IJBBE International Journal of Biophotonics & Biomedical Engineering Vol. 4, No. 1, Spring-Summer, 2024

Research Article

Simulation-based Optimization of Chemotherapeutic Drug Dosage: An Agent-based Q-learning Approach

H. Sadrian [a], P. Vafadoost [a], A. Hajipour [a] and H. Rokhsati [b]

a Biomedical Engineering Department, Electrical and Computer Faculty, Hakim Sabzevari University, Sabzevar, Iran
b Department of Computer, Control and Management Engineering, Sapienza University of Rome, Italy

***Corresponding Author Email: peymanvafadoost@gmail.com**

DOI: 10.71498/ijbbe.2024.1127216

ABSTRACT

Received: Jul. 26, 2024, Revised: Sep. 20, 2024, Accepted: Oct. 8, 2024, Available Online: Jan. 19, 2025

Cancer is indeed a growing concern worldwide for human health and existence, with its prevalence and impact on individuals and society increasing. The main objective of this article is to control and optimize drug dosage in order to prevent the uncontrollable growth of cancer cells and also restore the patient's immune cells to normal levels at the end of the training process, in such a way that the disease can be controlled in the early days of treatment. Reinforcement learning methods are widely applied in many domains nowadays and have attracted researchers' interest in conducting studies in this field. Therefore, in this article, specifically we also use the Q-learning method, one of the most famous model-free reinforcement learning methods, as well as the four-state nonlinear dynamic model called depillis, to simulate and design the proposed controller. The proposed controller's performance was evaluated in the presence of noise in three stages (training, simulation, and both stages simultaneously) as well as in the presence of uncertainty in one of the parameters of the depillis model. In a state of uncertainty, a combination therapy of chemotherapy and immunotherapy has been suggested as a treatment approach. Results indicate the significant impact of the proposed controller in determining the optimal drug dosage, improved accuracy, reduced side effects, and faster convergence compared to previous studies.

KEYWORD

Cancer, Control, Reinforcement Learning, Q-Learning.

I. INTRODUCTION

Cancer is recognized as one of the biggest threats and growing concerns worldwide, with various types of it being characterized by a phenomenon called cellular state change with the loss of control over cell division and proliferation. However, cancer is an abnormal, irregular, uncontrolled, and deadly growth of cells in the body's tissues, leading to the formation of a mass called a tumor. The American Cancer Society (ACS) collects the latest information and reports each year on the incidence, mortality, and outcomes of cancer in collaboration with two centers called the Central Cancer Registry and the National

Center for Health Statistics [1]. According to the World Health Organization (WHO), the projected data from 2030 to 2040 is concerning. It suggests that around 11.2 to 13.4 million individuals may die from this illness by 2030, and by 2040, approximately 27.5 million people will be affected by it [2]. The treatment plan and amount of medication given depend on the tumor stage (the stage of the tumor refers to how advanced it is and whether it has spread to other parts of the body), patient's weight, immunity level (white blood cell count), any existing illnesses, organ function, drug interactions, and the patient's age [3]. Based on these factors, the healthcare provider will determine the most appropriate treatment plan, which may include surgery, radiation therapy, chemotherapy, targeted therapy, immunotherapy, or a combination of these approaches. The dosage and type of medication given will also depend on these factors. It is important for the healthcare team to assess these factors and create an individualized treatment plan for each patient to optimize their chances of successful treatment while considering their specific circumstances. Given the severity of cancer, any method that improves the effectiveness of treatment, leading to decreased harm to organs and lower rates of morbidity, is highly sought after.

Implementing reinforcement learning techniques can help mitigate complications and address time constraints associated with administering chemotherapy in cancer treatment [4].

Chemotherapy is a crucial component of cancer treatment, but it is not without its challenges. In addition to targeting cancer cells, chemotherapy can also affect healthy cells, leading to various side effects such as fatigue, nausea, hair loss, and a weakened immune system. Furthermore, there may be limitations in terms of the duration and frequency of chemotherapy sessions, as well as the tolerance level of patients to the treatment. Despite these complexities and drawbacks, chemotherapy remains a valuable therapeutic tool in the fight against cancer, often used in conjunction with other treatments to achieve the best possible outcomes for patients [5]. Control theory, a branch of mathematics and engineering that deals with the behavior of dynamical systems, has recently been proposed as a potential tool to improve the efficacy of cancer chemotherapy [6].

In the modern world, mathematical models play a crucial role in understanding and optimizing cancer treatment strategies. These models help researchers and clinicians simulate the complex dynamics of tumor growth, drug interactions, and treatment response, allowing for personalized and precise approaches to therapy. By incorporating data-driven mathematical simulations, healthcare professionals can tailor cancer treatments to individual patients, predict outcomes, optimize drug dosages, and explore novel therapeutic interventions. Overall, mathematical modeling has revolutionized the field of cancer treatment by providing valuable insights and guiding decision-making processes to improve patient outcomes and quality of care [7,8]. A cancer dynamics model needs to take into consideration the growth of the tumor, the response of the immune system to the tumor growth, and the impact of chemotherapy on immune cells, normal cells, and tumor growth [3]. In summary, utilizing mathematical models to control and optimize chemotherapy drug dosage enables precision medicine, predicts drug response, optimizes treatment schedules, reduces trial and error, and enhances safety and efficacy in cancer therapy.

Reinforcement learning has shown promise in the field of chemotherapy drug control and optimization. Chemotherapy treatment often involves dosing medications at specific intervals and monitoring the patient's response to determine the effectiveness of the treatment. This process can be complex and time-consuming, requiring constant adjustments to ensure the patient is receiving the right dosage and that side effects are managed effectively.

Reinforcement learning algorithms can be used to optimize chemotherapy drug control by learning from past treatment outcomes and adjusting dosages in real time based on patient response. These algorithms can also take into

account individual patient characteristics, such as age, weight, and genetic factors, to tailor treatment plans to each patient's unique needs.

By using reinforcement learning in chemotherapy drug control, healthcare providers can potentially improve treatment outcomes, reduce side effects, and optimize drug dosages more efficiently. This can result in better patient outcomes, reduced healthcare costs, and a more personalized approach to cancer treatment. The study shows promising results in optimizing drug dosing for cancer chemotherapy treatment using RL, which could potentially improve patient outcomes and reduce side effects.

Modeling and controlling the growth of cancer cells as well as determining the optimal drug dosage in cancer patients are challenging and complex subjects in the field of cancer.

