International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(3):1643-1657

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i03.04789

Intelligent Spectrum Access Control in Cognitive Radio Networks: A Q-Learning and MDP Approach Intelligent CR

Anilkumar Dulichand Vishwakarma[1]*, Gajanan Uttam Patil[2], Tushar Hrishikesh Jaware[3], Priti Subramanium[4]

[1]Department of Artificial Intelligence and Data science Engineering, Godavari Foundation's, Godavari College of Engineering, Jalgaon, India, [2]Department of Electronics and Telecommunication Engineering, Hindi Seva Mandal's, Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, India, [3]Department of Electronics and Telecommunication Engineering, S.E.S's, R. C. Patel Institute of Technology, Shirpur, India, [4]Department of Computer Science and Engineering, Hindi Seva Mandal's, Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, India. *Corresponding Author's Email: vishwakarma67anil@gmail.com, anil18vishwakarma@gmail.com

Abstract

Cognitive radio (CR) technology improves frequency resource usage through unlicensed users' opportunistic use of unused spectrum bands without disrupting licensed ones. With growth in wireless communication needs, dynamic spectrum access (DSA) has emerged as a fundamental concept in enhancing spectral efficiency. New CR systems are projected to outgrow traditional artificial intelligence (AI) models, adopting reconfigurable network infrastructures with the ability to manage autonomously elements to provide uninterrupted service quality. To aid this development, a metacognitive level providing self-monitoring learning and adaptation is necessary to fine-tune AI-based decision-making in real time. A new threshold optimization approach for cognitive radio networks, highlighting detection based on the Maximum-Minimum Eigenvalue (MME) criterion, is the theme of this work. The method combines Markov Decision Processes (MDPs) and Q-Learning to support smart spectrum allocation and adaptive spectrum sensing. By adaptively varying parameters based on feedback from the environment, the method enhances decision-making in uncertain and varying network conditions. Simulation outputs show that the model provides enhanced spectrum efficiency, shorter convergence time, and less interference, while maintaining Quality of Service (QoS) for secondary users. This research advances CR systems by marrying signal detection precision with smart learning paradigms to create the potential for strong, autonomous communication networks that can adapt to dynamic spectral conditions.

Keywords: Cognitive Engine, Cognitive Radio, Markov Decision Process (MDP), Maximum-Minimum Eigenvalue (MME), Q-Learning.

Introduction

Recent technology developments, especially in the field of programmable integrated circuits and distributed artificial intelligence, have brought intelligent, autonomous, and interactive devices to a new level. They promise ubiquitous applications within and across several network domains such as cognitive radio networks. As their ubiquity increases, network traffic volume has experienced explosive growth, causing spectrum scarcity and resource famine problems. This increased level of network traffic has motivated researchers and scholars to investigate dynamic and opportunistic access of underused radio frequency bands, both in licensed and unlicensed spectral bands. Although fixed spectrum allocation efficiently avoids interference and collision among various users, it tends to lead to inefficient utilization of the spectrum. This is because spectrum utilization varies with time, and the variations occur on a scale of hours and geographical regions. To address this problem, researchers have proposed a solution that gave birth to cognitive radio technology. Cognitive radio is an innovative technological solution aimed at maximizing the optimization of spectrum use. It does this by facilitating the sharing of white spaces, which are underused parts of the spectrum, among secondary users (non-spectrum licensed users) and primary users (spectrum licensed users). This sharing requires dynamic spectrum access, which is enabled by a variety of methods, which range from auctions, Markov chains, multi-agent systems (MAS), and game theory. Cognitive radio systems attempt to achieve a balance between effective use of the spectrum and coexistence between primary and secondary users in the radio spectrum, thus

(Received 04th April 2025; Accepted 02nd July 2025; Published 30th July 2025)

mitigating the challenges brought about by the dynamic nature of wireless communication. The principal aim of this research is to formulate a dynamic spectrum allocation strategy that maximizes the utilization of the accessible spectrum while minimizing interference to primary (licensed) users within a CRN. In addressing these challenges, numerous studies have explored the application of machine learning techniques, particularly reinforcement learning, for intelligent spectrum access and decision making in cognitive radio networks. By generating a mapping between rewards and actions, reinforcement learning techniques can be used to develop action selection strategies that optimize the return of the environment. A specific advantage of applying reinforcement learning is in contexts where agents have little or no experience or knowledge about the capabilities and objectives of other agents.

The idea here is that the reinforcement learning approach to the problem may be employed as a novel coordination strategy for situations where the currently known coordination mechanisms are ineffectual, since they all rely on knowledge of the environment and information exchanged among agents. Although communication, control, and coordination between agents are frequently valuable and important as an aid to group activities, it does not ensure coordinated behavior (1). It may be time-consuming, and it might impede other problem-solving activities if not well regulated (2). Furthermore, agents that rely heavily on this communication will suffer greatly if its quality is compromised. At times, this connection can be dangerous or even fatal, such as in combat scenarios where adversaries can intercept transmitted messages. Even when such communication is viable and safe, it should be used only when essential.

In the independent kind of learning addressed here, each agent learns to optimize its reinforcement from the environment without explicitly modeling the other agents. However, the environment's response to the actions taken is immediate, and the solution to the channel selection problem contains a large set of optimal solutions due to the environment’s dynamic nature. As a result, neither prior knowledge of the environment's properties nor an explicit model of other agents' capabilities is required. This approach's limitation is its inability to build effective coordination when agent behaviors are strongly interdependent, when responses from the environment are delayed, or when only limited action combinations lead to optimal results. Reinforcement learning algorithms have been proposed as contemporary remedies for the dynamic channel selection problem involving multiple secondary users in uncoordinated environments (3–7).

A mechanism employing Q-learning was introduced to enable dynamic channel selection and optimize metrics such as channel utilization by primary users and packet error rate when compared to random channel selection strategies (4). This approach was later implemented in GNU Radio and yielded satisfactory results (3). Empirical findings have been reported characterizing the convergence behavior of reinforcement learning methods in multi-agent systems, especially in scenarios where the actions of other agents are not observable, which is a common limitation in real-world applications (5–7). Most agents must rely on sensor-based inputs to interact with their environment, and coordination becomes significantly more challenging if information about other agents’ actions or rewards is unavailable.

