**مجله کنترل**
ISSN (print) 2008-8345
ISSN (online) 2538-3752
**مقاله پژوهشی**
جلد ۱۵، شماره ۴، زمستان ۱۴۰۰، صفحه ۲۳-۱۳

# کنترل فیدبک مبتنی بر یادگیری تقویتی رشد تومور با محدودسازی دوز داروی شیمی‌درمانی با استفاده از منطق فازی

هدی مشایخی[۱]، مصطفی نظری[۲]

[۱] استادیار، دانشکده مهندسی کامپیوتر ،دانشگاه صنعتی شاهرود، شاهرود، ایران، hmashayekhi@shahroodut.ac.ir

[۲] استادیار، دانشکده مهندسی مکانیک و مکاترونیک، دانشگاه صنعتی شاهرود، شاهرود، ایران، nazari_mostafa@shahroodut.ac.ir

دریافت: ۱۳۹۹/۰۲/۱۹ ویرایش: ۱۳۹۹/۰۵/۰۹ پذیرش: ۱۳۹۹/۰۹/۱۰

**چکیده:** در این مقاله از یک روش کنترلی غیروابسته به مدل برای ارائه پروتکل درمانی استفاده شده است؛ چراکه استفاده از روش‌های وابسته به مدل به دلیل ماهیت به شدت غیرخطی دینامیک سرطان و وجود عدم قطعیت‌های فراوان با مشکلاتی مانند تضمین پایداری و سختی در طراحی روبرو هستند. در این مقاله، برای تعیین و بهینه‌سازی میزان دوز دارو، از روش کنترل حلقه بسته برمبنای یادگیری تقویتی استفاده شده است. برای ارائه کنترل‌کننده بهینه از روش یادگیری Q استفاده شده است. در این روش یادگیری، هر مدخل جدول Q نشان‌دهنده میزان مطلوب بودن یک عمل انتخابی یا همان دوز داروی شیمی‌درمانی نسبت به یک حالت بیمار می‌باشد. این جدول با استفاده از اطلاعات دریافت شده از حالت سیستم، عمل و پاداش، به روز می‌شود. برای نشان دادن موثر بودن روش کنترلی از یک مدل ریاضی که دارای چهار متغیر حالت سلول‌های ایمنی، سلول‌های سرطانی، سلول‌های سالم و غلظت داروی شیمی-درمانی در خون است، استفاده شده است. سه بیمار جوان، پیر و باردار با شرایط متفاوت و پارامترهای متفاوت درنظر گرفته شده اند، و برای محدود کردن دوز داروی شیمی‌درمانی بر مبنای سن بیمار از یک سیستم فازی استفاده شده است. در بیمار پیر به دلیل ضعف سیستم ایمنی علاوه بر شیمی‌درمانی از ایمنی-درمانی هم استفاده شده است که منجر به تقویت ماندگار سیستم ایمنی می‌شود. نتایج شبیه‌سازی بر روی سه بیمار با شرایط متفاوت، نشان دهنده موثر بودن الگوریتم کنترلی بهینه ارائه شده در درمان سرطان و قابل اعمال بودن آن برای بیماران با شرایط مختلف است. در تمامی بیماران، سرطان در زمان محدودی درمان و دارودهی نیز قطع شده است. همچنین نشان داده شده است که ایمنی‌درمانی در بیماران دارای سیستم ایمنی ضعیف، جهت درمان زمان محدود ضروری می‌باشد.

**کلمات کلیدی:** سرطان، شیمی‌درمانی، ایمنی‌درمانی، کنترل، یادگیری تقویتی.

# Reinforcement learning based feedback control of tumor growth by limiting maximum chemo-drug dose using fuzzy logic

**Hoda Mashayekhi, Mostafa Nazari**

**Abstract:** In this paper, a model-free reinforcement learning-based controller is designed to extract a treatment protocol because the design of a model-based controller is complex due to the highly nonlinear dynamics of cancer. The Q-learning algorithm is used to develop an optimal controller for cancer chemotherapy drug dosing. In the Q-learning algorithm, each entry of the Q-table is updated using data from states, action, and reward. The action is the chemo-drug dose. The proposed controller is implemented on a four states mathematical model including immune cells, tumor cells, healthy cells, and chemo-drug concentration in the bloodstream. Three different treatment strategies are proposed for three young, old, and pregnant patients considering his/her age. Chemotherapy is used in all cases. In the older patient, immunotherapy is also used for modifying the dynamics of cancer by reinforcing his/her weak immune system. A Mamdani fuzzy inference system is designed to limit the maximum chemo-drug dose by regarding the age of the patients. Simulation results show the effectiveness of the proposed treatment strategy. It is also shown that immunotherapy is necessary for finite duration cancer treatment in patients with a weak immune system. The used strategy is a model-free method which is the main advantage of this method.

**Keywords:** Cancer, Chemotherapy, control, Reinforcement learning.

## ۱- مقدمه

دومین عامل مرگ و میر در دنیا پس از بیماری‌های قلبی، سرطان می‌باشد [۱]. در طول دهه اخیر تحقیقات گسترده‌ای برای تشخیص، درمان و اجتناب از سرطان انجام شده است [۲]. به رغم چنین تحقیقات گسترده‌ای، مرگ به واسطه سرطان در حال افزایش است [۲]. پیش‌بینی می‌شود که ۱۳.۲ میلیون نفر به واسطه سرطان تا سال ۲۰۳۰ فوت کنند. از این رو، مدلسازی سرطان و درمان آن مورد توجه بسیاری از دانشمندان در زمینه‌های علوم ریاضی، مهندسی کنترل، بیولوژی و ... قرار گرفته است [۳]. مدلسازی ریاضی سرطان و ارائه روش‌های بهینه درمانی مبتنی بر چنین مدل‌هایی نه تنها موجب صرفه‌جویی در وقت و هزینه خواهد شد، بلکه موجب ارائه روش‌های درمانی جدیدی خواهد شد که تا کنون مورد توجه قرار نگرفته‌اند.

روش‌های گوناگونی برای درمان سرطان مورد استفاده قرار می‌گیرد. جراحی، پرتودرمانی، شیمی‌درمانی و ایمنی‌درمانی از روش‌هایی هستند که برای درمان مورد استفاده قرار می‌گیرند. از روش‌هایی که هنوز به طور گسترده مورد استفاده قرار می‌گیرد، شیمی‌درمانی می‌باشد. روش درمانی ارائه شده باید علاوه بر حذف سلول‌های سرطانی، به کمینه کردن دوز داروی مصرفی نیز بپردازند [۴]. داروهای درمانی مورد استفاده برای درمان سرطان مانند داروهای شیمی‌درمانی علاوه بر کشتن سلول‌های سرطانی، به سلول‌های سالم نیز لطمه وارد کرده و موجب مرگ آن‌ها نیز می‌شوند. همچنین، مصرف زیاد داروها موجب ایجاد مقاومت دارویی در بدن بیمار شده و اثر دارو در طول زمان کاهش می‌یابد.