In [3, 9, 10, 11, 12] and [13] predominantly reinforcement learning methods have been used for cancer control and treatment. Specifically, in [3] Padmanabhan and her colleagues have suggested a closed-loop controller based on reinforcement learning in their work. They utilize Q-learning with a four-state mathematical model for cancer chemotherapy, which includes immune cells, normal cells, tumor cells, and drug concentration. The simulation of three disease ranges shows that the injected drug dose effectively eliminates the tumor. One notable advantage of their method is that it does not require a system model to create a controller. In another study by Padmanaban and colleagues [13], mentioned in Chapter 9 of the book "Control Applications for Biomedical Engineering Systems" [14], researchers have focused on investigating reinforcement learning-based control of drug dosing with applications in anesthesia and cancer treatment. The main goal is to determine and control the intravenous dosage of the anesthetic drug using a reinforcement learning algorithm called Q-learning. The drug used in this study is propofol for patients in the ICU, which is regulated by the Q-learning algorithm. The study demonstrated the efficacy of the Q-learning algorithm in regulating the dosage of propofol for patients undergoing treatment. The authors demonstrate the effectiveness of the proposed approach through simulations and experiments, showing promising results in terms of improved treatment outcomes and reduced drug toxicity. In [9], researchers have proposed an optimal switching control strategy for drug therapy process in cancer chemotherapy. The proposed control algorithm dynamically adjusts the dosage and type of drugs administered based on real-time patient response data, tumor progression, and toxicity levels. The objective of the switching control is to maximize the therapeutic benefits by targeting the tumor cells while minimizing the detrimental effects on healthy tissues. A mathematical model of the tumor growth dynamics and drug pharmacokinetics is developed to simulate the patient's response to the treatment. The control algorithm incorporates a multi-objective optimization framework to simultaneously consider the trade-offs between tumor regression, toxicity reduction, and drug resistance. Simulation results demonstrate that the optimal switching control strategy outperforms traditional fixed-dose protocols in terms of tumor suppression and patient survival rates.

In [10], researchers have proposed a novel approach to optimizing dose-finding strategies using integral reinforcement learning. The aim is to develop a control algorithm that can adaptively adjust drug dosages based on patient responses to maximize efficacy while minimizing side effects. In particular, the use of integral reinforcement learning allows the algorithm to incorporate past experiences and account for the long-term effects of drug dosing decisions. This helps in fine-tuning the dosing strategy over time to achieve the best possible outcomes for patients. In [11], a novel approach for controlling tumor growth under anti-angiogenic therapy using reinforcement learning algorithms (RL) has been proposed. Anti-angiogenic therapy is a promising strategy for cancer treatment that aims to inhibit the growth of blood vessels that supply nutrients to tumors. However, this therapy is often plagued by the development of resistance and rebound

effects, leading to tumor regrowth. Overall, this study highlights the potential of using reinforcement learning techniques to optimize cancer treatment strategies and improve outcomes for patients undergoing anti-angiogenic therapy. In [12] authors have presented a supervised offline reinforcement learning approach for personalizing chemotherapy regimens for cancer patients. Offline reinforcement learning is a machine learning technique that allows for the optimization of treatment strategies based on historical data without the need for real-time feedback. First, a Markov Decision Process (MDP) framework is constructed for modeling the chemotherapy treatment process. The state space of the MDP includes patient and tumor characteristics, while the action space represents the chemotherapy drugs and doses that can be administered. The reward function captures the efficacy and toxicity of the treatment, with the goal of maximizing the former while minimizing the latter. Next, a deep Q-network (DQN) was trained using a dataset of historical patient records and treatment outcomes. The DQN learns to predict the optimal chemotherapy regimen for a given patient based on their individual characteristics and tumor type. By leveraging the rich information contained in the dataset, the model is able to generalize well to new patients and make personalized treatment recommendations. Overall, this study showcases the promise of supervised offline reinforcement learning for personalizing chemotherapy treatment decisions.

Melanoma is a type of skin cancer that can be challenging to treat due to its aggressive nature and tendency to spread rapidly. Traditional cancer therapies often have toxic side effects that can be detrimental to patient health. In [15], Noori et al introduced the use of an eligibility traces algorithm to determine the optimal dose for controlling the population of cancer cells in melanoma patients. The eligibility traces algorithm is a reinforcement learning technique that allows for efficient learning from past experiences by assigning credit to actions that lead to positive outcomes. By applying this algorithm to the problem of determining the optimal dosage for cancer treatment, aim to identify a treatment regimen that maximizes anti-cancer effects while minimizing the occurrence of side effects.

In [16], the authors also presented a novel approach to controlling cancer cells in a nonlinear model of melanoma by incorporating the uncertainty factor using the Q-learning algorithm under the Case-Based Reasoning (CBR) policy. The use of CBR policy allows us to make decisions based on past experiences and cases, enabling us to leverage the knowledge gained from previous treatments and outcomes to improve our current control strategies. By combining the Q-learning algorithm with the CBR policy, we can develop a robust and adaptive approach to controlling cancer cells in a nonlinear model of melanoma.

Fuzzy logic is used to model the uncertainty and imprecision in the feedback signals from the tumor growth dynamics. The fuzzy logic controller provides a flexible and adaptive strategy for adjusting the chemotherapy drug dose based on the tumor's current state. Based on this, authors in [17] have proposed a new feedback control strategy for regulating tumor growth by limiting the maximum dose of chemotherapy using fuzzy logic. The proposed control system uses reinforcement learning to learn the optimal dose of chemotherapy drug to administer at each time step based on feedback from the tumor growth dynamics. The system is designed to minimize tumor growth while also limiting the maximum dose of chemotherapy drugs to prevent harmful side effects on the patient.

There are many reasons to use reinforcement learning methods, some of which include:

1. Flexibility: Reinforcement learning methods can be applied to a wide variety of tasks and environments, making them flexible and adaptable for different scenarios.

2. Ability to learn from interactions: Reinforcement learning algorithms learn from trial and error by interacting with an

environment, enabling them to improve performance over time through experience.

3. Autonomous decision-making: Reinforcement learning methods enable machines to make autonomous decisions without the need for explicit programming, allowing them to adapt to changing conditions and learn from their mistakes.

4. Handling complex, dynamic environments: Reinforcement learning methods are well-suited for addressing problems in complex, dynamic environments where traditional algorithms may struggle, such as in robotics, autonomous driving, and game playing.

5. Scalability: Reinforcement learning algorithms can be scaled up to handle large amounts of data and complex tasks, making them suitable for real-world applications in fields like healthcare, finance, and transportation.

6. Continuous learning: Reinforcement learning algorithms can continuously learn and adapt to new information and changing conditions, allowing them to improve performance over time.

7. Model-free learning: Reinforcement learning methods do not require explicit models of the environment, making them suitable for situations where the underlying dynamics are unknown or difficult to model accurately.

Reinforcement learning can be used in the application of cancer chemotherapy drug dosage to optimize treatment outcomes and minimize side effects for patients. In this scenario, the chemotherapy dosage would be considered as the action taken by the system, and the outcome of the treatment, such as tumor size reduction and patient's quality of life, would be the reward signal. The reinforcement learning algorithm would learn from the feedback of previous treatments to adjust the dosage levels in subsequent rounds, aiming to find the optimal dosage that maximizes the treatment benefits while minimizing the negative side effects. By utilizing reinforcement learning in cancer chemotherapy drug dosage, oncologists can personalize treatment strategies for individual patients based on their response to the treatment, ultimately leading to better outcomes and improved patient care.

This article is a review on the application of using Q-learning method, one of the reinforcement learning methods, in determining and controlling the dosage of chemotherapeutic drugs. In the following sections of this article, we will delve into a comprehensive examination of reinforcement learning concepts, particularly focusing on the Q-learning method.

This article is structured in 3 sections: materials and methods, Results and discussion and conclusions.

