IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 August 2023, accepted 15 October 2023, date of publication 20 October 2023, date of current version 27 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3326432

RESEARCH ARTICLE

Channel Allocation to GAA Users Using Double Deep Recurrent Q-Learning Based on Double Auction Method

WASEEM ABBASS[1], RIAZ HUSSAIN[2], NASIM ABBAS[3], SHAHZAD A. MALIK[2], MUHAMMAD AWAIS JAVED[2], (Senior Member, IEEE), MUHAMMAD ZUBAIR KHAN[4], RAYAN HAMZA ALSISI[5], ABDULFATTAH NOORWALI[6], (Senior Member, IEEE), AND PRIYADARSHINI PATTANAIK[7]

[1]Department of Electrical and Computer Engineering, Capital University of Science and Technology (CUST), Islamabad 45750, Pakistan
[2]Department of Electrical and Computer Engineering, COMSATS University, Islamabad 45550, Pakistan
[3]Department of Computer Science, Muslim Youth University, Islamabad 45710, Pakistan
[4]Department of Computer Science and Information, Taibah University, Medina 42353, Saudi Arabia
[5]Department of Electrical Engineering, Islamic University of Madinah, Madinah 41411, Saudi Arabia
[6]Department of Electrical Engineering, Umm Al-Qura University, Makkah 21961, Saudi Arabia
[7]Faculty of Computer Science and Informatics, Berlin School of Business and Innovation (BSBI), 12043 Berlin, Germany

Corresponding authors: Mohammad Zubair Khan (mkhanb@taibahu.edu.sa) and Rayan Hamza Alsisi (ralsisi@iu.edu.sa)

This work was supported by the Deputyship for Research & Innovation, Ministry of Education, Saudi Arabia, under Project IFP22UQU4290235DSR257.

ABSTRACT The SAS-CBRS framework is being tested to share the federally held spectrum with licensed users and opportunistic users to maximize the underutilized spectrum's utility and overcome spectrum scarcity. In the SAS-CBRS framework, radio resources are assigned to the incumbent access (IA), primary access licensees (PAL), and general authorized access (GAA) users according to the given priority. The SAS-CBRS three-tier framework is different from the conventional cognitive radio networks (CRN) as it involves a central entity that acts as a server called a spectrum access system (SAS). The methods to assign the resources using the SAS are still in the research phase. Yet, no standard method is defined by the FCC for resource allocation. The current CRN methods cannot be directly applied because of the addition of the third tier and a central server. Moreover, strict rules are defined for using the 3.5 GHz spectrum band for communication. In this paper, a novel DDRQ-SAS algorithm integrated with the double auction (DA) algorithm is proposed that uses deep recurrent double Q-learning. The DDRQ-SAS is used by the SAS to hold a spectrum auction and create a spectrum pool to get information on PAL channels. PAL operators use the DA algorithm to generate the asking prices intelligently for their available idle channels and the GAA users will use the DA algorithm to intelligently bid for their preferred channels. The DDRQ-SAS-DA algorithm allows the GAA users to get the guaranteed QoS offered by the PAL operators in an auction. GAA users maintain the preference list of the PAL reserved idle channels and bid intelligently based on the available QoS. SAS completes the transaction by allocating the channels to the winning GAAs. The defined problem is also modeled using the double auction multi-winner multi-channel technique and the TDSA-PS algorithm. Numerical results show that the proposed DDRQ-SAS-DA algorithm provides up to 20% better QoS at higher loads for GAA users, generates 24% more revenue for PAL operators, and is 1.6 times more efficient in assigning 500 GAA users.

INDEX TERMS SAS-CBRS, double auction algorithm, deep learning, Q-learning, channel allocation.

I. INTRODUCTION

The rapid increase in the wireless dynamic interactive applications, services, and the astronomical growth of the

The associate editor coordinating the review of this manuscript and approving it for publication was Tiago Cruz.

internet of things IoT devices caused the enormous transmission of data over conventional cellular networks, creating severe spectrum scarcity [1]. Moreover, the current static allocation techniques are unable to meet the overwhelming bandwidth requirements of bandwidth-hungry applications [2]. The radio frequency spectrum is a finite and strictly bounded resource and most of the spectrum available for commercial users is underutilized because of inefficient static allocation techniques [3]. In response to the President's council of advisors on science and technology (PCAST) report [4], the Federal communications commission (FCC) identified 1473 MHz federally-held spectrum, which can be shared with commercial users to cope with the challenge of spectrum scarcity [5]. The 2.2 GHz spectrum band is used to test 5G services in Europe, known as the licensed spectrum access (LSA) system, while the 150 MHz federally held spectrum band (3550 MHz-3700 MHz) called the citizens broadband radio service (CBRS) band. This band is currently used for the testing of 5G services in the United States [6]. The CBRS radio spectrum band comprises a spectrum access system (SAS) based centralized server to share the available 150 MHz spectrum with commercial users in presence of the federal military users including the Naval radars and the Satellite earth stations [7].

The three-tier SAS-based CBRS system is a priority-based system in which the highest priority is given to the federal users known as the incumbent access (IA) users, the second tier exists with the primary access licensees (PAL) users. In the CBRS framework, opportunistic users are also provided with a dedicated spectrum band to communicate. These opportunistic users are the least priority users known as general authorized access (GAA) users [8]. The 150 MHz spectrum band is shared with commercial users and allocated through competitive auction. 70 MHz radio frequency spectrum band is reserved for the PAL operator's use and the rest of the 80 MHz spectrum band is opened for opportunistic use dedicated to the GAA users [9]. The IA users may use the complete 150 MHz radio frequency band and the SAS will provide free channels to the IA users with strict protection. Moreover, the licensed PALs are kept safe from harmful interference caused by opportunistic GAA users. The SAS will provide the imperative quality of service (QoS) to high-priority users. The SAS will not provide the GAA users with the services as per their requirements. The GAA users have been given the ability to access the PAL-reserved 70 MHz channels opportunistically if a transmission opportunity is available [10]. The 70 MHz frequency spectrum for PAL use is split into seven fragments of 10 MHz each. A PAL operator can apply for up to four bands of 10 MHz in a census tract i.e., a geographical location [6].

There are numerous challenges associated with the implementation of the recently proposed SAS-CBRS framework. Some of the important challenges are the radio frequency channel assignment, maintaining the priority of the users, interference management, heterogeneous co-existence of users, operational security of the IA users, and the protocols for the FCC to standardize the framework [11]. In this paper, we consider the centralized dynamic spectrum access (DSA) architecture with different PAL operators providing services to PAL users in presence of GAA users. The SAS can detect the states of a channel, i.e., either the channel is busy or unused at a given interval. This information is used by the SAS to assign channels to the asking users. The PAL users cannot perceive the channel states, so the chance of collisions of PAL users transmission from the PAL users is not possible. The GAA users have the cognitive ability and access the PAL reserved channels opportunistically. The transmission of the GAAs will be successful only if the net interference remains under the threshold limit after the assignment. The transmission of the GAAs will be failed if both users transmit on the same channel. This scenario will be refrained by the SAS and it will ensure that the transmission from the GAA users will not get affected by the GAA users. The information on PAL-reserved idle channels, that are to be auctioned is accessed by the SAS by managing a pool of spectrum.

The Markov decision process (MDP) [12] is used in these situations because of its efficient decision power that is based on reinforcement learning (RL). The purpose of using RL is to map the set of state spaces to the action spaces in an environment where the environmental characteristics are unexplored by the agent. However, in the scenarios of the large state spaces, this method is not used because most of the states remain unexplored due to which a generalized solution is not achieved. Moreover, to solve the shortcomings of the traditional reinforcement learning based techniques, the deep learning method i.e., google Deep Mind combined with the reinforcement learning known as the Deep Q-Networks (DQN) is used [13]. The DQN solutions are based on the approximate value function to choose an optimal policy. However, the DQN overestimates the value function by utilizing the correlative value function for selecting the particular action and estimating the value function. The inadequacies of the DQN lead to the derivation of the double DQN [14] to eliminate the overestimation of the optimal learning policy. Along with the double DQN, the deep recurrent DQN (DRQN) [15] also stabilizes the time sequence problems.

In this article, we proposed the DDRQ-SAS algorithm that uses both the double DQN for stabilizing the overestimation and the DRQN for fetching particular sequential information. We also proposed a novel double auction DA algorithm based on reinforcement learning to efficiently bid for the available radio frequency channels and then allocate the channels to the GAA users using the proposed DDRQ-SAS algorithm. We model the SAS-CBRS framework as a network of intelligent agents i.e., PAL operators and the GAAs that can sense the environment having multiple states and choose actions according to the particular state of the environment. By taking the benefits of reinforcement learning, the objective is to find an efficient as well as a

stable policy to assign the channels to SAS-CBRS users by maximizing the reward function received after each action. The DA algorithm helps to converge the channel allocation of SAS-CBRS users efficiently by discovering the actions that result in getting a maximum reward. The DA algorithm allows the PAL operators to auction the available idle spectrum dynamically according to the demand of the GAA users. Moreover, it enhances the GAA's ability to bid efficiently by exploring the environment states. The scenario of the SAS-CBRS framework is categorized into licensed and unlicensed portions in which a licensed portion is accessible by the opportunistic users but the unlicensed portion can only be accessed by the unlicensed users. Hence, the on-off policy of GAA users in the PAL reserved portion makes it difficult to increase the capacity of the system as well as the management of the spectrum to protect the QoS requirements of PAL users. The DA algorithm will ensure that only GAA users that need to get guaranteed QoS will be engaged in the auction process which will also automate the tasks of SAS to authenticate the users. Hence, the DDQR-SAS-DA algorithm maximizes the spectrum utilization of the CBRS band as per the rules proposed by the FCC.

Above all, the key objectives achieved in this paper are listed below.

1) We proposed the DDRQ-SAS model based on the double deep Q-network and the deep recurrent Q-network to evade the conflicts between the GAA users while considering the probability of every channel unoccupied channel and available environment states.
2) Double auction DA algorithm based on reinforcement learning is proposed to improve the spectral efficiency and the channel capacity while considering the optimal bidding strategy.
3) The proposed DDRQ-SAS-DA algorithm uses long short-term memory (LSTM) network instead of evaluating the lookup table of value function to enhance the temporal channel allocation policies. LSTM in the DDRQ-SAS-DA algorithm enhances the capacity to handle temporal dependencies and its potential to enhance policy learning and decision-making in the context of temporal channel allocation problems. By using LSTM instead of a lookup table, we achieve more efficient, adaptive, and effective channel allocation policies, leading to improved overall performance.
4) The scenario is also modeled using the double auction framework and the TSDA-PS algorithm to assign channels.
5) Comprehensive simulation analysis and investigations are represented to show the performance of the proposed DDRQ-SAS-DA algorithm compared with the double auction framework and the TSDA-PS algorithm.

The rest of this paper is organized as follows: The related work is summarized in Section II. The SAS architecture and the detailed problem formulation of the DDRQ-SAS algorithm and the DA algorithm are discussed in Section III. In Section IV, the proposed solution DDRQ-SAS is discussed in detail with the DA algorithm, Double auction multi-winning-framework, and the TSDA-PS algorithm. Section V presents a detailed comparison of the proposed DDRQ-SAS and DA algorithm with the competing algorithms and shows that it outperforms other algorithms. Finally, the conclusion of our findings is discussed in Section VI

II. RELATED WORK

In recent years, the problem of channel assignment to opportunistic users in presence of the licensed commercial users and federally administered users has been the subject of intensive investigations. Machine learning (ML) based solutions for dynamic spectrum access are massively recognized in wireless communications because of their decision power for unknown environments. The main emphasis of using ML techniques is to derive the adaptive mechanisms to fairly distribute the radio resources. Moreover, significant work is being done in the domains of interference management, admission control, the coexistence of heterogeneous radio access technologies, spectrum pricing, the privacy of the incumbents, and the protocols for the 5G and beyond communications [16].

The authors in [17] and [18] used the ML-based MDP process to model the coexistence of radar communications with commercial users to investigate the channel allocation problem by applying the policy iteration method. The policy iteration methods exploit the transition probabilities with the reward functions. On the contrary, the Q-learning algorithm is a model-free algorithm that learns from trial and error without depending on the environment model. The authors in [19], [20], [21], and [22] use the off-policy algorithms to achieve the optimal solutions using the state-action pairs by maintaining the Q-table. There are numerous advantages of using the off-policy methods i.e., reduced computational complexity and the convergence to an optimal policy without prior knowledge. However, in the scenario of large state spaces, the computational complexity to manage the Q-table is directly proportional to the scale of state spaces. To eliminate the issues of Q-learning based on reinforcement learning, the DQN method based on the neural networks was used to improve the approximation of the value table.