همچنین، روش درمانی ارائه شده باید به گونه‌ای باشد که بتواند سلول‌های سرطانی را در طول یک دوره محدود نابود کند؛ زیرا همانطور که گفته شد با طولانی شدن طول دوره درمان، علاوه بر ضعیف شدن سیستم ایمنی بدن، اثرات مقاومت دارویی نیز در بدن بیمار ایجاد می‌شود. روش‌های درمانی و دوز داروی اعمالی که توسط پزشکان برای هر بیمار ارائه می‌شود به فاکتورهای مختلفی از جمله سن بیمار، سابقه بیماری، جنسیت و فاکتور BSA و ... بستگی دارد. در این مقاله از یک سیستم فازی برای اعمال حداکثر دوز دارو بر مبنای سن بیمار استفاده شده است.

مدل‌های فراوانی برای ارائه دینامیک سرطان ارائه شده است. یک مدل باید شامل سلول‌های سرطانی، سلول‌های ایمنی و اثرات داروهای درمانی باشد [۵]. روش‌های بهینه‌سازی برای شیمی‌درمانی سرطان در کارهای [۶-۹] ارائه شده است. در مقاله چن و همکاران [۶] روش کنترل پیش‌بین مورد استفاده قرار گرفته است. در این مقاله، مدل با اندازه‌گیری متغیرهای حالت سیستم تعدیل می‌گردد. روش درمان ترکیبی شیمی-درمانی و ایمنی‌درمانی توسط کایران و همکاران [۷] مورد استفاده قرار گرفته است. در این مقاله، روش بهینه‌سازی چند هدفه مورد بررسی قرار گرفته است. نوبل و همکاران [۸] از روش کنترل پیش‌بین با تخمین پارامترهای سیستم استفاده کرده‌اند.

مسئله غیرخطی بودن دینامیک سرطان یکی از چالش‌های تحلیل و کنترل آن می‌باشد. غفاری و همکاران از روش کنترل بهینه وابسته به حالت به دلیل وجود ویژگی‌های خاص آن مانند انعطاف‌پذیری در طراحی و مقاوم بودن برای کنترل بهینه شیمی‌درمانی استفاده کردند [۱۰]. ترکیب روش کنترل بهینه وابسته به حالت و کنترل مدل مرجع برای ارائه یک روش درمانی وابسته به بیمار در [۱۱, ۱۲] ارائه شده است. سایر روش‌های بهینه‌سازی مانند الگوریتم ژنتیک برای بهینه‌سازی دوز داروی شیمی‌درمانی نیز مورد استفاده قرار گرفته‌اند [۱۳, ۱۴]. حجم محاسباتی بالا و سختی در انتخاب جمعیت اولیه و تنظیم پارامترهای اولیه از مشکلات این روش‌هاست.

یادگیری تقویتی یکی از روش‌های پرکاربرد در زمینه یادگیری ماشین می‌باشد [۱۵]. از این روش به طور گسترده برای کنترل بسیاری از سیستم‌ها مانند ربات‌ها، سرعت توربین باد و هلیکوپتر خودران استفاده شده است [۱۶-۲۰]. همچنین از این روش در برخی سیستم‌های بیولوژیکی نیز استفاده شده است [۲۱-۲۶]. با الهام گرفتن از روانشناسی، یادگیری تقویتی امکان یادگیری از تجربیات را برای عاملی که با محیط تعامل دارد، فراهم می‌سازد [۲۷]. عامل برای یافتن استراتژی بهینه، فضای استراتژی‌های ممکن را مورد کاوش قرار داده و بعد از انتخاب‌هایی که انجام می‌دهد بازخورد دریافت می‌کند. با تلاش برای بیشینه کردن کارآیی و رساندن سیستم به حالت مطلوب، سیاستی (استراتژی یا کنترلری) که به صورت ایده‌آل بهینه است، استخراج خواهد شد.

روش مبتنی بر یادگیری تقویتی فاقد مدل است و استراتژی یادگرفته شده می‌تواند برای کنترل مقدار دارو بدون استفاده از مدل بیمار مورد استفاده قرار گیرد. به جای استخراج جزئیات پاسخ، طراح این فرایند کنترلی بایستی پاداش‌های مناسب که نشان‌دهنده تناسب اعمال انتخاب شده توسط عامل باشند را ارائه نماید. استفاده از این روش نه تنها نیاز به دانستن مدل سیستم ندارد بلکه دارای پیچیدگی کمتر و حجم محاسباتی پایین‌تری در مقایسه با الگوریتم‌های تکاملی است.

هدف اصلی در مقاله حاضر، ارائه یک استراتژی درمانی برای درمان سرطان بدون نیاز به مدل سیستم می‌باشد. به این منظور، مسئله کنترل دوز داروی شیمی‌درمانی به صورت یک مسئله بهینه‌سازی ارائه شده و برای حل آن از روش مبتنی بر یادگیری تقویتی استفاده می‌شود. بنابراین نیازی به استفاده از مدل ریاضی سیستم سرطان نمی‌باشد و این روش دارای قابلیت پیاده‌سازی بیشتری در آزمایشات عملی را داراست. به منظور ارزیابی سیستم، برای سه بیمار جوان، پیر، و باردار سه پروتکل درمانی متناسب با سن و شرایط آن‌ها ارائه شده است. در هر سه بیمار از شیمی-درمانی استفاده شده و در بیمار پیر برای اصلاح دینامیک سیستم و تقویت سیستم ایمنی از ایمنی‌درمانی نیز استفاده شده است. با استفاده از یک سیستم فازی ممدانی، دوز داروی شیمی‌درمانی بر مبنای سن هر بیمار محدود می‌شود. تعیین دوز داروی شیمی‌درمانی توسط کنترل حلقه بسته مبتنی بر یادگیری تقویتی انجام می‌شود. به این منظور از الگوریتم تقویتی

یادگیری Q استفاده شده است. یک مدل ریاضی غیرخطی برای نشان دادن رفتار متقابل سیستم ایمنی و سرطان و آموزش سیستم مورد استفاده قرار می‌گیرد. نتایج شبیه‌سازی نشان دهنده موثر بودن الگوریتم کنترلی بهینه ارائه شده در درمان سرطان است.

در ادامه این مقاله در بخش دوم مدل ریاضی رفتار متقابل سیستم ایمنی و سرطان مورد بررسی قرار می‌گیرد، و کنترل مبتنی بر یادگیری تقویتی توضیح داده می‌شود. سپس، نتایج حاصل از شبیه‌سازی بر روی سه بیمار مختلف مورد بررسی قرار می‌گیرد. در انتها نتیجه‌گیری بیان می‌شود.

