needs, respectively. Then, according to these requirements, the appropriate solution approach with justifications are explained.

It is assumed that the targeted application already contains a temporal structure (i.e., monitoring and/or control). Thus, we

TABLE III
CATEGORIZATION TOOL

Category	Type
Action space	• Continuous • Discrete
State space	• Continuous • Discrete
Transition model	• Known - Deterministic • Known - Stochastic • Unknown
Sample complexity	• Low (simulators) • Medium (simulators with complex models) • High (real-world samples)
Objective	• Discounted rewards • Average rewards

are not only interested in finding the optimal decision variables of a well-defined problem, but rather interested in finding a policy (i.e., mapping) that describes the agent behavior in uncertain states. Hence, after categorizing the application's needs based on the above-mentioned points, the below recommendations can be used as initial proposals for the design of an RL-powered solution.

1) *Discrete Action/State Spaces With Known Models (Whether Stochastic or Deterministic):* DP approach (i.e., GPI or VI algorithms) are recommended since they can utilize the known dynamics to reach speedily to the optimal policy.
2) *Discrete, Limited Action Space With Continuous State-Space:* Value-based approaches (i.e., Q-learning and variants) have proved their efficiency in the literature as long as the number of actions is suitable to be cast as an NN output layer.
3) *Continuous Space for Both States and Actions:* Actor–critic approaches (i.e., DDPG and variants) are recommended since they directly optimize the continuous variables of the policy and do not need to represent each action independently, e.g., as an independent neuron in the NN output layer.
4) *High Sample Complexity:* Adding a planning component is recommended, specifically, when the interaction with the environment is expensive (e.g., due to lack of

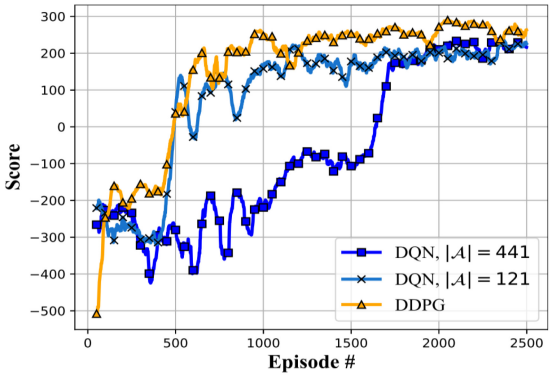
a simulator). In that case, agents can build an internal model of the environment and use it to obtain approximated samples to be used along with the true ones. This approach is known as planning, and it can significantly enhance the performance with a much less number of actual interaction samples if the model is reasonably accurate.

F. Quantitative Comparison

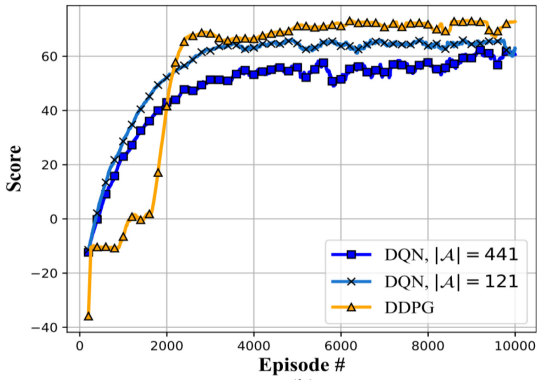
In this section, we empirically demonstrate the difference between the value-based methods with function approximation represented by DQN and actor–critic methods represented by the DDPG variant, which is specialized in an environment with continuous actions. The objective of these experiments is to investigate the performance of DRL in two main scenarios that can be followed in monitoring applications. In the first one, we discretize the action space according to different resolutions and then deploy DQN. In the second approach, we directly deploy the DDPG algorithm for the testing. The testing environments used are Lunar [67], a benchmark environment from open AI. Lunar simulates the control of a landing of an aerial vehicle through controlling two engines. The action vector consists of two real values vector from -1 to $+1$. The first controls the main engine's throttle (from off to full power), while the second value controls the orientation -1.0 to -0.5 for the left engine, $+0.5$ to $+1.0$ for the right engine, and off otherwise. We also use another specialized health monitoring environment, presented in [68], for secure and energy-efficient medical data transmission scheme. The considered action space therein consists of vectors of two elements, each in the range 0 to 1, in order to control the transmission power and compression ratio of a secure remote health monitoring system, respectively.

Q-learning algorithms require discretization of the action space. Too dense discretization is prone to the curse of dimensionality (exponential number of actions) and leads to slower learning and convergence to policies with moderate performance. On the other hand, discretization with low resolution does converge faster, but the performance is also limited as expected since the actions are not accurate enough for the environments (i.e., engine thrust actions in the testing environments cannot be tuned freely but according to a step size).

In Fig. 5(a), the performance of DDPG is expectedly better as the continuous action domains are natively supported, as explained earlier. However, this comes at the cost of more parameters to tune. The used parameters are shown in Table IV. The same superior performance is maintained in the health-monitoring environment in Fig. 5(b). However, we notice DQN is faster to reach good action at the initial phase. This is due to a smaller range of actions in this monitoring environment (i.e., 0 to 1 as opposed to -1 to 1 in Lunar), making more actions similar to each other, and the exploration requirements less. In conclusion, these empirical experiments suggest that when the action range is limited (i.e., action precision requirements are not high), then discretization and deep Q-learning can be used to deliver satisfactory results with a relatively small number of hyperparameters to tune.



(a)



(b)

Fig. 5. DQN with discretization and DDPG methods performance on: (a) lunar benchmark environment and (b) custom health monitoring environment.

TABLE IV
DDPG PARAMETERS

Parameter	Value
Discount factor γ	0.9
Exploration	Ornstein Uhlenbeck process $\theta_{OU} = 0.2$, $\sigma = 0.15$
Actor-Network neurons/layer	$24, 64, 64, 2$
Critic-Network neurons/layer	$24, 64, 64, 1$
learning rates (actor, critic)	$10^{-4}, 10^{-3}$
Activation function	Leaky ReLU, 0.01 -ve slope
Optimizer	ADAM [69]
Replay buffer size $\lvert\mathcal{D}\rvert$	10^5
Batch size $\lvert\mathcal{F}\rvert$	64
Soft update factor τ	10^{-3}
Soft update period $target$	1

On the other hand, if the precision is not known, or the action is a high dimensional vector, then DDPG can deliver higher performance but requires more parameters to tune per design.

V. APPLICATIONS AND SCENARIOS OF RL IN I-HEALTH SYSTEMS

With the increasing number of data-intensive applications in healthcare domain, many challenges are emerged regarding the stringent requirements of such applications (such as ultralow latency, high data rates, energy consumption, etc.) [70]. Toward enabling such applications, we argue that RL can

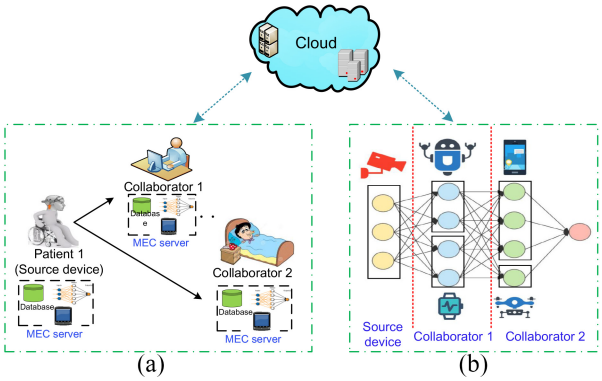


Fig. 6. Data/model parallelization in collaborative learning. (a) Data parallelization. (b) Model parallelization.

provide efficient solutions and algorithms to cope with these challenges. Indeed, EC and smart core network architecture, as well as HetNet infrastructure will play pivotal role in supporting diverse requirements of healthcare applications. In this section, we discuss the state-of-the-art applications of RL in three main areas related to I-health system, i.e., edge intelligence, smart core network, and dynamic treatment regimes.

A. Edge Intelligence

Recently, edge intelligence (or intelligent edge) has emerged as a promising 5G technology that enables integrating EC capabilities with AI techniques to provide fast processing with real-time analytics. Google, IBM, and Microsoft proposed the use of intelligent edge in many applications, including agriculture precision using cognitive assistance, smart health systems [39], and spanning from live video analytics [71]. However, developing intelligent edge is still facing several challenges due to the computational and data-intensive nature of AI models. AI software requires powerful computation, networking, and storage capabilities to support resource-intensive deep learning algorithms, such as convolutional NN (CNN) and recurrent NN (RNN), which were initially designed for natural language processing and computer vision applications. The complex structure, multiple layers, large data sets, and significant training resources needed for DNNs contribute to their high performance.

To decrease the overhead caused by the learning and inference processes, collaborative learning approaches have been proposed in the literature. In particular, two collaborative modes were designed (see Fig. 6). The first mode, namely, data parallelization, consists of distributing the training task over multiple collaborators [multiaccess EC (MEC) servers]. Each collaborator participates in the training phase using its private data set and creates its individual model. Then, different collaborators exchange their parameters to design a final model. Such approaches, called also federated learning [72], [73]. The second mode, namely, model parallelization, is related to the inference phase (i.e., execution or testing phase), where the collaborative executions are applied by dividing the complex AI models into small tasks and distribute them among the edge nodes [71]. Specifically, the DNN is divided into segments, and each segment (i.e., one or multiple layers) is allocated to a helper. Then, each helper shares the output of its segment execution to the next participant until generating the final result.

On the other hand, bringing the edge intelligence close to the users/patients, using MEC, along with transmitting the data over a heterogeneous 5G network is a key for supporting smart health applications [70]. Leveraging ubiquitous sensing, HetNet, and intelligent edges, the proposed MEC platform in Fig. 7 can real-timely monitor people's daily life and provide intelligent healthcare services, such as emergency detection and notification, epidemic prediction, and occupational safety. In Fig. 7, the edge node can:

1) acquire the medical and nonmedical data from various monitoring devices;
2) implement advanced AI techniques using the acquired data for data compression, event-detection, and emergency notification;
3) run RL-based solution for optimizing medical data delivery;
4) forward the paramount data or extracted features of interest to the cloud/healthcare service providers.

Interestingly, various health-related applications (apps) can be implemented at the edge level for supporting real-time patient–doctors' interactions. Hence, the patients can participate in their treatment regimes while interacting with their doctors anywhere and anytime. Also, specialized context-aware processing schemes can be implemented near the patients to allow for optimizing medical data processing and transmission based on the context (i.e., data type, supported application, and patient's state) and wireless networks' dynamics [74], [75].

In what follows, we will focus on some use cases that implement RL-based solutions at the network edge for enabling efficient healthcare services.

1) Secure Medical Data Transmission: Secure medical data transmission from a patient to healthcare service providers is mandatory for smart healthcare systems. Traditionally, for supporting secure wireless communication, different security schemes have been implemented either at the upper layers of the open system interconnection (OSI) model using cryptographic techniques, or leveraging physical-layer security (PLS). The latter refers to the fact that data transmission in a wireless channel is subjected to random channels' effects, e.g., noise, attenuation, and multipath fading. PLS leverages such channels' effects to secure the wireless transmission by utilizing the secrecy capacity notion [68]. Indeed, secrecy capacity refers to the secure wireless communication rate that can be transmitted through the legitimate channel without being wiretapped by the eavesdropper (see Fig. 8).