II. Materials And Methods

This section outlines the mathematical model of depillis pharmacology, which is used to analyze the effectiveness of chemotherapy in treating cancer. It introduces the concept of reinforcement learning and describes how a controller is created using Q-learning to calculate and regulate the best dosage of medication for chemotherapy.

A. *Mathematical Model*

So far, a large number of mathematical models have been proposed for the growth of cancer cells, each of which has its advantages and disadvantages, and in fact, there is no correct answer as to which model is more realistic [7, 8]. Mathematical models serve as valuable instruments in grasping the underlying mechanics of dynamic processes within cancer and are essential for investigating a wide range of scientific inquiries. The human body can be represented by a mathematical model, which can efficiently simulate complex systems at low costs. These models are useful for predicting the growth and spread of cancer cells, understanding the immune system's response, evaluating the impact of different cancer treatments, and assessing drug toxicity on healthy tissues. They can also help in studying

the interactions between various factors that contribute to tumor formation and predicting tumor size. By developing control models based on these mathematical models, we can improve drug prescription for cancer patients. A well-fitted mathematical model of cancer cell growth can provide valuable insights for analyzing the system accurately.

Mathematical modeling can be applied to different aspects of cancer research, including tumor growth, mutations, metastasis, treatment methods like chemotherapy and immunotherapy, and the diversity of tumors. This is typically done through the use of differential equations for analytical simulation and modeling purposes [18,19].

The depillis model is a mathematical model used in epidemiology to simulate the spread of infectious diseases within a population. This model is one of the most comprehensive models proposed in the field of chemotherapy; because the reason and importance of using this model is the addition and impact of the drug on the expression of immune cells. In depillis mathematical model, the dynamics of normal cells, tumor cells, immune cells and drug concentration can be represented by a system of differential equations [3].

Let $N(t)$ be the population of normal cells at time t, $T(t)$ be the population of tumor cells at time t, $I(t)$ be the population of immune cells at time t, and $D(t)$ be the concentration of drug at time t. The model can be described by the following equations:

$$\begin{aligned}\frac{dI}{dt}(t) &= s + \frac{\rho I(t)T(t)}{\alpha + T(t)} - d_1 I(t) \\ &\quad - c_1 I(t)T(t) - a_1(1 \\ &\quad - e^{-D(t)})I(t)\end{aligned}$$

$$\begin{aligned}\frac{dT}{dt}(t) &= r_1 T(t)\big(1 - b_1 T(t)\big) \\ &\quad - c_2 I(t)T(t) \\ &\quad - c_3 N(t)T(t) - a_2(1 \\ &\quad - e^{-D(t)})T(t)\end{aligned} \qquad (1)$$

$$\begin{aligned}\frac{dN}{dt}(t) &= r_2 N(t)\big(1 - b_2 N(t)\big) \\ &\quad - c_4 N(t)T(t) - a_3(1 \\ &\quad - e^{-D(t)})N(t)\end{aligned}$$

$$\frac{dD}{dt}(t) = -d_U D(t) + u(t)$$

This model takes into account various factors such as the growth rate of the cancer cells, the effectiveness of the chemotherapy drugs, and the impact on the immune system. By varying the parameters in the model, researchers can predict how different treatment strategies will affect tumor growth and the overall outcome for the patient. The values of the parameters of this model are also shown in table 1.

Depillis' model has been used to study optimal drug dosing regimens, the development of drug resistance, and the potential for combination therapies to improve treatment outcomes. By incorporating mathematical modeling into cancer research, scientists can better understand the complex interactions between cancer cells, chemotherapy drugs, and the immune system, ultimately leading to more effective treatment options for patients.

B. *Reinforcement learning*

Reinforcement learning, an exciting concept in machine learning, is rapidly progressing and is set to be a major advancement in artificial intelligence in the coming years. Reinforcement learning involves a machine learning method where an agent learns to make decisions through receiving feedback from its actions within an environment. The schematic of reinforcement learning operation is shown in Fig. 1.

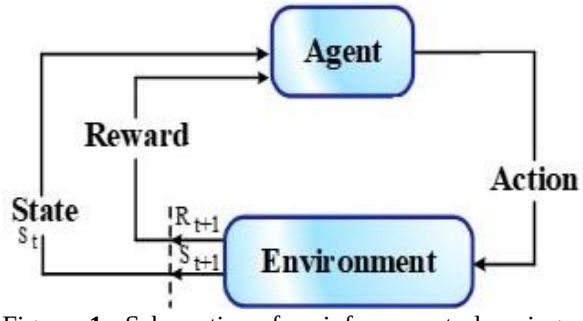


Fig. **1** Schematic of reinforcement learning operation.

The agent performs actions within the environment, receives either rewards or penalties as feedback, and adapts its behavior to maximize its rewards gradually. This process is rooted in the concept of trial-and-error learning, as the agent gains knowledge through practice and by engaging with its surroundings. This approach has shown promising results in various domains, such as game-playing, robotics, and autonomous driving [20,21]. By allowing agents to learn directly from interacting with their environment, reinforcement learning has the potential to create more autonomous and intelligent systems that can adapt to new situations and learn complex tasks without explicit programming.

Types of reinforcement learning methods include Value-based methods, Policy-based methods, Model-based methods, Model-free methods, Actor-critic methods, multi-agent reinforcement learning methods and Hierarchical reinforcement learning methods. The method proposed in this article is among value-based methods. These methods involve estimating the value of state-action pairs to make decisions on which actions to take. The value of an action is typically defined as the expected cumulative reward that an agent can achieve by taking that action and following a certain policy thereafter. Overall, value-based methods in reinforcement learning are powerful techniques for learning optimal policies in environments with discrete or continuous action spaces. They provide a fundamental framework for understanding the trade-offs between different actions and guiding the agent towards achieving its goals. An example of value-based methods used is Q-learning.

TABLE **1** PARAMETER VALUES OF THE DEPILLIS CHEMOTHERAPEUTIC MODEL [22].

Parameter	Value	Description
a_1	0.2	Fractional immune cell kill rate
a_2	0.3	Fractional tumor cell kill rate
a_3	0.1	Fractional normal cell kill rate
b_1	1	Reciprocal carrying capacity of tumor cells
b_2	1	Reciprocal carrying capacity of normal cells
c_1	1	Immune cell competition term (competition between tumor cells and immune cells)
c_2	0.5	Tumor cell competition term (competition between tumor cells and immune cells)
c_3	1	Tumor cell competition term (competition between normal cells and tumor cells)
c_4	1	Normal cell competition term (competition between normal cells and tumor cells)
r_1	1.5	Per unit growth rate of tumor cells
r_2	1	Per unit growth rate of normal cells
d_1	0.2	Immune cell death rate
dU	1	Decay rate of injected drug
s	0.33	Immune cell influx rate
ρ	0.01	Immune response rate
α	0.3	Immune threshold rate

C. *Q-learning*

In Q-learning, an agent learns the value of each action in each state by updating a Q-table with learned rewards from interactions with the environment. The agent then selects actions based on the values in the Q-table [23]. Some of the key features of the Q-learning method include:

1. Model-free: Q-learning is a model-free reinforcement learning algorithm, meaning that it does not require knowledge of the underlying dynamics of the environment. Instead, the agent learns through trial and error by interacting with the environment and updating its Q-values based on the rewards it receives.
2. Q-values: In Q-learning, the agent maintains a Q-table that stores the expected rewards for each action in each state. The Q-value for a

particular state-action pair represents the expected cumulative reward that the agent will receive if it takes that action in that state and follows the optimal policy thereafter.