## ۲- موادها و روش‌ها

### ۲-۱- مدل‌سازی ریاضی سرطان

مدل‌های بسیاری برای نشان دادن دینامیک تقابل بین سلول‌های ایمنی و سرطانی ارائه شده است [۲۸, ۲۹]. در این مقاله از مدل مرجع [۲۸] برای شبیه‌سازی دینامیک سرطان استفاده می‌کنیم. این مدل از چهار متغیر حالت سلول‌های ایمنی ($I(t)$)، سلول‌های سرطانی ($T(t)$)، سلول‌های سالم ($N(t)$) و غلظت داروی شیمی‌درمانی در خون ($M(t)$) تشکیل شده است. با تغییر نام متغیرها به صورت $x_1(t)=N(t)$، $x_2(t)=T(t)$ ،$x_3(t)=I(t)$ و $x_4(t)=M(t)$ مدل به صورت زیر می‌باشد.

$$\dot{x}_1 = r_2 x_1(t)\left(1-b_2 x_1(t)\right) - c_4 x_1(t)x_2(t) - a_3 x_1(t)\left(1-e^{x_4(t)}\right) \quad (۱)$$

$$\dot{x}_2 = r_1 x_2(t)\left(1-b_1 x_2(t)\right) - c_2 x_3(t)x_2(t) - c_3 x_1(t)x_2(t) - a_2 x_2(t)\left(1-e^{x_4(t)}\right) \quad (۲)$$

$$\dot{x}_3 = s + \frac{\rho x_3(t)x_2(t)}{\alpha + x_2(t)} - c_1 x_3(t)x_2(t) - d_1 x_3(t) - a_1 x_3(t)\left(1-e^{x_4(t)}\right) \quad (۳)$$

$$\dot{x}_4 = -d_2 x_4(t) + u(t) \quad (۴)$$

در این مدل رشد سلول‌های سالم و سرطانی به صورت لجستیک، به ترتیب با نرخ رشدهای $r_1$ و $r_2$ در نظر گرفته شده است. پارامترهای $b_1$ و $b_2$ نشان دهنده معکوس حداکثر ظرفیت جمعیت سلول‌های مورد نظر می‌باشند. تقابل بین سلول‌ها به صورت حاصلضرب با نرخ‌های متفاوت درنظر گرفته شده است. سلول‌های ایمنی با نرخ $s$ تولید می‌شوند و با نرخ $d$ به صورت طبیعی می‌میرند. سلول‌های ایمنی همچنین به دلیل وجود سلول‌های سرطانی تحریک شده و تکثیر می‌شوند که با عبارت $\frac{\rho x_3(t)x_2(t)}{\alpha+x_2(t)}$ نشان داده شده است.

عبارت $u$ نشان دهنده دوز داروی اعمالی به داخل بدن بیمار است. داروی شیمی‌درمانی نه تنها سلول‌های سرطانی، بلکه سلول‌های سالم را نیز از بین می‌برد که اثر آن به صورت یک تابع نمایی بر روی هر سلول نشان داده شده است. هدف کنترلر، انتخاب دوز بهینه داروی شیمی‌درمانی برای رسیدن به این هدف است: کمینه کردن دوز داروی شیمی‌درمانی و حذف سلول‌های سرطانی.

سیستم در حالت بدون درمان دارای سه نوع نقطه تعادل می‌باشد [۳۰]. آنچه برای داشتن درمان زمان محدود دارای اهمیت است، وجود نقطه تعادل بدون تومور پایدار است. نقطه تعادل بدون تومور سیستم به صورت زیر می‌باشد:

$$E_{TF} = \left(\frac{1}{b_2}, 0, \frac{s}{d_1}\right) \quad (۵)$$

این نقطه تعادل پایدار است اگر و تنها اگر [۳۱]:

$$r_1 < c_3 + \frac{c_2 s}{d_1} \quad (۶)$$

بنابراین مقادیر پارامترهای رابطه (۶) نقش مهمی در رسیدن به درمان زمان محدود ایفا می‌کنند. چنانچه سیستم ایمنی دارای قدرت کافی باشد (پارامترهای سمت راست رابطه (۶))، آنگاه سیستم دارای یک نقطه تعادل بدون تومور پایدار خواهد بود. برای بیمارانی که دارای سیستم ایمنی ضعیفی هستند، نقطه تعادل بدون تومور ناپایدار خواهد بود و برای رسیدن به یک پروتکل درمانی زمان محدود باید در کنار شیمی‌درمانی از ایمنی‌درمانی نیز بهره جست. ایمنی‌درمانی بر روی پارامترهای سیستم اثر گذاشته، باعث تقویت سیستم ایمنی بیمار می‌گردد [۳۲, ۳۳]. ایمنی‌درمانی باید به گونه‌ای باشد که دینامیک سیستم را اصلاح نماید. برای تغییر در دینامیک سیستم، باید ورودی بر روی پارامترهای سیستم اثر گذاشته و باعث تغییر ماندگار در آن شوند. از این رو، مدل‌سازی باید به صورت زیر باشد [۱۰, ۳۳, ۳۴]:

$$\begin{cases} \dot{x} = f(x,\mu) \\ \dot{\mu} = g\big(u(t)\big) \quad t_1 < t < t_2 \end{cases} \quad (۷)$$

که $\mu$ بخشی از پارامترهای سیستم است که توسط ورودی زمان محدود $u(t)$ اثر می‌پذیرد. در این صورت، پس از قطع ورودی برای $t > t_2$ داریم: $\dot{x} = f(x,\mu') \neq f(x,\mu)$. به عبارت دیگر، پس از قطع ورودی، دینامیک سیستم نسبت به قبل از اعمال ورودی برای $t < t_1$ تغییر کرده است.

بنابراین، ایمنی‌درمانی به صورت زیر مدل‌سازی می‌شود:

$$\frac{ds}{dt} = \mu_s v_v(t)\left(1-\frac{s}{k_s}\right) \quad (۸)$$

در رابطه اثر ایمنی‌درمانی به صورت جمله $v_v(t) \geq 0$ نشان داده شده است. فرض می‌شود که نرخ تغییرات این پارامترها با اندازه بزرگی ورودی $v_v(t)$ متناسب است. مقادیر $\mu_s$ به دینامیک پارامتر $s$ بستگی دارد. این ضریب به صورت اشباعی به حد نهایی $k_s$ که مربوط به محدودیت‌های زیستی عضوهای بدن و انباشتگی اثرات خارجی هستند

اشباع می‌شوند [۳۳, ۳۵].