Q-learning and DRL schemes have been recently utilized for PLS to develop a learning-based model for secure medical data transmission in a dynamic wireless environment [76]. For instance, Xu et al. [77] exploited stochastic game, i.e., considers transmitter, receiver, and multiple attackers as players, to tackle the PLS problem, while leveraging the Q-Learning algorithm to control the power allocation and enhance secrecy capacity. Xiao et al. [78] leveraged RL to optimize the authentication policy for controller area networks' bus authentication. In particular, the proposed solution uses RL

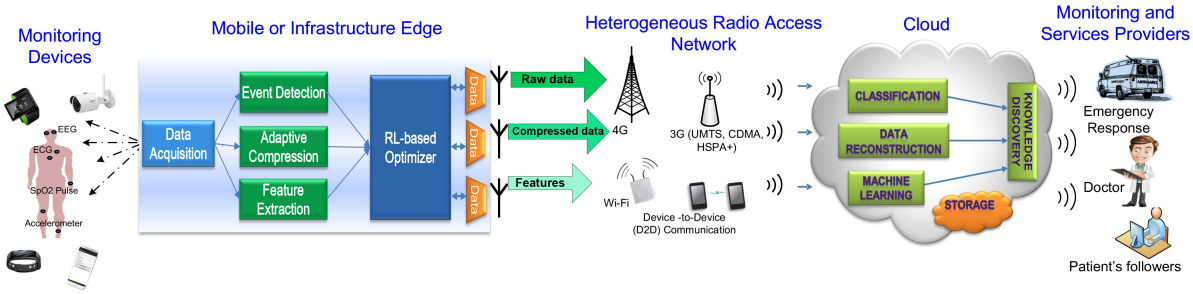


Fig. 7. Implementation of edge functionality within the I-health system.

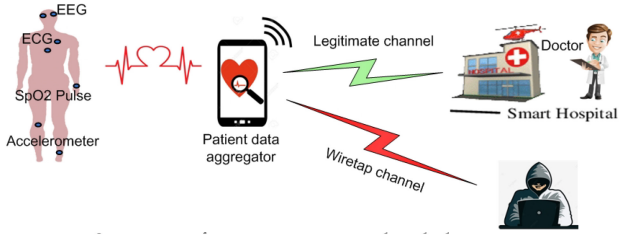


Fig. 8. PLS system for securing medical data transmission.

to select the authentication mode that captures the physical-layer features (i.e., the arrival intervals and signal voltages of the received messages), while detecting spoofing attacks and improving authentication accuracy. In [79], a physical-layer anti-eavesdropping scheme is proposed for multiple-input-single-output visible light communication wiretap channel. Indeed, a DRL-based smart beamforming solution is proposed to reduce the eavesdropped signal level while enhancing the received signal level at the legitimate receiver. Also, an actor–critic algorithm is used to improve the learning rate, while utilizing the information of the high-dimensional structure of the beamforming policy domain. Allahham et al. [68] exploited RL to provide secure and optimized medical data delivery over highly dynamic healthcare environment. Indeed, they show that RL-based solution could significantly reduce the computational overhead resulting from reoptimizing the solution when changes happen, compared to the optimization-based solutions in [80] and [81].

2) Privacy-Aware Health Systems: MEC architecture can reduce the energy consumption at the IoMT devices by processing the acquired information at the edge nodes (e.g., base stations, access point, and laptops) that have sufficient computational and energy resources. However, offloading such private data from the IoMT devices to the edge nodes comes at the risk of violating privacy conditions. An eavesdropper, near to the IoT devices, can detect the location and data pattern of the patients during data offloading. Thus, efficient privacy-preserving schemes should be implemented near to the patients.

In this context, Min et al. [82] presented a privacy-aware offloading scheme that aims at obtaining the offloading rate of the acquired medical data in energy-harvesting powered healthcare systems. Specifically, the RL-based algorithm is utilized to define the optimal offloading policy while accounting for the energy harvesting, radio channel states, size and priority of the acquired data, as well as the battery level. Moreover, a transfer learning mechanism, post-decision state (PDS) method, and Dyna architecture are used to accelerate the learning process. The problem of decentralized real-time sequential clinical decisions problem has been tackled in [72], considering MEC architecture. In particular, a double DQN (DDQN) algorithm is implemented, at each edge node, to define the optimal clinical treatment policy, while using a decentralized federated framework for DDQN models aggregation from all edge nodes. Moreover, two additively homomorphic encryption schemes have been developed to guarantee the privacy of the acquired medical data during the training process. A lightweight privacy-preserving scheme is presented in [83], based on Q-learning. In particular, the additive secret sharing and EC are used to enhance the efficiency of data encryption for diabetic patients. Indeed, the edge servers are used to decrease the needed model's updates and drug dose calculation times. Hence, the proposed scheme allows for reducing the demand for computing power, while maintaining the required efficiency and privacy protection.

3) Network Selection Over Heterogeneous Health Systems: 5G HetNet is considered as an important feature that can provide high spectral efficiency, low latency, and high throughput for diverse types of applications. Thanks to the availability of several cellular, D2D communication [84], WiFi, and fixed access technologies, the performance of I-health system can be significantly enhanced by enabling data transfer from edge nodes to the core network in an energy-efficient manner, while maintaining strict QoS requirements. However, this calls for designing efficient network selection/association algorithms to allow for connecting different users to the optimal network(s), while satisfying user preferences and reducing the access failure rate. Indeed, developing low-latency, highly reliable, and cost-effective network association schemes can significantly help in fulfilling diverse requirements of the I-health systems, which include the following.

1) Fulfilling the explosive traffic demand in remote monitoring applications.
2) Providing ubiquitous connectivity and smooth experience for different patients.
3) Supporting ultralow latency applications, particularly in case of adverse patient's events.

By using a patient edge node (PEN), e.g., smartphone, equipped with different radio access technologies (RATs), a patient (or user) would be able to forward its data to the healthcare service providers through multi-RAT [85]. However, the real-time process of obtaining which RAT (or RAN) to use

is not an easy task, given the RANs' dynamics and traffic variations. In this context, RL has been used to allow each user to learn from his/her previous experience and define the optimal RAN-selection policy. RL-based network selection solutions in [86] and [87] have depicted the efficiency of RL in achieving swift convergence to near-optimal performance. Specifically, Chkirbene et al. [87] presented a DRL-based solution that leverages dense HetNet within the smart health system to provide seamless connectivity for healthcare applications. In particular, a multiobjective optimization problem was formulated to minimize, for each user, the transmission energy consumption, monetary cost, latency, and signal distortion, while satisfying the applications' QoS requirements. The objective of this problem was obtaining the optimal compression ratio and selected RAN(s) for each user. Then, this problem was solved using DRL. In particular, the set of states is formulated as $S = \{T_1^r, T_2^r, \ldots T_R^r, PEN_i\}$, where T_j^r refers to the available time slots at RAN j. The set of actions A is defined as $A = \{P_{i,1}, P_{i,2}, \ldots P_{i,R}, k_i\}$, where P_{ij} is the RAN indicator that refers to the fraction of data of PEN i that will be sent through RAN j, and T_j is the available fraction of time over RAN j that can be used by different PENs (i.e., resource share). Then, due to the continuous action space of the formulated problem, a DDPG algorithm is proposed to solve the formulated problem.

The existence work for network selection has been also extended for large-scale systems by leveraging MARL techniques [88]. Given that multiple users are competing on the available resources over different RANs, MARL can provide a decentralized solution that optimizes the individual users' policies. For instance, Yan et al. [89] presented an optimized solution, leveraging MARL techniques, for maximizing the average system throughput over multi-RAT network, while maintaining the users' QoS constraints. However, this work ignores the energy consumption constraint and the high-level applications' requirements, such as the reliability, which opens the door for future research directions.

B. Smart Core Network

RL has been extensively used as an efficient tool for optimizing core network functions and dynamic treatment regimes at the cloud. In particular, with the rapid evolution of I-heath systems along with cloud computing, RL techniques gained much interest for optimizing policy making and treatment strategies. In what follows, we present some use cases that consider RL techniques at the core network for optimizing the performance of I-health system.

1) Resource Allocation and Task Offloading: RL is considered as one of the promising solutions for dynamic resource allocation in wireless communication systems [90], [91]. The effectiveness of DRL-based solutions for dynamic resource allocation are twofold. First, the obtained actions (or decisions) are taken with the help of DNN, which provides accurate solutions for the whole operational period [8]. Second, leveraging DRL allows for considering the state transition overheads, e.g., power consumption related to the transition from one transmission mode to another. Such transition states have been shown to be significant [92], however they are neglected in most of the resource allocation schemes that provide optimal/suboptimal solutions for the current time slot only. In this context, Xu et al. [90] have leveraged the power of DRL in solving complex control problems to develop a DRL-based solution for dynamic resource allocation in a cloud-RAN architecture. The proposed framework aims at minimizing the total power consumption while fulfilling the requested demand for different users. In particular, the state space is defined as a function of the binary states (active or sleep) of each remote radio heads (RRHs) and the required demand of each user. The action space refers to which RRH will be turned on or off, while the reward function is defined as a function of the maximum possible value of the total power consumption. Then, the DNN is used to approximate the action-value function, while formulating the resource allocation problem (in each decision epoch) as a convex optimization problem.

In [91], a centralized DRL-based downlink power allocation technique is proposed for a multicell system with the aim of maximizing the total network throughput. In particular, a DQN approach is applied to define a near-optimal power allocation policy. The obtained results depict that the proposed scheme outperforms the conventional power allocation schemes in a multicell scenario. Meng et al. [93] developed several DRL architectures, such as REINFORCE, DQN, and DDPG, for optimizing the power allocation in multiuser cellular networks while maximizing the overall sum-rate of the network. Naparstek and Cohen [94] leveraged DRL for solving the problem of dynamic spectrum access in wireless networks. The main goal of this work was to maximize the utility of each user in a distributed manner, i.e., without exchanging information. He et al. [95] presented a DRL-based link scheduling algorithm to mitigate the interference impact in wireless networks. The proposed algorithm applies MDP to model the network's dynamics, i.e., channel state information (CSI) and cache states, at the transmitter side. In particular, a DNN model is utilized to learn the states' variations. In [96], DRL is exploited to address the mobility impact in wireless networks, where an RNN and a CNN are used to extract the features from the received signal strength indicators.

Mobile cloud computing (MCC) has also gained much interest recently to enable pervasive healthcare services in an energy-efficient and cost-effective way [3], [97], [98], [99]. The main objective of MCC is to efficiently offload computation-intensive tasks from mobile devices (i.e., battery-operated devices) to the cloud, while guaranteeing continuous and uninterrupted healthcare services. For instance, a model-free RL-based scheduling approach, leveraging Q-learning, has been proposed in [98] to enhance the task scheduling and dynamic computation offloading from mobile devices to the cloud. Indeed, this work focuses on three main objectives for optimizing computation-intensive tasks offloading, i.e., battery lifetime, diagnostic accuracy, and processing latency. The study in [3] focuses on improving healthcare network quality by analyzing and predicting network traffic across Mobile, Cloudlet, Network, and Cloud layers. They propose a deep learning-based scheme that utilizes big data and high-performance computing resources to predict traffic,

which is then used by Cloudlets and network layers to optimize data rates, caching, and routing decisions to meet healthcare application communication needs.