The primary challenge lies in enabling proactive spectrum utilization while avoiding interference with primary users (PUs). This necessitates the implementation of Cognitive Radio (CR), a technology inherently equipped with the ability to learn from experience—an essential aspect of intelligence. A vision of CR as an intelligent wireless communication system was proposed with the dual objective of ensuring reliable communication and maximizing radio resource utilization (8). Achieving these objectives requires a level of intelligence comparable to human cognitive capabilities (9). Efficient spectrum access is facilitated by an architecture inspired by the human brain, termed the Cognitive Engine (CE) (10). This engine is crucial for enabling CRs to make informed decisions and adapt to varying spectrum conditions.

The CE functions as the core intelligence of CR, executing a range of cognitive tasks and enabling the system to complete the cognitive cycle through machine learning methods. CR technology is fundamentally supported by Software Defined

Radio (SDR), which allows radios to adapt dynamically to environmental conditions through software-driven control. In recent years, increased attention has been given to integrating machine learning techniques into CR networks (11, 12). A comprehensive review of existing work reveals the wide application of various learning algorithms across CR network functionalities (13). Machine learning has been specifically applied to areas such as spectrum sensing and decision-making processes (14).

Cognitive radios are expected to function under unpredictable conditions, with either full or partial Channel State Information (CSI). They must also anticipate the behavior of other cognitive wireless systems (CWS) to coordinate actions effectively. Machine learning approaches have been used to support decision-making and feature classification across a broad set of environmental scenarios. Learning becomes essential when the input–output relationships in the system are vague, which is common in the context of CWS due to channel unpredictability.

Thus, learning becomes a powerful strategy for estimating channel characteristics and minimizing error probability (9). Multiple parameters in CRs, such as spectrum availability (15), transmit power (16), adaptive coding and modulation (9), antenna selection, rate control (17), and spectrum handoff (18, 19), must be jointly optimized. Because it is impractical to manually configure all these factors simultaneously, learning algorithms provide an effective means to automate and optimize CR functionalities.

Research Gap

In the modern world of research on cognitive radio networks, a major gap can be sensed in the area of optimizing threshold levels for the spectrum. Although there has been investigating various techniques and algorithms for dynamic spectrum sensing and allocation in previous research works, a comprehensive investigation into how the optimal threshold values must be optimized effectively, with special focus on factors such as noise dependence and performance criteria, remains an unexplored area. This gap leaves room for further research to develop and evaluate innovative solutions that can enhance the effectiveness and flexibility of cognitive radio networks in wireless environments subject to dynamism and uncertainty.

Problem Formulation

In cognitive radio networks (CRNs), efficient and agile spectrum allocation is of utmost significance in fulfilling the growing demand for wireless communication, while at the same time ensuring the best possible utilization of the available spectrum resources. The major stumbling block is in adapting to the constantly shifting and heterogeneous environment of the radio frequency spectrum, where licensed and unlicensed users share space.

Components of the Problem

State Space (S): In our problem statement, state space refers to the set of possible spectrum states, each corresponding to the occupancy status of separate frequency bands in the CRN. The state can be distinguished based on factors such as channel availability, interference levels, and previous information regarding spectrum usage.

Action Space (A): The action space comprises the available spectrum allocation decisions that can be made by the cognitive radio nodes. This includes decisions on which frequency bands to transmit or receive on, power levels, and modulation schemes.

Reward Function (R): The reward function establishes the instant benefit or usefulness linked to undertaking a specific action within a given state. It quantifies the balance between enhancing data transmission and steering clear of interference with primary users. The primary goal is to maximize the cumulative reward over an extended period.

Transition Probabilities (P): The transition probabilities model the stochastic nature of the CRN, representing how the system state evolves. These probabilities capture the dynamics of channel availability, primary user activity, and spectrum variations.

Methodology

System Model

Consider $X_c(t)$ as the continuous-time received signal, where it is the sum of $S_c(t)$ representing the detected primary signal and $W_c(t)$ representing the modeled noise signal. The noise signal is modelled as a fixed procedure having zero mean as well as the variance of σ_η^2. Two fundamental signal detection assumptions are developed, and incoming continuously timed signals undergo sampling. H_0 Signifies the non-existence of the signal, whereas H_1 signifies its presence.

$$H_0: X(n) = W(n) \quad [1]$$
$$H_1: X(n) = S(n) + W(n) \quad [2]$$

The gathered samples of the signal display feature, such as path attenuation, multipath fading, and temporal spreading. In addition to energy detection, our proposed hybrid system integrates eigenvalue spectrum sensing. Notably, the threshold values, denoted as ξ_1 and ξ_2, of the energy detector are contingent on the noise factor, with superior performance observed as the noise factor decreases. Nevertheless, it is crucial to acknowledge that the efficacy of energy detection diminishes in the face of uncertain noise, resulting in the emergence of an SNR threshold and an increased susceptibility to false alarms. Utilizing methodologies delineated in the current body of literature, we propose two strategies grounded in the sample covariance matrix obtained from signals received at the sensing node. These strategies pivot on the evaluation of the maximum-to-minimum eigenvalue (MME) ratio and the ratio of average signal power to the minimum eigenvalue (EME). In the realm of detection, MME takes precedence, demonstrating superior performance compared to EME consistently.

MME (Maximum-Minimum Eigenvalue) Detection

Using N_s samples, the following equation is employed to compute the covariance matrix for the received signal samples

$$R_x(N_s) = \frac{1}{N_s} \sum_{n=L-1}^{L-2+N_s} x(n)x\dagger(n) \quad [3]$$

Here, the Hermitian (transpose-conjugate) operation is denoted by †.

The matrix$R_x(N_s)$'s maximum and minimum eigenvalues, designated as $\lambda_{max}$and λ_{min}, are then calculated.

Predetermined threshold value γ is used to compare the ratio of $\lambda_{max}/\lambda_{min}$ before making an ultimate detection of signals. The signal occurs if max/ min>γ_1; else, no signal is present.

High spectrum sensing accuracy is required in cognitive radio systems to balance spectrum utilization and protect the primary user. Two important measures involved here are the Probability of Detection (P_d) and the Probability of False Alarm (P_{fa}). High P_d prevents interference with active primary users, whereas low P_{fa} prevents underutilization of available spectrum resources. In this manuscript employ the Maximum-Minimum Eigenvalue (MME) approach, which is resistant to low-SNR conditions and independent of noise power estimation. For further improvement in sensing reliability, propose a threshold optimization method based on the MDP-Q learning approach. An adaptive method that adjusts the detection threshold dynamically according to environment feedback is utilized for maintaining the optimal balance between P_d and P_{fa}under a non-stationary spectrum. In addition, the detection threshold is calculated analytically by Tracy-Widom distribution and random matrix theory, enabling the system to have controlled false alarm probabilities even in noise uncertainties. The design of the system makes the cognitive radio sensing model both practical and dependable for actual deployment in cognitive radio systems.