## ۲-۲- تعیین حداکثر دوز داروی شیمی‌درمانی بر مبنای سن بیمار

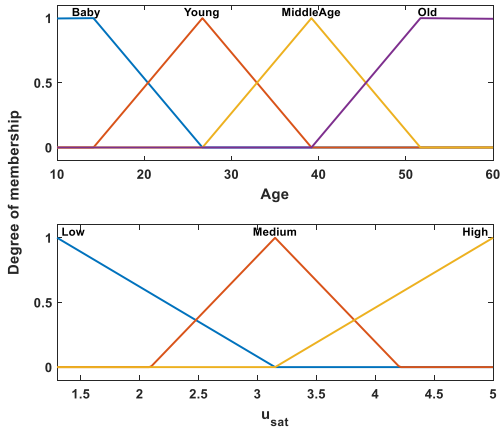
یکی از مواردی که در درمان سرطان باید به آن توجه داشت، درنظر گرفتن شرایط خاص بیمار از جمله شرایط سنی اوست. برای یک بیمار جوان می‌توان دوز داروی بیشتری را نسبت به یک بیمار کهنسال تجویز کرد، زیرا بدن یک بیمار جوان توانایی بیشتری برای بازسازی سلول‌های سالم دارد. درحالیکه برای یک بیمار کهنسال باید دوز کمتری را پیشنهاد داد، زیرا بدن یک بیمار کهنسال توانایی بازسازی سلول‌های سالم را ندارد و باید تا جای امکان به حفظ سلول‌های سالم در بدن او اهتمام ورزید. به طور کلی، با افزایش سن، با توجه به کاهش توانایی بدن در بازتولید سلول‌های سالم باید ازدوز داروی شیمی‌درمانی کم کرد و بر حفظ سلول‌های سالم اهتمام ورزید. فقط در بیمار کودک به دلیل اینکه سیستم ایمنی بدن هنوز مانند بیمار جوان کامل نشده است، باید از دوز کمتری نسبت به بیمار جوان استفاده کرد. برای درنظر گرفتن این مسئله، یک سیستم فازی برای استخراج حداکثر دوز داروی اعمالی $u_{sat}$ پیشنهاد گردیده است که قوانین آن بر مبنای مفاهیم فیزولوژیکی بیان شده است. توابع عضویت مربوط به سن بیمار و حداکثر دوز دارو در شکل ۱ آورده شده است.همانطور که در شکل ۱ نشان داده شده است حداکثر دوز دارو برابر به ۵ درنظر گرفته شده است [۳۰]. قوانین فازی برای سیستم استنتاج ممدانی نیز در جدول ۱ نشان داده شد است.

سیستم فازی ارائه شده از نوع ممدانی می‌باشد و از موتور استنتاج ضرب استفاده شده است. از فازی‌ساز سینگلتون برای فازی سازی و از میانگین مراکز برای فازی‌زدایی استفاده شده است. خروجی سیستم فازی، دامنه تغییرات دوز دارو را برای الگوریتم آموزش تعیین می‌کند.

## ۲-۳- کنترل مبتنی بر یادگیری تقویتی

مفاهیم اصلی یادگیری تقویتی در پیوست آ شرح داده شده‌اند. به منظور به کارگیری کنترل مبتنی بر یادگیری تقویتی، لازم است که مجموعه متناهی حالات $\mathcal{S}$، مجموعه متناهی اعمال $\mathcal{A}$ و تابع پاداش $\mathcal{R}$ متناسب با مساله تعریف گردند. در شیمی درمانی سرطان، هدف انتخاب بهترین میزان داروی شیمی درمانی است، به نحوی که سلول‌های تومور که در ابتدا تعداد آنها بیش‌تر از صفر است، به سمت حالت نهایی مطلوب که همان صفر است متمایل شوند. بنابراین مجموعه حالات $\mathcal{S}$ در حالت کلی همان تعداد سلول‌های سرطانی $(x_2(t))$ در نظر گرفته شده و این مقدار بنا به ویژگی‌های بیمار همانگونه که در بخش بعد شرح داده می‌شود، گسسته‌سازی خواهد شد. حالت مطلوب همان $x_2(t)=0$ می‌باشد. در حالتی که ایمنی درمانی در کنار شیمی درمانی استفاده شود، حالت سیستم وابسته به سلول‌های ایمنی و سرطانی بوده و به صورت $\Gamma\, x_2(t)+(1-\Gamma)x_1(t)$ تعریف می‌شود. در این حالت می‌توان با تغییر ضریب $\Gamma$ میزان حفظ سلول‌های ایمنی و از بین بردن سلول‌های سرطانی را تنظیم نمود. عمل انتخاب شده در هر حالت، دوز داروی شیمی درمانی است که ۲۰ مقدار متفاوت برای آن به صورت $\{$۱، ۰/۹، ۰/۸، ۰/۷، ۰/۶، ۰/۵، ۰/۴، ۰/۳۵، ۰/۲۵، ۰/۲، ۰/۱۵، ۰/۱، ۰/۰۸، ۰/۰۶، ۰/۰۴، ۰/۰۳، ۰/۰۲، ۰/۰۱، ۰$\}=\mathcal{A}$ در نظر گرفته شده است. بسته به شرایط بیمار، می‌توان ضریبی برای مجموعه دوز دارو در نظر گرفت.

عامل یادگیر در زمان یادگیری، با شروع از حالت اولیه تصادفی، عمل تصادفی را انتخاب کرده و مطابق مدل‌سازی ریاضی صورت گرفته به حالت بعدی منتقل می‌شود. در این زمان، با دریافت پاداش از محیط، جدول Q به روز می‌گردد. سپس انتخاب عمل تصادفی تکرار می‌گردد. این امر تا زمان رسیدن به حالت مطلوب ادامه دارد. کل فرآیند یادگیری نیز حداقل ۱۰ بار تکرار می‌شود.




شکل ۱: توابع عضویت مربوط به سن بیمار و حداکثر دوز دارو

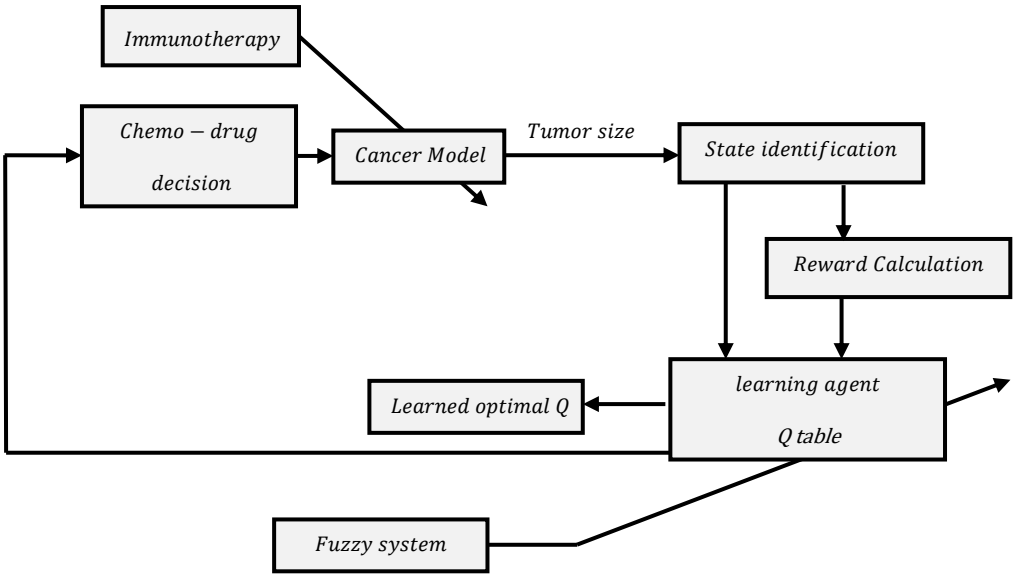
جدول ۱: قوانین فازی

| **سن** | کودک | جوان | میانسال | کهنسال |
|---|---|---|---|---|
| $u_{sat}$ | متوسط | زیاد | متوسط | کم |