Authors in [23] used the concept of deep reinforcement learning for the power allocation problem in the D2D domain to tackle the non-cooperative problem scenario. We used a double deep recurrent Q- network based on reinforcement learning techniques to model the non-cooperative problem. Authors in [24], [25], and [26] considered the auction scenarios in which the network operators auctioned the idle channels to generate revenue. However, the common practice in realistic scenarios considers both the auctioneer and bidder in an auction. Consequently, a double auction is required to enable a complete auction process in which a network operator specifies its cost preference at a particular time while the bidders take the QoS requirements offered by the

network operators. The McAfee auction proposed in [27] is an example of a double auction method that sells similar items generating profits for the sellers. However, it does not allow spectrum reuse. In [28] the authors proposed decoupled auction for the operators and the bidders for spectrum allocation. The graph-theory-based solution is used on the bidder's side and calculates the price for each subgraph. The traditional auction mechanism to sell the items is used on the seller's side. The purpose of using this technique is to achieve truthfulness and high profits. However, the solution is not viable in heterogeneous items like in the SAS-CBRS constrained environment in which the GAA users are of different types and directly competing with each other in the absence of IA and PAL users. Moreover, authors in [28], [29], [30], and [31] considered the double auction scenarios where only a single channel was to be auctioned by the PAL operators but if there will be multiple idle channels available then PAL operators need to lease all the available idle channels. Furthermore, the throughput of the conventional cognitive networks was improved using multi-channel allocation techniques. In the scenarios where a PAL user wants to take the assigned channel back and SU has to vacate the channel, then the communication of SUs gets terminated. To evade this severe switching overhead at the operator's end, the availability time of channels to the SUs for the lease must be taken into account. However, there is no distinguished work published that considered the double auction algorithm while taking these constraints into an account. Thus, the network performance is degraded remorselessly.

Deep reinforcement learning-based mechanisms are utilized by the authors in [32] and [33] to assign the channel to the single opportunistic user in multiple correlated licensed channels. However, the states' information is not completely available to the opportunistic users instead they learn through the deep Q-network to map the action space with the available environment states. It is not feasible in the scenario of the SAS-based CBRS architectures with very large state space. Authors in [34] proposed the concept of the Recurrent neural network (RNN) to solve the computational games for partial observations based on MDPs called (POMDP). The authors in [35], [36], and [37] integrated the LSTM with RNN to solve allocation problems in the DSA to maintain the sequence information along with the internal states. The authors investigated the proposed DSA scenario in absence of the primary users and developed the DRQN for the opportunistic users to learn only good policies which is not a practical approach adopted by the DSA in 5G and beyond communications. In this work, we considered a scenario with many GAA users in presence of the multiple PAL users. Moreover, the information is not exchanged between GAA and PAL users. However, the SAS has been equipped with the cognitive capability to sense the information from the PAL operators. Furthermore, the SAS-CBRS framework is considered the POMDP in which every GAA is capable of sensing state information but in real scenarios, the GAA users can sense the states of multiple channels in each time slot. The POMDP scenarios were discussed in detail in [38] and [39].

Moreover, significant research has been done in recent years to assign the channels in a heterogeneous environment where the coexistence of PAL and GAA users is considered. The authors in [40] proposed the concept of game theory with RL to develop the hybrid MAC to reduce collisions between secondary users. However, the proposed techniques cannot be implemented in the SAS-CBRS-based system because of its inefficiency in a real dynamic environment. The authors published a survey [41] that investigates the use of the double deep recurrent Q-network in the domain of 5G and beyond 5G. The authors evaluated the ML techniques in the context of RAN slicing. This work can be further extended to be used in the 3.5 GHz CBRS-SAS framework and LSA 2.2 GHz framework.

The authors in [42] proposed the concepts of graph theory integrated with reinforcement learning but the solutions are only acceptable for simple static allocation scenarios with no PAL users. In [43] multi-slot sensing mechanism based on Bayesian fusion integrated with RL is investigated. The proposed method achieves higher detection probability with decreased error probability. Thompson sampling is used to find the state information efficiently by integrating reinforcement learning for all channels. Furthermore, we modeled the channel allocation problem as a multi-objective optimization problem in [44] and proposed the SAS-QLA algorithm in [45] based on reinforcement learning to assign the channels to the GAA users through a competitive auction process where the SAS acts as auctioneer. The field trials and the hardware experiments are investigated in [46] and [47] without considering the collision scenarios of GAA-to-GAA users and GAA-to-PAL users. The concept of decentralized SAS using blockchain technology is proposed in [11]. However, the computational complexity of separating the entities of the SAS will make it difficult to scale the networks. Authors [48] allocated the radio resources to heterogeneous technologies i.e., cellular networks, the WiMax, and Wi-Fi. The authors in [28], [29], [30], and [49] proposed the concept of radio resource allocation to secondary users through spectrum trading using the concepts of spectrum auction. However, the operators offer a single licensed channel to be part of the auction process. In more realistic scenarios the GAA users can access more than one PAL channel based on their QoS requirements. Neural networks are considered in many studies of SAS-CBRS. The authors in [10] proposed the concept of getting extra information about the PAL users apart from the sensing information of the GAA users. The authors modeled the PAL user's features as the multivariable unordered time series to predict the spectrum information. The authors in [50] proposed to detect the spectrum resources using ML-based clustering techniques. The authors also investigated the use of Q-learning, deep learning, kernel-based learning, and transfer-based learning.

The use of double-deep recurrent reinforcement learning is preferred over non-learning methods including dynamic programming, approximation algorithms, divide and conquer algorithms, and the integer linear programming algorithm because of its ability to handle large state spaces, and learning algorithms easily deal with continuous state and action space. The adaptability to give better results in complex environments like the scenario of the CBRS-SAS system that includes three tiers. The choice of using the double deep learning algorithm also depends on the requirements of the scenario where it is to be implemented and to get the user-centric results or to improve the system for network operators, learning algorithms will be the first choice to be used because of the adaptability to the complex environment. The time complexity of non-learning algorithms in a scenario of dynamic environments is not very impressive and computational time grows exponentially as the size of the input increases. Hence, to get adapted to the complex dynamic environments, it will be always preferred to use the learning mechanisms.

As mentioned above, significant research has been published in the domain of spectrum allocation and management for three-tier SAS-CBRS systems. Moreover, the DQN-based solutions were widely accepted for spectrum allocation but the overestimation of value tables does not make it a perfect choice for three-tier spectrum allocation. So we proposed the double deep recurrent reinforcement learning for the spectrum allocation to users in the three-tier SAS-CBRS framework which stabilizes the overestimation of the value table obtained for optimal policy.

III. SYSTEM MODEL

A. NETWORK SCENARIO

The CBRS SAS architecture protects current users while ensuring quick and safe access to the shared spectrum. The dynamic management of spectrum allocation, enforcement of operational parameters, and continuous monitoring and coordination mechanisms contribute to maximizing the utilization of the CBRS band for various wireless communication applications. To ensure efficient and interference-free utilization of the spectrum, the CBRS utilizes a SAS architecture. The SAS serves as a central entity responsible for managing and coordinating access to the CBRS spectrum. The detailed SAS architecture is presented in Figure 1. The CBRS radio spectrum 3.5 GHz band is governed by the centralized SAS, which is responsible for authorizing the allocation of radio frequency channels. GAA users are given an opportunity to use the licensed channels, and the SAS has cognitive abilities to sense transmission opportunities in these channels. Environmental sensing capability (ESC) sensors detect incumbent users, and if IA user activity is detected, the SAS is notified and will free up channels from low-priority users. Every user in the CBRS band transmits and receives data through citizen broadband radio service devices (CBSD) known as enodeBs, which are capable of transmitting and receiving data in the 3.5 GHz frequency band. The SAS registers CBSDs and maintains information for registered CBSDs, PAL users, and GAA users in external databases. To manage large networks, the SAS can utilize either the network management system NMS or the element management system in conjunction with domain proxies.

The SAS maintains a comprehensive record of all users in FCC-supported databases. Of the 150 MHz radio spectrum band, 70 MHz is reserved for PAL users who are responsible for operating this spectrum band. The activity of the licensed users in the licensed band is very limited, this makes an opportunity for the network operators to improve the spectrum utilization and increase revenue. Moreover, the reserved band for the GAA user is overcrowded due to the provisioning of free services by operators. If the GAA users want to get guaranteed QoS for the required services, they need to get access to the licensed channels. To ensure a successful transaction, the interference caused by GAA users to PAL users and other GAA users should be within the defined threshold limits. The PAL reserved channel is modeled using a two-state Markov chain where the channel is either occupied or idle. The state of the GAA users taking part in an auction process is taken as active as they have data to transmit.

B. PROBLEM FORMULATION

SAS-based CBRS environment is considered in which z PAL-reserved channels represented by

$$p = \{1, 2, \ldots..z\},$$

are available. There are μ PAL users in the system are represented by

$$u = \{1, 2, \ldots..\mu\},$$

The idle channels are considered non-overlapping. There are ν GAA users given by

$$g = \{1, 2, \ldots.., \nu\},$$

and the GAA users always have some data to transmit. The system model is depicted in Figure 2. The data transmission of the GAA users will be considered successful if the ν^{th} GAA user will send the data using the PAL reserved channel allocated to it and the particular channel is idle i.e., it is not occupied by the PAL user. The factor $\psi_g h_g$ represents the signal strength received by the GAA users and the factor $\psi_c h_c$ represents the interference experienced by the PAL users caused by the addition of GAA users. Where the factors ψ_g and psi_c show the transmit power of GAA users and the GAA CBSDs respectively. The factors h_g and h_c represent the particular channel gains. The transmission rate received by the PAL users is modeled as given in Equation 1.

$$T_p = \beta_p^{\nu} log_2(1 + \frac{\psi_p h_p}{\psi_c h_c + \sigma^2}) \quad (1)$$

The noise in the system is considered the additive white Gaussian noise (AWGN) with zero mean and variance σ^2.

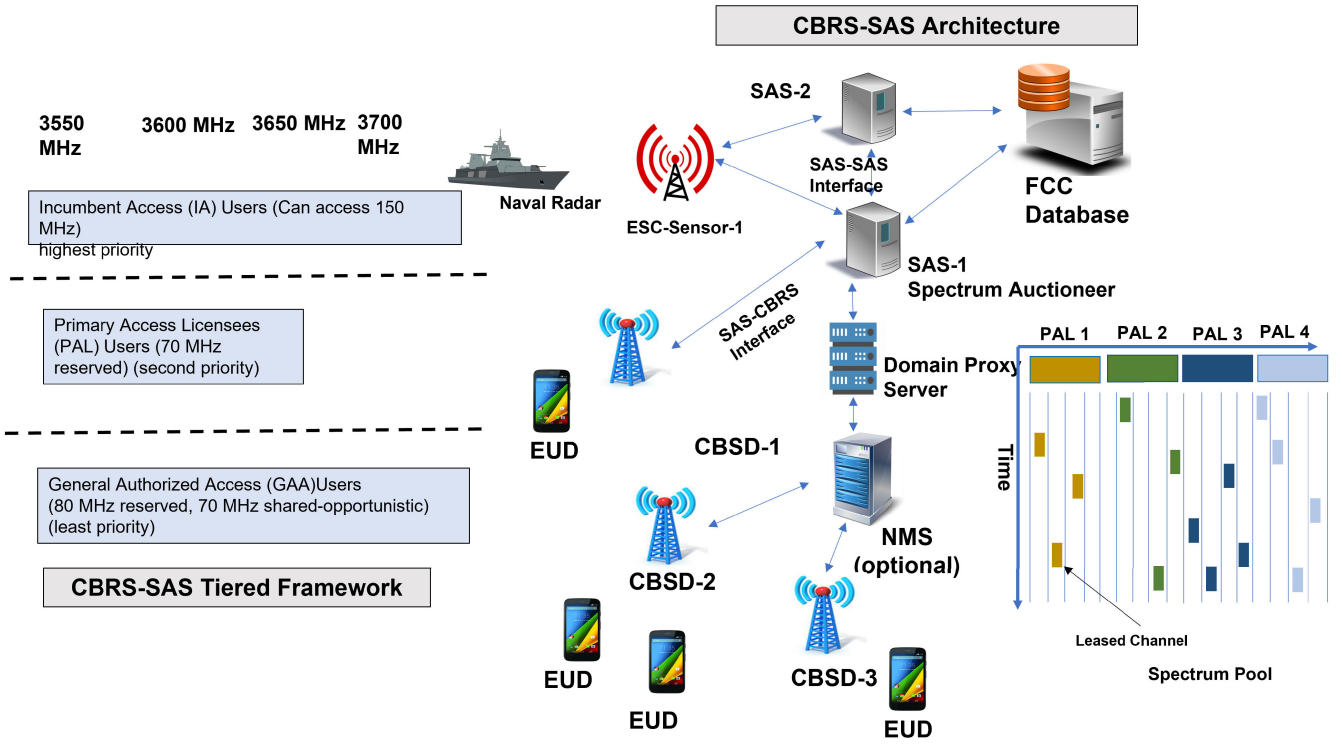


FIGURE 1. CBRS-SAS architecture.

The factor $\psi_c h_c$ represents the interference to the PAL users and the factor $\psi_p h_p$ shows the signal strength received by the PAL users. If there are fewer channels and more GAA users then the chance of the collisions of the GAA user's packet will be increased so the condition $1 \leq \nu \leq z$. The SAS will collect the information of all z PAL channels at the beginning of each time slot. The SAS will randomly allocate the ν out of z channels to the GAA users so that each user is assigned a channel. After each transmission at the end of every time slot τ, the acknowledgment signal α is sent to the SAS to see whether the transmission of the GAA users is successful or not. SAS records the observations $o(\tau)$ for each GAA user which is a binary response represented as.

$$o(\tau) = \begin{cases} 1 & \text{if the channel is idle,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The scenario of the SAS-CBRS framework is based on the MDP model to achieve the maximum transmission of packets in each time slot. To define the MDP model, the states of the given environment, actions related to the particular states, rewards associated with the concerned actions, and the discount factor for the dependence of the optimal policy on immediate or future reward should be defined.