Several research studies have also tackled the human activity recognition in the field of smart health. Indeed, different practical applications, such as gym physical activity recognition and fall detection have been investigated. AI schemes play a major role in such studies through automatically detecting and identifying human activities and behavior patterns by analyzing the acquired data from various IoMT devices. The recent studies that tackled the human activity recognition problems can be categorized into two groups: 1) ambient sensor-based and 2) wearable sensor-based schemes. The former refers to the techniques that utilize surveillance camera, temperature, and other indoor sensors to acquire the environment-related data to detect people's daily activities within smart-assisted environment, such as smart homes and recovery centers [100], [101]. The latter refers to the techniques that leverage wearable devices or smartphones to acquire and monitor vital signs or on-body physiological signals using accelerometer, magnetometer, and gyroscope sensors [102]. For instance, Zhou et al. [103] presented a semi-supervised deep learning framework that incorporates an auto-labeling technique with a long short-term memory (LSTM)-based classification scheme. The main goal of this framework is to efficiently leverage the large amount of weakly labeled data to train a classifier in order to enhance the classification accuracy of human activity recognition applications. Indeed, the proposed auto-labeling technique is developed based on DQN to address the problem of inadequately labeled data and enhance the learning accuracy.

2) Network Slicing: Network slicing is one of the major technologies in 5G networks that aims at enabling the realization of the growth in the Industrial IoT market. It refers to the ability of the network operators to split the physical network components into multiple logical slices (i.e., subnetworks) with different characteristics. Hence, diverse IoT applications with different demands can be supported using various dedicated network slices [104]. Recently, network slice as a service has been emerged as a new concept that allows network operators to create a customized network slice for each user/application as a service. In the context of I-health systems, network slicing can help in the following.

1) Maximizing resource utilization by dynamically adapting available resources for each slice [105].
2) Enhancing the system scalability while fulfilling diverse QoS requirements.
3) Providing high security and high privacy for sensitive healthcare applications.

Specifically, remote monitoring applications, smart hospitals, and remote surgeries demand for high data rates, supreme reliability, and ultralow latency communication, which can be supported by novel 5G technologies such as network slicing [88]. Indeed, each health-related application can be allocated a customized network slice to support its distinctive requirements (see Fig. 9). Thus, Pries et al. [106] have investigated the potential of network slicing for supporting efficient connection between smart wearable devices and the cloud. In [107], network slicing and 5G network have been leveraged for enabling remote surgeries applications with ultrareliable requirements, ultralow latency, and guaranteed bandwidth allocation.

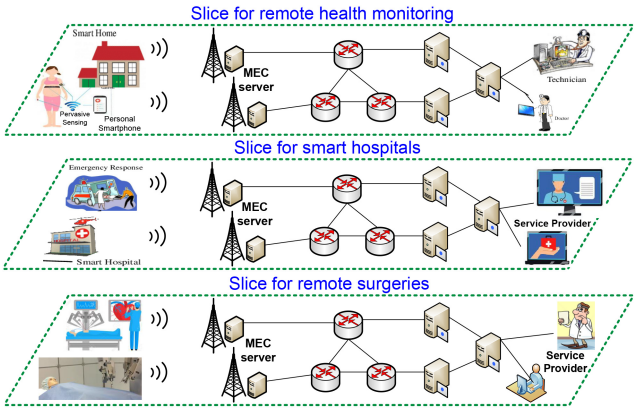


Fig. 9. Network slicing for I-health systems.

Conventional optimization models for network slicing and queueing theoretic modeling turn to be intractable, especially for highly dynamic environments with strict reliability, bandwidth, and latency requirements. Thus, learning techniques, such as deep learning and RL, can help enhance the dynamicity of heterogeneous IoT systems, such as I-health systems, through dynamic slice allocation and resource allocation between slices. Predicting users' demands using learning techniques can enable dynamic adjustment of resource allocation based on the predicted demands [108]. Leveraging DRL for dynamic network slicing and resource management in 5G network has gained much research interest recently [109], [110]. DRL has proved its efficiency in enhancing the overall system performance in terms of throughput, latency, and reliability. For instance, the formulated network slicing problem in [110] has been decomposed into a master problem and several slave problems. The master problem is solved using convex optimization, while the DDPG algorithm is used to solve the slave problems. Herein, the DDPG algorithm is used to learn the optimal policy for allocating users' resources without requiring closed-form expressions for the users' utility functions in the slave problems. Koo et al. [111] leveraged DRL to simultaneously consider varying traffic arrival rates, dynamic applications' requirements, and limited resources availability. In particular, the network slicing resource allocation problem is mathematically formulated as an MDP, then a policy-gradient method, using the REINFORCE algorithm [112], is utilized to solve this problem and learn the optimal policy for network slicing resource allocation. The network slicing resource trading process between a slice provider and various tenants is also formulated as an MDP in [113]. In particular, the MDP state space is defined as a function of the QoS satisfaction parameters. Then, a Q-learning-based dynamic resource allocation strategy is used to maximize tenants' profit, while fulfilling diverse QoS requirements of end users in each slice.

Interestingly, strict security and privacy requirements for sensitive IoT applications, such as I-health, can be maintained by exploiting the concept of slice isolation [114], [115], [116].

TABLE V
SUMMARY OF THE IMPORTANT STUDIES THAT CONSIDERED RL-BASED SECURITY SCHEMES

Ref	Technique	Objectives	DRL algorithm	States	Actions	Rewards
[126]	Secure data offloading	Create a policy for a secure data offloading to avoid jamming attacks	DQN	User density, battery levels channel bandwidth and jamming strength	selected edge node, transmission power, offloading rate and time	Defined based on secrecy rate, communication efficiency, and power consumption
[127]	Secure mobile edge caching	Maximizing the offloading traffic	A3C	Network conditions and signals characteristics, i.e., user requests and information	Caching policy (i.e., caching or not)	Computed based on the the amount of traffic at each edge node
[128], [76]	Spoofing detection	Create an optimal authentication model	Q-learning and Dyna-Q	False alarm rate, missing spoofing rate, and detection rate	Authentication discrete levels	Defined using an utility function that is computed based on the Bayesian risk
[129]	Secure network resources allocation	Assign resources to each user	Double dueling DQN	Base stations' status, caching contents, and MEC servers' conditions	Users to base station association, caching locations, and selected MEC servers for computing	Computed based on the SNR of the wireless access link, cache status, and the computation ability
[123]	Secure CPS	Detecting false inputs	DDQN and A3C	System output	Select the next input value from a set of constant inputs signals	Defined based on the past dependent life long property, time and output signal

For instance, fog computing and network slicing have been used in [114] to develop a secure service-oriented authentication scheme for IoT systems over 5G network. This work aims at isolating users that have suspicious behavior in quarantine slices until executing the necessary actions. De Brito Gonçalves et al. [115] integrated the network slicing concept with blockchain technology to support privacy isolation for a hospital network. Thus, we envision that implementing slice isolation concept within the I-health system can be beneficial in addressing several security issues.

3) Security Schemes for Health Systems: Security is one of the fundamental concerns in any healthcare system. Typically, in I-health systems, the acquired medical data from the patients (e.g., blood pressure, blood sugar, etc.) are forwarded to the cloud servers for processing and storage, where sophisticated data mining and AI models are implemented to provide prediagnosis decisions. The healthcare service providers are then notified with these decisions in order to take actions and provide the appropriate responses. However, enabling secure data exchange, remote monitoring, data processing, and real-time diagnosis services without revealing patients' information and privacy is still challenging [117], [118]. Indeed, classical healthcare systems usually rely on weak security schemes to protect their data processing and management, which results in serious security problems in these systems. For example, the number of recorded security attacks in healthcare systems increased between 2016 and 2017 by 89% as reported in [119]. Moreover, many reports depicted that a wave of cyberattacks have threaten several hospitals during the recent COVID-19 pandemic [120]. These attacks can shut down hospitals' services and hinder healthcare facilities. For example, an attacker can steel user identification by obtaining login credentials via the use of phishing emails. This attacker who has access to the medical system can falsify patients' reports as well as attacked the control system.

Different studies have tackled such problems using RL schemes in order to protect patients' personal information and healthcare systems. For instance, Nguyen and Reddi [121] presented a mathematical model for cyber state dynamics as $x(t) = f(t, x, u, w; \theta(t, a, d))$, where the physical layer disturbances are denoted by w, cyber attack is denoted by a, and the defense strategy is denoted by d. The attack-defense dynamics are presented by $\theta(t, a, d)$ at time t, and x defines the physical state. Then, an actor–critic DRL algorithm is used to present the defense strategy against the hackers. The obtained results depict the efficiency of the proposed DRL model with its ability to obtain the optimal strategy that could enhance the system performance in terms of attacks detection [121]. Liu et al. [122] developed a privacy-preserving RL-based framework for patient-centric dynamic treatment regime. This framework has been deployed at the cloud computing environment for providing secure multisource processing and data storage, as well as secure RL training without leaking the data of the patients. In [123], DDQN and asynchronous advantage actor–critic (A3C) algorithms have been used to solve the robustness guided falsification problem of cyber-physical system (CPS). This work has shown that using traditional methods, such as simulated annealing [124] and cross entropy [125] are inefficient due to the infinite state space of CPS models. On the contrary, formulating the problem as an RL problem could obtain better results compared with the existing state-of-the-art techniques for detecting any falsified inputs of CPS.

Table V summarizes some of the existing DRL models for network security, which can be easily extended to protect healthcare systems.

C. Dynamic Treatment Regimes

RL-based solutions have proved their efficiency in different aspects of healthcare systems. In particular, different RL

schemes have been applied for diverse dynamic treatment regimes, including chronic diseases, mental diseases, and highly infectious diseases [130], [131], [132], [133], [134], [135]. Indeed, dynamic treatment regimes map individualized treatment plans into sequences of decision rules for each stage of clinical intervention, while considering current patients' state to support them with a recommended treatment. Such a sequential decision-making process is perfectly fit with RL concept, since it can be easily modeled as an MDP, while finding the optimal treatment regime using several RL algorithms. However, relying only on observational data[2] to learn and evaluate dynamic treatment regimes may face some safety challenges. Indeed, the observational data may be subject to selection bias, confounding, and other sources of bias. Selection bias occurs when patients are not randomly assigned to treatment groups, but instead are assigned based on factors that are related to their prognosis or likelihood of receiving a certain treatment. Confounding occurs when there are other factors that are associated with both the treatment and the outcome, making it difficult to determine whether the treatment is causing the outcome or whether the association is due to other factors.

Despite these challenges, there have been many studies in epidemiology and biostatistics that have developed methods to address these issues when learning and evaluating dynamic treatment regimes from observational data [43], [136]. These methods typically involve adjusting for confounding factors, using propensity scores to balance treatment groups, and employing various statistical models to estimate treatment effects. Thus, while using purely retrospective data to learn and evaluate dynamic treatment regimes may be challenging, there are well-established methods in epidemiology and biostatistics that can help address these challenges and improve the validity of the resulting treatment recommendations.

Table VI summarizes some of the e-health systems that exploited RL for addressing the challenges of different diseases. It is not the objective of this article to provide an in-depth technical comparison on different proposed e-health systems. However, we investigate the practical benefits of leveraging RL in such applications. In what follows, we group these e-health systems under two main categories: 1) remote monitoring applications and 2) hospital-based applications.