به منظور ارزیابی مطلوبیت عمل انتخابی، از تابع پاداش $(\mathcal{R})$ زیر استفاده می‌شود:

$$r_{t+1}=\begin{cases}\dfrac{e(tT)-e\big((t+1)\,T\big)}{e(tT)}, & e\big((t+1)\,T\big)<e(tT)\\ 0 & e\big((t+1)\,T\big)\geq e(tT)\end{cases} \qquad (۹)$$

پارامتر $T$ نشان دهنده زمان نمونه‌برداری از سیستم است که در این مقاله ۰/۱ درنظر گرفته شده است. $e(t)$ مقدار خطا است که بسته به شرایط بیمار در بخش نتایج شرح داده خواهد شد. بعد از اتمام یادگیری، می‌توان از جدول Q برای تعیین دوز داروی شیمی درمانی استفاده نمود. دیاگرام بلوکی سیستم در طول درمان در شکل ۲ نشان داده شده است. همچنین شبه کد الگوریتم در شکل ۳ آمده است.




شکل ۲: طرح کلی آموزش برای یافتن Q بهینه، سیستم فازی تعیین کننده بیشینه دوز داروی شیمی‌درمانی است و ایمنی درمانی، در صورت لزوم قبل از شیمی‌درمانی، برای اصلاح دینامیک سیستم اعمال می‌شود.

## ۳- نتایج

برای نشان دادن تاثیر روش پیشنهادی در درمان سرطان، سه نوع بیمار جوان، پیر و یک خانم باردار درنظر گرفته می‌شود. برای هرکدام از این افراد، استراتژی درمانی مخصوص به خود بیمار را ارائه می‌دهیم.

جدول ۲: مقادیر پارامترهای مدل برای شبیه‌سازی بیمار [۲۲, ۲۴, ۳۲]

| پارامتر | توصیف | مقدار | واحد |
|---|---|---|---|
| $\boldsymbol{a_1}$ | نرخ مرگ سلول‌های ایمنی | ۰/۲ | $\boldsymbol{mg^{-1}\, l\, day^{-1}}$ |
| $a_2$ | نرخ مرگ سلول‌های سرطانی | ۰/۳ | $mg^{-1}\, l\, day^{-1}$ |
| $a_3$ | نرخ مرگ سلول‌های سالم | ۰/۱ | $mg^{-1}\, l\, day^{-1}$ |
| $b_1$ | عکس ظرفیت نهایی سلول‌های سرطانی | ۱ | $\boldsymbol{cell^{-1}}$ |
| $b_2$ | عکس ظرفیت نهایی سلول‌های سالم | ۱ | $cell^{-1}\, day^{-1}$ |
| $c_1$ | نرخ رقابت بین سلول‌های ایمنی و سلول‌های سرطانی | ۱ | $cell^{-1}\, day^{-1}$ |
| $c_2$ | نرخ رقابت بین سلول‌های ایمنی و سلول‌های سرطانی | ۰/۵ | $cell^{-1}\, day^{-1}$ |
| $c_3$ | نرخ رقابت بین سلول‌های سالم و سلول‌های سرطانی | ۱ | $cell^{-1}\, day^{-1}$ |
| $c_4$ | نرخ رقابت بین سلول‌های سالم و سلول‌های سرطانی | ۱ | $cell^{-1}\, day^{-1}$ |
| $d_1$ | نرخ مرگ سلول‌های ایمنی | ۰/۲ | $day^{-1}$ |
| $d_2$ | نرخ نابودی داروی تزریقی در بدن | ۱ | $day^{-1}$ |
| $r_1$ | نرخ رشد سلول‌های سرطانی | ۱/۵ | $day^{-1}$ |
| $r_2$ | نرخ رشد سلول‌های سالم | ۱ | $\boldsymbol{day^{-1}}$ |
| $s$ | نرخ تولید سلول‌های ایمنی | ۰/۳۳ | $\boldsymbol{cell^{-1}\, day^{-1}}$ |
| $\alpha$ | نرخ حدی سیستم ایمنی | ۰/۳ | $\boldsymbol{cell}$ |
| $\rho$ | نرخ پاسخ سیستم ایمنی | ۰/۰۱ | $\boldsymbol{day^{-1}}$ |

جدول ۳: پارامترهای یادگیری تقویتی

| پارامتر | توصیف | مقدار |
|---|---|---|
| $\gamma$ | فاکتور تنزیل در یادگیری تقویتی | ۰/۷ |
| $\eta$ | نرخ یادگیری در یادگیری تقویتی | ۰/۲ |
| $\delta$ | پارامتر آستانه در یادگیری تقویتی | $۱۰^{-۵}$ |
| $\epsilon$ | پارامتر یادگیری حریصانه در یادگیری تقویتی | ۱ |

**Learning Algorithm:**

**Input:** States $S$, Actions $A$, Reward Function $R$, parameters $\{\gamma, \eta, \epsilon, \delta\}$

**Output:** $Q: S \times A \rightarrow \mathbb{R}$

Initialize $Q$ to all zero or arbitrarily

**Repeat** while Q is not converged or for the desired episodes

$s \leftarrow$ initial random state from $S$

**While** $s$ is not the terminal state

$a \leftarrow$select action from $A$ according to $(\epsilon)$-greedy policy

$s' \leftarrow$ next state according to cancer model based on $s$ and $a$

$r \leftarrow$ calculate reward using $R$

Update $Q$ using equation 2

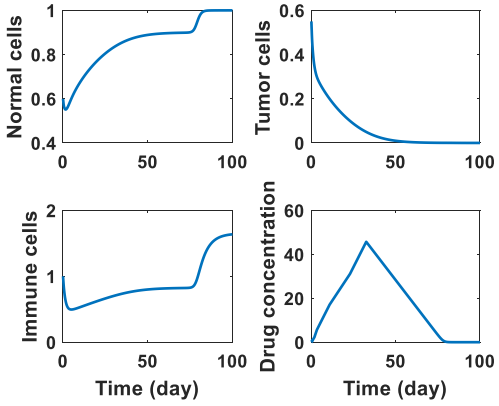
$s \leftarrow s'$

**Return** Q

شکل ۳: الگوریتم کنترل مبتنی بر یادگیری تقویتی

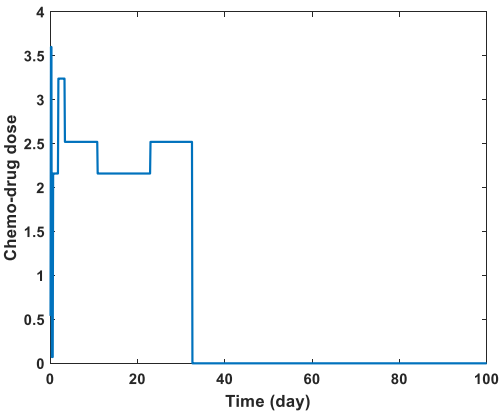
## حالت اول: بیمار جوان ۲۰ ساله؛ شیمی‌درمانی