3. Exploration vs exploitation: To balance exploration (trying new actions to discover the optimal policy) and exploitation (taking the best-known action to maximize rewards), Q-learning uses an epsilon-greedy strategy. This means that the agent will choose a random action with probability epsilon and the best-known action with probability 1-epsilon.

4. Bellman equation: Q-learning updates its Q-values using the Bellman equation, which states that the expected cumulative reward for a state-action pair should be equal to the immediate reward obtained by taking that action plus the expected cumulative reward of the next state-action pair.

5. Convergence: Q Learning is guaranteed to converge to the optimal policy under certain conditions, such as having a finite state and action space, and that the agent visits all state-action pairs infinitely often.

6. Off-policy: Q-learning is an off-policy algorithm, meaning that it can learn from any policy, not just the one it follows. This property allows the agent to learn from past experience and improve its policy over time.

7. Suitable for discrete actions: Q-learning is best suited for environments with a discrete action space, as it requires calculating Q-values for each possible action in each state. For continuous action spaces, techniques such as Deep Q-learning are typically used.

Overall, Q-learning is a versatile and effective reinforcement learning algorithm that has been successfully applied to a wide range of problems, including game playing, robotics, and autonomous driving [23].

In the Q-learning algorithm process, it learns the Q-value function $Q(s,a)$; meaning how to take action "a" in a specific state "s". The goal of the agent in Q-learning is to maximize the Q-value, ultimately leading to the most optimal decision-making process. The value of Q-learning is obtained using the Bellman equation. The Bellman equation is expressed as follows:

$$V(s) = \max\left[R(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')\right] \quad (2)$$

In Fig. 2, we have shown the steps of executing the Q-learning algorithm in the form of a flowchart:

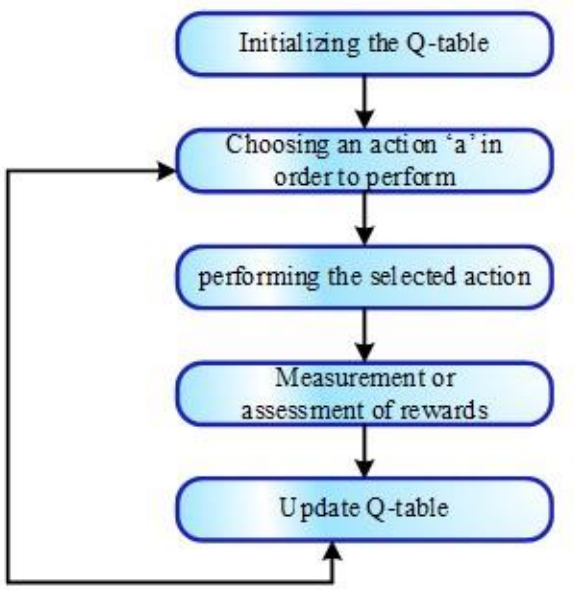


Fig. **2** Flowchart of the steps of implementing the Q-Learning algorithm.

Updating the Q-table is done by equation 3:

$$Q(s,a) = Q(s,a) + \alpha[R(s,a) + \gamma maxQ'(s',a')] \quad (3)$$

In the above equation, α is the learning rate, γ is the discount factor, and $maxQ'(s',a')$ represents the maximum Q value of the next state. Following the flowchart of the Q-learning algorithm in Fig. 3, in the second step, the agent must choose an action according to a policy that balances exploitation (using previously known information to maximize reward) and exploration (finding more information about the environment). This policy selection is done through two common methods: empirical search and random policy. Typically, the idea and method used in this algorithm are based on empirical search and the epsilon-greedy strategy; this is because the agent increases its confidence in finding the optimal solution by exploring the environment more. This method

usually selects the action that is estimated to have the highest reward in most cases. In the next section, we investigate the design of a Q-learning-based controller for determining the optimum drug dosage in cancer and then present the relevant results.

D. *Depilis model review and controller design*

Understanding the dynamics of the model in order to determine the equilibrium points of the system and their stability is necessary; therefore, equilibrium points without tumor that are stable in the limited treatment time are defined as follows [17]:

$$(\frac{1}{b_2}, 0, \frac{s}{d_1}) \quad (4)$$

As it is evident, at this point, there are zero tumor cells and immune and normal cells present. The point achieved will be under our ultimate goal at the end of chemotherapy [17]:

$$(N, T, I) \rightarrow (1,0,1.65) \quad (5)$$

Using the dynamic model of depillis and with the help of the Q-learning algorithm, we design a controller to minimize or eliminate tumor cells (reduce tumor volume or eradicate it) and also maintain immune cells at a certain level.

Our environment is an optimal drug dosing control system in cancer patients, where tumor cells and immune cells are considered as system states, and the amount of prescribed drug dose is considered as the system action. The drug dose amount in the current research is specified in the range [0,1]. Therefore, our state space in this research is a two-dimensional state space. In the training stage for determining states, we define a range where this range is defined as 0+0.3×rand() for tumor cells and 0.15+(2-0.15)×rand() for immune cells; meaning that for tumor cells, a random number between 0 to 0.3 and for immune cells, a random number between 0.15 to 2 is chosen. Then, based on the division of state space for these two variables and the formation of a Q-table in each generation of current state values, one of the states in the table is selected. In order to evaluate the reward, the following method was used:

$$\begin{cases} if\ tumor\ cell = 0, r = 10, \\ \quad otherwise, r = 1. \end{cases} \quad (6)$$

In this way, in each iteration, the values of Q-table are completed and updated, and eventually, the complete table is obtained. After completing the table and also determining the specific initial states for normal, tumor, and immune cells, the training process continues until the end of the training, where the number of cancer cells reaches a minimum or zero and the immune and normal cells also reach a specified desired level.

Initial conditions are also defined for three normal, tumor, and immune cells as 1, 0.25, and 0.15, respectively. The values of Q-learning parameters are defined according to the table below:

Table **2** Reinforcement learning parameters.

Parameter	Value	Description
γ	0.8	Discount factor
η	0.2	Learning rate
ε	0.05	Greedy learning parameter

III. SIMULATION

For simulation purposes, we first examine the model without the impact of the drug; for this purpose, using the mathematical equations of the depillis model and also the parameters in Table 1, we observe the changes obtained in the absence of drug effect, and then by entering the drug, we focus on controlling and determining the optimal drug dose in the chemotherapy process for cancer patients using the Q-learning method. Therefore, in order to display the chart of the change, we first convert the depillis model equations according to the relationships

(1) to (4), which are continuous, to discrete. With high precision in discretization, the output results of the graph should be similar to the original model results [3]. The chart of changes in the discretized model in the absence of drug influence is shown in Fig. 3.