Parameters for Probability and Threshold Setting in MME Detection

The formulations for the probability of false alarms and the threshold value are established employing random matrix theory and particular distribution functions. We present approximate formulas for the performance metrics and threshold value by treating $R_w(N_s)$ as a Wishart random matrix and utilizing Tracy-Widom distributions to describe its Eigenvalues.

False Alarm Probability (P_{fa}) in the context of MME detection

$$P_{fa} = 1 - F_1\left[\frac{\gamma\left(\sqrt{N_s} - \sqrt{ML}\right)^2 - \mu}{v}\right] \quad [4]$$

Where values of Tracy-Widom distribution function $F_1(t)$ are known.

Threshold: We find following formula for threshold

$$\gamma = \frac{\left(\sqrt{N_s} + \sqrt{ML}\right)^2}{\left(\sqrt{N_s} - \sqrt{ML}\right)^2} \cdot \left[1 + \frac{\left(\sqrt{N_s} - \sqrt{ML}\right)^{-\frac{2}{3}}}{(N_s ML)^{\frac{1}{6}}} \cdot F_1^{-1}(1 - P_{fa})\right] \quad [5]$$

Where *L* is smoothing factor, *M* is sampling factor, as well as N_s is no. of samples taken. The threshold setting has no relationship with noise power. The threshold may be calculated in advance using simplyN_s, L and P_{fa}.

The probability of detection (P_d)

When a signal is present, sample matrix of covariance $R_x(N_s)$ is ceased to be a Wishart matrix. In this scenario, eigenvalue distributions were unknowable. The following are approximate formulas for the chance of detection:

$$P_d = 1 - F_1 \frac{\left[\gamma N_s + \frac{N_s(\gamma\rho ML - \rho)}{\sigma_n^2} - \mu\right]}{v} \quad [6]$$

P_d is affected by no. of samples N_s, signal covariance matrix's greatest as well as least eigenvalues.

In this study, the optimization of the threshold is accomplished through the utilization of Markov Decision Process (MDP) as well as Q-Learning-based reinforcement learning models. This approach is elaborated upon in the subsequent subsection.

Threshold Optimization Using Reinforcement Learning Models

Markov Decision Process (MDP)

This framework capable of resolving the majority of discrete action reinforcement learning challenges. An agent can use the Markov decision process to arrive at an optimum strategy for maximum rewards over time (20).

An MDP can be denoted by the tuple $< S, A, T, R >$ with the following components:

• S: A finite collection of states $\{s^1, \dots, s^N\}$ with a state space size of N. A state $s \in S$ is a detailed representation of all the elements constituting a state in the simulated problem (21, 22).

• A: A finite assortment of actions $\{a^1, \dots, a^K\}$ with an action space size of K is defined. The system's condition can be controlled by executing actions. $A(s)$ Signifies the collection of actions permissible in a specific state$s \in S$, where$A(s) \subseteq A$. In the state$s \in S$, an action $a \in A$ is deemed appropriate (21, 22).

• T:Upon executing the action, $a \in A$ in a given state $s \in S$, a probability distribution encompassing the array of feasible transitions is employed to determine the system's shift from state s to a new state $s' \in S$. Following the definition of the transition function T, denoted as $T: S \times A < S \rightarrow [0,1]$, the likelihood of being in state s' while undertaking action a in state s is represented as $T(s, a, s')$. For any action$a$, and any states s and s', $T(s, a, s') \geq 0$ and $T(s, a, s') \leq 1$ are prerequisites. Naturally, it is essential to ensure that $\sum_{s' \in S} T(s, a, s') = 1$ for all states and actions, thereby establishing T as a valid probability distribution across potential succeeding states (21, 22).

• R: The reward function defines the benefits for being in a state or executing an activity while in a state. The reward acquired within the states is determined by the state reward function, denoted as $R: S \rightarrow R$. There are, however, two different definitions. We can define $R: S \times A \rightarrow R$ as a reward for executing an action in a state, or $R: S \times A \times S \rightarrow R$ as a reward for certain transitions between states (21, 22).

The MDP model is characterized by the reward function R and the transition function T. MDPs are commonly illustrated as graphs depicting state transitions, where nodes symbolize states, and directed edges signify transitions.

The Apprenticeship Policy

In the context of an MDP $< S, A, T, R >$, we define a policy as a function created for each state $s \in S$ and action $a \in A$. Formally, we express the policy as $\pi = S \times A \rightarrow [0,1]$.

To apply a policy to an MDP, several steps are undertaken. Initially, the initial state distribution I is utilized to generate an initial states_0. Subsequently, the policy advises the action$a_0 = \pi(s_0)$, which is then executed. Thereafter, a transition to the state s_1 is made with a probability of$T(s_0, a, s_1)$, and a reward r_0 is determined using the reward function R along with the transition function$T$.

This sequence persists, producing $s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, \dots.$, concluding when the target is achieved.

The policy is an integral element of the agent, aiming to influence the environment, typically represented by an MDP (21, 22).

Criteria for Optimization and Updating

The objective of learning within a MDP is to acquire benefits. If agent is just concerned with the immediate reward, maximizing $E[r_t]$ is a simple optimality criterion. However, there are numerous methods to consider the future in order to know how to behave in the present. In the MDP, there are essentially three optimality models, which are adequate to encompass most techniques in the literature (21, 22):

- Finite horizon model: It takes a limited horizon having length h as well as specifies that agent must optimise their anticipated rewards to that horizon; it may be described as $E[\sum_{t=0}^{h} r_t]$. Nevertheless, the challenge with this model lies

in the uncertainty regarding the pre-determination of the optimal horizon length (h).

- Infinite horizon model: In this model, long-term rewards are considered, then future benefits are discounted based on the period from which they will be received, as represented by$E[\sum_{t=1}^{\infty} \gamma^t r_t]$, γ is the discount factor with a value of $0 \leq \gamma < 1$. This approach is more practical technically, but logically it is comparable to the finite horizon model.
- The average reward model: It is the model that maximizes the long-term average reward, is signified by the equation $\frac{1}{h} E[\sum_{t=0}^{h} r_t]$. This is sometimes referred to as the optimal gain policy, and when the discount factor approaches one, it equals the infinite horizon discounted model.

The choice of these optimality criteria may be associated with the particular learning challenge under consideration. Finite horizon model is the best choice if the duration of the episode is known. However, because this is frequently unknown or the activity persists, the infinite horizon approach is preferable.