مقادیر پارامترهای مدل برای بیمار جوان در جداول ۲ و ۳ نشان داده شده اند. برای پارامترهای داده شده در جدول ۲، بیمار دارای یک نقطه تعادل بدون تومور پایدار می‌باشد. به عبارت دیگر سیستم ایمنی این بیمار دارای این قدرت می‌باشد که با کوچک شدن حجم تومور، بتواند بدون دخالت درمان خارجی سلول‌های سرطانی را نابود کند. حداکثر دوز دارو طبق سیستم فازی طراحی شده برای این بیمار $4.4 \frac{\text{mg}}{\text{day L}}$ می‌باشد. دامنه حالات برای این بیمار و همچنین بیمار سوم که در ادامه بحث خواهد شد، با شروع از صفر با نقاط تفکیک {۰/۰۰۶۳، ۰/۰۱۲۵، ۰/۰۲۵، ۰/۰۳۵، ۰/۰۵، ۰/۱، ۰/۲، ۰/۲۵، ....، ۰/۷، ۰/۸، ۰/۹} گسسته‌سازی می‌شود. همانطور که بیان شد، برای بیمار جوان به دلیل قوی بودن سیستم ایمنی، از بین بردن سلول‌های سرطانی در اولویت قرار دارد. رفتار سیستم سرطان در بیمار جوان در حین شیمی‌درمانی در شکل ۴ نشان داده شده است. چون حالت مطلوب همان $x_{2d} = 0$ است، داریم $e(t) = x_2(t) - x_{2d} = x_2(t)$. همانطور که در این شکل مشاهده شده است، تعداد سلول‌های سرطانی به سمت صفر رفته است و غلظت داروی شیمی‌درمانی در خون نیز پس از چند روز صفر شده است. سلول‌های ایمنی و سالم نیز پس از قطع دارودهی با سرعت بیشتری رشد کرده‌اند. تعداد سلول‌های سالم و سلول‌های ایمنی در ابتدا به دلیل عوارض شیمی‌درمانی دچار یک کاهش می‌شوند که در ادامه با طی شدن روند درمان، جمعیت آن‌ها اصلاح می‌شود.




شکل ۴: رفتار سیستم سرطان برای بیمار جوان ۲۰ ساله. غلظت داروی شیمی‌درمانی در خون پس از مدتی صفر شده و به دلیل پایدار بودن نقطه تعادل بدون تومور، سیستم به سمت نقطه تعادل بدون تومور خواهد رفت.

دوز داروی شیمی‌درمانی برای بیمار جوان در شکل ۵ نشان داده شده است. همانطور که مشخص است، به دلیل پایدار بودن نقطه تعادل بدون تومور، دارودهی در کمتر از ۴۰ روز به پایان رسیده است و سیستم به سمت نقطه تعادل بدون تومور درحال حرکت است.




شکل ۵: دوز داروی شیمی‌درمانی اعمالی به بیمار جوان ۲۰ ساله. به دلیل پایدار بودن نقطه تعادل بدون تومور، دارودهی در زمان محدودی انجام شده است.

## حالت دوم: بیمار پیر ۶۰ ساله؛ درمان ترکیبی ایمنی‌درمانی-شیمی‌درمانی

مقادیر پارامترهای سیستم مشابه با بیمار جوان می‌باشد و فقط به دلیل کهولت سن و عدم توانایی در تولید سلول‌های ایمنی، نرخ رشد سلول‌های ایمنی برابر با $s = 0.2$ درنظر گرفته می‌شود. گسسته سازی حالات برای این بیمار با شروع از صفر و با نقاط تفکیک {۰/۰۳، ۰/۱، ۰/۲، ...، ۱، ۱/۲، ...، ۲/۲، ۲/۵، ۳} به ۲۰ حالت انجام می‌شود. با توجه به رابطه (۶)،

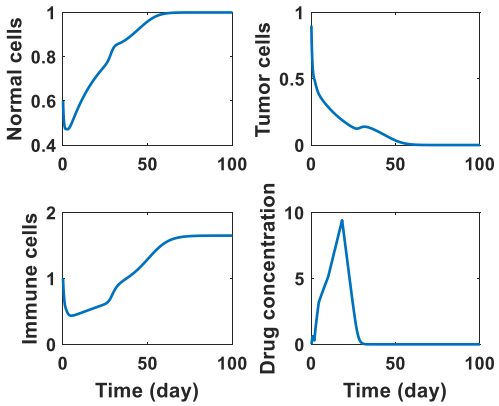
نقطه تعادل بدون تومور ناپایدار بوده و باید با استفاده از ایمنی‌درمانی تقویت شود. ایمنی‌درمانی مطابق با رابطه (۸) بر روی پارامتر نرخ تولید سلول‌های ایمنی اثر می‌گذارد و باعث افزایش آن می‌شود. ایمنی‌درمانی باید تا جایی ادامه یابد که مقدار پارامتر $s$ در رابطه (۶) صدق کند. به عبارت دیگر، سیستم ایمنی باید تا جایی تقویت شود که اگر تومور تا اندازه‌ای کوچک شد، سیستم ایمنی بتواند بدون دخالت درمان خارجی، مبادرت به حذف سلول‌های سرطانی کند.

پس از تقویت سیستم ایمنی با استفاده از ایمنی‌درمانی، شیمی‌درمانی با حداکثر دوز $1.9\frac{\text{mg}}{\text{day L}}$ اعمال می‌شود. با توجه به اینکه مقدار مطلوب برای سلول‌های ایمنی $x_{1d}=1$ و مقدار مطلوب برای سلول‌های سرطانی $x_{2d}=0$ است، برای مصالحه میان حفظ سلول‌های ایمنی و حذف سلول‌های سرطانی در بیمار پیر، پارامتر خطا به صورت $e(t)=\Gamma x_2(t)+(1-\Gamma)x_1(t)$ تعریف می‌شود. در این حالت می‌توان با تغییر ضریب $\Gamma$ میزان حفظ سلول‌های ایمنی و از بین بردن سلول‌های سرطانی را تنظیم نمود. در این مقاله $\Gamma=0.9$ درنظر گرفته شده است.