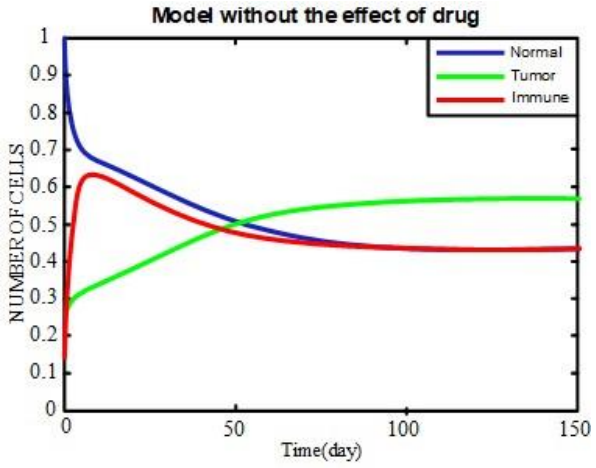
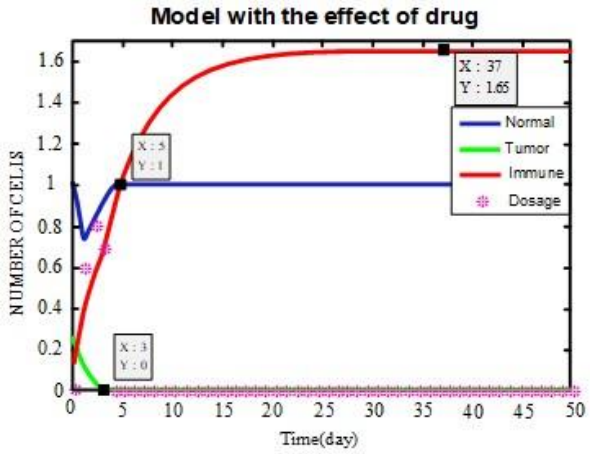


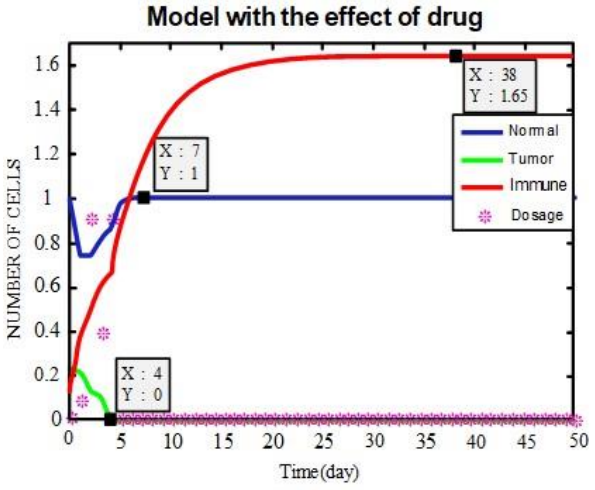
Fig. **3** Discretization model in the absence of drug effect.

As seen in Fig. 3, in the early stages of cancer cell growth in the patient's body, immune cells try to fight against cancer cells, but over time, due to the weakening of the immune cells, cancer cells grow and multiply more and more until they cover the entire surface of the curve and system; furthermore, immune and normal cells quickly reach their lowest vital limit and, in this situation, the patient will lose their life [22].

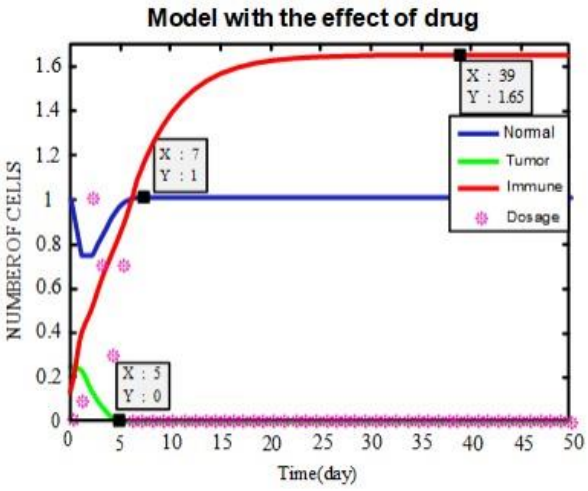
Now we need to simulate the model in the presence of drug input using the q-learning method. As mentioned earlier, we use the epsilon-greedy method to determine the optimal drug dose, in this method, due to the use of the rand() function, we will observe different results in the output. In this method, usually the experiment is conducted for 10 different scenarios and then the average of these 10 scenarios is calculated. However, the important point is that taking 10 runs and averaging the results is only valid under the assumption that we can fully complete table Q; whereas in implementation, with a higher number of training iterations (assuming a total of 10,000 episodes), there will be no need for multiple runs. When we want to use the learned Q-table for simulation, there is no longer a random state; this is because the Q-table has been ideally or close to ideally trained during the training phase. Therefore, the small difference observed in each execution of the program is the result of using the same ε-greedy method in the training phase. The following figures display various types of simulated charts using the Q-learning method:



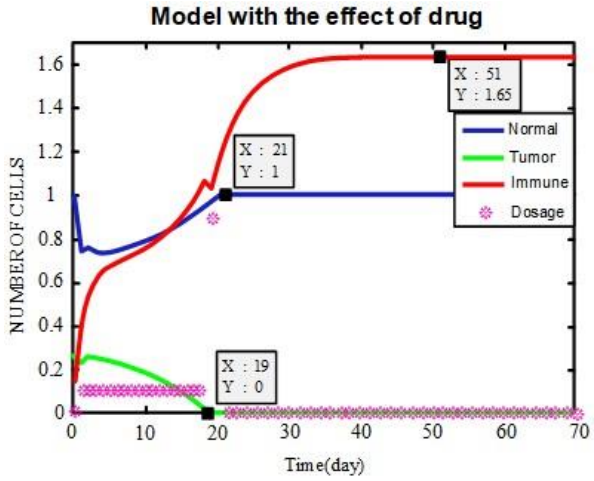
(a)



(b)



(c)



(d)

Fig. **4** (a,b,c,d) Optimal control of model with Q-learning algorithm in the presence of drug influence.

Figure 4 diagrams all represent optimal control of cancer drug dosage using the q-learning method; however, as it is evident, a slight difference in each execution can be seen in the diagrams, which is the effect of using the ε-greedy method in the training phase. Furthermore, changing the input range of tumor and immune cells at the beginning of forming the q-table also results in noticeable changes. It should also be noted that different values for the gamma variable do not lead to significant changes.

As you can see in Fig. 4(d), the tumor cells reach zero at a longer period, which is the 19th day. The reason for the change in the behavior of the graph, in this case, is that only data with very low tumor amounts have been given to the model, so the model does not see any data with higher tumor amounts to make decisions about cases with more tumors. For more details on the implementation of the Q-learning algorithm, refer to [3,24]. In the next section, we will examine the results of implementing this method to control the dosage of chemotherapeutic drugs and delve into the discussion.

IV. Result And Discussion

In this part, we provide examples with numbers to show how effective the suggested RL method is for controlling cancer chemotherapy drug doses in a closed-loop system. Researchers conducted three experiments on three groups of young patients, elderly patients, and pregnant women to investigate the proposed controller, here only two of them have been studied [3,17].