1) States: The state of each channel at the given time slot τ can be defined as ψ_τ. The states for each channel can be either idle state or busy state. The states for all channels are represented as:

$$\psi_\tau = \{\psi_{\tau 1}, \psi_{\tau 2}, \psi_{\tau 3} \ldots .., \psi_{\tau z}\},$$

where $\psi_{\tau z} \in \{0, 1\}$. The SAS is aware of all the states of PAL-reserved channels i.e., how many channels are idle. The actions γ defined for each state ψ at time slot τ is represented as.

$$\gamma_\tau(\psi) = \{\gamma_{\tau 1}, \gamma_{\tau 2}, \gamma_{\tau 3} \ldots .., \gamma_{\tau z}\},$$

where the $\gamma_{\tau z}$ shows that whether the channel z^{th} is selected or not. The SAS selects ν number of channels from z channels. Accordingly, the action for ν GAA users is given by.

$$A(\nu, z) = \frac{z!}{\nu!(z-\nu)!} \tag{3}$$

2) Actions: The number of actions for allocating the z channels will be much greater than z in some scenarios that degrade the performance while making decisions for each action. The action space must be within limits to get stable and optimal rewards. Hence, we assume a condition that the action space must be equal to or less than z. Every GAA user must occupy the channel from the vector p. The SAS will choose the ν number of actions using our proposed technique. For each time slot, the SAS will assign the channels to ν GAA users. Our proposed algorithm assigns the channels very efficiently in case of multiple GAA users. If there

are enough idle channels available then each GAA user can get a slot to transmit data.

3) Reward Function: The purpose of the reward function in reinforcement learning is to give directions to the agent to get a reward while exploring the states of an environment. Every action has associated rewards so we need to define the rewards for the action space. Let $\omega_g(\tau)$ be the reward of the GAA user in time slot τ. The reward of the system is defined as the total number of accessible channels to the GAA users. Mathematically it is represented as $\omega_g(\tau) = \sum_{g=1}^{\nu} \omega_g(\tau) = \sum_{g=1}^{\nu} o_g(\tau)$. Where $o_g(\tau)$ is the observations of getting the idle PAL channels defined in Equation. 2.
4) Optimal Learning Policy: The optimal policy in reinforcement learning is to get the maximum reward. In this scenario, SAS will find the optimal policy π^* of the GAA users for which the maximum sum of the rewards is discounted by a factor σ i.e.,

$$\pi^* = argmax_{\pi} \Gamma \tag{4}$$

where Γ is the accumulated sum of rewards discounted by factor σ. The optimal policy must satisfy the following constraint.

$$\Gamma = E\left[\sum_{\tau=1}^{T} \sigma^{\tau}.\omega(\tau)\right] \tag{5}$$

The σ is the discount factor that varies within 0 to 1, $0 < \sigma < 1$. The average reward for optimal policy during the finite time duration T is calculated as.

$$\Gamma = \frac{1}{T}\sum_{\tau=1}^{T} \sigma(\tau) = \frac{1}{T}\sum_{\tau=1}^{T}\sum_{g=1}^{\nu} \sigma_g(\tau) \tag{6}$$

Hence, the optimal policy for the defined problem scenario is represented as.

$$\pi^* = argmax\left(\frac{1}{T}\sum_{\tau=1}^{T}\sum_{g=1}^{\nu} \sigma_g(\tau)\right) \tag{7}$$

In this paper, we considered the SAS-CBRS spectrum access framework in which there are multiple PAL operators having idle channels and there are numerous GAA users. The channel is only considered idle if the noise power of the idle channel is less than the defined threshold limits. we proposed a model-free solution for this allocation problem that is adaptable to the available environment to solve the dynamic problems.

C. THE REINFORCEMENT LEARNING MODEL FOR AUCTION PROCESS

Q-learning is an off-policy RL-based algorithm that works without prior knowledge. The agent in this scenario takes actions for the defined states of the environment to explore the rewards. The optimal policy defined relies on the maximized rewards, the dependence on the immediate or future reward, and the speed of learning. The agent explores the new states of the environment by taking the defined actions and receiving the associated rewards. The received maximized reward shows how well the action it took to improve the situation. The agent uses this reward signal to update its policy and improve its future decision-making. Reinforcement learning methods are broadly categorized into policy-based methods and value-based methods. The Value-based methods learn a state-value function that estimates the expected reward for being in a given state, while policy-based methods learn a policy that maps states directly to actions.

Q-Learning is a popular RL-based method that is based on the idea of using a Q-table to store the maximum expected reward for executing a particular action for a defined state. The Q-table is initialized with arbitrary values, and over time it is updated with more accurate estimates of the expected reward based on the agent's experiences in the environment. In Q-Learning, the agent selects actions based on an exploration-exploitation trade-off. At first, the agent selects actions randomly to explore the environment and gather information about the rewards associated with different actions. Over time, as the estimates in the Q-table become more accurate, the agent becomes more confident in its estimates and begins to select actions based on the highest expected reward. The Q-Learning algorithm is summarized as follows:

- First of all, initialize the Q-table with arbitrary values.
- Repeat each step of the episode and observe the states of the environment.
- Select an action based on the exploration-exploitation trade-off.
- Collect the reward.
- Update the Q-table using the observed reward and the maximum expected reward.
- Iterate the algorithm until convergence.

The detailed Q-learning formulation is discussed in the following Section III-D.

D. DOUBLE AUCTION (DA) ALGORITHM FORMULATION

The GAA users in a competitive environment looking for guaranteed QoS will take part in an auction managed by SAS. The GAAs evaluate their requirements and available QoS offered by PAL operators to bid accordingly. The bid value may vary based on the traffic classification based on differentiated services. Moreover, the DA algorithm enables the GAA users to learn from its moves without having prior knowledge of an environment. A bidding vector is also stored at the GAA user's end in which the cost factor for each available channel in an auction pool is calculated. To get an optimal solution a function F_{π}^{*} is defined, which returns the optimal policy while observing the actions related to particular states. The F_{π^*} is defined as

$$F_{\pi^*} = Q_{g,p}^{\tau+1}(\psi_{g,p}(\tau), \gamma_{g,p}(\tau)) \tag{8}$$

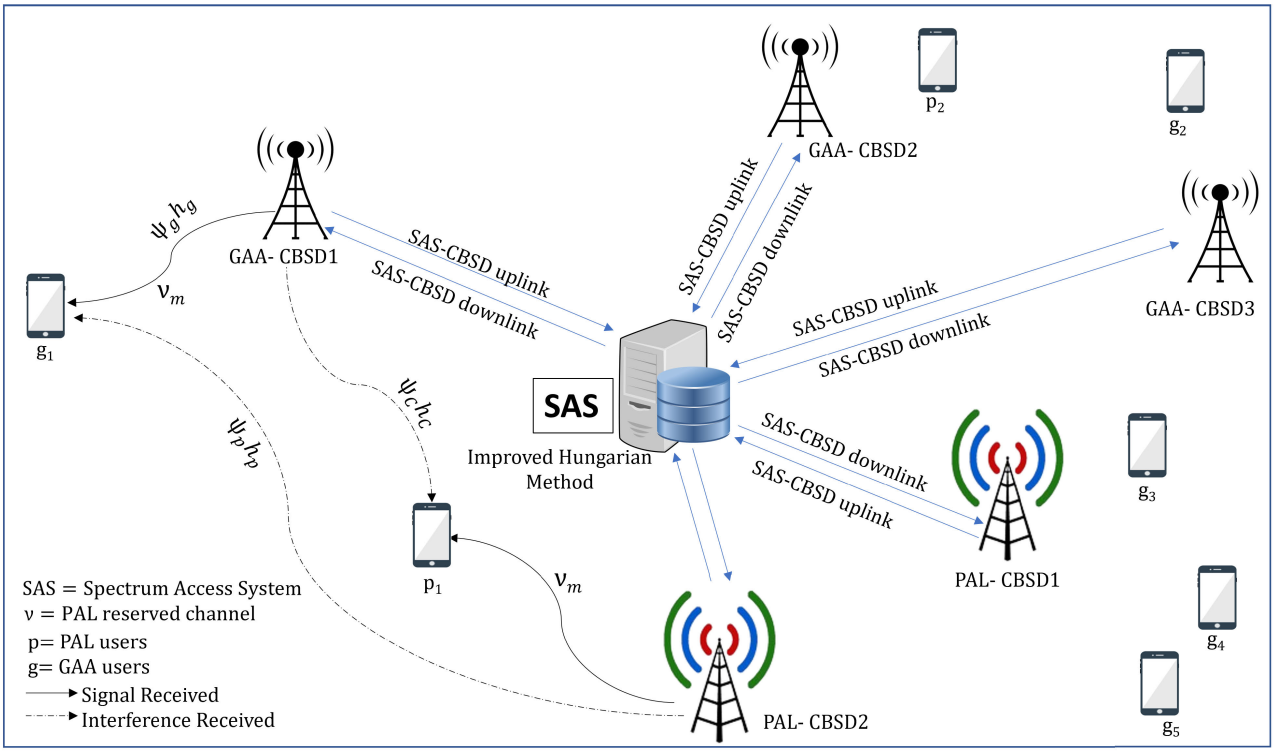


FIGURE 2. System model.

The $Q_{g,p}^{\tau+1}(\psi_{g,p}(\tau), \gamma_{g,p}(\tau))$ is the Q-function that shows the optimal policy for the GAA user g accessing the channel p at time t. The function observes the state $\psi_{g,p}(\tau)$ for the GAA user g accessing the channel p at time t and performs the action $\gamma_{g,p}(\tau)$. The agent gets the reward $R_{g,p}^{\tau}$ in the next iteration at time $\tau + 1$. In our proposed scenario, the optimal policy for every user may variate because of the real-time traffic dynamics. To formulate the optimal policy for GAA users the states, actions, reward, and learning policy plays an important role in the convergence of the Q-learning algorithm. The detailed formulation is presented in the following sections.

1) STATES

The channel is either occupied by licensed PAL operators or available to be auctioned. So, the states of the environment can be defined by the stochastic events whether the transmission opportunity is available for the channel or not. The state of an environment can be defined as

$$\psi_{u,p}(\tau) = \{\zeta_{u,p}(\tau)\} \in \psi$$

where $\zeta_{u,p}(\tau) = \gamma_{u,p}(\tau) \times a_{u,p}(\tau)$. The $\gamma_{u,p}(\tau) \in \{0, 1\}$. If $u = p$ then the $\gamma_{u,p}(\tau) = 1$ means that the PAL users want to occupy the channel and are not available for the auction and $\gamma_{u,p}(\tau) = 0$ shows that the PAL users have no data to send. The factor $a_{u,p}(\tau)$ shows the transition from a state of transmission opportunity available or not. If the transmission opportunity is available then the state will shift from the $\psi_{u,p}(\tau)$ to $\psi_{u,p}^{\tau+1}$ is considered to be a two-state Markov chain model defined in Section III-A as

$$\zeta_{u,p}(\tau) = \zeta_{u,p}^{\tau-1}.(1 - x_{u,p}) + (1 - \zeta_{u,p}^{\tau-1}).y_{u,p} \quad (9)$$

Similarly, for the scenario of GAA users the environment states will be defined accordingly, whether the transmission opportunity available or not along with the case of GAA users have data to send or there is no data which is defined as

$$\psi_{g,p}(\tau) = \{\gamma_{g,p}(\tau) \times a_{g,p}(\tau)\} \quad (10)$$

In the scenario of GAA users $\gamma_{g,p}(\tau) = y_{g,p} \times a_{g,p}(\tau)$ where $y_{g,p} \in \{0, 1\}$.

2) ACTIONS

The agent selects the environment states defined in section III-D1. Based on each state, an action is performed both on the PAL operator's end and the GAA user's side. The action defined in our scenario for PAL operators is to select the GAA users at the PAL operators asking price given by

$$a_p(\tau) = \psi_{u,p}(\tau), \gamma_{u,p}(\tau).B_{u,p}(\tau) \quad (11)$$

where $\psi_{u,p}(\tau)$ is the selection of bidder at asking price and the $\gamma_{u,p}(\tau).B_{u,p}(\tau)$ shows the asking price. In case of the GAA user's action, it is given by.

$$a_g(\tau) = \psi_{g,p}(\tau), \gamma_{g,p}(\tau).B_{g,p}(\tau) \quad (12)$$

where $\psi_{g,p}(\tau)$ is to choose a bidder at the offered price and the $\gamma_{g,p}(\tau).B_{g,p}(\tau)$ is the offered bidding price. In each cycle,

the PAL operators and the GAA users select the action based on the environment state $\psi_{u,p}(\tau)$, $u \in \{p, g\}$. Right after taking an action an immediate reward $\omega_{u,p}(\tau)$ will be released and the current state moves from $\psi_{u,p}(\tau)$ to $\psi_{u,p}^{\tau+1}$ with some transition probability.