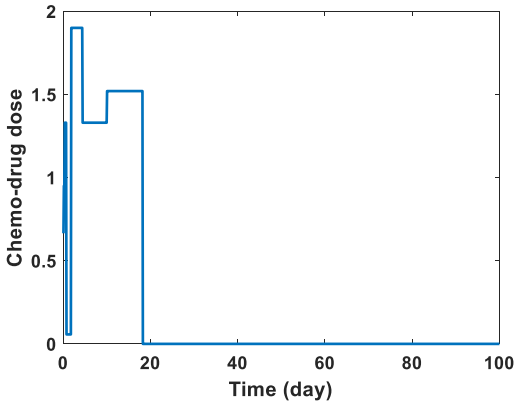
رفتار سیستم برای بیمار ۶۰ ساله پس از اعمال ایمنی‌درمانی و اصلاح پارامتر $s$، در حین شیمی‌درمانی در شکل ۶ نشان داده شده است. تعداد سلول‌های سرطانی و غلظت داروی شیمی‌درمانی در خون در زمان محدود به سمت صفر رفته است. همانطور که بیان شد، کاهش اولیه در سلول‌های سالم و ایمنی ناشی از اثرات شیمی‌درمانی بوده که در ادامه درمان اصلاح می‌گردد. دوز داروی شیمی‌درمانی در شکل ۷ نشان داده شده است.

## حالت سوم: یک خانم باردار جوان؛ شیمی‌درمانی با دوز متغیر

تا زمانیکه جنین در شکم مادر است، دوز داروی شیمی‌درمانی کمی اعمال می‌شود تا هم جلوی رشد تومور گرفته شود و هم کمترین اثربخشی بر روی نوزاد را داشته باشد. پس از تولد نوزاد، دوز دارو افزایش پیدا می‌کند تا تومور از بین برود. رفتار سیستم در شکل ۸ نشان داده شده است. در این حالت فرض شده است که ۲۰ روز تا تولد نوزاد باقی مانده باشد. در طول این ۲۰ روز، حداکثر دوز $1\frac{\text{mg}}{\text{day L}}$ اعمال می‌شود تا اثرات داروی شیمی‌درمانی بر روی جنین کم باشد. پس از تولد نوزاد دوز دارو تا از بین رفتن سلول‌های سرطانی افزایش پیدا می‌کند. پارامتر خطا در این حالت مشابه با حالت ۱ می‌باشد. همانطور که در شکل ۸ مشاهده می‌شود، تا ۲۰ روز سعی بر کنترل رشد سلول‌های سرطانی است. در ابتدا به دلیل کم بودن دوز داروی شیمی‌درمانی، تومور تا حدی شروع به رشد می‌کنند، ولی پس از تولد نوزاد با افزایش دوز دارو، سرطان کنترل می‌شود. همانطور که مشاهده می‌شود، تا تولد نوزاد، غلظت داروی شیمی‌درمانی در خون بسیار پایین می‌باشد.

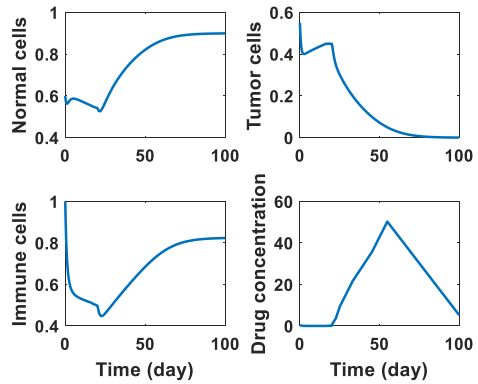



شکل ۶: رفتار سیستم سرطان برای بیمار پیر ۶۰ ساله. غلظت داروی شیمی‌درمانی در خون پس از مدتی صفر شده و به دلیل پایدار شدن نقطه تعادل بدون تومور با استفاده از ایمنی‌درمانی، سیستم به سمت نقطه تعادل بدو تومور خواهد رفت.

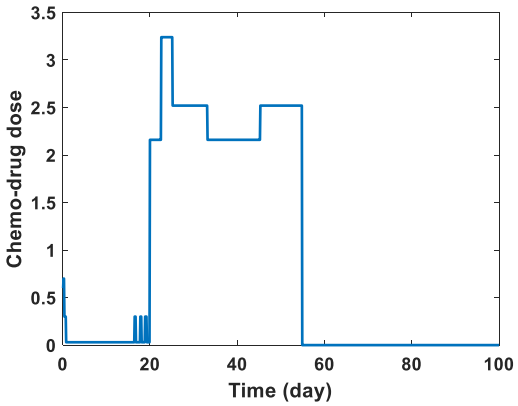



شکل ۷: دوز داروی شیمی‌درمانی اعمالی به بیمار پیر ۶۰ ساله. به دلیل پایدار شدن نقطه تعادل بدون تومور با استفاده از ایمنی‌درمانی، شیمی‌درمانی در زمان محدودی انجام شده است.

دوز داروی شیمی‌درمانی اعمال شده در این حالت در شکل ۹ نشان داده شده است. پس از ۲۰ روز، دوز دارو زیاد شده تا بتواند سلول‌های سرطانی را از بین ببرد. به دلیل جوان بودن بیمار و پایدار بودن نقطه تعادل بدون تومور، دارودهی پس از مدتی قطع می‌شود و خود سیستم ایمنی قادر به تقابل با سلول‌های سرطانی می‌باشد.




شکل ۸: رفتار سیستم سرطان برای خانم جوان بارداری که ۲۰ روز تا تولد نوزاد باقی مانده است. تا تولد نوزاد غلظت دارو بسیار پایین می‌باشد و پس از تولد نوزاد، غلظت دارو افزایش می‌یابد.




شکل ۹: دوز داروی شیمی‌درمانی اعمالی به بیمار خانم بارداری که ۲۰ روز تا تولد نوزاد باقی مانده است. تا ۲۰ روز دوز دارو بسیار محدود شده و پس از آن محدودیت تقلیل می‌یابد.

## ۴- بحث و جمع‌بندی

ارائه یک روش درمانی مناسب، از اصلی‌ترین چالش‌های درمان سرطان می‌باشد. در این مقاله برای سه بیمار جوان، پیر، و باردار سه پروتکل درمانی متناسب با سن و شرایط آن‌ها ارائه شده است. در هر سه بیمار از شیمی‌درمانی استفاده شده است. به دلیل اینکه دینامیک سرطان بسیار غیرخطی می‌باشد و دارای عدم قطعیت‌های فراوانی است، در این مقاله برای اعمال مقدار بهینه دوز داروی شیمی‌درمانی از روش کنترل حلقه بسته مبتنی بر یادگیری تقویتی که غیروابسته به مدل می باشد، استفاده شده است. در این روش، استراتژی بهینه دارودهی با یک استراتژی اولیه شروع شده و در طول زمان با تقابل با سیستم و درنظر گرفتن یک تابع پاداش، بهینه می‌گردد.