A. *Young patient*

When treating a young patient with cancer, an oncologist prioritizes reducing the number of cancer cells quickly to prevent metastasis, even though this may result in damage to normal cells and immune cells. The body's ability to regenerate normal cells, which may be decreased as a side effect of chemotherapy, is more robust in young patients [3]. The body of a young patient can eliminate tumor cells due to the strength of the immune system. Therefore, if we consider the parameter s to be 0.33 in the depillis model, the output graph would appear as shown in Fig. 4(a).

B. *Elderly patient*

If we consider an elderly patient with weak immune cells, meaning we assume the value of s to be 0.2, in this condition the output graph will be transformed into Fig. 5.

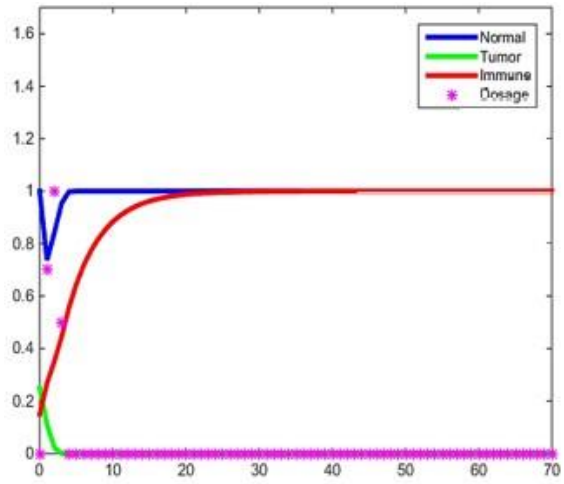


Fig. **5** Optimal control model for untreated elderly patient.

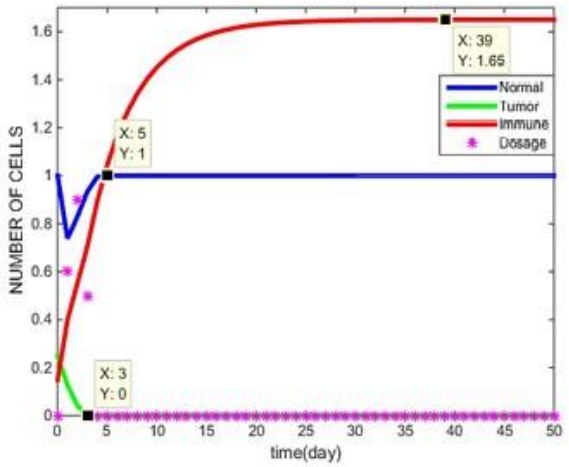


Fig. **6** Optimal control model for treated elderly patient.

As can be seen, tumor cells have reached zero at the end of treatment, while the rate of immune cells has decreased to 1. In such conditions, to treat, immunotherapy should be used in addition to chemotherapy to improve these lost immune cells. The immunotherapy used is modeled by the equation below [17]:

$$\frac{ds}{dt}(t) = \mu_s v_v(t)(1 - \frac{s}{k_s}) \qquad (7)$$

The therapeutic effect of immunotherapy with $v_v\ (t) \geq 0$ has been shown. The values of μ_s depend on the dynamics of the parameter s. This coefficient saturates to the final limit k_s, which is related to the biological constraints of body organs and the accumulation of external effects [17,25].

The output chart after chemotherapy treatment along with immunotherapy is shown using relationship 9 in Fig. 6.

Chemotherapy involves using drugs to kill cancer cells, while immunotherapy works by boosting the body's natural defenses to help the immune system recognize and attack cancer cells. When these two treatments are combined, they can have a synergistic effect and offer a more effective way to treat cancer. By combining chemotherapy with immunotherapy, doctors can potentially lower the dosage of chemotherapy drugs needed to achieve the desired effect, reducing the likelihood of side effects and toxicity. One way that doctors can control the optimal drug dosage when using a combination of chemotherapy and immunotherapy is through careful monitoring and adjustment of treatment plans.

The use of both chemotherapy and immunotherapy allows for a more personalized treatment plan for the elderly patient. Combining chemotherapy with immunotherapy can enhance the overall effectiveness of the treatment by targeting cancer cells through different mechanisms. This can potentially lead to better outcomes for the elderly patient. By carefully monitoring and controlling the dosage of each drug, doctors can potentially reduce the risk of side effects associated with chemotherapy and immunotherapy. This is especially important for elderly patients, who may be more susceptible to adverse reactions. Studies have shown that combining chemotherapy with immunotherapy can lead to improved survival rates for cancer patients, including elderly individuals. By controlling

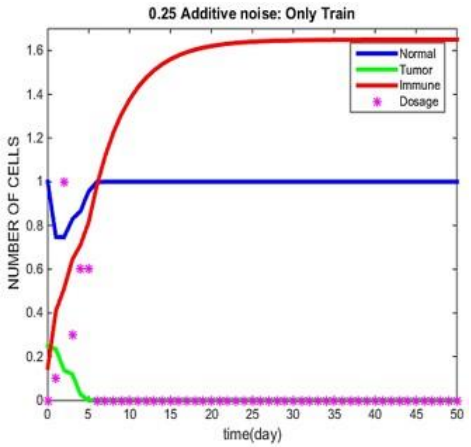
the optimal drug dosage, doctors can potentially increase the chances of successful treatment outcomes for elderly patients.

In this article, three scenarios were considered in order to investigate the effect of noise on the controller: the first scenario is adding noise only during the training phase, the second scenario is adding noise only during the simulation phase, and the third scenario is adding noise in both phases. Three values (0.25, 0.5, 1) were chosen for the percentage of noise. The noise is generated as a percentage of a randomly generated normal value that has been scaled to the desired variable value. The method of adding noise in the simulation stage was performed as follows:

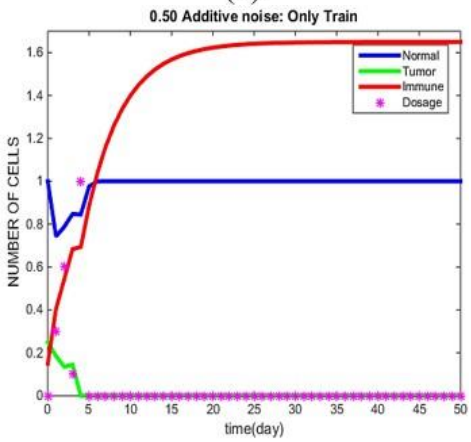
After determining the initial state to start the simulation, a certain percentage of noise is added based on the range of the variable of interest, and the index of the Q-table is selected based on the noisy variable for updating. However, in the stage of obtaining the next state variables of the problem (calling the depillis model function), noise-free data is used. In this way, the added noise can be considered a measurement error in the simulation stage. (This means we have an error in measurement but in reality, the levels of different cells in the patient will change regardless of our measurement, so the noise-free value is given to the model).

Also, in the training phase, noise is added similar to the simulation phase, but with the difference that in the training phase, the current state is randomly selected in each iteration. The experiment conducted for different noise levels did not result in any significant difference in the performance of the model, and even when adding noise in both phases, a suitable and stable performance was observed.