3) REWARD FUNCTION

The purpose of the reward function in reinforcement learning is to get either an immediate advantage or long-term advantage by choosing a particular action defined in Section III-D2 according to the available environment states defined in Section III-D1. Hence, the reward of the system is based on the actions and the states represented by a function $\omega_{u,p}(\tau)(\psi_{u,p}(\tau), a_{u,p}(\tau))$. In the SAS-CBRS framework the PAL operators are looking to lease the idle channels to generate surplus revenue and the GAA users will try to get the PAL channels at the minimum cost that meets their required QoS. In this scenario, the reward for both PAL operators, and the GAA users will be defined separately. The reward function for the PAL operators is defined as.

$$\omega_{u,p}(\tau) = \sum_{p=1}^{z} \psi_{u,p}(\tau)\gamma_{u,p}(\tau).B_{u,p}(\tau) \tag{13}$$

The reward function for the GAA users is defined as.

$$\omega_{g,p}(\tau) = \sum_{p=1}^{z} \psi_{g,p}(\tau)\gamma_{g,p}(\tau).B_{g,p}(\tau) \tag{14}$$

To meet the requirements of the reward functions, the following constraints are considered.

$$\psi_{u,p}(\tau) \in \{0, 1\}, \quad \psi_{g,p}(\tau) \in \{0, 1\} \tag{15}$$

$$\sum_{p} \psi_{u,p}(\tau) \leq \sum_{p=1}^{z} p, \quad \sum_{g} \psi_{g,p}(\tau) \leq \sum_{p=1}^{z} p \tag{16}$$

Here, z represents the PAL reserved channels assigned to PAL and GAA users p and g respectively. The equations 15 and 16 are the constraints that show the user's action space. It is also assumed that the payoff by the GAA users is transferred to the PAL operator as it is without any bargain factor or deduction. Hence the reward for the PAL operators is actually the payoff by the GAA users

4) LEARNING POLICY

The learning policy in reinforcement learning based models depends on the choice of actions it takes while exploring the environment states to update the Q-table. Thus, the probability of selected actions $a_{\psi,p}(\tau)$ and $a_{\psi,g}(\tau)$ against particular states $\psi_p(\tau)$ and $\psi_g(\tau)$ respectively with received utilities received for both PAL and GAA users i.e., $Q_{g,p}^{\tau+1}(\psi_{u,p}(\tau), \gamma_{u,p}(\tau))$ and $Q_{g,p}^{\tau+1}(\psi_{g,p}(\tau), \gamma_{g,p}(\tau))$ are taken into account to map the learning policy with the current choice of action. The learning policy for PAL operators and the GAA users differ from each other because of their different service requirements and for each state they will receive different reward utilities. Hence the learning policy for both the PAL operators and the GAA users will be defined separately. The learning policies are defined as.

$$\begin{aligned} Q_{u,p}^{\tau+1}(\psi_{u,p}^{\tau}, \gamma_{u,p}^{\tau}) &= (1-\eta_p)Q_{u,p}^{\tau}(\psi_{u,p}^{\tau}, \gamma_{u,p}^{\tau}) \\ &\quad + \eta_p(\omega_p\sigma_p.max\zeta_p^{\tau}(\psi_p^{\tau+1}, \gamma_p^{\tau+1})) \end{aligned} \tag{17}$$

$$\begin{aligned} Q_{g,p}^{\tau+1}(\psi_{g,p}^{\tau}, \gamma_{g,p}^{\tau}) &= (1-\eta_g)Q_{g,p}^{\tau}(\psi_{g,p}^{\tau}, \gamma_{g,p}^{\tau}) \\ &\quad + \eta_g(\omega_g + \sigma_g.max\zeta_g^{\tau}(\psi_g^{\tau+1}, \gamma_g^{\tau+1})) \end{aligned} \tag{18}$$

The optimal policies for the PAL operators and the GAA users $(\gamma_{u,p}^{\tau+1})^*$, $(\gamma_{g,p}^{\tau+1})^*$ respectively are the probabilistic sum of the collective rewards i.e., ω_p and ω_g discounted by factors σ_p and σ_g respectively. The DA function updates the Q-table and approaches to the true values defined as:

$$(\gamma_{u,p}^{\tau+1})^* \longleftarrow argmax(Q_{u,p}^{\tau+1}(\psi_{u,p}^{\tau}, \gamma_{u,p}^{\tau})) \tag{19}$$

$$(\gamma_{g,p}^{\tau+1})^* \longleftarrow argmax(Q_{g,p}^{\tau+1}(\psi_{g,p}^{\tau}, \gamma_{g,p}^{\tau})) \tag{20}$$

Learning rate factor η and the discount rate factor σ play an important role in the convergence of the DA algorithm. The values of η and σ remain in between 0 and 1 given by $0 \leq \eta \leq 1$ and $0 \leq \eta \leq 1$. The learning rate η is a hyper-parameter that decides the speed of learning, if the value is close to 0 then it shows the Q values are not updated. While if the value of η is close to 1, that shows the DA will update the Q-values with the updated values. Usually, the value of η is set in between 0.1 - 0.5. The optimal value depends on the type of problem and the detailed investigations. So, this is a regulatory factor between exploitation and exploration. η with higher values makes the DA converge quickly as it exploits the agent to depend on its current knowledge instead of exploring the new information. The major risk involved with this is getting non-optimal solutions.

The discount factor σ is also a hyper-parameter that helps the agent to decide how much importance is given to the future reward. The value of σ remains within 0 to 1. The factor σ is typically set to the maximum value so that the maximum reward should be taken into account to get the maximum sum of rewards for the entire session. Moreover, it also helps the agent to depend on the long-term future reward or short-term immediate reward. The DA multiplies the discount factor with next state's highest Q-value. If the discount factor is too less or approaches 0, then it shows that the Q-value after multiplication with the discount factor σ will be negligible and the DA will depend only on the immediate reward. Hence, the DA algorithm needs two values to update the Q-table i.e, the projected maximum Q-value of the next state $max\zeta^{\tau}(S^{\tau+1}, \gamma^{\tau+1})$ and the instantaneous reward value $\omega(\tau)$.

IV. PROPOSED SOLUTIONS

A. DOUBLE DEEP RECURRENT Q-NETWORK FOR SAS DDRQ-SAS

The DDRQ-SAS algorithm complexity depends on the number of available PAL-reserved idle channels. If there are a huge number of idle channels detected by SAS then the states defined for the available channels and the actions

Algorithm 1 Double Deep Q-Network (DDQN) Algorithm

Initialize parameters
state_dim ← Dimensionality of the state space
action_dim ← Dimensionality of the action space
learning_rate ← Learning rate for the Q-network
discount_factor ← Discount factor (gamma) for future rewards

epsilon ← Exploration probability for epsilon-greedy
max_memory_size ← Maximum size of the replay memory
batch_size ← Batch size for training
C ← Target network update frequency
Function QNetwork(*state_dim*, *action_dim*, *learning_rate*):
Define neural network architecture with input size *state_dim* and output size *action_dim*
Define optimizer using learning rate *learning_rate*
return Q-network and optimizer **end function Function** epsilon_greedy(*q_values*, *epsilon*):
With probability ϵ, select a random action
Otherwise, select the action with the highest Q-value from *q_values*
return selected action **end function**
Initialize Q-networks
online_network, *online_optimizer* ← QNetwork(*state_dim*, *action_dim*, *learning_rate*)
target_network, *target_optimizer* ← QNetwork(*state_dim*, *action_dim*, *learning_rate*)
Initialize replay memory
replay_memory ← Empty deque
for $t \leftarrow 1toT$ **do**
  state ← Observe current state from the environment
  q_values ← online_network.predict(*state*)
  action ← Perform epsilon-greedy action selection with ϵ
  next_state, *reward*, *done* ← Take action *action* and observe the reward and next state
  replay_memory.append((*state*, *action*, *reward*, *next_state*, *done*))
  if length of replay_memory > max_memory_size **then**
    Remove the oldest entry from replay_memory
  end if
  if length of replay_memory > batch_size **then**
    minibatch ← random sample from replay_memory of batch_size
    for (x, a, r, x_next, d) *in* minibatch **do**
      target_q_values ← target_network.predict(x_next)
      target_q_value ← r *if* d *is True else* $r +$ *discount_factor* $\cdot$ max(*target_q_values*)
      q_values ← online_network.predict(x)
      q_values[a] ← *target_q_value*
      online_network.update(x, a, *q_values*)
    end for
  end if
  if $t \mod C = 0$ **then**
    Update the target network parameters
    target_network_parameters ← $\tau \times$ online_network_parameters $+ (1 - \tau) \times$ target_network_parameters
  end if
end for

associated with the states make a larger state, actions space. This will lead to increased computational complexity i.e., the computational complexity is affected by the number of available idle channels. So, it is difficult to devise an optimal policy for large-scale networks. We proposed the double deep recurrent Q network (DDRQ-SAS) to solve the resource allocation problem for numerous GAA users in the SAS-CBRS framework without any prior knowledge of the environment. The architecture of the proposed DDRQ-SAS algorithm is discussed in detail in the following section.

B. DDRQ-SAS ARCHITECTURE

The DDRQ-SAS architecture is shown in the Figure 3. In the first step, we use the states of the available channels detected by the SAS to process it in form of sequential information. Long short-term memory (LSTM) is then used to compute the sequential information from the previous step over time to predict the right scenario for the network. Then the value of each state-action pair is estimated to decrease the computational complexity of fetching and retrieving information from the value table. The sub-optimal policies are estimated and combined to get an optimal output for the given input of idle channels and the users. The double Q-learning algorithm is used to stabilize the overestimation of the value table to get an optimal solution. The detailed architecture is discussed below.

1) DDRQ-SAS INPUT LAYER

The input layer in the DDRQ-SAS architecture is depicted in the Figure 3. depends on the values of the previous two time slots for the given state. The input layer is defined as $\psi(\tau) = \{p_{(\tau-1)1}, \ldots p_{(\tau-1)n}, p_{(\tau)1}, \ldots p_{(\tau)n}\}$. The value $p_{(\tau)n} \in 0, 1$ shows the state of the n^{th} channel in time slot τ. The purpose of getting the previous state information is to get larger and dense data in sequence form. It will eventually help us to optimize the value table more accurately. The accurate estimation of the value function of the system depends on the number of input states' information.

2) LSTM LAYER

A long short-term memory is used in deep learning to process sequential data i.e., time-series data obtained using the input layer. The LSTM layer has a memory cell to store or forget the information over time using the previous state information and the input value. The memory cell consists of three gates input gate, the output gate, and the forget gate. These gates are interlinked. The input gate process the sequential input information to the memory cell and the output gate controls the flow of the output information from the memory cell. The forget memory cell is designed to decide whether the old data need to be kept or discarded. The backpropagation method is used to train the LSTM layer over time. The gradient is calculated over the entire sequence of input data, allowing the network to learn to process and remember long-term dependencies.

3) HIDDEN LAYERS

Two hidden layers are proposed to be added after LSTM layers which are fully connected with each other and integrated with the LSTM to estimate the true value function

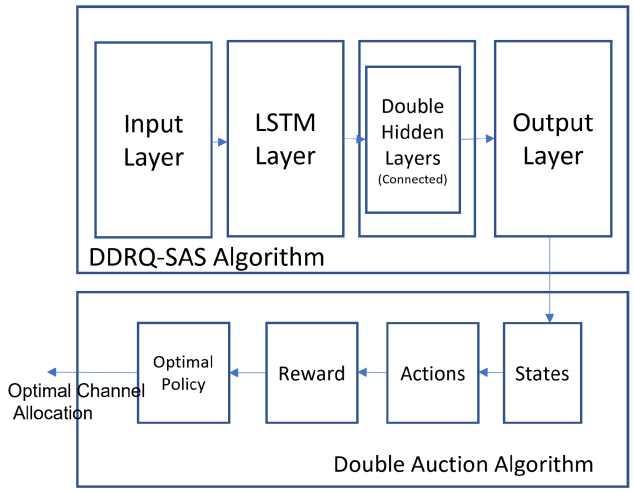


FIGURE 3. DDRQ-SAS-DA architecture.

of the state-action pairs. The hidden layer uses the RNN to decrease the time complexity. The hidden layer does not consider the look-up table instead it uses the RNN. It will also help the DDRQ-SAS algorithm to leverage the prior knowledge to derive the sub-optimal policies rather than generating the random policies.

4) BLOCK OUTPUT LAYER

The output of the DDRQ-SAS is an estimated Q-value generated for time slot *tau* for the transmission of data over the channel z. The out is generated in a form of a vector of size equal to the number of GAA users asking for the PAL reserved idle channel.

5) DOUBLE Q-LEARNING

To alleviate overestimation, we implemented double Q-learning to refrain the SAS from the selection of actions from Q-value evaluation. This is achieved by exploiting two neural networks, the target networks, and the online networks with matching structures for the estimation of value functions and action selection, respectively. Double Q-learning is a reinforcement learning algorithm that extends the standard Q-learning algorithm by addressing the issue of overestimation of action values. In standard Q-learning, the action-value function is updated using the maximum estimated action value over all possible actions in the next state. However, this approach can lead to over-optimistic estimates, especially when the estimates are noisy or inaccurate. Double Q-learning addresses this issue by using two separate Q-functions, often referred to as Q_1 and Q_2. The idea is to use one Q-function for action selection and the other for action evaluation.

C. DDRQ-SAS ONLINE LEARNING

We assume that the SAS operates in an autonomous environment in the CBRS band without any prior knowledge of the environment's characteristics. The SAS uses the neural network to devise an optimal policy that makes independent decisions that are centralized and as well as online by using the ACK signal. To begin each time slot (τ), the agent retrieves the latest channel state data from the previous two-time slots (ψ_τ). The SAS uses the epsilon-greedy approach to select the n channels with the highest values based on the target network output. The epsilon-greedy strategy typically favors selecting actions that have the highest estimated reward. However, it also aims to strike a balance between exploration and exploitation. Exploration is crucial because it enables us to experiment with new ideas, even if they contradict what we have already learned.