1) Remote Monitoring Applications: Chronic diseases are long-lasting health conditions requiring ongoing medical attention, which account for a significant number of deaths each year [137]. The main challenge of chronic diseases is that they require continuous monitoring for the patients' state, in addition to a sequence of medical intervention to avoid adverse effects of persistent treatment. Patients' conditions (such as treatment duration and dosage, as well as patients' response to certain medications) have to be continuously revised and updated to provide sufficient treatment. To relieve the pressure on healthcare facilities, it is important to widely rely on remote monitoring applications, which allow for moving a large number of patients with mild symptoms into home care [138]. In this context, RL schemes have been widely applied for supporting dynamic treatment regimes, inside and outside healthcare facilities, which helps in many chronic diseases and mental illnesses, as will be shown below.

Diabetes is one of the main chronic diseases in the world that requires continuous-remote monitoring without limiting patients' activities. According to [139], 451 millions of people are living with diabetes in 2017. RL for patient-specific glucose regulation in artificial pancreas (AP) [140] has attracted much interest recently. For instance, Zhu et al. [141] developed a DRL model for optimizing single-hormone (insulin) and dual-hormone (insulin and glucagon) delivery. Specifically, the proposed model is designed using double Q-learning with dilated RNNs to provide effective closed-loop control of blood glucose levels for Type 1 diabetic patients. The presented results depict the efficiency of the proposed model in obtaining an effective glucose control strategy compared to standard basal-bolus therapy with low-glucose insulin suspension. In [142], a model-free Q-learning algorithm is leveraged to control the blood glucose levels for Type 1 diabetic patients, who undergo to an intensive insulin treatments, by adjusting insulin regulation rate.

Diverse RL models have been also used for Anemia [143] and mental diseases, such as epilepsy[3] [144], depression, and different kinds of brain disorders. For instance, an RL-based approach is leveraged in [145] for dynamically learning an optimal neurostimulation strategy for the treatment of epilepsy. The presented results show that the RL-based solution outperforms the related schemes by reducing the incidence of seizure by 25% and total amount of electrical stimulation to the brain by a factor of 10. In [144], the EEG recordings are classified into normal or preseizure baseline patterns using a K-nearest-neighbor (KNN) classifier. Then, an RL approach is used to adaptively update the normal and preseizure baseline patterns according to the feedback from the prediction result.

2) Hospital-Based Applications: DRL has been also used for life-threatening diseases, such as cancer.[4] For instance, a model-free TD method with Q-learning is used in [148] for agent dosage in cancer chemotherapy. DRL-based solutions have proved their considerable potential in clinical research since it could optimize the taken actions to maximize the rewards even if the relationship between the actions and rewards is not fully known. In [149], a natural actor–critic (NAC) approach [153] is proposed for optimizing the cancer drug scheduling policy. The main goal of the proposed approach is minimizing the tumor cell population and the drug dose while obtaining sufficient population levels of normal cells and immune cells. The obtained control policy by the proposed NAC approach suggests the drug should be injected continuously from the beginning until an appropriate time. This policy depicts better performance than the traditional pulsed chemotherapy strategy.

[2]Observational data refers to data that is collected without randomization, such as data from medical records or surveys.

[3]Epilepsy is one of the major severe neurological diseases that affects around 1% of the world population.

[4]Cancer is one of the most serious chronic diseases that cause death. Statistics show that 15.7% of total deaths in the world is caused by cancer.

TABLE VI
SUMMARY OF THE RELEVANT E-HEALTH SYSTEMS

Application	Medical Data	Disease	Description	RL Benefits
Remote monitoring [141]	insulin and glucagon levels	Type 1 diabetes	DRL model is proposed for optimizing basal insulin and glucagon delivery	The proposed method clearly enhanced glycemic outcomes in adult population
Remote monitoring [146]	Glucose level	Type 1 diabetes	The proposed method optimizes the dynamic treatment regime for patients with type 1 diabetes	RL could improve patients' state by enabling frequent treatment adjustments based on the evolving health conditions of each patient
Remote monitoring [143]	HGB level	Anemia	It addresses the challenges of individual response variations to the treatment and changing patients' response over time	The proposed approach generates adequate dosing strategies for representative individuals from different response groups
Remote monitoring [144]	EEG	Epilepsy	An adaptive learning approach is proposed by integrating RL with online monitoring to enhance personalized seizure prediction	RL prove its efficiency in improving prediction accuracy of the system
Remote monitoring [147]	EEG	Epilepsy	It obtains the optimal deep-brain stimulation strategy as a function of the acquired EEG signal	The proposed RL method could minimize the frequency and duration of seizures
Hospital-based [148]	Dosage in cancer chemotherapy	Cancer	A model-free TD method with Q-learning is applied for discovering individualized treatment regimens	RL could obtain the optimal treatment strategies directly from clinical data without identifying any accurate mathematical models
Hospital-based [149]	Drug dose in cancer chemotherapy	Cancer	Drug dose is fed as an input to the RL scheme, while defining the reward as a function of drug dose and cell populations	The obtained results showed that RL yield better performance than conventional policy of pulsed chemotherapy
Hospital-based [150]	Drug dose in cancer radiotherapy	Cancer	Q-learning algorithm is proposed to optimize dose calculation in cancer radiotherapy	This work states the power of RL in optimizing complex biological problems such as radiotherapy in cancer treatment
Hospital-based [151]	EHR	Sepsis	RL-based model is developed to dynamically obtain the optimal treatments for patients with sepsis in the ICU	Mortality rate decreased in the patients that follow the treatment decisions obtained by the RL
Hospital-based [152]	EHR	anesthesia	RL is utilized to design a closed-loop anesthesia controller monitor and control the infusion of anesthetics	RL show comparable performance with the recent clinical trials conducted

In [150], an agent-based simulation model along with a Q-learning algorithm is proposed to optimize dose calculation in cancer radiotherapy. Indeed, RL is used to optimize the main two factors of radiotherapy, i.e., radiation dose and fractionation scheme. In particular, each fraction is considered as a state variable, and the intensity of radiation is considered as an action variable, while presenting the reward as a function of the positive effect on tumor and negative effect on healthy tissue. Padmanabhan et al. [154] presented a Q-learning algorithm that aims to implement an optimal controller for cancer chemotherapy drug dosing. A DRL-based framework has been proposed in [155] to automate the adaptive radiotherapy decision for nonsmall cell lung cancer patients. Three NN components have been considered.

1) Generative adversarial network (GAN), which is used to generate sufficiently large synthetic data from historical small-sized real clinical data.
2) DNN, which is employed to learn how states would transit under different actions of dose fractions leveraging the synthesized data and the available real clinical data.
3) DQN, which is responsible for mapping different states into possible dose strategies while optimizing future radiotherapy outcomes.

In [156], a DRL-based solution is presented to detect abnormalities in medical images. In particular, the proposed scheme is used to optimize the lymph node bounding boxes, hence enhancing the lymph node segmentation performance, which is crucial for quantitatively accessing disease progression.

RL is gaining also much interest for intensive care (IC) diseases [151], [157], [158], [159]. Early detection and description of critical diseases significantly help doctors in the treatment, which can save many lives. Indeed, RL has been widely applied in the treatment of sepsis, regulation of sedation, and some other diseases that require IC unit (ICU), such as mechanical ventilation and heparin dosing. Sepsis is one of the main causes of death in hospitals and stands in the third place of mortality worldwide, however the optimal treatment strategy is still ambiguous [160], [161]. There is no universally agreed-upon treatment for sepsis due to the varying patients' responses to medical interventions. To tackle such a complex decision-making problem, RL has been used in [151] to learn the optimal treatment policy by analyzing the massive amount of information extracted from different patients. This work depicts that the obtained treatment decisions by RL are on average more reliable than the decisions of human clinicians. In [158] and [162], a DRL with a continuous state-space model is used to deduce the optimal treatment strategies from the available training samples that do not represent the optimal behavior. Indeed, the patient's physiological state is represented as a continuous vector (using physiological data from the ICU) in order to define the suitable actions with Deep-Q Learning. The application of RL for sepsis treatment has been also considered in [159]. However, the authors there focus on the reward learning problem of RL. In particular, a deep inverse RL with mini-tree (DIRL-MT) model is presented to define the best reward function from a set of possible treatment

strategies using real-world medical data. In this scheme, the Mini-Tree model is used to learn the main components that affect the death rate during sepsis treatment, while the deep inverse RL model being leveraged to obtain the complete reward function as a function of the weights of those components. In [163], sepsis challenges have been tackled using mixture-of-experts framework to personalize sepsis treatment. Specifically, the proposed mixture model automatically swaps between neighbor-based (kernel) learning and DRL based on the current history of the patients. The obtained results in [163] depict the superior performance of the proposed mixture model compared with applying the strategies of physicians, Kernel learning only, and DRL only. Petersen et al. [164] leveraged the DDPG scheme to deal with the continuous state and action spaces of the sepsis environment, hence defining an effective treatment strategy for sepsis.

It is defined as a state of controlled, temporary loss of sensation. Critically ill patients, who are supported by mechanical ventilation, require adequate sedation for several days to guarantee safe treatment in the ICU [165]. RL methods have been applied for the regulation of sedation in ICUs. The problem of anesthesia control using RL has been studied in several works, such as [152], [166], and [167]. A closed-loop anesthesia controller model is presented in [152]. This model is able to regulate the bispectral index (BIS) and mean arterial pressure (MAP) within the required range. In particular, the proposed RL model has considered the weighted combination of the error of the BIS and MAP signals to decrease the computational complexity and processing time. In [167], an adaptive NN filter (ANNF) is proposed to improve the RL model for closed-loop anesthesia control. In particular, the authors propose a method for smoothing the acquired BIS measurements while minimizing time delay. This leads to enhancing the patient state estimation, which allows for improving the performance of anesthesia controller. In [166], an RL-based fuzzy controllers architecture is proposed for an automation of the clinical anesthesia. The authors presented a multivariable anesthetic mathematical model to compute the anesthetic state using two anesthetic drugs of Atracurium and Isoflurane.

D. Lessons Learned

In this section, we explore how RL is currently being applied in three key areas relevant to the I-health system: 1) edge intelligence; 2) smart core network; and 3) dynamic treatment regimes. Several conclusions can be drawn in this context.

1) Despite the great potential of RL in enhancing clinical decision making, relying solely on RL approaches that utilize observational data may pose safety challenges, especially for critical cases (such as those in the ICU). Thus, to address these challenges and improve the reliability of the treatment recommendations derived from RL, it is crucial to implement carefully designed methods. These may include involving clinicians in the iterative learning process, ensuring balanced treatment groups, and utilizing different statistical models to estimate treatment effects.
2) RL can help healthcare professionals make better decisions and generate personalized treatment plans by analyzing massive amounts of data and providing optimized treatment recommendations.
3) Deploying RL-based solutions at the network edge can offer real-time analytics with rapid processing, enabling healthcare providers to make faster, more informed decisions.
4) RL has been widely utilized as an effective tool to optimize both core network functions and dynamic treatment regimes. It can help allocate healthcare resources efficiently, optimizing costs while still maintaining quality of care.

VI. Open Challenges and Future Research Directions

In this section, we propose several open research directions along with their inherent challenges that can be tackled in future studies for implementing an efficient I-health system. In particular, we investigate the opportunities of integrating the emerging RL models in new scenarios/services related to I-health system, which are envisioned to emerge in future healthcare systems.