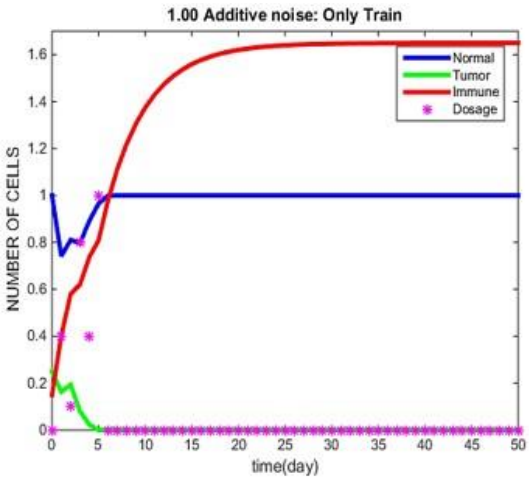
In Figs. 7-9, charts related to adding noise to the model are shown



(a)

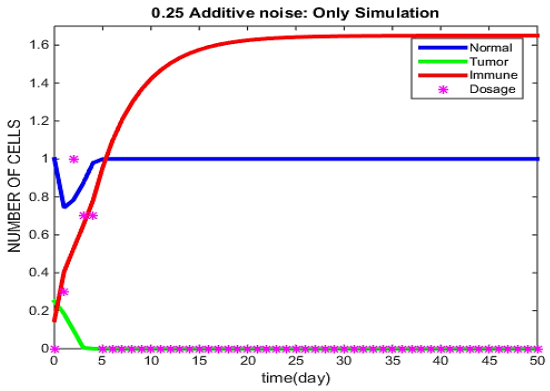


(b)

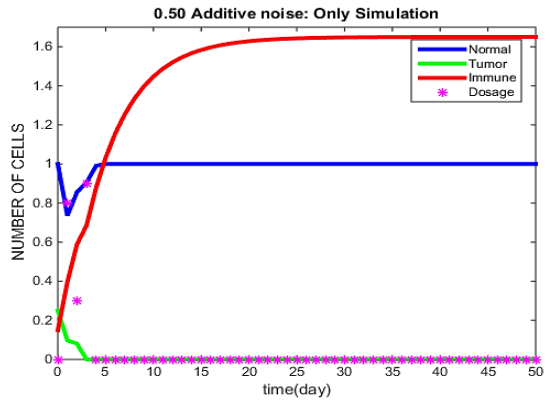


(c)

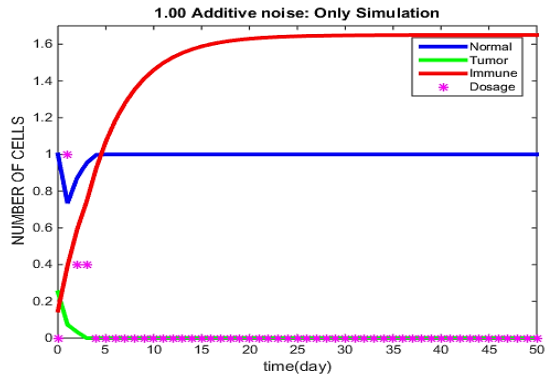
Fig. **7** Optimal control model in the presence of noise (only during the training phase).



(a)

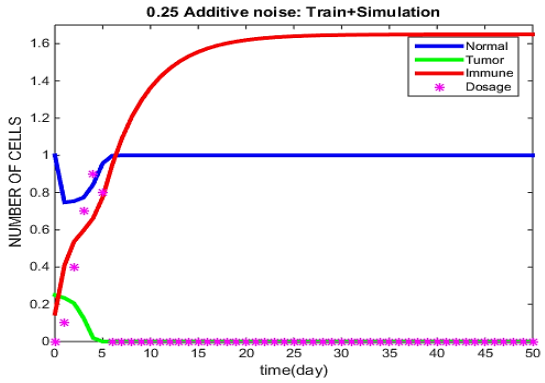


(b)

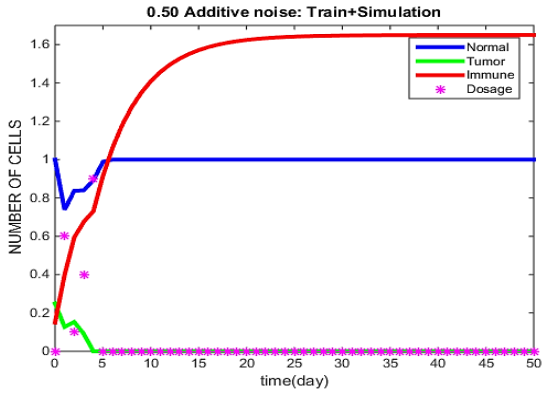


(c)

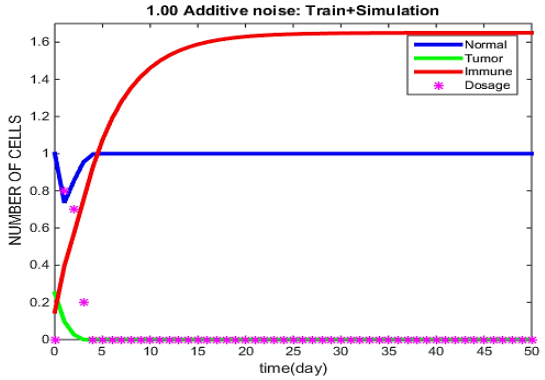
Fig. **8** Optimal control model in the presence of noise (only during the simulation phase).



(a)



(b)



(c)

Fig. **9** Optimal control model in the presence of noise (In both stages of training and simulation).

Overall, the effect of adding noise to the Q-learning controller for determining cancer drug dosages will depend on the specific implementation and the balance struck between exploration and exploitation. Proper tuning and optimization of the noise parameters will be crucial in achieving the desired balance between exploration and exploitation while ensuring the safety and effectiveness of the treatment.

The benefit of the RL-based approach is its capability to develop a controller without the need for a system model.

Traditional chemotherapy dosage determination often relies on a trial-and-error approach, leading to suboptimal treatment outcomes and increased risk of side effects. Reinforcement learning-based controllers can expedite the dosage optimization process by continuously learning from patient responses and making data-driven decisions, reducing the need for trial-and-error approaches. Reinforcement learning algorithms can continuously learn and adapt to new information and patient responses, leading to ongoing improvements in treatment outcomes over time. This continuous learning process can help in refining the dosage regimen to achieve better long-term outcomes for cancer patients. When comparing our simulation findings to those reported in [3], we observe that both approaches yield very similar outcomes. In both instances, the tumor is eliminated through the use of optimal chemotherapy dosages, and the controllers prove to be resilient to changes in parameters. Simulation and differential equations described have been implemented using MATLAB software.