The value of ϵ in the epsilon-greedy approach ranges between 0 and 1 i.e., $0 < \epsilon < 1$, and it determines the trade-off between exploration and exploitation. When ϵ is high, the algorithm tends to explore more, whereas when ϵ is low, it focuses more on exploiting existing knowledge. Thus, the larger the value of ϵ, the more the algorithm emphasizes exploration and vice versa. The SAS will assign one of n channels to each GAA using a random allocation method. The reward for each channel will be obtained from the ACK signals. A tuple consisting of the current state $\psi(\tau)$, the next state $\psi(\tau + 1)$, the highest-scoring action, and its corresponding reward will be stored in the replay buffer. Later, a random tuple will be selected from the replay buffer to update the neural network based on the mean squared error.

D. DOUBLE AUCTION (DA) ALGORITHM

The double auction DA algorithm works on the principle of a trade market, where an auctioneer arranges an auction in which a seller asks a price for a product to sell and purchasers submit a bid to the auctioneer to purchase a product. In our scenario, SAS is an auctioneer who manages a pool of radio spectrum, each PAL operator acts as seller and adds their idle available channels in the pool with the asking price $\gamma^t_{u,p}.B^\tau_{g,p}$. The GAA users act as potential bidders to buy the available channels for a particular time at a bidding price $B^\tau_{g,p}$. The SAS will complete the auction trading by assigning the required PAL reserved idle channels to the GAA users and releasing the cost to the PAL operators.

The cost reserved by the PAL operators may vary according to the environment or market fluctuations, while the Pay off price by the GAA users depends on the available QoS. The double auction DA algorithm maintains the competitive environment by helping the PAL operators to set the reserved cost or base price dynamically to maximize their earnings and also it helps the GAA users to bid according to the available QoS requirement. Thus, it is important aspect to explore the cost setup by the PAL operators and the GAA users in detail.

1) RESERVED COST FOR PAL OPERATORS

The PAL operators paid the license to get dedicated access to the 70 MHz spectrum band. They take advantage of PAL users' absence to generate additional revenue to increase their profit revenue by leasing the unused spectrum to the GAA users. So, to ensure profitability, it is important to devise

a pricing strategy that maximizes the profit margins in an auction market. Therefore a reserved cost r_p^τ is set for the idle channel that is to be auctioned by an auctioneer i.e., SAS. The SAS will ensure that no channel will be auctioned to GAA users below the reserved cost. This lucrative offer will create interest in the PAL operators to join the SAS auction framework. The reserve cost r_p^τ for the PAL users by the DA algorithm is defined as the possible future reward with $Q_{u,p}^{\tau+1}(S_p^l, \gamma_p^\tau)$, where $Q_{u,p}^{\tau+1}(S_p^\tau, \gamma_p^\tau) \in Q_p^{\tau+1}(S_p^\tau, \gamma_p^\tau)$, $u \in \{1, 2, \ldots., \mu\}$.

$$r_{u,p}^\tau = Q_{u,p}^{\tau+1}(S_p^\tau, \gamma_p^\tau) \tag{21}$$

Moreover, the PAL operators can set the reserve price using the Q-learning algorithm derived from the equation. 17. Using this Q-learning process the PAL operators can set the price while considering the future reward and the history of states visited.

2) GAA BIDDING STRATEGY

The SAS as auctioneer maintains a spectrum pool and ensures fairness by giving equal opportunity to both PAL operators and GAA users. The information on the available channels is made available to all purchasers taking part in an auction. The GAA users as a purchaser generate a list $l_{g,p}^\tau$ of preferred channels according to the QoS available, and the related cost for each channel p at time τ. The preference list $l_{g,p}^\tau$ is defined as

$$l_{g,p}^\tau = \Psi\beta_{g,p}^\tau + Q_{g,p}^{\tau+1}(S_g^\tau, \gamma_g^\tau) \tag{22}$$

where $Q_{g,p}^{\tau+1}(S_g^\tau, \gamma_g^\tau)$ is the future reward, Ψ is the trade-off regulatory factor between the current packet and future market expectations, and the $\beta_{g,p}^\tau$ is the number of total packets in the buffer.

A random variable $\kappa_{g,p}^\tau$ independent of time is defined to store the number of packets following the Poisson distribution with the arrival rate of κ packets per second. The buffer capacity is said to be $\lambda_{g,p}$. Hence, the buffer state for the GAA user will be calculated as.

$$\beta_{\psi,g}^\tau = min\{(\beta_{\psi,g}^{\tau-1} - K_{\psi,g}^{\tau-1})^+ + \kappa_g^\tau, \lambda_g\} \tag{23}$$

The factor $(\beta_{\psi,g}^{\tau-1} - K_{\psi,g}^{\tau-1})^+ = max(0, (\beta_{\psi,g}^{\tau-1} - K_{s,g}^{\tau-1}))$ is the immediate gain received.

Thus, the bid offered by the GAA users is according to the learning process and the accumulated packets to ensure the successful bid compilation that meets the dynamic practical requirements.

3) DA ALGORITHM IMPLEMENTATION

The DA algorithm allocates the primary channels for a time period T. The cost offered by the GAA users after acceptance from the SAS will remain the same for this time period T. The resources will be released to GAA users once the SAS received the r_p^τ reserve price, it will complete the transaction after the λ_g constraint is satisfied. Once the transaction is completed, the SAS calculates the reward defined as.

$$r_{g,p}^\tau = \Psi.\beta_{g,p}^\tau \tag{24}$$

The objective of the SAS is to maximize the profit for PAL operators by leasing the idle channels to GAA users at a minimum of the defined reserved price. Accordingly, it will be defined as an optimization problem with the objective to maximize the profit defined as:

$$\Upsilon(\psi_{g,p}, r_g^\tau) = argmax(r_{u,p}^\tau) \tag{25}$$

subject to:

$$r_p^\tau > \lambda_{u,p}^\tau \tag{26}$$

and

$$r_{u,p}^\tau = r_{g,p}^\tau \tag{27}$$

4) DA ALGORITHM CONVERGENCE

Following are the conditions for the convergence of the DA dual auction algorithm based on the time-varying learning factors η_p and η_g that utilize the results obtained from the Robbins-Monro theory [51] must meet the following conditions for convergence of equations 17 and 18. The DA method learning policies defined in equations 17 and 18 converges to its optimal point defined $(Q_{g,p}^{\tau+1}(\psi_g^\tau, \gamma_g^\tau))^*$. The values for $(Q_{g,p}^{\tau+1}(\psi_g^\tau, \gamma_g^\tau))^*$ must be uniformly distributed for all the states $s_{g,p}$, and the actions $\gamma_{g,p}$ with the probability of 1, if the following conditions are satisfied.

- All the states and actions defined for both PAL operators and the GAA users i.e.,ψ_p,ψ_g,γ_p, and γ_g must be finite.
- The reward factor for both PAL operators and the GAA users $\omega_p(\psi_p, \gamma_p)$, and $\omega_g(\psi_g, \gamma_g)$ must be finite.
- The factor $\sum_{\tau=0}^{+\infty} \eta_p, \eta_g = \infty, \sum_{t=0}^{+\infty} (\eta_p)^2, (\eta_g)^2 = \infty$
- If the factors σ_p, and σ_g approach 1, then it shows that the policies will converge to a cost-free terminal with probability 1.

These conditions are satisfied in our defined scenario and the DA algorithm fulfills the convergence requirements. In Section III-D, the states and the actions for the environment and agents respectively are defined and there is a finite set of states and actions that proves the first condition. The reward functions defined in equations 13 and 14, we defined that the $\omega_{p,g}^{min} \leq \omega_{p,g} \leq \omega_{p,g}^{max}$. Hence the factor $\omega_{p,g}^2$ will also be a finite value that shows that the $Var\{\omega_{p,g}\} = E(\omega_{p,g})^2 - (E(\omega_{p,g}))^2$ is also finite, This shows that the second condition is also satisfied.

The factor η is defined in Section. III-D4 i.e.,

$$\eta_{g,p} = \begin{cases} \frac{1}{T}, & \tau > 0 \\ 0, & \tau = 0 \end{cases} \tag{28}$$

It is proved from equation 28 that the value of η never approaches 1. Thus, the third condition is also satisfied. The DA double auction algorithms maximize the reward functions

instead of achieving maximum gains. If the discount factor σ_p or σ_g approaches 1, it violates the DA algorithm objective. Therefore, the last condition also holds in our defined scenario.

Consequently, the convergence of the DA double auction algorithm is guaranteed in the defined scenario as it maximizes the rewards for the PAL operators and the GAA users.

E. DOUBLE AUCTION FRAMEWORK FOR MULTI-CHANNEL MULTI-WINNER ALLOCATION

The double auction algorithm for multi-channel multi-winner allocation with heterogeneous channel conditions is proposed in [52]. This algorithm implements the dual auction and allows to collect the bids from the purchaser by an auctioneer and pay off to the seller after completing the transactions. The algorithms considered the dynamic channel spectrum opportunities and channel variations. The authors proposed the concept of grouping the users to get the group bids. The channel allocation pattern depends on the group bids from each group. The algorithm maintains a preference list at a secondary users end. The SUs are grouped and each group is individually considered. The proposed work focused on the profits for both sellers and auctioneer. The double auction algorithm is an iterative algorithm that assigns the channel to SUs in each round. If any SU remains unassigned then it will be assigned a channel in next round. The working of the algorithm is given below:

There are n primary channels available for an auction. The channels are considered heterogeneous in nature and are stored in a vector K. The channels available from each PAL operators are grouped individually i.e., if there are n primary operators then there will n groups given by.

$$K = \{K_1, K_2, \ldots .., K_n\}$$

The channel variations and the interference between the PAL and GAA users are also considered. The capacity of the channel is calculated using the Shannon capacity theorem defined in [52]:

$$\Psi_{i,j} = Wlog_2\left(1 + \lambda_j \frac{P_{L_{(i)}}}{I_i + \sigma^2}\right) \tag{29}$$

The authors maintained the preference list at the SU end by calculating the difference between the availability time and the requirement time for the channel. The bid will only be calculated if the available time is greater than the required time for channel given by $T_r > T_a$. Once the bidders and the channel preference lists are finalized then the problem is formulated as a linear assignment algorithm with NxN Matrix and defined an objective function for channel assignment to SU as:

$$SU = \sum_{j=1}^{Y}\sum_{i=1}^{N} b_{ji}^{(b)}.a_{ij} \tag{30}$$

The auction method proposed consists of three main steps.

- Winner determination
- Payment method
- New auction round

In winner determination, the auctioneer finds the strategy to find the winning bidder. Groups are formed for non-interfering SUs for each channel considering the heterogeneous constraints. Accordingly a group bid is calculated using the individual bids offered by SUs. After this step the channel allocation strategy is applied to assign the channels to the winning bidders. The group bid is calculated from the following equation.

$$\mu_z^j = min\{b_{jh}^{(b)} | h \in g_z^j\}.|g_z^j| \tag{31}$$

where g_z^j is the group made for channel j, and mu_z^j is a a group bid. h is a secondary user with a lowest bid and $b_{jh}^{(b)}$ is the bid value offered by user h for channel j.

In the payment section, price payoff is calculated for all winner SUs who are assigned the channels. The price calculated is payable to auctioneer after successful assignment. The auctioneer in this scenario earns profit in terms of the group valuation denoted by δ_z^j defined as.

$$\delta_z^j = min\{v_{jh}^{(b)} | h \in g_z^j\}.|g_z^j| \tag{32}$$

where $v_{jh}^{(b)}$ is equal to the bid value $b_{jh}^{(b)}$. Once the profit and bid value is calculated then the revenue for the licensed operator is calculated as:

$$r_{qj} = \delta_z^j - v_{qj}^s \tag{33}$$

The factor r_{qj} shows the revenue of the licensed operator q for selling the channel j. If the auctioneer is unable to sell the channel then the value of r_{qj} will remain 0.

Final step of the this algorithm is the new auction round. In this step auctioneer looks for the unassigned channels during the last iteration in which the channels were auctioned altogether. The unassigned SUs are allowed to resubmit the bids for the new available channels in the current iteration. Similarly, if any channel remains unassigned in the previous iteration the licensed operators will decrease the asking price of the channel.