A. Federated RL

Toward 6G networks evolution, edge intelligence is foreseen to be a crucial component in 6G network architecture. The concepts of AI for EC and EC for AI are emerging for the next-generation networks. The former refers to implementing distributed computing applications (i.e., AI or deep learning techniques) at the MEC servers to enhance the performance of EC [168], while the latter refers to leveraging EC capabilities to run sophisticated AI models. Despite the potentials of AI for EC, there are pivotal challenges that need to be address.

1) Complex structure of DL models and the need for a large amount of data and resources for the training.
2) Computational capabilities of the mobile edge nodes are typically limited.
3) Traditional learning algorithms, that run independently at each node, cannot provide the required system-level scalability and optimal performance.
4) Transferring a large amount of data between the edges while considering the high-dynamic nature of wireless networks and restricted bandwidth is not an easy task.

Indeed, most of the existing systems that leverage deep/ML algorithms assume easy availability of the data. However, for example in E-health systems, the data is locally generated and stored across different devices/nodes in the network, e.g., PDAs, hospitals, etc. Moreover, the locally generated data at each node cannot be gathered in a central node due to the privacy issues and rising demand for network bandwidth.

To address these challenges, collaborative learning approaches (such as federated learning, distributed training over multiple helpers, and distributed inference among multiple participants) have been recently proposed [169], [170], [171]. However, how to efficiently integrate collaborative learning approaches with MEC architecture in

ultradense 5G network is a critical problem in the future development of 5G [172]. Indeed, efficient utilization and collaboration between multiple edge nodes within 5G ultradense network should be considered in order to guarantee.

1) Efficient utilization of diverse resources over 5G HetNet, such as computation, communication, and storage resources [173].
2) Minimizing the overhead of decision-making and resource allocation processes.
3) Privacy and security protection for the acquired data.

The main question now is how to obtain the global optimal strategy for all collaborative edges? Unfortunately, due to the dynamic nature of the MEC systems and diverse characteristics of the collaborative edges, the generated optimization problems are always nonconvex and NP-hard [174]. Hence, it will be hard to rely on the traditional vanilla optimization methods for solving the joint MEC collaboration problems.

Federated RL begins to attract more attention for optimizing the distributed/collaborative learning approaches, where an agent learns and obtains its optimal policy leveraging its local observations and cooperation with other agents for optimizing the same system targets [175], [176], [177], [178], [179]. In Federated RL, a DRL agent build its own learning model by interacting with its environment. Then, it uploads the generated local model to a central node that is responsible for aggregating the local models from different agents and generating a global model for all agents. This model is then broadcast to all DRL agents that exchange their previous local models with the updated global model and iteratively reconstruct their local models. The work in [72] is the only study that considered federated RL in E-health system, where a DDQN model is implemented at each edge node to define a stable and sequential clinical treatment policy, while extracting the knowledge from electronic health records (EHRs) across all edge nodes by using a decentralized federated framework. Thus, we envision that the potentials of federated RL in healthcare systems is still worth further investigation. In particular, with the rapid interest in enhancing home care services, federated RL can be an efficient candidate for locally processing the patients' data and identify their states. Indeed, leveraging federated RL solution within the large-scale healthcare systems is a key for detecting and managing urgent outbreaks, since it enables a swift and portable emergency detection through identifying and monitoring infected individuals at the edge, without the need of transferring patients' data to a centralized server.

B. Dynamic Network Slicing for I-Health Systems

Strict latency and reliability requirements needed for diverse healthcare applications cannot be accomplished through the traditional telecommunication systems. The demanding need for higher data rates and extremely fast response time enable swift and accurate clinical decisions needed for patients' remote monitoring. Network slicing is a promising technology that can fulfil these strict requirements for healthcare application via dynamic slice allocation and isolation.

Different from the related work in the literature, leveraging the intelligence at the edge for optimizing the slice allocation decisions in ultradense HetNets is still an open research direction. Although few research attempts have been presented for dynamic network slicing using the Q-learning method or RL [180], we are still at the beginning level. Thus, a potential future direction can advance the state-of-the-art by the following.

1) Leveraging RL algorithms with feature-assisted schemes to obtain swift and efficient slice allocation decisions in high-dynamic environments.
2) Considering the impact of the resources' quality, users' behavior, and applications' characteristics on dynamic network slicing, hence avoiding inefficient utilization of network slices.

Moreover, building reliable prediction models to foresee users' demand and network dynamics, based on historical data and real-time gathered data, can significantly improve network slicing design through: 1) stabilizing wireless connections; 2) optimizing radio resource management; and 3) implementing proactive network slicing strategies.

We argue that network dynamics are not the only pivotal factor for designing optimal, dynamic network slicing schemes. Predicting users' state (or behavior) can also be crucial for stabilizing wireless connections and fulfilling strict QoS requirements. Specifically, in E-health remote monitoring applications, the acquired medical data of the patients must be sent to the health cloud every 5–10 min, in normal conditions. On the contrary, in robotic telesurgery, the acquired physical vital signs must be reported with a latency less than 250 ms [181]. By knowing such information in advance, the network operator can always allocate the slice with low energy consumption in normal case, while allocating the slice with high data rates in case of emergency. Thus, predicting users' behavior or state at the edge can significantly changing slice allocation decisions.

By accounting for the available slice characteristics, users' behavior and context, energy efficiency, and applications' requirements, efficient RL-based approaches can be developed to address the typical challenges of dynamic network slicing in I-health systems, using the following.

1) *Historical Information:* The edge nodes can build efficient RL models for the surrounding-complex environment to accelerate the decision process of slice allocation based on their previous experiences.
2) *Prediction Models:* By interacting with the surrounding environment, an edge node can utilize diverse performance indicators, such as the received power, received signal quality, and computational capabilities to predict the characteristics of different slices, hence connecting to the optimal slice.
3) *Network Dynamics:* We can estimate at the edge the obtained reward before associating to a specific slice.
4) *Applications' Requirements:* The edge node can associate to a customized slice based on the characteristics/requirements of the running applications.

Accordingly, we envision that the effective solution should comprise two main steps (see Fig. 10): 1) estimating users' demand, resources availability and cost, and traffic flows using accurate RL models and 2) developing MARL algorithms to be run at different edges for dynamic network slicing.

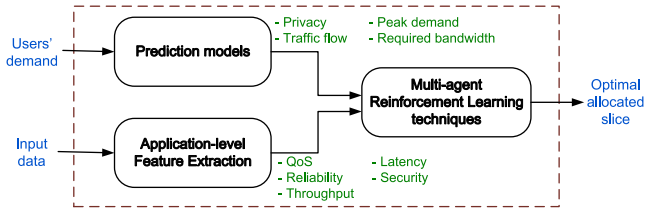


Fig. 10. Example for efficient network slicing solution over I-health system.

C. Active RL

Home care applications have gained much interest recently. The goal of home care is to continuously monitor a large number of patients, at their homes, without limiting their daily activities. We envision that, patients, equipped with wearable devices, sensors, and PDAs, can learn, detect, and prevent complex situations occurring anytime anywhere, thanks to RL techniques. However, the challenge is that patients may be exposed to unexpected situations every day, such as emergency conditions, for which limited historical data is available. Moreover, most of RL techniques depend on a long training phase, where each agent has to interact and learn from the environment the optimal policy. However, we cannot rely on such an assumption in many real-time applications, such as critical healthcare applications. To address such a crucial problem, developing an effective solution that integrates RL with active leaning concept [182], will be needed. The main idea of AL is that an agent can actively interact with its neighbors to select over time the most informative data to be considered in the training phase in order to improve its learning performance [183]. Such a solution that integrates the RL with the active leaning can comprise the following.

1) Leveraging D2D communication to increase the amount of data acquired at a PDA.
2) Investigating different types of data that can be exchanged between the PDAs, through D2D communication [184], and their impact on the obtained classification performance and communication load.
3) Proposing efficient algorithms for data quality assessment and improvement.
4) Developing efficient RL algorithms at the edge nodes. Specifically, upon the occurrence of unexpected conditions, connected PDAs do, not only interact with their environment, but also cooperate with their neighbors to optimize their rewards.

D. Optimized and Secure Data Exchange

Interactions and data exchange across distributed entities are essential to provide accurate control and management, and improve the response time in emergency conditions. However, critical challenges have emerged with data sharing and data access across distributed I-health systems.

1) Data management in untrusted cloud servers, with risks for the users' privacy.
2) Fulfilling diverse security and privacy requirements, while tackling the challenges of data processing and transfer.
3) Remote accessibility of acquired data by various authorized entities.

Thus, improving the accessibility and information sharing among diverse entities in I-heath system is mandatory to provide secure services. Indeed, we envision that a secure, trusted, and decentralized intelligent platform tackling aforementioned challenges can be designed and realized by integrating EC and blockchain technologies. Blockchain is a decentralized ledger of transactions that are shared between several entities while maintaining the integrity and consistency of the data [185]. Thus, being decentralized, it is perfectly consistent with the potentiality of EC, which supports data processing/storage at diverse entities.

Although blockchain promises to provide adjustable and distributed security protection for diverse healthcare applications [186], [187], it poses also a number of challenges that need to be addressed. These include allowing a timely access to data for e-Health services, limiting latency, required storage space, computational power, and cost, and ensuring data privacy protection. Thus, developing an efficient RL-based solution to address these challenges is needed. Indeed, RL-based solutions can adapt to different blockchain states and system's transitions in order to customize and optimize the blockchain configuration's parameters (e.g., number of verifiers, block size, transaction size, and block generation time out). Hence, such solutions would be able to establish the best tradeoff among diverse conflicting objectives in I-health systems (e.g., latency, storage space, and cost) to support secure data sharing and storage services.

E. Communication-Efficient MARL

The main wireless communication challenges, such as delay, noise, failure, and time-varying typologies, have been ignored in most of the studies that considered MARL. Although some of the recent studies have investigated the impacts of bandwidth and multiple-access techniques on the performance of emerging policies of MARL [188], [189], it is still missing how to design effective MARL-based solutions under realistic wireless network constraints.

Designing communication-efficient MARL solutions is still challenging. On the one hand, algorithms that rely on training their models at centralized servers while running the execution tasks in a decentralized manner can reduce the communication overheads at the execution phase. Meanwhile, they will be able to obtain a near-optimal joint policy at the centralized training phase. However, the adaptability of the obtained solution to the high-dynamics environments is not guaranteed. Moreover, retraining may be needed to adapt to the major changes in the environment. On the other hand, fully decentralized agents may be able to learn good policies, but this comes at the expense of communicating more to react to their joint actions. Thus, it is important to design adaptable, yet communication efficient, MARL-based solutions, which still needs further investigation.

F. Critical Care

Critical care is mainly related to seriously impacted patients who need swift and special medical treatments. Typically, intensive monitoring and fast reaction are needed for such

patients, unlike the normal treatment for elders or chronic disease patients, who are usually less critical and need constant monitoring and medication for a long period of time. Remote surgery is one of the mission-critical healthcare applications that allows for providing a remote service from a set of consultants, who are in different places around the world, leveraging augment reality (AR), and virtual reality (VR). Since such type of applications directly deal with the patients' lives, they require communication services with ultrareliability and ultralow latency. Thus, medical data transmission and management for such critical care applications are very challenging research topics in healthcare. Much attention of future research topics should pave the way to provide high-quality healthcare-data transmission that satisfies diverse strict QoS requirements. RL algorithms can fulfil this gap by efficiently processing the acquired data in MEC servers, while minimizing the disturbance resulting from the variation of network resources utilization from other applications.