Bellman's Equations and Value Functions

MDP is defined and the optimality criteria employed to acquire optimal policies in earlier sections. The value functions are defined here as a mechanism connects optimality criteria to policies. The majority of MDP learning methods compute optimum policies by learning a value function. The latter is an evaluation of the agent's quality in a given condition: $V^{\pi}(s) = E_{\pi}\{s_t = s\}$ or of the quality of the execution of a certain action in this state: $Q^{\pi}(s, a) = E_{\pi}\{s_t = s, a_t = a\}$. We use E_{π} for the expectation under the policy π (21, 22).

The objective of an MDP is to discover the optimal policy, i.e., the policy yielding the highest return. This pertains to the act of maximizing the value function$V^{\pi}(s)$ for all $s \in S$ states. Optimum policy π^* is one in which $V^{\pi^*}(s) = V^{\pi}(s)$ for every $s \in S$ and all policies. In this situation, the optimality equation of Bellman is defined as follows.

$$V^{\pi^*}(s) = \sum_{s' \in S} T(s, a, s') \left(R(s, a, s') + \gamma V^{\pi^*}(s')\right) \quad [7]$$

As per Equation (7), the value of a state under an optimal policy equals the anticipated return for the optimal action within that state. The optimal action value function is defined as follows:

$$Q^*(s, a) = \sum_{s' \in S} T(s, a, s')(R(s, a, s') + \gamma Q^*(s', a')) \quad [8]$$

$V^*(s) = Q^*(s, a)$ is the connection between Q^* and V^*. In other words, the optimal action refers to the option with the greatest anticipated utility, considering the future possible states that may arise (21, 22).

We can calculate the optimum policies now that we've described the policies, MDPs, value functions, and optimality criteria. Solving a certain MDP is the same as determining the best policy π^*. The Q-learning algorithm developed to solve MDPs is presented in the next subsection.

Q-Learning

Q-learning falls under the category of a reinforcement learning algorithm designed to identify the most favorable course of action based on the current situation. It is regarded non-policy compliant because the Q-learning feature learns behaviors that are against current policy, such as performing random actions, and so no policy is necessary. Q-learning, in particular, attempts to create a policy that maximizes overall reward (23). At the core of Q-learning lies the essential idea of iteratively estimating Q-value functions for actions, taking into account rewards and the agent's existing Q-value function. The updating rule encapsulates a variation of this learning concept, wherein Q_t is advanced to Q_{t+1} through the utilization of Q-values and an inherent max operator applied to the Q-values of the states specified below:

$$Q_{k+1}(s_t, a_t) = Q_k(s_t, a_t) + \alpha(r_t + \gamma Q_k(s_{t+1}, a_t) - Q_k(s_t, a_t)) \quad [9]$$

With$\alpha \in [0,1]$, which is just a degree of acceptance of the new value in comparison to the previous one?

The agent transitions from state s_t to state s_{t+1} and simultaneously acquires the reward r_t. The modification occurs on the action a_t value Q in the state s_t where this particular action was executed. The Q-Table data structure is employed to discern the maximum expected future rewards for the activities in each state; this table guides us to the optimal action in every state.

The cognitive radio environment under consideration here is non-stationary, with time-varying activity of primary users, varying channel conditions, and time-varying levels of interference and noise. These changes cause both the transition probabilities and reward structure of the system to change over time. In order to counter this non-

stationarity, the reinforcement learning-based approach followed here uses Q-learning to update its Q-values in real-time based on environmental feedback. The algorithm does not depend on a fixed model of the environment; rather, it employs an exploration-exploitation approach to learn and adapt optimal policies dynamically. The learning rate parameter keeps the model responsive to recent changes, allowing effective adaptation to changing spectrum conditions. The system is made robust in dynamic and unpredictable radio environments by using this ongoing learning mechanism.

In real-world cognitive radio environments, misidentification of primary user (PU) activity can lead to interference that constitutes a violation of regulatory spectrum usage policies. To avoid this, the proposed solution includes robust sensing and learning mechanisms. The MME detection method achieves stable performance under low-SNR and noise-uncertainty conditions that prevail in practical deployments. Besides, the detection threshold used is not fixed; it is optimized by an MDP-Q learning-based policy dynamically responding to changes in the environment. In order to prevent interference, the reward function structures severe negative rewards for actions that cause collisions with primary users. This reward structure induces conservative spectrum access behavior in situations of uncertainty. The adoption of Tracy-Widom-based threshold modeling also ensures false alarms and misdetections are statistically regulated. Combined, these components ensure real-time regulatory limits are met and the integrity of licensed use of the spectrum is maintained. The methodology of the Q-learning algorithm is illustrated in Figure 1.

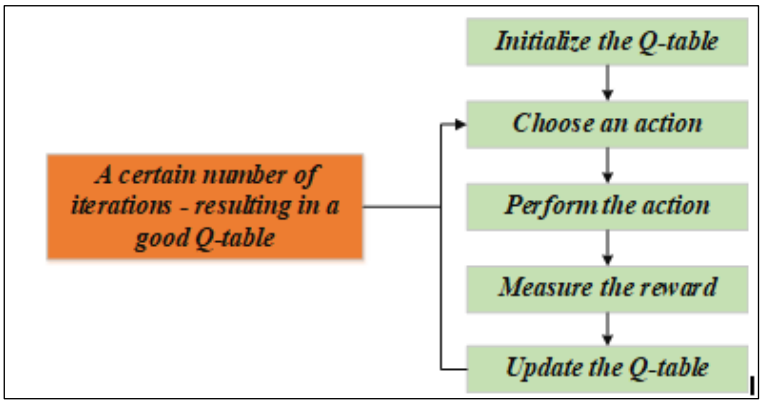

Figure 1: Q-Learning Algorithm

Reward Function Definition and Justification

In the proposed reinforcement learning model, the reward function $R(s,a)$ is designed to guide the cognitive radio agent towards optimal behaviour in a dynamic spectrum environment. The reward function is defined as follows,

- $R(s,a) = +1$ if the secondary user selects an idle channel and successfully transmits without causing interference.
- $R(s,a) = -10$ if the action causes interference with a primary user.
- $R(s,a) = 0$ if the selected channel is idle, but transmission fails due to environmental noise or weak signal conditions.

They were selected to strongly deter damaging interference against primary users while encouraging efficient use of the spectrum. The high penalty for interference represents regulatory adherence and maintains the QoS for licensed users. The neutral reward for failed but non-interfering transmissions enables the agent to learn from uncertainty without early penalization, facilitating exploration. Reward weights were empirically adjusted through iterative simulation runs and were mapped to best practice in reinforcement learning-based spectrum access architectures. Reward space shaping facilitates better convergence and stability in learning.