F. TRUTHFUL DOUBLE AUCTION FOR CRN: TRANSMITTING AND SHARING

The truthful double auction for cognitive radio networks: transmitting and sharing (TDSA-PS) proposed in [30]. The authors proposed the TDSA-PS algorithm to auction the channels occupied by the primary users and proved that their proposed algorithm is budget balanced, truthful, efficient and individually rational. the authors in formulated the CRN model with auction model. In the CRN model PU_m primary users are considered and each PU holds a single channel CH_m. n secondary users are considered denoted by SU_n, while each SU is equipped with a transmitter T_j and a receiver Γ_j. The secondary users is provided with an indicator signal to check whether the SU is allowed to transmit at CH_i or not. The

interference temperature limit (ITL) is calculated to see, if the transmission is successful or not i.e.,

$$\sum_{y_i^j=1} \frac{\epsilon_j}{E(T_j, L_i^l)^\alpha} \leq \gamma_i \tag{34}$$

where α is the path loss exponent, L_i^l is the location of user i at location l. For each SU the authors calculated the SINR [newrc] defined as:

$$SINR_i^j = \frac{\frac{\epsilon_j}{E(T_j, L_i^l)^\alpha}}{\frac{\rho_i}{E(T_i, \Gamma_j)^\alpha} + N_0 + \sum_{j' \neq j, y_i^{j'}=1} \frac{\epsilon_{j'}}{E(T_j, L_i^l)^\alpha}} \tag{35}$$

In the proposed scenario, FCC acts as an auctioneer to conduct the double auction in which the primary licensed users sell their channel to the SUs at a minimum cost defined as c_i. The secondary users may submit a different bid that may not necessarily be equal to or greater than the primary users asking price. When the auctioneer received the complete information of the seller's asking price and the buyer's bid, then the auctioneer calculates the winning and allocation indicators represented by x and y respectively. Let's the payment paid to the primary users is P and the payment received by the secondary users is represented by q. Then the profit utility of an auctioneer is calculated as the difference in payment received by the secondary users and submitted to the primary users.

$$u = \sum_{j=1}^{n} q_j - \sum_{i=1}^{m} p_i \tag{36}$$

In implementing the TDSA-PS, the auctioneer sorts the asking price from the primary users in non-decreasing order and sorts the bids received from the secondary users against the primary channels in non-increasing order. The author proposed the concept to find the smallest index k PU channel to be assigned to the largest index l secondary user. Truthfulness is guaranteed in this scenario. In the second step, the channels are assigned to the secondary users and in the last step the TDSA-PS is applied which gives the winning vector, allocation vector, and payment vectors as an output. The auctioneer implemented a double-auction algorithm for both the primary user's and the secondary user's side. One of the objectives of TDSA-PS is to yield a good profit ratio for an auctioneer too. The algorithm was proved to be efficient good and the budget is balanced for the primary as well as secondary users.

V. NUMERICAL RESULTS

A. SIMULATION SETUP

The performance of the DDRQ-SAS algorithm with the DA algorithm is evaluated for four PAL operators offering different numbers of channels in the presence of 10-500 GAA users. The SAS as auctioneer holds an auction where up to a maximum of four PAL operators from a particular region can take part as proposed by the FCC. The PAL operators send the information including the asking price of available idle channels to the SAS to take part in the auction. SAS manages the spectrum pool where all the channels' information is listed and shared with the competing GAA users. The least priority GAA users take advantage of this service to get guaranteed QoS as per their requirements at a competing cost. The SAS monitors the activity of IA users using the ESc sensor to pull out the channels from the PAL and GAA users. In our proposed scenario, the PAL users as sellers and the GAA users as purchasers use the proposed DA algorithm to compute the asking price and the bidding price respectively and the SAS uses the DDRQ-SAS algorithm to allocate the resources. The Shannon capacity theorem is used to model the transmission rate of GAA users defined as

$$T_g = \beta_g^p log_2(1 + \frac{S_g^p}{N_g^p}) \tag{37}$$

where β shows the accessible secondary user's g bandwidth for the available PAL reserved channel p. $\frac{S_g^p}{N_g^p}$ is the SNR received at GAA's boundary by the other GAA users accessing the PAL reserved channel p at a distance d is defined as:

$$SNR_g = \frac{max(\rho^p)}{\sigma^2} \left(\frac{d}{d_0}\right)^\alpha \tag{38}$$

The ρ^p represents the transmission power of PAL operators which is defined as 30 dB. The value α is adjusted at 4 and d_0 to 1. The urban propagation model for macro-cells is used for simulation along with Rayleigh multi-path fading model. The simulation parameters defined in our scenario are presented in Table 1. The simulations are performed to 100 times to reduce the randomness to achieve steady outcomes.

The SAS is the central entity that is considered an intelligent agent that uses the DDRQ-SAS algorithm. The SAS comprises two neural networks i.e., the target neural network and the online neural network. Both neural networks have three layers, with the first layer being an LSTM layer of the same size as the state size. The second layer has 100 neurons with ReLU activation, and the last layer has p neurons, where p represents the total number of channels. To ensure adequate training of the proposed algorithm, the target network's parameters are updated every hundred timeslots by the online network, and the memory size is set to 1,000. During each time slot, a minibatch of 32 samples is randomly selected from memory to train the neural network using the mean squared error function as the loss function, with the Adam algorithm optimizing the neural network's parameters by minimizing the loss function.

The Deep Q-Learning (DQN) algorithm includes two hidden layers and a memory of size 1,000. During each time slot, a random minibatch of 32 samples is extracted from memory to update the neural network's parameters. When compared to the DDRQN, the DQN has the same fully connected hidden layer structure. Specifically, in our

TABLE 1. Simulation parameters.

Attribute	Value
Bandwidth of Subcarriers	10 KHz
Power constraint	1W
Timing period	0.2 Sec
Learning Rate	0-1
Discount Factor	0.5
Propagation model	$128.1 + 37.6Log_{10}(R)$
Fading Model	Rayleigh multi-path
Cell radius	1 Km

implementation, the DQN also comprises two fully connected hidden layers of the same size.

B. PERFORMANCE EVALUATION

In this section, we compared and evaluated our proposed DA algorithm based on reinforcement learning with the double auction multi-channel multi-winner algorithm and the TDSA-PS algorithm for channel assignment to the GAA users.

The cumulative distribution function (CDF) for channel allocation to 500 GAA users by four PAL operators is depicted in Figure 4. The CDF is calculated for one and more PAL operators. when there is a single PAL operator, the probability to assign channels to GAA users is 0.34, which is 47% higher than its competing algorithm. If a single PAL operator is taking part in the auction it means that less number of channels are available to be auctioned that is the reason that, the number of users getting access to PAL channels is less as compared to the results when there are more than one PAL operator. This is the reason that out of 500 GAA users, only 170 users get access to the PAL reserved channels. Meanwhile, when there are all four PAL operators are taking part in an auction, then the CDF approaches 1. which is 7% better than the double auction multi-winning algorithm and 10% better than the TDSA-PS algorithm. The reinforcement learning-based double auction algorithm is more efficient as compared to the available static double auction algorithms without learning capabilities.

In the defined DDRQ-SAS and DA algorithms, the speed of the convergence of an algorithm depends on the step size or the learning factor η_p and η_g defined in equations 17 and 18 respectively. The Figure. 5 depicts the effect of the learning rate on the asking price for the PAL operators and on the bidding price for the GAA users. It is evident from the results that the GAA user's bidding price is always greater than the PAL operator's reserved price for all the values of η between 0 and 1. Hence, we can say that the factor η is not helpful in the convergence of the DDRQ-SAS-DA algorithm but it controls the speed of convergence of the DA algorithm that is shown in Figure. 7. So, we can use any values of η for the learning policies defined for both PAL operators and the GAA users.

The impact of the discount factor is very much interesting in the study of reinforcement learning as it shows the dependence of the optimal Q-learning policy on the reward

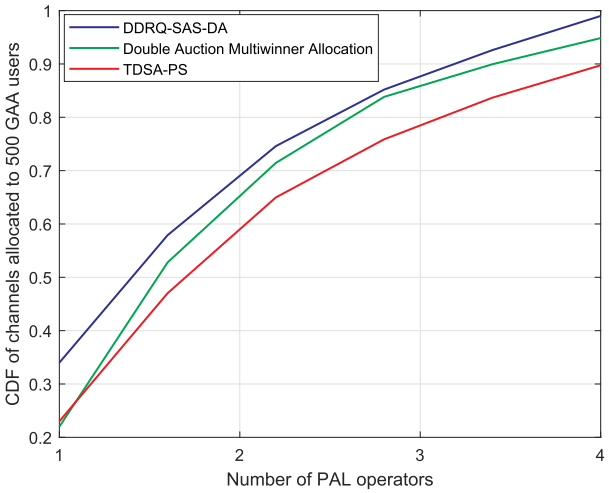


FIGURE 4. CDF of channels allocated to 500 users.

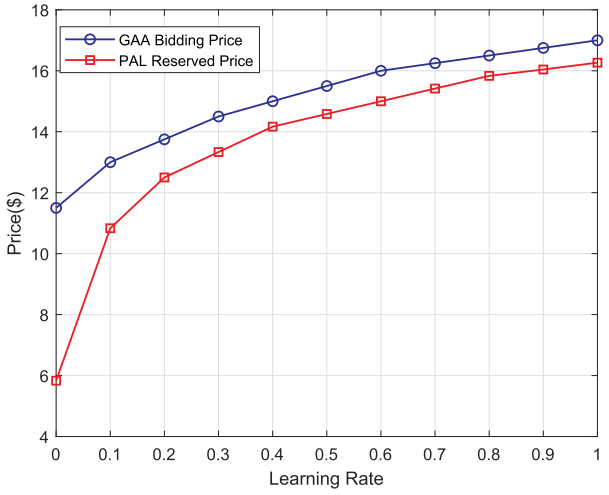


FIGURE 5. Learning rate in DA algorithm.

values i.e., whether to get the immediate reward or give weight to the future reward. Figure. 6 shows the comparison of discount factor σ defined in equations 17 and 18. The GAA bidding price remains higher than the PAL operators reserved price for the values between 0.02-0.5. The bidding price and the reserved price grow much faster when the value of σ is set to 0.6 or greater values. This pattern shows that when the value of σ approaches 1, it considers both the future reward and the immediate reward equally. For a successful trade, the bidding price of the GAA users must be greater than the PAL operator's reserved price. From the graph, it is clear that the proposed optimal learning policy is considering immediate rewards, and less weightage is assigned to future rewards. If the GAA user's optimal policy is more dependent on future rewards then they will always predict a lesser price than the reserved price which is not a feasible solution for an auction problem. Therefore, we considered the default value for discount factor σ in our simulations which assures that

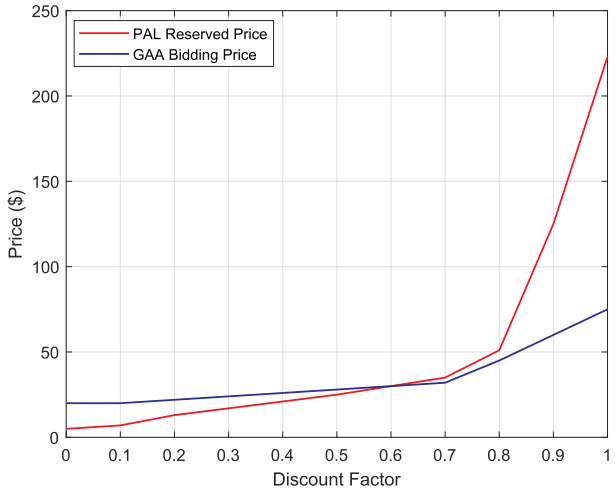


FIGURE 6. Effect of a discount factor in DA algorithm.

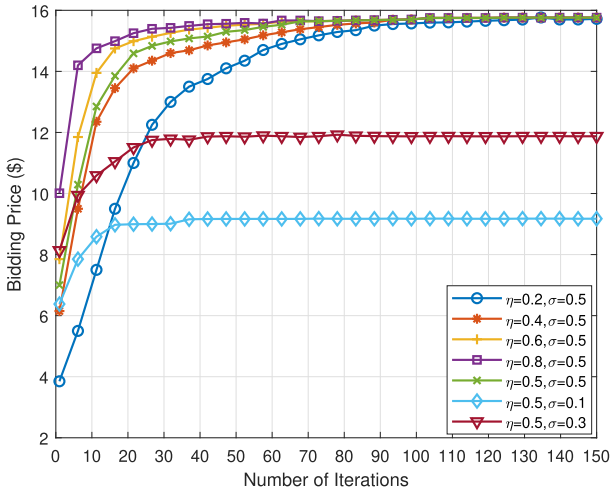


FIGURE 7. DA convergence of GAA bidding price.

difference between the bidding price and the reserved price is less enough to be acceptable by both seller and purchaser.

In Figure 7, the convergence of the DDRQ-SAS and DA algorithm for the bidding price of GAA users is shown for numerous values of learning rate η and discount factor the σ. It is clear from the graph that the algorithm converges in almost 20 iterations. It can be seen from the graph that, for the values of η greater than 0.5, the algorithm converges very fast i.e., 95% of the results are achieved in just less than 10 iterations. so this factor controls the speed of the convergence of the proposed algorithm. While the discount factor σ decides the dependence of DA on the immediate and future reward. It is also obvious from the Figure. 6 that for a successful trade, the value of the discount factor should be in the range of 0.02 - 0.5. We can see the effects of discount factor σ for the values of 0.1, 0.3, and 0.5 in Figure. 7. The less value shows that the dependence on the future reward is less. That is the reason that a value should be selected that helps the DA algorithm to converge at a price acceptable to the seller and purchaser. If the value of the σ is 0.1, then more weightage is given to the immediate reward and the price converges at 9$. It is also evident from the convergence of DA verifies the effects of learning rate η and the discount factor σ.

The Figure. 8 depicts the GAA users received data rate per unit cost. When there are fewer GAA users, the overall competition to get the channel is decreased. This makes the GAA users to get the best QoS at minimum cost. While in the scenarios, where the number of GAA users taking part in an auction is high, then the competing users will get the compromised QoS to fulfill their requirement, that is why the average data rate per unit cost is decreased as the number of GAA users increased. The DA algorithm in comparison with the double auction multi-winning algorithm and the TDSA-PS algorithm performs quite well. When the overall load is less, the DA is performing 10% better than the double auction winning algorithm and 30% better than the

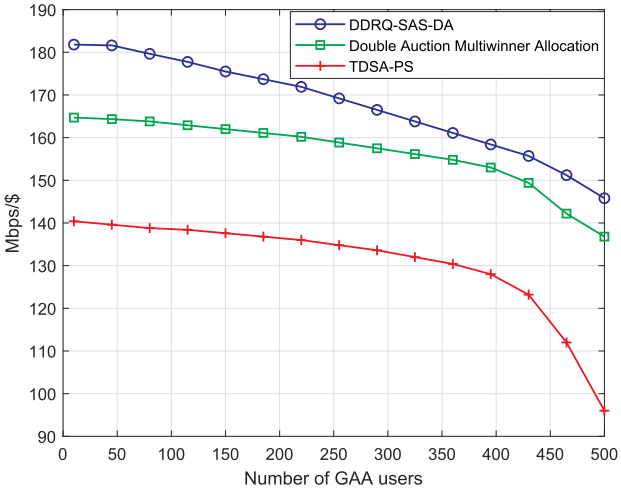


FIGURE 8. Received data rate per unit cost.