VII. Conclusion

RL is becoming increasingly popular across various domains due to its effectiveness in addressing complex optimization problems, particularly those that are large scale and highly dynamic. Thus, this article presents a comprehensive review of the recent advances in RL models and their potential applications in I-health systems to fulfil diverse healthcare QoS requirements. Specifically, we first presented the major challenges in I-health systems, while proposing our I-health system architecture that aims at addressing these challenges. Then, we discussed the fundamentals and background of diverse RL, DRL, and MARL models, while highlighting the major benefits that can be obtained by incorporating such RL models within the proposed I-health architecture. In addition, we reviewed the state-of-the-art applications of RL in three main areas: 1) edge intelligence; 2) smart core network; and 3) dynamic treatment regimes. In this context, many crucial challenges have been identified along with the RL-based solutions that have been proposed for, to mitigate these challenges in I-health systems. Finally, we presented our vision for the open challenges and problems that need to be tackled in the future, along with some research directions for innovation.

References

[1] M. S. Hossain and G. Muhammad, "Deep learning based pathology detection for smart connected Healthcare," *IEEE Netw.*, vol. 34, no. 6, pp. 120–125, Nov./Dec. 2020.

[2] A. A. Abdellatif, A. Z. Al-Marridi, A. Mohamed, A. Erbad, C. F. Chiasserini, and A. Refaey, "ssHealth: Toward secure, blockchain-enabled healthcare systems," *IEEE Netw.*, vol. 34, no. 4, pp. 312–319, Jul./Aug. 2020.

[3] T. Muhammed, R. Mehmood, A. Albeshri, and I. Katib, "UbeHealth: A personalized ubiquitous cloud and edge-enabled networked healthcare system for smart cities," *IEEE Access*, vol. 6, pp. 32258–32285, 2018.

[4] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath. "A brief survey of deep reinforcement learning." 2017. [Online]. Available: http://arxiv.org/abs/1708.05866

[5] H. Saad, A. Mohamed, and T. ElBatt, "Distributed cooperative *Q*-learning for power allocation in cognitive femtocell networks," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, 2012, pp. 1–5.

[6] K. I. Ahmed, H. Tabassum, and E. Hossain. "Deep learning for radio resource allocation in multi-cell networks." 2018. [Online]. Available: http://arxiv.org/abs/1808.00667

[7] N. Casas. "Deep deterministic policy gradient for urban traffic light control." 2017. [Online]. Available: http://arxiv.org/abs/1703.09035

[8] D. Silver et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[9] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal Policy Optimization Algorithms*, OpenAi, San Francisco, CA, USA, 2017.

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

[12] A. Alharin, T. N. Doan, and M. Sartipi, "Reinforcement learning interpretation methods: A survey," *IEEE Access*, vol. 8, pp. 171058–171077, 2020.

[13] 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.

[14] W. Chen, X. Qiu, T. Cai, H.-N. Dai, Z. Zheng, and Y. Zhang, "Deep reinforcement learning for Internet of Things: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1659–1692, 3rd Quart., 2021.

[15] L. Lei, Y. Tan, K. Zheng, S. Liu, K. Zhang, and X. Shen, "Deep reinforcement learning for autonomous Internet of Things: Model, applications and challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1722–1760, 3rd Quart., 2020.

[16] A. Uprety and D. B. Rawat, "Reinforcement learning for IoT security: A comprehensive survey," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8693–8706, Jun. 2021.

[17] 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.

[18] Y. Rizk, M. Awad, and E. W. Tunstel, "Decision making in multiagent systems: A survey," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 3, pp. 514–529, Sep. 2018.

[19] 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.

[20] A. Alwarafy, M. Abdallah, B. S. Ciftler, A. Al-Fuqaha, and M. Hamdi. "Deep reinforcement learning for radio resource allocation and management in next generation heterogeneous wireless networks: A survey." 2021. [Online]. Available: https://doi.org/10.36227/techrxiv.14672643.v1

[21] 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.

[22] F. Tang, B. Mao, Y. Kawamoto, and N. Kato, "Survey on machine learning for intelligent end-to-end communication towards 6G: From network access, routing to traffic control and streaming adaption," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1578–1598, 3rd Quart., 2021.

[23] B. R. Kiran et al., "Deep reinforcement learning for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4909–4926, Jun. 2022.

[24] O. B. Sezer, E. Dogdu, and A. M. Ozbayoglu, "Context-aware computing, learning, and big data in Internet of Things: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 1–27, Feb. 2018.

[25] F. Zantalis, G. Koulouras, S. Karabetsos, and D. Kandris, "A review of machine learning and IoT in smart transportation," *Future Internet*, vol. 11, no. 4, p. 94, 2019.

[26] T. K. Rodrigues, K. Suto, H. Nishiyama, J. Liu, and N. Kato, "Machine learning meets computation and communication control in evolving edge and cloud: Challenges and future perspective," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 38–67, 3rd Quart., 2019.

[27] D. Ravì et al., "Deep learning for health informatics," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.

[28] Y. Zhang, J. Yao, and H. Guan, "Intelligent cloud resource management with deep reinforcement learning," *IEEE Cloud Comput.*, vol. 4, no. 6, pp. 60–69, Nov./Dec. 2017.

[29] Q. Chen et al., "A survey on an emerging area: Deep learning for smart city data," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 3, no. 5, pp. 392–410, Oct. 2019.

[30] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE J. Power Energy Syst.*, vol. 4, no. 3, pp. 362–370, 2018.

[31] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019.

[32] A. Gupta and V. K. Chaurasiya, "Reinforcement learning based energy management in wireless body area network: A survey," in *Proc. IEEE Conf. Inf. Commun. Technol.*, 2019, pp. 1–6.

[33] A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola, "Reinforcement learning for intelligent healthcare applications: A survey," *Artif. Intell. Med.*, vol. 109, Sep. 2020, Art. no. 101964.

[34] Y. Li, "Deep reinforcement learning: An overview," 2017, *arXiv:1701.07274*.

[35] 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.

[36] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Machine learning for wireless networks with artificial intelligence: A tutorial on neural networks," 2017, *arXiv:1710.02913*.

[37] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. "Benchmarking deep reinforcement learning for continuous control." 2016. [Online]. Available: https://arxiv.org/pdf/1604.06778.pdf

[38] C. Yu, J. Liu, and S. Nemati, "Reinforcement learning in healthcare: A survey," 2019, *arXiv:1908.08796*.

[39] A. A. Abdellatif et al., "Medge-chain: Leveraging edge computing and blockchain for efficient medical data exchange," *IEEE Internet Things J.*, vol. 8, no. 21, pp. 15762–15775, Nov. 2021.

[40] M. Ienca and E. Vayena, "On the responsible use of digital data to tackle the COVID-19 pandemic," *Nat. Med.*, vol. 26, pp. 463–464, Mar. 2020.

[41] A. A. Abdellatif, A. Emam, C.-F. Chiasserini, A. Mohamed, A. Jaoua, and R. Ward, "Edge-based compression and classification for smart healthcare systems: Concept, implementation and evaluation," *Exp. Syst. Appl.*, vol. 117, no. 1, pp. 1–14, 2019.

[42] R. Vincent, "Reinforcement learning in models of adaptive medical treatment strategies," Ph.D. dissertation, Scool Comput. Sci., McGill Univ. Lib., Montreal QC, Canada, 2014.

[43] C. X. Ji, M. Oberst, S. Kanjilal, and D. Sontag, "Trajectory inspection: A method for iterative clinician-driven design of reinforcement learning studies," *AMIA Summits Transl. Sci. Proc.*, vol. 2021, p. 305, Jan. 2021.

[44] R. Liu, Y. Rong, and Z. Peng, "A review of medical artificial intelligence," *Global Health J.*, vol. 4, no. 2, pp. 42–45, 2020.

[45] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 9, no. 1, p. 7567, 2017.

[46] F. Ghesu et al., "Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 176–189, Jan. 2019.

[47] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

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

[49] J. A. Bagnell and J. G. Schneider, "Autonomous helicopter control using reinforcement learning policy search methods," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, vol. 2, 2001, pp. 1615–1620.

[50] G. Konidaris and A. Barto, "Skill discovery in continuous reinforcement learning domains using skill chaining," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 22, 2009, pp. 1015–1023.

[51] A. Gosavi, "Reinforcement learning: A tutorial survey and recent advances," *INFORMS J. Comput.*, vol. 21, no. 2, pp. 178–192, 2009.

[52] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2014.

[53] E. Hazan, "Introduction to online convex optimization," 2019, *arXiv:1909.05207*.

[54] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, 1988.

[55] C. J. Watkins and P. Dayan, "*Q*-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[56] D. P. Bertsekas, *Reinforcement Learning and Optimal Control*. Belmont, MA, USA: Athena Sci., 2019.

[57] T. P. Lillicrap et al. "Continuous control with deep reinforcement learning." 2015. [Online]. Available: https://arxiv.org/pdf/1509.02971v1.pdf

[58] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "A survey and critique of multiagent deep reinforcement learning," *Auton. Agents Multiagent Syst.*, vol. 33, no. 6, pp. 750–797, 2019.

[59] W. Du and S. Ding, "A survey on multi-agent deep reinforcement learning: From the perspective of challenges and applications," *Artif. Intell. Rev.*, vol. 54, pp. 3215–3238, Jun. 2021.

[60] X. Liu, J. Yu, Z. Feng, and Y. Gao, "Multi-agent reinforcement learning for resource allocation in IoT networks with edge computing," *China Commun.*, vol. 17, no. 9, pp. 220–236, 2020.

[61] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2145–2153.

[62] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep Decentralized multi-task multi-agent reinforcement learning under partial observability," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 2681–2690.

[63] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor–critic for mixed cooperative-competitive environments," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6382–6393.

[64] R. E. Wang, M. Everett, and J. P. How, "R-MADDPG for partially observable environments and limited communication," 2020, *arXiv:2002.06684*.

[65] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "QMix: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4295–4304.

[66] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi, "QTran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5887–5896.

[67] OpenAI. "Gym: A toolkit for developing and comparing reinforcement learning algorithms." Accessed: Sep. 1, 2021. [Online]. Available: https://gym.openai.com

[68] M. S. Allahham, A. A. Abdellatif, A. Mohamed, A. Erbad, E. Yaacoub, and M. Guizani, "I-SEE: Intelligent, secure, and energy-efficient techniques for medical data transmission using deep reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6454–6468, Apr. 2021.

[69] K. Chen, H. Ding, and Q. Huo, "Parallelizing ADAM optimizer with blockwise model-update filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2020, pp. 3027–3031.

[70] A. A. Abdellatif, A. Mohamed, C. F. Chiasserini, M. Tlili, and A. Erbad, "Edge computing for smart health: Context-aware approaches, opportunities, and challenges," *IEEE Netw.*, vol. 33, no. 3, pp. 196–203, May/Jun. 2019.

[71] Y. Kang et al., "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," *ACM SIGARCH Comput. Architect. News*, vol. 45, no. 1, pp. 615–629, 2017.

[72] Z. Xue et al., "A resource-constrained and privacy-preserving edge computing enabled clinical decision system: A federated reinforcement learning approach," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 9122–9138, Jun. 2021.