V. CONCLUSIONS AND FUTURE WORK

In this paper, the application and efficiency of a reinforcement learning-based controller in determining the optimal dosage of chemotherapy drugs were investigated. One of the primary goals of determining the dosage of chemotherapeutic drugs using Q-learning is to maximize the efficacy of the treatment. Another important goal is to minimize the toxicity of the treatment. Chemotherapeutic drugs can have significant side effects, so it is important to find the right dosage that will effectively treat the cancer while minimizing harm to the patient's healthy tissues. Q-learning can help doctors find the optimal balance between efficacy and toxicity. In the context of chemotherapy dosage optimization, the environment could be the patient's body, where the agent (in this case, the Q-learning algorithm) needs to determine the optimal dosage of the chemotherapeutic drug to administer based on factors such as the type of cancer, the patient's age and overall health, and the drug's pharmacokinetics. The reward in this case could be a combination of factors, such as the reduction in tumor size, the patient's overall health, and the absence of severe side effects. The Q-learning algorithm would learn from experience by iteratively adjusting the dosage of the chemotherapeutic drug based on the feedback it receives from the environment. Over time, the algorithm would converge on an optimal dosage that maximizes the reward while minimizing the side effects. In order to evaluate the RL-based method, noise was applied to the controller. One of the main advantages of RL-based controllers is their ability to learn and adapt to new environments and tasks without the need for extensive hand-coding or pre-programming. RL algorithms work by interacting with the environment and receiving feedback in the form of rewards, which allows them to learn optimal actions that maximize long-term rewards. This adaptability makes RL controllers well-suited for complex tasks or environments where traditional control methods may struggle to find a solution. Additionally, RL controllers can handle non-linear, high-dimensional, and uncertain systems, making them a versatile and powerful tool in the field of control systems.

RL-based controllers are trained on a specific dataset and may not generalize well to new or unseen data. This could lead to suboptimal dose determination in real-world scenarios. One of the future works could be to investigate the use of deep reinforcement learning algorithms for determining drug doses in clinical settings. This could involve developing more sophisticated Q-learning algorithms that can handle the complexities of individual patient responses and varying drug interactions. Examining the integration of reinforcement learning with other machine learning techniques, such as deep learning or natural language processing, to improve the accuracy and efficiency of drug dosage determination algorithms.

REFERENCES

[1] R.L. Siegel, K.D. Miller, and N.S. Wagle, "Cancer statistics," Ca Cancer J Clin, vol. 73, pp. 17-48, 2023.

[2] World Health Organization, *WHO report on cancer: setting priorities, investing wisely and providing care for all*, 2020.

[3] Padmanabhan, R., N. Meskin, and Wassim M. Haddad, "Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment." Mathematical biosciences, vol. 293, pp. 11-20, 2017.

[4] C.Y. Yang, C. Shiranthika, C.Y. Wang, K.W. Chen, and S. Sumathipala, "Reinforcement learning strategies in cancer chemotherapy treatments: A review," Computer Methods and Programs in Biomedicine, vol. 229, pp.107280, 2023 .

[5] Perry MC, *The chemotherapy source book. Lippincott Williams & Wilkins*, 2008.

[6] P. Lecca, "Control theory and cancer chemotherapy: How they interact. Frontiers in Bioengineering and Biotechnology," vol. 14, pp.621269, 2021.

[7] R. Padmanabhan, N. Meskin, and AE. Al Moustafa, *Mathematical models of cancer and different therapies.* Singapore: Springer, 2021.

[8] Schättler H, and Ledzewicz U. *Optimal control for mathematical models of cancer therapies*, An application of geometric methods. 2015.

[9] Wu X, Liu Q, Zhang K, Cheng M, and Xin X. "Optimal switching control for drug therapy process in cancer chemotherapy," European Journal of Control. vol. 42, pp. 49-58, 2018.

[10] Padmanabhan R, Meskin N, and Haddad WM. "Optimal adaptive control of drug dosing using integral reinforcement learning," Mathematical biosciences. vol. 309, pp.131-142, 2019.

[11] Yazdjerdi P, Meskin N, Al-Naemi M, Al Moustafa AE, and Kovács L. "Reinforcement learning-based control of tumor growth under anti-angiogenic therapy," Computer methods and programs in biomedicine. vol. 173, pp. 15-26, 2019.

[12] Shiranthika C, Chen KW, Wang CY, Yang CY, Sudantha BH, and Li WF. "Supervised optimal chemotherapy regimen based on offline reinforcement learning," IEEE Journal of Biomedical and Health Informatics. vol.26, pp. 4763-4772, 2022.

[13] R. Padmanabhan, N. Meskin, and W.M. Haddad, "Reinforcement learning-based control of drug dosing with applications to anesthesia and cancer therapy," Control applications for biomedical engineering systems, Academic Press. vol.1, pp. 251-297, 2020.

[14] A.T. Azar, *Control Applications for Biomedical Engineering Systems*, Academic Press; 2020.

[15] Kalhor E, Noori A, and Saboori Rad S, "Sadrnia MA. Using Eligibility Traces Algorithm to Specify the Optimal Dosage for the Purpose of Cancer Cell Population Control in Melanoma Patients with a Consideration of the Side Effects," Journal of Soft Computing and Information Technology. vol.10, pp. 72-92, 2021.

[16] A. Noori, E. Kalhor, and S. Saboori Rad, "Controlling the Cancer Cells in a Nonlinear Model of Melanoma by Considering the Uncertainty Using Q-learning Algorithm Under the Case Based Reasoning Policy," Journal of Iranian Association of Electrical and Electronics Engineers, vol. 17, pp. 25-37, 2020.

[17] Mashayekhi H, and Nazari M. "Reinforcement learning based feedback control of tumor growth by limiting maximum chemo-drug dose using fuzzy logic," Journal of Control. vol. 15, pp. 13-23, 2022.

[18] H. Tourajizadeh, Z.G. Zarandi, Z. Farbodi, and E.S. Ghasemabadi, "Modelling and Control of Mutation Dynamics of the Cancer Cells Employing Chemotherapy," International Journal of Advanced Design & Manufacturing Technology. vol. 15, pp. 73-83, 2022.

[19] Z.G. Zarandi, H. Tourajizadeh, Z. Farbodei, E.S. Ghasemabad, "Dynamic Modeling of the Cancer Cell Mutation with the Capability of Control Using Chemotropic Injection," International Conference on Robotics and Mechatronics (ICRoM),vol. 26, pp. 1-8, 2020.

[20] Agarwal A, Jiang N, Kakade SM, and Sun W. "Reinforcement learning: Theory and algorithms," CS Dept. UW Seattle, Seattle, WA, USA, Tech. Rep. vol. 32, pp. 1-83, 2019.

[21] Winder P. *Reinforcement learning*, O'Reilly Media, 2020.

[22] De Pillis LG, and Radunskaya A. "The dynamics of an optimally controlled tumor model: A case study," Mathematical and

computer modeling. Vol. 37, pp. 1221-1244, 2003.

[23] Clifton J, and Laber E. "Q-learning: Theory and applications," Annual Review of Statistics and Its Application. vol. 7, pp. 279-301, 2020.

[24] Padmanabhan R, and Meskin N, "Haddad WM. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. Biomedical Signal Processing and Control. Vol. 22, pp. 54-64, 2015.

[25] Nazari M, and Ghaffari A. "The effect of finite duration inputs on the dynamics of a system: Proposing a new approach for cancer treatment," International Journal of Biomathematics. vol. 8, pp. 1550036, 2015.

THIS PAGE IS INTENTIONALLY LEFT BLANK.