TDSA-PS algorithm. In higher loads, the proposed algorithm outperforms the other two algorithms and performs 7% better than the double auction winning algorithm and achieves 53% better results in comparison with the TDSA-PS algorithm.

The Figure. 9 shows the revenue achieved by the PAL operators by auctioning the available idle PAL reserved channels. It is evident from the graph that the PAL revenue using the DA algorithm is better in comparison with the other algorithms. When there are fewer PAL reserved channels available the net revenue collected by the operators is almost the same. when there is a high number of channels available, the DA algorithm performs almost 12% better than the double auction winning algorithm and 24% better than the TDSA-PS algorithm.

The execution efficiency of the three algorithms is shown in Figure. 10. The DA algorithm outperforms the double auction multi-winning algorithm and TDSA-PS algorithm. The total time taken by DA to assign idle PAL reserved channels to 500 GAA users is 15 ms which is almost 1.6 times

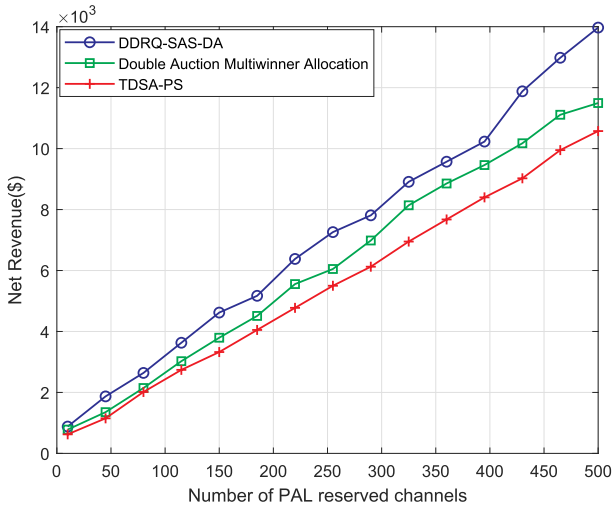


FIGURE 9. Average PAL revenue.

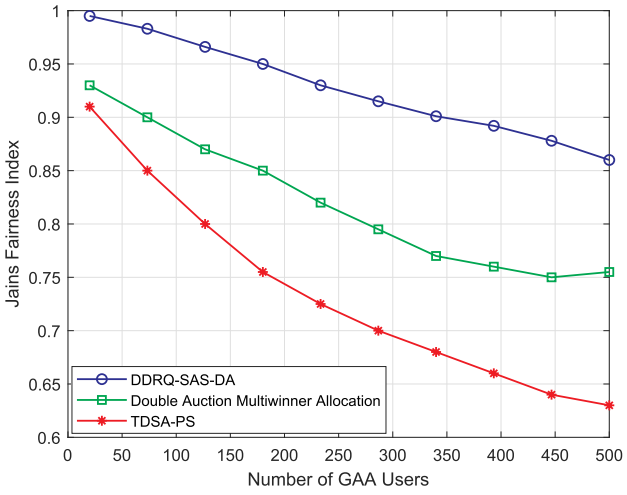


FIGURE 11. Jain's fairness index.

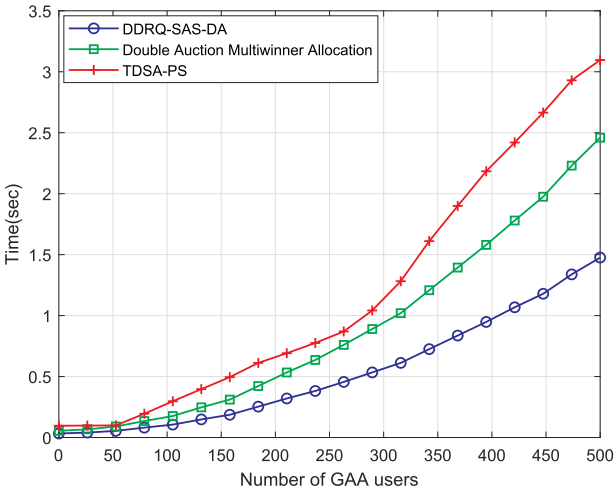


FIGURE 10. Comparison of execution efficiency.

better than the double auction multi-winning algorithm that assigns the channels to 500 GAA users in 25 ms and almost 2 times better than the TDSA-PS algorithm which assigns the channels to 500 GAA users in 31ms. When to load of GAA users is less than 100 users the difference between the three algorithms is minimal but at a higher load, the computational complexity of engaging a higher number of users increases which causes a delay in assigning the channels. Hence, the execution efficiency of the DA algorithm shows that the DA algorithm is much more efficient in comparison to the double auction multi-winning algorithm and TDSA-PS algorithm.

Jain's fairness index (JFI) is shown in Figure. 11. The JFI is a parameter used to measure the fairness of competition in a game. The JFI is defined in the equation.

$$J = 1 - \frac{\sum (R_i - R_{avg})^2}{n * R_{avg}^2} \tag{39}$$

where J is Jain's fairness index, R_i is the reward earned by the i^{th} participant and the R_{avg} is the average reward of all the participants who took part in the game. The value of the JFI is varied from 0 to 1. The value 0 shows complete unfairness and 1 shows complete fairness. It is evident from the Figure. 11 that when there is less number of GAA users, the JFI approaches to 1 for DA algorithm and 0.94 and.0.92 for the double auction multi-winning algorithm and TDSA-PS algorithm respectively. when the number of users varied from 50 to 500 the JFI value decreased to 0.86 for DA. The JFI value for the dual auction multi-winning algorithm for 500 GAA users is 0.75 and for the TDSA-PS algorithm, the value shows a downtrend with a value of 0.63. The fairness index of the DA algorithm is almost 17% better than other competing algorithms.

The simulation results for the algorithm show that it is an efficient algorithm for assigning the PAL reserved channels to the GAA users. The DDRQ-SAS algorithm integrated with the DA algorithm automates the GAA user's bidding and the PAL operator's asking price. The DDRQ-SAS with DA algorithm is efficient in comparison with its competing algorithms. It also ensures providing the required QoS to the GAA users at the maximum available data rate at a rational cost while generating handsome revenue for PAL operators. Maximum GAA users are accommodated with the proposed DA algorithm without degrading the overall performance.

VI. CONCLUSION AND FUTURE WORK

To cope with the challenges of spectrum scarcity and meet the requirements of 5G, FCC allowed sharing of the 3.5 GHz federally held spectrum with commercial users for both licensed and opportunistic use. The PAL-licensed operators can share the unused spectrum with opportunistic users to generate additional revenue. In this paper, we proposed a DDRQ-SAS algorithm integrated with the double auction DA algorithm to assign the unused channels held by PAL operators to the GAA opportunistic users. The SAS as a

central entity acts as an auctioneer to conduct an auction and manages a spectrum pool. The GAA users who need to get guaranteed services will take part in an auction.

The DDRQ-SAS algorithm uses the double deep Q-network integrated with the deep recurrent network while the DA algorithm uses the Q-learning algorithm that is based on the reinforcement learning method. The Sellers and purchasers are defined agents and the states for both actors are defined according to their environment, i.e., whether there is an opportunity to transmit the data or not, and whether there is data to send or not. The actions are defined for each environment state. The reward is defined for each action. Finally, the optimal learning policy is defined based on the learning rate factor, discount factor, and reward factor. The DDRQ-SAS algorithm is an off-policy model-free algorithm in which the SAS learns by exploring the new environment states. The DA algorithm allows the PAL operators and GAA users to intelligently select the asking price and the bidding price respectively. For a successful trade, the bidding price must be greater than the PAL operators asking reserved price. Finally, the problem is also solved using the double auction multi-winning algorithm and the TDSA-PS algorithm and the results are compared.

The DDRQ-SAS-DA algorithm is validated through an extensive set of simulations. The simulation results proved the practicality of the DDRQ-SAS-DA algorithm. It is evident from the results that the DDRQ-SAS-DA algorithm is much more efficient in comparison with the double auction algorithm and the TDSA-PS algorithm. The numerical results show that the DA algorithm achieves up to 20% higher data rate per unit cost at a higher user load. The DDRQ-SAS-DA algorithm is up to 1.6 times more efficient in assigning the PAL reserved idle channels to the 500 GAA users. The Proposed algorithm also ensures that a handsome profit is generated which makes it a lucrative offer to take part in an auction process by the SAS. The fairness index is proved using Jain's fairness index which shows that, even at a higher load of users, the GAA user's satisfaction level is higher in comparison with the double auction multi-winning algorithm and the TDSA-PS algorithm.

REFERENCES

[1] D. Gomez-Barquero, J. J. Gimenez, G.-M. Muntean, Y. Xu, and Y. Wu, "IEEE transactions on broadcasting special issue on: 5G media production, contribution, and distribution," *IEEE Trans. Broadcast.*, vol. 68, no. 2, pp. 415–421, Jun. 2022.

[2] A. Chakraborty, M. Kumar, N. Chaurasia, and S. S. Gill, "Journey from cloud of things to fog of things: Survey, new trends, and research directions," *Softw., Pract. Exp.*, vol. 53, no. 2, pp. 496–551, Feb. 2023.

[3] B. T. Maharaj, B. S. Awoyemi, B. T. Maharaj, and B. S. Awoyemi, "Introduction to cognitive radio networks," in *Developments in Cognitive Radio Networks: Future Directions for Beyond 5G*, 1st ed. Cham, Switzerland: Springer, 2022, pp. 3–12.

[4] *Report to the President Realizing the Full Potential of Government-Held Spectrum to Spur Economic Growth*, PCAST Spectrum Policy Invited Experts, Office Sci. Technol. Policy, Washington, DC, USA, 2012.

[5] R. Frieden, "The evolving 5G case study in United States unilateral spectrum planning and policy," *Telecommun. Policy*, vol. 44, no. 9, Oct. 2020, Art. no. 102011.

[6] M. D. Mueck, S. Srikanteswara, and B. Badic, "Spectrum sharing: Licensed shared access (LSA) and spectrum access system (SAS)," Intel, Santa Clara, CA, USA, White Paper, 2015, pp. 1–26.

[7] K. Mun, "CBRS: New shared spectrum enables flexible indoor and outdoor mobile solutions and new business models," CBRS Alliance, OnGo Alliance, Beaverton, OR, USA, White Paper, Mar. 2017, p. 25.

[8] M. M. Sohul, M. Yao, T. Yang, and J. H. Reed, "Spectrum access system for the citizen broadband radio service," *IEEE Commun. Mag.*, vol. 53, no. 7, pp. 18–25, Jul. 2015.

[9] C. W. Kim, J. Ryoo, and M. M. Buddhikot, "Design and implementation of an end-to-end architecture for 3.5 GHz shared spectrum," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Sep. 2015, pp. 23–34.

[10] X. Ying, M. M. Buddhikot, and S. Roy, "SAS-assisted coexistence-aware dynamic channel assignment in CBRS band," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6307–6320, Sep. 2018.

[11] Y. Xiao, S. Shi, W. Lou, C. Wang, X. Li, N. Zhang, Y. T. Hou, and J. H. Reed, "Decentralized spectrum access system: Vision, challenges, and a blockchain solution," *IEEE Wireless Commun.*, vol. 29, no. 1, pp. 220–228, Feb. 2022.

[12] M. Van Otterlo and M. Wiering, "Reinforcement learning and Markov decision processes," in *Reinforcement learning: State-of-the-Art*. Berlin, Germany: Springer, 2012, pp. 3–42.

[13] Y. Huang, "Deep Q-networks," in *Deep Reinforcement Learning: Fundamentals, Research and Applications*. Singapore: Springer, 2020, pp. 135–160.

[14] M. Sewak and M. Sewak, "Deep Q network (DQN), double DQN, and dueling DQN: A step towards general artificial intelligence," in *Deep Reinforcement Learning: Frontiers of Artificial Intelligence*. Singapore: Springer, 2019, pp. 95–108.

[15] J. Baek and G. Kaddoum, "Heterogeneous task offloading and resource allocations via deep recurrent reinforcement learning in partial observable multifog networks," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 1041–1056, Jan. 2021.

[16] B. Tezergil and E. Onur, "Wireless backhaul in 5G and beyond: Issues, challenges and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2579–2632, 4th Quart., 2022.

[17] E. Selvi, R. M. Buehrer, A. Martone, and K. Sherbondy, "On the use of Markov decision processes in cognitive radar: An application to target tracking," in *Proc. IEEE Radar Conf. (RadarConf)*, Apr. 2018, pp. 537–542.

[18] E. Selvi, R. M. Buehrer, A. Martone, and K. Sherbondy, "Reinforcement learning for adaptable bandwidth tracking radars," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 5, pp. 3904–3921, Oct. 2020.

[19] Y. Yao and Z. Feng, "Centralized channel and power allocation for cognitive radio networks: A Q-learning solution," in *Proc. Future Netw. Mobile Summit*, Jun. 2010, pp. 1–8.