Results and Discussion

We evaluate the efficacy of the proposed RADDPG learning algorithm through extensive simulation studies. Tables 1 to 4 depict in the underlay scenario, a primary network coexists on a single channel with the primary user (PU) and a secondary system, with an interference ratio set to 1.2 dB. Each secondary user (SU) in the sequence

moves at a random speed of 0.99 m/s. The transmission power and PU Gaussian noise power are both configured at 1nW and 10mW, respectively. SUs and PUs are distributed randomly within a 300-meter radius of their respective base stations. The primary and cognitive base stations are situated at a distance of 2.5 kilometers. The path loss exponent for the channel gain is set at 2.9.

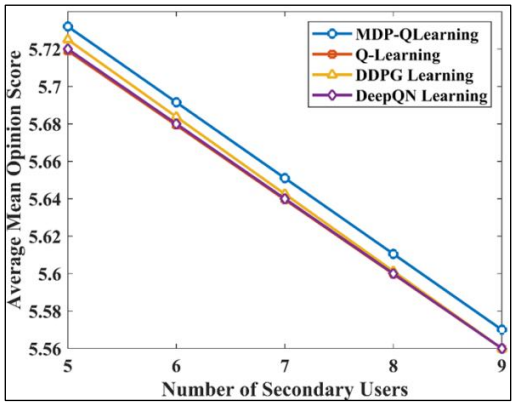

Figure 2: Comparative Graph for Mean Opinion Score (MOS) for Two Sets of SUs

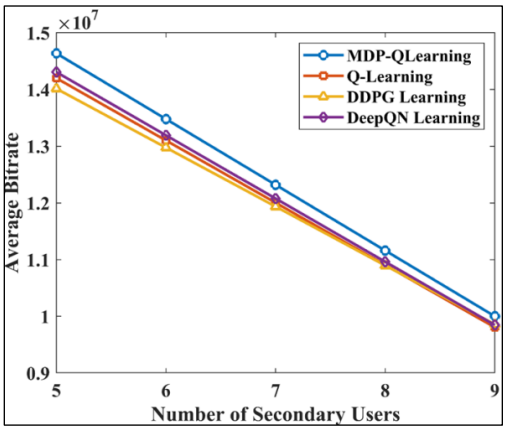

Figure 3: Comparative Graph for Average Bitrate for Two Sets of SUs

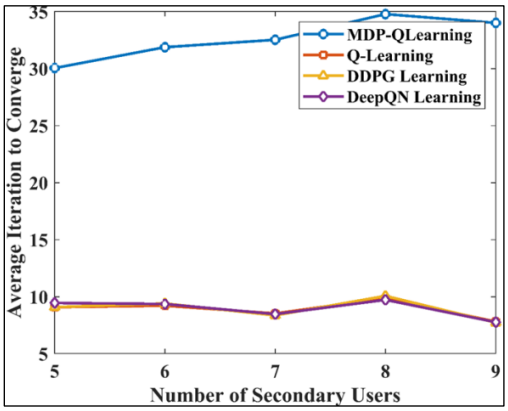

Figure 4: Comparative Graph for the Average Number of Iterations vs the Number of Convergences for Two Sets of SUs

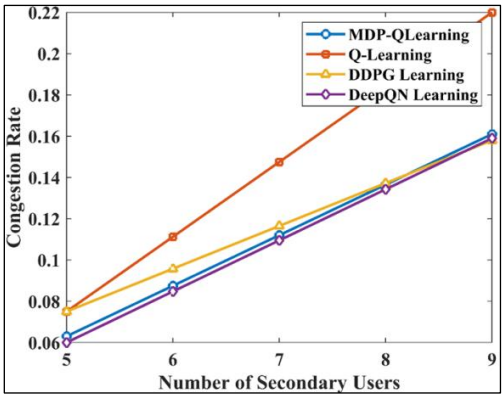

Figure 5: Comparative Graph for Congestion Rate for Two Sets of SUs

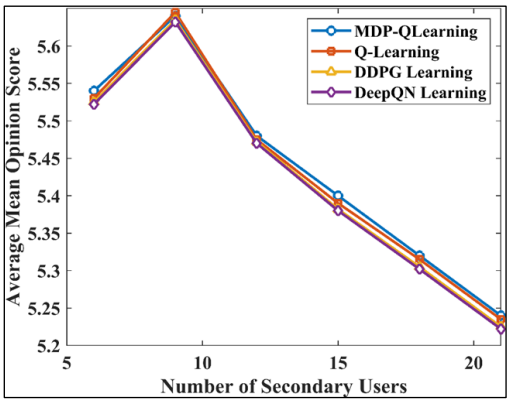

Figure 6: Comparative Graph for Mean Opinion Score (MOS) for Five Sets of SUs

Table 1: Tabular Results for Mean Opinion Score (MOS) vs. the Number of Secondary Users with Five Sets of SUs

Number of Secondary Users	Mean Opinion Score (MOS) for Various Methods			
	MDP-Q Learning	Q learning	Deep Deterministic Policy Gradient (DDPG) Learning	Deep Q network (Deep QN) Learning
5	5.5197	5.5102	5.4995	5.5112
9	5.6310	5.6206	5.6234	5.6215
13	5.4188	5.3989	5.4058	5.3985
17	5.3410	5.3316	5.3357	5.3378
21	5.2473	5.2466	5.2470	5.2462

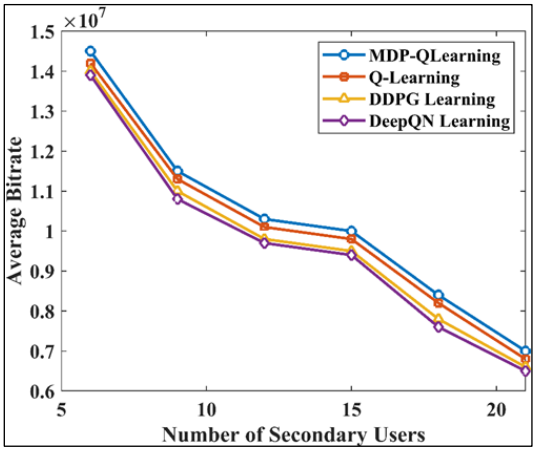

Figure 7: Comparative Graph for Average Bitrate for Five Sets of SUs

Table 2: Tabular Results for Average Bitrate vs. No. of Secondary Users with Five Sets of SUs