[73] N. Mhaisen, A. Awad, A. Mohamed, A. Erbad, and M. Guizani, "Optimal user-edge assignment in hierarchical federated learning based on statistical properties and network topology constraints," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 1, pp. 55–66, Jan./Feb. 2022.

[74] A. Emam, A. A. Abdellatif, A. Mohamed, and K. A. Harras, "EdgeHealth: An energy-efficient edge-based remote mHealth monitoring system," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2019, pp. 1–7.

[75] A. A. Abdellatif, M. G. Khafagy, A. Mohamed, and C. Chiasserini, "EEG-based transceiver design with data decomposition for Healthcare IoT applications," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3569–3579, Oct. 2018.

[76] L. Xiao, Y. Li, G. Han, G. Liu, and W. Zhuang, "PHY-layer spoofing detection with reinforcement learning in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10037–10047, Dec. 2016.

[77] Y. Xu, J. Xia, H. Wu, and L. Fan, "*Q*-learning based physical-layer secure game against Multiagent attacks," *IEEE Access*, vol. 7, pp. 49212–49222, 2019.

[78] L. Xiao, X. Lu, T. Xu, W. Zhuang, and H. Dai, "Reinforcement learning-based physical-layer authentication for controller area networks," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2535–2547, 2021.

[79] L. Xiao, G. Sheng, S. Liu, H. Dai, M. Peng, and J. Song, "Deep reinforcement learning-enabled secure visible light communication against eavesdropping," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6994–7005, Oct. 2019.

[80] B. E. ElDiwany, A. A. Abdellatif, A. Mohamed, A. Al-Ali, M. Guizani, and X. Du, "On physical layer security in energy-efficient wireless health monitoring applications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–7.

[81] L. Samara, A. Gouissem, A. A. Abdellatif, R. Hamila, and M. O. Hasna, "On the performance of tactical communication interception using military full duplex radios," in *Proc. IEEE 30th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2019, pp. 1–6.

[82] M. Min et al., "Learning-based privacy-aware offloading for healthcare IoT with energy harvesting," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4307–4316, Jun. 2019.

[83] Z. Ying, Y. Zhang, S. Cao, S. Xu, and X. Liu, "OIDPR: Optimized insulin dosage based on privacy-preserving reinforcement learning," in *Proc. IFIP Netw. Conf. (Netw.)*, 2020, pp. 655–657.

[84] A. Awad, A. Mohamed, C. Chiasserini, and T. Elfouly, "Network association with dynamic pricing over D2D-enabled heterogeneous networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2017, pp. 1–6.

[85] A. A. Abdellatif, A. Mohamed, and C. Chiasserini, "User-centric networks selection with adaptive data compression for smart health," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3618–3628, Dec. 2018.

[86] D. D. Nguyen, H. X. Nguyen, and L. B. White, "Reinforcement learning with network-assisted feedback for heterogeneous RAT selection," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6062–6076, 2017.

[87] 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. 2021.

[88] M. S. Allahham, A. A. Abdellatif, N. Mhaisen, A. Mohamed, A. Erbad, and M. Guizani, "Multi-agent reinforcement learning for network selection and resource allocation in heterogeneous multi-RAT networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 1287–1300, Jun. 2022.

[89] M. Yan, G. Feng, J. Zhou, and S. Qin, "Smart multi-RAT access based on multiagent reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4539–4551, May 2018.

[90] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2017, pp. 1–6.

[91] K. I. Ahmed and E. Hossain, "A deep Q-learning method for downlink power allocation in multi-cell networks," 2019, *arXiv:1904.13032*.

[92] C. Xu, F. X. Lin, Y. Wang, and L. Zhong, "Automated OS-level device runtime power management," *ACM SIGPLAN Notices*, vol. 50, no. 4, pp. 239–252, 2015.

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

[94] 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.

[95] Y. He et al., "Deep-reinforcement-learning-based optimization for cache-enabled opportunistic interference alignment wireless networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10433–10445, Nov. 2017.

[96] G. Cao, Z. Lu, X. Wen, T. Lei, and Z. Hu, "AIF: An artificial intelligence framework for smart wireless network management," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 400–403, Feb. 2018.

[97] L. Hu, M. Qiu, J. Song, M. S. Hossain, and A. Ghoneim, "Software defined healthcare networks," *IEEE Wireless Commun.*, vol. 22, no. 6, pp. 67–75, Dec. 2015.

[98] X. Wang, W. Wang, and Z. Jin, "Context-aware reinforcement learning-based mobile cloud computing for telemonitoring," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Inform. (BHI)*, 2018, pp. 426–429.

[99] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, "FedHome: Cloud-edge based personalized federated learning for in-home health monitoring," *IEEE Trans. Mobile Comput.*, vol. 21, no. 8, pp. 2818–2832, Aug. 2022.

[100] P. Gupta and T. Dallas, "Feature selection and activity recognition system using a single triaxial Accelerometer," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1780–1786, Jun. 2014.

[101] L. Wang, X. Zhao, Y. Si, L. Cao, and Y. Liu, "Context-associative hierarchical memory model for human activity recognition and prediction," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 646–659, Mar. 2017.

[102] K. Doddabasappla and R. Vyas, "Statistical and machine-learning based recognition of coughing events using tri-axial accelerometer sensor data from multiple wearable points," *IEEE Sensors Lett.*, vol. 5, no. 5, Jun. 2021, Art. no. 6001104.

[103] X. Zhou, W. Liang, K. I.-K. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-learning-enhanced human activity recognition for Internet of Healthcare Things," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6429–6438, Jul. 2020.

[104] S. Wijethilaka and M. Liyanage, "Survey on network slicing for Internet of Things realization in 5G networks," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 957–994, 2nd Quart., 2021.

[105] V. P. Kafle, Y. Fukushima, P. Martinez-Julia, T. Miyazawa, and H. Harai, "Adaptive virtual network slices for diverse IoT services," *IEEE Commun. Stand. Mag.*, vol. 2, no. 4, pp. 33–41, Dec. 2018.

[106] R. Pries, H.-J. Morper, N. Galambosi, and M. Jarschel, "Network as a service—A demo on 5G network slicing," in *Proc. Int. Teletraffic Congr. (ITC 28)*, vol. 1, 2016, pp. 209–211.

[107] "How 5G mobile networks are opening the door to remote surgery." Accessed: May 8, 2018. [Online]. Available: https://www.medicaldevice-network.com/features/5g-remote-surgery

[108] A. A. Abdellatif, A. Mohamed, A. Erbad, and M. Guizani, "Dynamic network slicing and resource allocation for 5G-and-beyond networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 262–267.

[109] Q. Liu and T. Han, "When network slicing meets deep reinforcement learning," in *Proc. Int. Conf. Emerg. Netw. Exp. Technol.*, 2019, pp. 29–30.

[110] 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. (GLOBECOM)*, 2020, pp. 1–6.

[111] J. Koo, V. B. Mendiratta, M. R. Rahman, and A. Walid, "Deep reinforcement learning for network slicing with heterogeneous resource requirements and time varying traffic dynamics," in *Proc. 15th Int. Conf. Netw. Service Manag. (CNSM)*, 2019, pp. 1–5.

[112] D. Silver et al., "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[113] Y. Kim, S. Kim, and H. Lim, "Reinforcement learning based resource management for network slicing," *Appl. Sci.*, vol. 9, no. 11, p. 2361, 2019.

[114] J. Ni, X. Lin, and X. S. Shen, "Efficient and secure service-oriented authentication supporting network slicing for 5G-enabled IoT," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 644–657, Mar. 2018.

[115] J. P. De Brito Gonçalves, H. C. De Resende, E. Municio, R. Villaça, and J. M. Marquez-Barja, "Securing E-health networks by applying network slicing and blockchain techniques," in *Proc. IEEE 18th Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2021, pp. 1–2.

[116] A. Mathew, "Network slicing in 5G and the security concerns," in *Proc. 4th Int. Conf. Comput. Methodol. Commun. (ICCMC)*, 2020, pp. 75–78.

[117] W. Xiangyu, J. Ma, M. Yinbin, X. Liu, and Y. Ruikang, "Privacy-preserving diverse keyword search and online pre-diagnosis in cloud computing," *IEEE Trans. Services Comput.*, vol. 5, no. 2, pp. 710–723, Apr. 2022.

[118] X. Yao, Y. Lin, Q. Liu, and J. Zhang, "Privacy-preserving search over encrypted personal health record in multi-source cloud," *IEEE Access*, vol. 6, pp. 3809–3823, 2018.

[119] "Healthcare report for 1st half of 2018." Accessed: Dec. 5, 2020. [Online]. Available: https://www.cryptonitenxt.com/resources

[120] "Note to nations: Stop hacking hospitals." Accessed: Dec. 5, 2020. [Online]. Available: https://foreignpolicy.com/2020/04/06/coronavirus-cyberattack-stop-hacking-hospitals-cyber-norms

[121] T. T. Nguyen and V. J. Reddi, "Deep reinforcement learning for cyber security," 2019, *arXiv:1906.05799*.

[122] X. Liu, R. H. Deng, K.-K. R. Choo, and Y. Yang, "Privacy-preserving reinforcement learning design for patient-centric dynamic treatment regimes," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 1, pp. 456–470, Jan.–Mar. 2021.

[123] Y. Yamagata, S. Liu, T. Akazaki, Y. Duan, and J. Hao, "Falsification of cyber-physical systems using deep reinforcement learning," *IEEE Trans. Softw. Eng.*, vol. 47, no. 2, pp. 2823–2840, Dec. 2021.

[124] H. Abbas and G. Fainekos, "Convergence proofs for simulated annealing falsification of safety properties," in *Proc. IEEE 50th Annu. Allerton Conf. Commun. Control Comput. (Allerton)*, 2012, pp. 1594–1601.

[125] S. Sankaranarayanan and G. Fainekos, "Falsification of temporal properties of hybrid systems using the cross-entropy method," in *Proc. 15th ACM Int. Conf. Hybrid Syst. Comput. Control*, 2012, pp. 125–134.

[126] L. Xiao, X. Wan, C. Dai, X. Du, X. Chen, and M. Guizani, "Security in mobile edge caching with reinforcement learning," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 116–122, Jun. 2018.

[127] H. Zhu, Y. Cao, W. Wang, T. Jiang, and S. Jin, "Deep reinforcement learning for mobile edge caching: Review, new features, and open issues," *IEEE Netw.*, vol. 32, no. 6, pp. 50–57, Nov./Dec. 2018.

[128] L. Xiao, Y. Li, G. Liu, Q. Li, and W. Zhuang, "Spoofing detection with reinforcement learning in wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2015, pp. 1–5.

[129] Y. He, F. R. Yu, N. Zhao, V. C. Leung, and H. Yin, "Software-defined networks with mobile edge computing and caching for smart cities: A big data deep reinforcement learning approach," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 31–37, Dec. 2017.

[130] E. B. Laber, K. A. Linn, and L. A. Stefanski, "Interactive model building for *Q*-learning," *Biometrika*, vol. 101, no. 4, pp. 831–847, 2014.

[131] K. A. Linn, E. B. Laber, and L. A. Stefanski, "Interactive *Q*-learning for quantiles," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 638–649, 2017.

[132] P. J. Schulte, A. A. Tsiatis, E. B. Laber, and M. Davidian, "*Q*- and A-learning methods for estimating optimal dynamic treatment regimes," *Stat. Sci. Rev. J. Inst. Math. Stat.*, vol. 29, no. 4, p. 640, 2014.