[20] J. Lundén, S. R. Kulkarni, V. Koivunen, and H. V. Poor, "Multiagent reinforcement learning based spectrum sensing policies for cognitive radio networks," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 5, pp. 858–868, Oct. 2013.

[21] Y. Li, H. Ji, X. Li, and V. C. M. Leung, "Dynamic channel selection with reinforcement learning for cognitive WLAN over fiber," *Int. J. Commun. Syst.*, vol. 25, no. 8, pp. 1077–1090, Aug. 2012.

[22] J. Oksanen, J. Lundén, and V. Koivunen, "Reinforcement learning based sensing policy optimization for energy efficient cognitive radio networks," *Neurocomputing*, vol. 80, pp. 102–110, Mar. 2012.

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

[24] Z. Shi and G. Luo, "Multi-band spectrum allocation algorithm based on first-price sealed auction," *Cybern. Inf. Technol.*, vol. 17, no. 1, pp. 104–112, Mar. 2017.

[25] M. Devi, N. Sarma, and S. K. Deka, "Multi-winner heterogeneous spectrum auction mechanism for channel allocation in cognitive radio networks," in *Distributed Computing and Internet Technology*. Bhubaneswar, India: Springer, Jan. 2020, pp. 251–265.

[26] I. A. Kash, R. Murty, and D. C. Parkes, "Enabling spectrum sharing in secondary market auctions," *IEEE Trans. Mobile Comput.*, vol. 13, no. 3, pp. 556–568, Mar. 2014.

[27] R. P. McAfee, "A dominant strategy double auction," *J. Econ. Theory*, vol. 56, no. 2, pp. 434–450, Apr. 1992.

[28] W. Dong, S. Rallapalli, L. Qiu, K. K. Ramakrishnan, and Y. Zhang, "Double auctions for dynamic spectrum allocation," *IEEE/ACM Trans. Netw.*, vol. 24, no. 4, pp. 2485–2497, Aug. 2016.

[29] X. Zhou and H. Zheng, "TRUST: A general framework for truthful double spectrum auctions," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 999–1007.

[30] X. Zhang, D. Yang, G. Xue, R. Yu, and J. Tang, "Transmitting and sharing: A truthful double auction for cognitive radio networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[31] X. Feng, Y. Chen, J. Zhang, Q. Zhang, and B. Li, "TAHES: A truthful double auction mechanism for heterogeneous spectrums," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 4038–4047, Nov. 2012.

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

[33] C. Zhong, Z. Lu, M. C. Gursoy, and S. Velipasalar, "A deep actor-critic reinforcement learning framework for dynamic multichannel access," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 1125–1139, Dec. 2019.

[34] C. Schulze and M. Schulze, "ViZDoom: DRQN with prioritized experience replay, double-Q learning, & snapshot ensembling," 2018, *arXiv:1801.01000*.

[35] 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.

[36] Y. Xu, J. Yu, and R. M. Buehrer, "The application of deep reinforcement learning to distributed spectrum access in dynamic heterogeneous environments with partial observations," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4494–4506, Jul. 2020.

[37] A. Wang, L. Zhang, D. Chen, and J. Chen, "Deep reinforcement learning for dynamic multichannel access in multi-cognitive radio networks," *J. Phys., Conf. Ser.*, vol. 1550, no. 3, May 2020, Art. no. 032135.

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

[39] H. Q. Nguyen, B. T. Nguyen, T. Q. Dong, D. T. Ngo, and T. A. Nguyen, "Deep Q-learning with multiband sensing for dynamic spectrum access," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Oct. 2018, pp. 1–5.

[40] Z. Youssef, E. Majeed, M. D. Mueck, I. Karls, C. Drewes, G. Bruck, and P. Jung, "Concept design of medium access control for spectrum access systems in 3.5 GHz," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2018, pp. 1–8.

[41] Y. Azimi, S. Yousefi, H. Kalbkhani, and T. Kunz, "Applications of machine learning in resource management for RAN-slicing in 5G and beyond networks: A survey," *IEEE Access*, vol. 10, pp. 106581–106612, 2022.

[42] Ł. Kułacz, P. Kryszkiewicz, A. Kliks, H. Bogucka, J. Ojaniemi, J. Paavola, J. Kalliovaara, and H. Kokkinen, "Coordinated spectrum allocation and coexistence management in CBRS-SAS wireless networks," *IEEE Access*, vol. 7, pp. 139294–139316, 2019.

[43] X. Liu, C. Sun, M. Zhou, C. Wu, B. Peng, and P. Li, "Reinforcement learning-based multislot double-threshold spectrum sensing with Bayesian fusion for industrial big spectrum data," *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3391–3400, May 2021.

[44] W. Abbass, R. Hussain, J. Frnda, N. Abbas, M. A. Javed, and S. A. Malik, "Resource allocation in spectrum access system using multi-objective optimization methods," *Sensors*, vol. 22, no. 4, p. 1318, Feb. 2022.

[45] W. Abbass, R. Hussain, J. Frnda, I. L. Khan, M. A. Javed, and S. A. Malik, "Optimal resource allocation for GAA users in spectrum access system using Q-learning algorithm," *IEEE Access*, vol. 10, pp. 60790–60804, 2022.

[46] N. N. Krishnan, N. Mandayam, I. Seskar, and S. Kompella, "Experiment: Investigating feasibility of coexistence of LTE-U with a rotating radar in CBRS bands," in *Proc. IEEE 5G World Forum (5GWF)*, Jul. 2018, pp. 65–70.

[47] A. Kliks, P. Kryszkiewicz, Ł. Kułacz, K. Kowalik, M. Kołodziejski, H. Kokkinen, J. Ojaniemi, and A. Kivinen, "Application of the CBRS model for wireless systems coexistence in 3.6–3.8 GHz band," in *Cognitive Radio Oriented Wireless Networks*. Lisbon, Portugal: Springer, 2018, pp. 100–111.

[48] X. Dong, L. Cheng, G. Zheng, and T. Wang, "Network access and spectrum allocation in next-generation multi-heterogeneous networks," *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 8, 2019, Art. no. 1550147719866140.

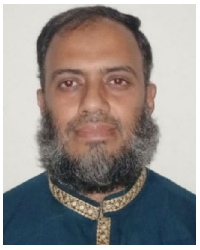
[49] P. Lin, X. Feng, Q. Zhang, P. Lin, X. Feng, and Q. Zhang, "Truthful double auction mechanism for heterogeneous spectrums," in *Auction Design for the Wireless Spectrum Market*. Cham, Switzerland: Springer, 2014, pp. 19–38.

[50] M. Flammini, M. Mauro, M. Tonelli, and C. Vinci, "Inequity aversion pricing in multi-unit markets," in *Frontiers in Artificial Intelligence and Applications*, vol. 325. Amsterdam, The Netherlands: IOS Press, 2020, pp. 91–98.

[51] D. B. Rokhlin, "Robbins–Monro conditions for persistent exploration learning strategies," in *Modern Methods in Operator Theory and Harmonic Analysis*. Rostov-on-Don, Russia: Springer, Apr. 2019, pp. 237–247.

[52] M. Devi, N. Sarma, and S. K. Deka, "A double auction framework for multi-channel multi-winner heterogeneous spectrum allocation in cognitive radio networks," *IEEE Access*, vol. 9, pp. 72239–72258, 2021.

WASEEM ABBASS received the B.Sc. degree in computer engineering from COMSATS University Islamabad (CUI), in 2011, the Master of Science degree in computer engineering from the University of Engineering and Technology (UET) Taxila, in 2014, and the Ph.D. degree in electrical engineering from CUI, in 2023. He is currently an accomplished scholar. He is also an Assistant Professor with the Department of Electrical and Computer Engineering, Capital University of Science and Technology (CUST), Islamabad. His research interests include wireless communication, cognitive radio networks, wireless sensor networks, and the Internet of Things.



RIAZ HUSSAIN received the B.S. degree (Hons.) in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, the master's degree in networks from North Carolina State University, Raleigh, NS, USA, and the Ph.D. degree from the COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2013. His dissertation was titled "Modeling, Analysis and Optimization of Vertical Handover Schemes in Heterogeneous Wireless Networks." He is currently an Assistant Professor with the Department of Electrical Engineering, COMSATS University Islamabad. His current research interests include cognitive radio networks, device-to-device communication, and the Internet of Things.

NASIM ABBAS received the B.Sc. degree in electrical engineering from COMSATS University Islamabad, Pakistan, in 2009, the M.S. degree in electronic engineering from Muhammad Ali Jinnah University, Islamabad, in 2013, and the Ph.D. degree in electrical engineering from the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China, in 2019. His research interests include wireless communication and multimedia wireless sensor networks.

SHAHZAD A. MALIK received the B.S. degree in electrical engineering from the University of Engineering and Technology Lahore, Lahore, Pakistan, in 1991, and the M.S. degree in communication systems and networks and the M.Phil. degree in digital telecommunication systems from Ecole National e Suprieur d'Electrotechnique, Toulouse, France, in 1997 and 1998, respectively. He was a Postdoctoral Research Fellow/Student Project Advisor with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada, from 2003 to 2004. He was an Assistant Professor with the College of Electrical Engineering and Mechanical Engineering, National University of Sciences and Technology, Rawalpindi, Pakistan, from 2004 to 2007. Since 2007, he has been with the Department of Electrical Engineering, COMSATS University Islamabad, Pakistan, where he is currently a Full Professor and the Chairperson of Electrical Engineering. His current research interests include wireless multimedia information systems, mobile computing, QoS provisioning and radio resource management in heterogeneous wireless networks (mobile cellular-2.5/3G/4G, HSPA, LTE, WLANs, WiMAX, MANETs, and WSN), modeling, simulation, performance analysis, network protocols, architecture and security, wireless application development, embedded system design, and the Internet of Things.

MUHAMMAD AWAIS JAVED (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Engineering and Technology Lahore, Pakistan, in August 2008, and the Ph.D. degree in electrical engineering from The University of Newcastle, Australia, in February 2015. He is currently an Assistant Professor with COMSATS University Islamabad, Pakistan. From July 2015 to June 2016, he was a Postdoctoral Research Scientist with the Qatar Mobility Innovations Center (QMIC) on SafeITS Project. His research interests include intelligent transport systems, vehicular networks, protocol design for emerging wireless technologies, and the Internet of Things.

MUHAMMAD ZUBAIR KHAN received the M.Tech. degree in computer science and engineering from Uttar Pradesh Technical University, Lucknow, India, and the Ph.D. degree in computer science and information technology from the Faculty of Engineering, M. J. P. Rohilkhand University, Bareilly, India. He was the Head and an Associate Professor with the Department of Computer Science and Engineering, Invertis University, Bareilly. He is currently an Associate Professor with the Department of Computer Science, College of Computer Science and Engineering, Taibah University. He has more than 15 years of teaching and research experience. He has published more than 60 journal articles and conference papers. His current research interests include the IoT, machine learning, parallel and distributed computing, and computer networks. He has been a member of the Computer Society of India, since 2004.

RAYAN HAMZA ALSISI received the Ph.D. degree in electrical and computer engineering from the University of Western Ontario, London, ON, Canada, in 2018. He is currently the Vice Dean of Development and Quality with the Faculty of Engineering, Islamic University of Madinah, Saudi Arabia, where he is also an Assistance Professor. He is also a Consultant with the Saudi Council of Engineers. He has more than 15 years teaching and research experience. He has authored many technical articles in journals and international conferences. His current research interests include wireless communications, digital communications, communication and information systems, information theory, signal processing, optical communications, the Internet of Things, and communication networks.

ABDULFATTAH NOORWALI (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Western Ontario, London, ON, Canada, in 2017. His thesis titled ''Modeling and Analysis of Smart Grids for Critical Data Communication.'' He is currently the Chairperson of the Electrical and Computer Engineering Department, Faculty of Engineering and Islamic Architecture, Umm Al-Qura University, where he is also an Assistant Professor. He is also a Senior Consultant with Umm Al-Qura Consultancy Oasis, Institute of Consulting Research and Studies (ICRS), Umm Al-Qura University, where he is also the Chairperson of Vision Office of Consultancy. He has authored many technical articles in journals and international conferences. His research interests include smart grid communications, cooperative communications, wireless networks, the Internet of Things, crowd management applications, and smart city solutions.

PRIYADARSHINI PATTANAIK is currently a Senior Researcher with the Faculty of Computer Science and Informatics, Berlin School of Business and Innovation (BSBI), Berlin, presents her area of expertise on medical image analysis and visualization, machine learning (deep learning), computer vision, applied mathematics, and robotics. Her recent project is dedicated to the idea of developing concepts and tools to address one of the great challenges of the musculoskeletal (MSK) field: Understanding and exploiting the link between the shape and the function of a joint. She has many publications at high-impact research journals and conferences. Her Ph.D. thesis was on machine-learning-based classification of microscopic blood smear images for early detection of malaria. During this period, she was partially supported through a research grant from Intel and developed image processing algorithms for deep learning-based digital microscopy and ultrasound imaging reference designs and systems. She was a Postdoctoral Scientist with collaboration with a range of academic, institutional, and industrial partners, such as Télécom SudParis, University of Saclay, a team from the Center for Mathematical Morphology of Mines ParisTech and the company TRIBVN. Her research interests include developing machine learning algorithms with deep neural networks and graphical models for visual computing including medical image analysis and disease detection.