Number of Secondary Users	Average Bitrate for Various Methods			
	MDP-Q Learning	Q learning	DDPG Learning	Deep QN Learning
5	14637351.09	14200078.56	14014694.99	14307305.25
9	10111856.96	9721093.81	9674766.18	9715688.91
13	9773130.36	9298987.07	9465149.72	9351236.03
17	8078443.14	7871497.66	7784489.83	7875037.08
21	6749612.67	6705425.35	6656853.24	6646357.72

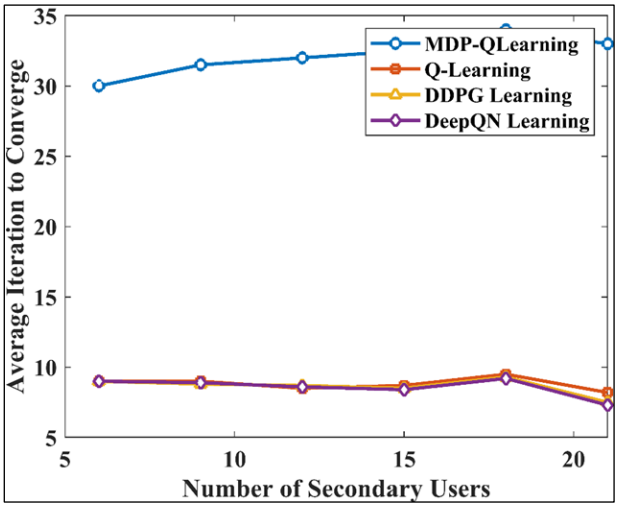

Figure 8: Comparative Graph for the Average Iteration Number of Convergence for Five Sets of SUs

Table 3: Tabular Results for Average Iteration Number of Convergence vs. the Number of Secondary Users with Five Sets of SUs

No of Secondary Users	Average Iteration No of Convergence for Various Methods			
	MDP-Q Learning	Q learning	DDPG Learning	Deep QN Learning
5	30.0472	9.0805	9.0841	9.4488
9	31.8776	9.2107	9.4021	9.3588
13	32.5273	8.5261	8.3594	8.4697
17	34.7807	9.8777	10.0675	9.7231
21	34.0074	7.8105	7.7252	7.7647

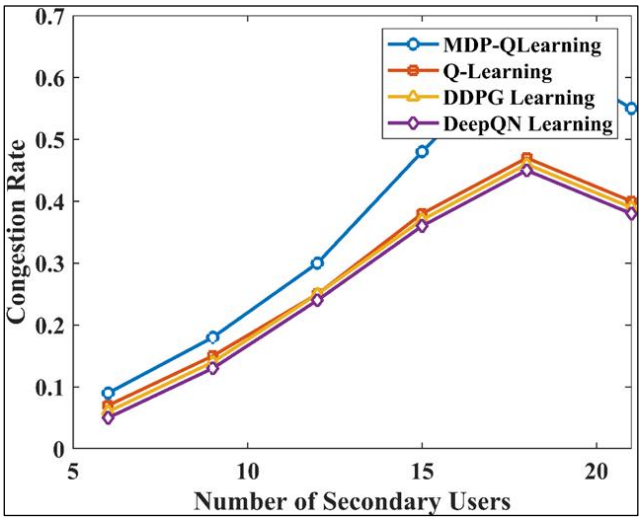

Figure 9: Comparative Graph for Congestion Rate for Five Sets of SUs

Table 4: Tabular Results for Congestion Rate vs. the Number of Secondary Users with Five Sets of SUs

Number of Secondary Users	Congestion Rate for Various Methods			
	MDP-Q Learning	Q learning	DDPG Learning	Deep QN Learning
5	0.0555	0.0660	0.0650	0.0610
9	0.1545	0.1535	0.1630	0.1710
13	0.2795	0.2950	0.2790	0.2720
17	0.4815	0.4690	0.4725	0.4690
21	0.3885	0.4120	0.3830	0.3900

The simulation settings employed in this manuscript were chosen to closely model real-world cognitive radio network (CRN) deployment situations. The 1 nW SU transmission power is used to represent ultra-low-power communication, typically seen in low-energy IoT or edge devices that operate in shared spectrum bands and have to limit interference to primary users (PUs). The PU Gaussian noise power at 10 mW is consistent with average noise floor values found in urban and suburban macro-cell networks. Mobility speed was set to 0.99 m/s to model slow-moving equipment like wearable for health monitoring or pedestrian IoT nodes. The 300-meter user distribution radius models local-area coverage like that of smart campuses or public safety areas. The 2.5 km spacing between primary and cognitive base stations models a realistic inter-cell distance in rural or semi-urban scenarios where spectrum reuse is possible. Finally, the path loss exponent of 2.9 simulates real outdoor propagation conditions with moderate obstructions, typical in suburban settings. These parameter values guarantee that the assessment of the MDP-Q learning algorithm proposed is rooted in realistic operating scenarios, thus providing added realism to the applicability of the results to actual CRN deployment contexts.

In terms of the quantity of Secondary Users (SUs) in the system, we contrast the proposed RADDPG learning algorithm with the classical Q-learning and DQN learning algorithms. Our simulation outcomes unveil three pivotal performance metrics: Mean Opinion Score (MOS), congestion rate, and average convergence cycle. As depicted in Figures 2 and 3, the MOS value experiences a reduction as the number of SUs in the network rises. Each SU is compelled to converge at a lower Signal-to-Interference-plus-Noise Ratio (SINR) value to adhere to the interference constraints stipulated in equation (2), resulting in a dip in the MOS value. However, it is noteworthy that the MOS value consistently surpasses the permissible threshold (MOS > 3).

The network congestion rate for 5 secondary users is shown in Figure 9. The network congestion rate may increase in circumstances where SINR restrictions are broken, even though all SUs are within acceptable MOS levels. The number of SUs that may be sustained by the system depends on the congestion percentage.

For only 9 SUs in our study, the congestion rate has practically reached 1. The congestion rate will be a major worry if the number of SU is raised further. The clustering strategy, in which a cluster head handles key communications, can be used to boost system scalability. However, it's essential to note that the scope for research work does not encompass the clustering of networks.

In order to assess scalability, the proposed framework tested different network conditions by varying the number of secondary users (SUs) from 5 to 21 over multiple available channels. The devised MDP-Q Learning model performed robustly in all measurements, MOS, average bitrate, and convergence rate, under higher network loads. The system is naturally distributed, with each SU acting autonomously and learning its spectrum access policy from local observations and rewards. This decentralized strategy minimizes the complexity of coordination and provides a guarantee of effective scalability of the system with higher densities of users or channels. In cases of extremely high numbers of channels or users, the framework can be extended via deep reinforcement learning to estimate Q-values to improve the scalability even more without changing decision-making logic.