[133] R. Song, W. Wang, D. Zeng, and M. R. Kosorok, "Penalized *Q*-learning for dynamic treatment regimens," *Statistica Sinica*, vol. 25, no. 3, p. 901, 2015.

[134] Y. Liu, Y. Wang, M. R. Kosorok, Y. Zhao, and D. Zeng, "Robust hybrid learning for estimating personalized dynamic treatment regimens," 2016, *arXiv:1611.02314*.

[135] K. Deng, R. Greiner, and S. Murphy, "Budgeted learning for developing personalized treatment," in *Proc. Int. Conf. Mach. Learn. Appl.*, 2014, pp. 7–14.

[136] P. W. Lavori and R. Dawson, "Adaptive treatment strategies in chronic disease," *Annu. Rev. Med.*, vol. 59, pp. 443–453, 2008.

[137] *Preventing Chronic Diseases: A Vital Investment: WHO Global Report*, World Health Org., Geneva, Switzerland, 2005.

[138] A. A. Abdellatif, A. Mohamed, C. F. Chiasserini, A. Erbad, and M. Guizani, "Edge computing for energy-efficient smart health systems: Data and application-specific approaches," in *Energy Efficiency of Medical Devices and Healthcare Applications*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 53–67.

[139] N. Cho et al., "IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, Apr. 2018.

[140] M. K. Bothe et al., "The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas," *Exp. Rev. Med. Devices*, vol. 10, no. 5, pp. 661–673, 2013.

[141] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Basal glucose control in type 1 diabetes using deep reinforcement learning: An in Silico validation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 4, pp. 1223–1232, Apr. 2021.

[142] S. Yasini, M. B. Naghibi-Sistani, and A. Karimpour, "Agent-based simulation for blood glucose," *Int. J. Appl. Sci. Eng. Technol.*, vol. 3, no. 9, pp. 1–8, 2009.

[143] A. E. Gaweda, M. K. Muezzinoglu, G. R. Aronoff, A. A. Jacobs, J. M. Zurada, and M. E. Brier, "Reinforcement learning approach to individualization of chronic pharmacotherapy," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 5, 2005, pp. 3290–3295.

[144] S. Wang, W. A. Chaovalitwongse, and S. Wong, "Online seizure prediction using an adaptive learning approach," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 12, pp. 2854–2866, Dec. 2013.

[145] A. Guez, "Adaptive control of epileptic seizures using reinforcement learning," Ph.D. dissertation, Appl. Sci. Artif. Intell., McGill Univ., Montreal, QC, Canada, 2010.

[146] D. J. Luckett, E. B. Laber, A. R. Kahkoska, D. M. Maahs, E. Mayer-Davis, and M. R. Kosorok, "Estimating dynamic treatment regimes in mobile health using *v*-learning," *J. Amer. Stat. Assoc.*, vol. 115, no. 530, pp. 692–706, 2020.

[147] A. Guez, R. D. Vincent, M. Avoli, and J. Pineau, "Adaptive treatment of epilepsy via batch-mode reinforcement learning," in *Proc. AAAI*, 2008, pp. 1671–1678.

[148] Y. Zhao, M. R. Kosorok, and D. Zeng, "Reinforcement learning design for cancer clinical trials," *Stat. Med.*, vol. 28, no. 26, pp. 3294–3315, 2009.

[149] I. Ahn and J. Park, "Drug scheduling of cancer chemotherapy based on natural actor-critic approach," *BioSystems*, vol. 106, nos. 2–3, pp. 121–129, 2011.

[150] A. Jalalimanesh, H. S. Haghighi, A. Ahmadi, and M. Soltani, "Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning," *Math. Comput. Simulat.*, vol. 133, pp. 235–248, Mar. 2017.

[151] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," *Nat. Med.*, vol. 24, no. 11, pp. 1716–1720, 2018.

[152] R. Padmanabhan, N. Meskin, and W. M. Haddad, "Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning," *Biomed. Signal Process. Control*, vol. 22, pp. 54–64, Sep. 2015.

[153] J. Peters and S. Schaal, "Natural actor–critic," *Neurocomputing*, vol. 71, nos. 7–9, pp. 1180–1190, 2008.

[154] R. Padmanabhan, N. Meskin, and W. M. Haddad, "Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment," *Math. Biosci.*, vol. 293, pp. 11–20, Nov. 2017.

[155] H.-H. Tseng, Y. Luo, S. Cui, J.-T. Chien, R. K. T. Haken, and I. E. Naqa, "Deep reinforcement learning for automated radiation adaptation in lung cancer," *Med. Phys.*, vol. 44, no. 12, pp. 6690–6705, 2017.

[156] Z. Li and Y. Xia, "Deep reinforcement learning for weakly-supervised lymph node segmentation in CT images," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 3, pp. 774–783, Mar. 2021.

[157] M. Komorowski, A. Gordon, L. Celi, and A. Faisal, "A Markov decision process to suggest optimal treatment of severe infections in intensive care," in *Proc. Neural Inf. Process. Syst. Workshop Mach. Learn. Health*, 2016, pp. 1–8.

[158] A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi, "Deep reinforcement learning for sepsis treatment," 2017, *arXiv:1711.09602*.

[159] C. Yu, G. Ren, and J. Liu, "Deep inverse reinforcement learning for sepsis treatment," in *Proc. IEEE Int. Conf. Healthcare Inf. (ICHI)*, 2019, pp. 1–3.

[160] L. Li, M. Komorowski, and A. A. Faisal, "The actor search tree critic (ASTC) for off-policy POMDP learning in medical decision making," 2018, *arXiv:1805.11548*.

[161] C. P. Utomo, X. Li, and W. Chen, "Treatment recommendation in critical care: A scalable and interpretable approach in partially observable health states," in *Proc. Int. Conf. Interact. Sci.*, 2018, p. 9.

[162] A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, and M. Ghassemi, "Continuous state-space models for optimal sepsis treatment—A deep reinforcement learning approach," 2017, *arXiv:1705.08422*.

[163] X. Peng et al., "Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning," in *Proc. AMIA Annu. Symp.*, 2018, p. 887.

[164] B. K. Petersen et al., "Precision medicine as a control problem: Using simulation and deep reinforcement learning to discover adaptive, personalized multi-cytokine therapy for sepsis," 2018, *arXiv:1802.10440*.

[165] W. M. Haddad et al., "Clinical decision support and closed-loop control for intensive care unit sedation," *Asian J. Control*, vol. 15, no. 2, pp. 317–339, 2013.

[166] N. Sadati, A. Aflaki, and M. Jahed, "Multivariable anesthesia control using reinforcement learning," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, vol. 6, 2006, pp. 4563–4568.

[167] E. C. Borera, B. L. Moore, A. G. Doufas, and L. D. Pyeatt, "An adaptive neural network filter for improved patient state estimation in closed-loop anesthesia control," in *Proc. Int. Conf. Tools Artif. Intell.*, 2011, pp. 41–46.

[168] L. Lovén et al., "EdgeAI: A vision for distributed, edgenative artificial intelligence in future 6G networks," in *Proc. 1st 6G Wireless Summit*, 2019, pp. 1–2.

[169] G. Paragliola and A. Coronato, "Definition of a novel federated learning approach to reduce communication costs," *Exp. Syst. Appl.*, vol. 189, Mar. 2022, Art. no. 116109.

[170] G. Paragliola, "Evaluation of the trade-off between performance and communication costs in federated learning scenario," *Future Gener. Comput. Syst.*, vol. 136, pp. 282–293, Nov. 2022.

[171] G. Paragliola, "A federated learning-based approach to recognize subjects at a high risk of hypertension in a non-stationary scenario," *Inf. Sci.*, vol. 622, pp. 16–33, Apr. 2023.

[172] S. Yu, X. Chen, Z. Zhou, X. Gong, and D. Wu, "When deep reinforcement learning meets federated learning: Intelligent Multitimescale resource management for multiaccess edge computing in 5G ultra-dense network," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2238–2251, Feb. 2021.

[173] B. Liu, C. Liu, and M. Peng, "Resource allocation for energy-efficient MEC in NOMA-enabled massive IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1015–1027, Apr. 2021.

[174] J. Zhao, Q. Li, Y. Gong, and K. Zhang, "Computation offloading and resource allocation for cloud assisted mobile edge computing in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7944–7956, Aug. 2019.

[175] Z. Zhu, S. Wan, P. Fan, and K. B. Letaief, "Federated multi-agent actor–critic learning for age sensitive mobile edge computing," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1053–1067, Jan. 2022.

[176] S. Lee and D.-H. Choi, "Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 488–497, Jan. 2022.

[177] M. Xu et al., "Multi-agent federated reinforcement learning for secure incentive mechanism in intelligent cyber-physical systems," *IEEE Internet Things J.*, vol. 9, no. 22, pp. 22095–22108, Nov. 2022.

[178] F. Majidi, M. R. Khayyambashi, and B. Barekatain, "HFDRL: An intelligent dynamic cooperate cashing method based on hierarchical federated deep reinforcement learning in edge-enabled IoT," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1402–1413, Jan. 2022.

[179] Y.-J. Liu, G. Feng, Y. Sun, S. Qin, and Y.-C. Liang, "Device association for RAN slicing based on hybrid federated deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15731–15745, Dec. 2020.

[180] Y. Sun et al., "Efficient handover mechanism for radio access network slicing by exploiting distributed learning," *IEEE Trans. Netw. Service Manag.*, vol. 17, no. 4, pp. 2620–2633, Dec. 2020.

[181] Q. Zhang, J. Liu, and G. Zhao, "Towards 5G enabled tactile robotic telesurgery," 2018, *arXiv:1803.03586*.

[182] A. A. Abdellatif, C. F. Chiasserini, and F. Malandrino, "Active learning-based classification in automated connected vehicles," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2020, pp. 598–603.

[183] A. A. Abdellatif, C. F. Chiasserini, F. Malandrino, A. Mohamed, and A. Erbad, "Active learning with noisy Labelers for improving classification accuracy of connected vehicles," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3059–3070, Apr. 2021.

[184] A. A. Abdellatif, A. Mohamed, and C.-F. Chiasserini, "Concurrent association in heterogeneous networks with underlay D2D communication," in *Proc. Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, 2017, pp. 56–61.

[185] B. Houtan, A. S. Hafid, and D. Makrakis, "A survey on blockchain-based self-sovereign patient identity in healthcare," *IEEE Access*, vol. 8, pp. 90478–90494, 2020.

[186] Y. A. Qadri, A. Nauman, Y. B. Zikria, A. V. Vasilakos, and S. W. Kim, "The future of healthcare Internet of Things: A survey of emerging technologies," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1121–1167, 2nd Quart., 2020.

[187] S. Shalaby, A. A. Abdellatif, A. Al-Ali, A. Mohamed, A. Erbad, and M. Guizani, "Performance evaluation of hyperledger fabric," in *Proc. IEEE Int. Conf. Inf. IoT Enabling Technol. (ICIoT)*, 2020, pp. 608–613.

[188] R. Wang, X. He, R. Yu, W. Qiu, B. An, and Z. Rabinovich, "Learning efficient multi-agent communication: An information bottleneck approach," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9908–9918.

[189] H. Mao, Z. Zhang, Z. Xiao, Z. Gong, and Y. Ni, "Learning agent communication under limited bandwidth by message pruning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 5142–5149.