Figures 4 to 8 display the average convergence iteration cycles for the number of SUs. The effectiveness of a learning algorithm is determined by its convergence rate and stability. The presented chart showcases the most demanding situation for the convergence of Q-learning. To

mitigate the intricacies associated with high-dimensional action spaces, Deep Q-Networks (DQN) employ deep neural networks (DNNs) to predict the intricate relationship between state and action. In contrast to the conventional Q-learning approach, DQN demonstrates enhancements in network convergence speed, congestion rate reduction, and improvements in the Mean Opinion Score (MOS) value. Furthermore, the RADDPG (Recurrent Advantage-Disadvantage Deep Deterministic Policy Gradients) method outperforms both Q-learning and DQN in terms of performance. This superiority arises from its ability to fine-tune the policy by amalgamating the policy gradient and value function learning processes.

Justification

For the comprehensive assessment of the effectiveness of the suggested MDP-Q Learning model, we have performed a comparative study with three popular reinforcement learning methods, namely normal Q-Learning, Deep Q-Network (DQN), and Deep Deterministic Policy Gradient (DDPG). These baseline models were chosen because they are very common in dynamic spectrum access (DSA) studies for cognitive radio networks. Q-Learning is a basic tabular reinforcement learning method and is traditionally employed as a baseline in decentralized CRN scenarios because of its ease of implementation and capacity for learning in discrete state-action environments. DQN is an extension of Q-Learning by the application of deep neural networks for function approximation so that it can be adapted for high-dimensional state spaces. It is also famous for its better generalization and convergence performance in complicated wireless scenarios. DDPG, which is a policy-gradient-based method, is especially suited to continuous action spaces and has been investigated in CRNs to address real-time spectrum control in adaptive environments.

Table 5: Overall Performance Comparison Table

Metric	MDP-Q Learning (proposed)	Q learning	DDPG Learning	Deep QN
Mean Opinion Score (MOS) (avg)	5.4316	5.4216	5.4223	5.4230
Average Bitrate (bps)	9,953,278.84	9,411,416.89	9,336,390.39	9,459,925.80
Average Iteration to Convergence	32.0480	8.9011	8.9157	8.9522
Congestion Rate (avg)	0.2719	0.2791	0.2725	0.2726

As Table 5 indicates, all these baseline methods have certain advantages. Q-Learning has quicker convergence cycles because of its straightforward update rules and minimal exploration overhead. DDPG and DQN provide relatively stable learning in dynamic settings but are hyper parameter-sensitive and need additional training samples to achieve useful policies. Although each has its own merits, the proposed MDP-Q Learning model performs better than all three baselines on all key performance metrics: Mean Opinion Score (MOS), average bitrate, congestion rate, and overall stability of learning. This sustained advancement proves that the integration of the Markov Decision Process and Q-Learning allows for smarter, adaptive, and interference-sensitive spectrum allocation. By better formulating the underlying state transitions and maximizing the long-term accumulation of rewards, the proposed approach is capable of achieving better performance in dynamic and uncertain CRN settings.

Constraints

Interference Constraints: Decisions on spectrum allocation should meet interference constraints to avoid harmful interference to primary users. The constraint guarantees regulatory compliance and protection of the integrity of licensed communication.

Quality of Service (QoS) Constraints: Quality of Service (QoS) requirements for secondary users, such as minimum latency or data rate, require satisfaction during optimization of spectrum allocation.

Approach

This paper proposed using Markov Decision Processes (MDPs) and Q-learning as a solution

framework. MDPs provide a formalism to describe the sequential decision-making process, and Q-learning gives a reinforcement learning mechanism to learn optimal policies for spectrum allocation in dynamic and uncertain environments.

Expected Outcome: It is anticipated that the new method will provide an adaptive and smart spectrum allocation mechanism, maximizing the throughput and efficiency of cognitive radio networks. At the same time, it should comply with regulatory needs and Quality of Service (QoS) restrictions.

Significance

The proposed dynamic spectrum allocation scheme has the potential to enhance the effectiveness and scalability of cognitive radio networks, making them stronger and more flexible to the constantly evolving wireless communication environment.

Contribution of Research Work

Threshold Optimization: In a significant contribution, this research provides an innovative approach for optimizing threshold values in cognitive radio networks. It addresses the critical problem of dynamic spectrum access by considering aspects of noise dependence, interference avoidance, and efficient use of the spectrum.

Integration of MDP and Q-Learning: The research integrates two effective reinforcement learning methods, Markov Decision Processes (MDP) and Q-Learning, for the optimization of spectrum access thresholds. This convergence of machine learning techniques presents an intelligent decision-making platform for dynamic spectrum allocation.

Noise Factor Dependency: It is a significant contribution to take noise factor dependency into account when setting threshold values. The adaptive method considers different noise conditions and thus performs better as noise factors are reduced.

Performance Metrics: The study formulates and assesses performance metrics for threshold optimization and provides an effective understanding of the various threshold settings affecting system performance, such as reliability and efficiency metrics.

Tracy-Widom Distribution: This paper utilizes Tracy-Widom distribution functions to simulate the Eigenvalues of the signal covariance matrix. In this way, it gives a strong mathematical platform for describing threshold values and false alarm probabilities.

Covering the Research Gap: The research points to and covers an important research gap in the domain of cognitive radio networks. It illustrates the need for more complex and responsive threshold optimization methods that bridge the knowledge gap available and provide insight into spectrum access in dynamic wireless networks.

Potential for Real-World Implementation: The suggested threshold optimization method has potential for real-world implementation in practical cognitive radio systems. It has the potential to optimize spectrum use and reliability and thus make a contribution to more efficient and adaptive wireless networks.

Contribution to Wireless Communications: Finally, this research contribution contributes to the overall wireless communications domain by providing a novel and smart solution for spectrum access management, which is critical in meeting the needs of contemporary wireless networks.

Conclusion

This manuscript introduces a strong and flexible spectrum sensing and allocation scheme for cognitive radio networks, based on Maximum-Minimum Eigenvalue (MME) detection combined with reinforcement learning with a Markov Decision Process and Q-learning. The suggested model allows threshold optimization in dynamic and uncertain wireless environments, overcoming noisy uncertainty, non-stationary utilization of the spectrum, and interference avoidance. Using extensive simulations, the method was tested on a variety of performance metrics such as MOS, average bitrate, convergence time, and congestion rate. The outcomes indicated that the suggested MDP-Q Learning strategy outperformed conventional methods like baseline Q-Learning, DQN, and DDPG when it came to utilizing spectrum and learning efficiency, especially when the secondary users increased from 5 to 21. The application of reinforcement learning for adaptive threshold control effectively minimized false alarms while maintaining primary user protection, ensuring adherence to real-time regulatory limits. The decentralized aspect of the learning agents also guarantees scalability and robustness in multi-user, multi-channel settings. In future

research, plan to generalize this framework with the use of deep reinforcement learning for massive-scale deployments and investigate cooperative multi-agent learning models for collaborative spectrum access in extremely dense network environments.

Abbreviations

AI: Artificial Intelligence, CE: Cognitive Engine, CR: Cognitive Radio, CRN: cognitive radio networks, CSI: Channel State Information, CWS: Cognitive Wireless Systems, DDPG: Deep deterministic policy gradient, Deep QN: Deep Q-Network, DSA: Dynamic Spectrum Access, EME: Minimum Eigenvalue, MAS: Multi-Agent Systems, MDP: Markov Decision Process, MME: Maximum-Minimum Eigenvalue, MOS: Mean Opinion ScorE, PER: Packet Error Rate, PU: Primary User, QoS: Quality of Service, SDR: Software Defined Radio, SNR: Signal to Noise Ratio, SU: Secondary User, RADDPG: Recurrent Advantage-Disadvantage Deep Deterministic Policy Gradients.

Acknowledgement

The authors would like to thank the Deanship of Godavari College of Engineering for supporting this work.

Author Contributions

AD Vishwakarma: data collection, conceptualization, methodology, data collection, writing the manuscript, GU Patil: analysis the dataset, conceptualization, TH Jaware: analysis the overall concept, writing, editing, P. Subramanium: analysis the paper, supervisor.

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethics Approval

No ethics approval is required.

Funding

No fund received for this project.

References

1. Halpern JY, Moses Y. Knowledge and common knowledge in a distributed environment. Journal of the ACM (JACM). 1990 Jul 1;37(3):549-87.
2. Agrawal SK, Samant A, Yadav SK. Spectrum sensing in cognitive radio networks and metacognition for dynamic spectrum sharing between radar and communication system: A review. Phys Commun. 2022;54:101673.
3. Ren Y, Dmochowski P, Komisarczuk P. Analysis and implementation of reinforcement learning on a GNU radio cognitive radio platform. In: Proceedings of the Fifth International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM). 2010:1–6. DOI: 10.4108/ICST.CROWNCOM2010.9170
4. Yau KLA, Komisarczuk P, Paul DT. Enhancing network performance in distributed cognitive radio networks using single-agent and multi-agent reinforcement learning. In: Proceedings of the IEEE Local Computer Network Conference (LCN). 2010:152–9. https://doi.org/10.1109/LCN.2010.5735689
5. Mishra N, Srivastava S, Sharan SN. RADDPG: Resource allocation in cognitive radio with deep reinforcement learning. In: 2021 International Conference on Communication Systems & Networks (COMSNETS). 2021:589–95.
6. Ling MH, Yau KLA, Qadir J, Ni Q. A reinforcement learning-based trust model for cluster size adjustment scheme in distributed cognitive radio networks. IEEE Trans Cogn Commun Netw. 2018;5(1):28–43.
7. Tan X, Zhou L, Wang H, Sun Y, Zhao H, Seet BC, Zhang X. Cooperative multi-agent reinforcement learning based distributed dynamic spectrum access in cognitive radio networks. IEEE Internet Things J. 2022;9(12):12232–47.
8. Khozeimeh F, Haykin S. Brain-inspired dynamic spectrum management for cognitive radio ad hoc networks. IEEE Trans Wirel Commun. 2012;11(10): 3509–17.
9. Zafari F, Gkelias A, Leung KK. A survey of indoor localization systems and technologies. IEEE Commun Surv Tutor. 2019;21(3):2568–99.
10. Dong X, Li Y, Wei S. Design and implementation of a cognitive engine functional architecture. Chin Sci Bull. 2012;57(28):3698–704.
11. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerg Artif Intell Appl Comput Eng. 2017;160(1):3–24.
12. Khamayseh S, Halawani A. Cooperative spectrum sensing in cognitive radio networks: A survey on machine learning-based methods. J Telecommun Inf Technol. 2020;(3):32–41.
13. Wang Y, Ye Z, Wan P, Zhao J. A survey of dynamic spectrum allocation based on reinforcement learning algorithms in cognitive radio networks. Artif Intell Rev. 2019;51(3):493–506.
14. Abbas N, Nasser Y, Ahmad KE. Recent advances on artificial intelligence and learning techniques in cognitive radio networks. EURASIP J Wirel Commun Netw. 2015;2015(1):1–20.
15. Yang P, Li L, Yin J, et al. Dynamic spectrum access in cognitive radio networks using deep reinforcement learning and evolutionary game. In: 2018 IEEE/CIC International Conference on Communications in China (ICCC). 2018:405–9.
16. Gulzar W, Waqas A, Dilpazir H, Khan A, Alam A, Mahmood H. Power control for cognitive radio networks: A game theoretic approach. Wirel Pers Commun. 2022;123(1):745–59.
17. Jiang X, Li P, Li B, Zou Y, Wang R. Secrecy performance of transmit antenna selection for

underlay MIMO cognitive radio relay networks with energy harvesting. IET Commun. 2022;16(3):227–45.
18. Thakur P, Kumar A, Pandit S, Singh G, Satashia SN. Spectrum mobility in cognitive radio network using spectrum prediction and monitoring techniques. Phys Commun. 2017;24:1–8.
19. Singh G, Thakur P. Spectrum sharing in cognitive radio networks: Towards highly connected environments. Chichester: John Wiley & Sons; 2021:336.
20. Wachi A, Sui Y. Safe reinforcement learning in constrained Markov decision processes. In: Proceedings of the 37th International Conference on Machine Learning (ICML). 2020;119:9797–806.
21. Gattami A, Bai Q, Aggarwal V. Reinforcement learning for constrained Markov decision processes. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS). 2021;130:2656–64.
22. van Otterlo M, Wiering M. Reinforcement learning and Markov decision processes. In: Wiering M, van Otterlo M, editors. Reinforcement Learning: State-of-the-Art. Berlin, Heidelberg: Springer; 2012:3–42.
23. Zhang J, Dong A, Yu J. Intelligent dynamic spectrum access for uplink underlay cognitive radio networks based on Q-learning. In: International Conference on Wireless Algorithms, Systems, and Applications (WASA). 2020:691–703.