برای پیاده سازی روش بیان شده، از مدل سلولی غیرخطی (۱)-(۴) استفاده شده است. در این مدل، هم سیستم ایمنی ذاتی بدن و هم سیستم ایمنی تطبیقی بدن درنظر گرفته شده است. مدل سلولی در مقاله حاضر و مرجع [۲۱]، از مرجع [۲۸] استخراج شده است، با این تفاوت که ما در این مقاله برای نمایش دقیق‌تر اثرات داروی شیمی‌درمانی از ترم نمایی مشابه با مرجع [۳۰] استفاده کرده‌ایم که خود منجر به غیرخطی‌تر شدن دینامیک سیستم می‌شود. نتایج شبیه‌سازی برای دو بیمار جوان و باردار نشان دهنده عملکرد مناسب روش کنترلی پیاده‌سازی شده است. در بیمار جوان، شیمی‌درمانی در کمتر از ۴۰ روز به پایان رسیده است و پس از این مدت روز، خود سیستم ایمنی بدن قادر به حذف سلول‌های سرطانی می‌باشد (پایدار بودن نقطه تعادل بدون تومور). در بیمار جوان باردار، تا مرحله تولد نوزاد، دوز داروی شیمی‌درمانی تا مقدار یک محدود شده است. پس از تولد نوزاد، دوز دارو افزایش داشته است. در بیمار پیر برای اصلاح دینامیک سیستم و تقویت سیستم ایمنی، از ایمنی‌درمانی نیز استفاده شده است. به عبارت دیگر، سیستم ایمنی بدن باید تا حدی تقویت گردد که اگر سلول‌های سرطانی تا میزان مشخصی کاهش یافتند، خود سیستم ایمنی بدن بدون دخالت درمان خارجی بتواند بر سلول‌های سرطانی غلبه کند. ایمنی درمانی مطابق با رابطه (۸) بر روی پارمتر S اثر ماندگار دارد و منجر به تغییر آن خواهد شد. تغییر پارامتر سیستم با استفاده از واکسن‌درمانی در مراجع [۳۰] و [۳۲] اشاره شده است.

پزشکان دوز داروی اعمالی بر هر بیمار را بر مبنای شرایط آن بیمار مانند جنسیت، حجم بدن، داشتن سابقه بیماری، و سن آن بیمار اعمال می‌کنند. در این مقاله، برای محدودسازی دوز داروی شیمی‌درمانی بر مبنای سن هر بیمار، یک سیستم فازی ممدانی طراحی شده است. در این سیستم فازی بر مبنای سن بیمار، دوز مشخصی استخراج می‌شود.

در مقایسه با کار انجام شده در [۳۱]، نتایج به طور مشابه می‌باشد. در هر دو تحقیق، کنترل‌کننده‌ها قادر به حذف سلول‌های سرطانی به صورت بهینه بودند. اما اهمیت روش استفاده شده در این است که نیازی به مدل ریاضی سیستم در طراحی کنترل‌کننده ندارد. در مقایسه با نتایج ارائه شده در [۲۱] نیز نتایج مشابه می‌باشد. در مقاله حاضر در مقایسه با [۲۱] عبارت مربوط به شیمی‌درمانی دقیق‌تر شده است و همچنین برای اعمال داروی شیمی‌درمانی از یک سیستم فازی استفاده شده است. همچنین، اهمیت ایمنی‌درمانی و تغییر در دینامیک سیستم برای داشتن درمان زمان محدود مورد بررسی قرار گرفته است.

## ۵- نتیجه گیری

در این مقاله یک استراتژی درمانی برای درمان سرطان بدون نیاز به مدل سیستم بر مبنای یادگیری تقویتی ارائه گردید. به این منظور، مسئله کنترل دوز داروی شیمی‌درمانی به صورت یک مسئله بهینه‌سازی ارائه و برای حل آن از یادگیری Q استفاده شد. از محاسن روش ارائه شده، عدم نیاز به دانستن مدل ریاضی سیستم می‌باشد که خود قدم بزرگی در طراحی کنترل‌کننده‌ها در سیستم غیرخطی سرطان است. برای سه بیمار جوان، پیر، و باردار سه پروتکل درمانی متناسب با سن و شرایط آن‌ها ارائه شد و از یک سیستم فازی ممدانی برای محدودسازی دوز داروی شیمی‌درمانی استفاده گردید. در این سیستم فازی، سن بیمار به عنوان ورودی و حداکثر دوز داروی مجاز خروجی آن می‌باشد. همچنین، در بیمار پیر برای رسیدن به درمان زمان محدود از ایمنی‌درمانی استفاده شده است. ایمنی‌درمانی باعث تقویت رشد سلول‌های ایمنی در بیمار می‌شود و از منظر ریاضی باعث پایدار شدن نقطه تعادل بدون تومور می‌گردد. در تمامی بیماران، هدف از شیمی‌درمانی، قرار دادن مسیر حرکت سیستم در ناحیه جذب نقطه تعادل بدون تومور می‌باشد. از این رو، وجود نقطه تعادل بدون تومور پایدار در درمان زمان محدود ضروری می‌باشد. نتایج شبیه‌سازی نشان دهنده موثر بودن الگوریتم کنترلی بهینه ارائه شده در درمان سرطان است. استفاده از روش ارائه شده با استفاده از درمان‌های ترکیبی و استفاده از توابع پاداش انتگرالی از کارهایی است که در آینده ارائه خواهد شد. همچنین استفاده از روش یادگیری Q به صورت پایه نیازمند گسسته‌سازی مجموعه حالات و اعمال است، که در کاربردهای جهان واقعی مبتنی بر نظر متخصص می‌باشد. همچنین انتخاب پارامترهای مدل به صورت تجربی کاری زمانبر می‌باشد. اعمال الگوریتم یادگیری در

فضای پیوسته، انتخاب خودکار پارامترها و همچنین استفاده از یادگیری تقویتی عمیق از جمله کارهای آینده می‌باشد.

## ۶- مراجع

[1] H. Ritchie and M. Roser, "Causes of Death," *Our World in Data*, 2020.

[2] F. Biemar and M. Foti, "Global progress against cancer—challenges and opportunities," *Cancer biology and medicine*, vol. 10, no. 4, pp. 183-186, 2013.

[3] R. P. Araujo and D. L. S .MCelwain, "History of the study of solid tumour growth: the contribution of mathematical modeling," *Bulletin of 2Mathematical Biology*, vol. 66, pp. 1039–1091, 2004.

[4] J.C. Doloff and D. J. Waxman, "Transcriptional profiling provides insights into metronomic cyclophosphamide-activated, innate immune-dependent regression of braintumor xenografts," *BMC Cancer*, vol. 15, no. 1, p. 375, 2015.

[5] L. G. De Pillis and A. E. Radunskaya, "A mathematical tumor model with immune resistance and drug therapy: an optimal control approach," *Journal of Theoretical Medicine*, vol. 3, no. 2, pp. 79–100, 2001.

[6] T. Chen, N.F.Kirkby, and R. Jena, "Optimal dosing of cancer chemotherapy using model predictive control and moving horizon state/parameter estimation," *Computer Methods Programs Biomedicine*, vol. 108, no. 3, pp. 1337-1340, 2012.

[7] K.L. Kiran, D. Jayachandran, and S. Lakshminarayanan, "Multi-objective optimization of cancer immuno-chemotherapy," presented at the 13th International Conferenceon Biomedical Engineering, 2009 .

[8] S.L. Noble, E. Sherer, R.E.Hannemann, D.Ramkrishna, T. Vik, and A.E.Rundell, "Using adaptive model predictive control to customize maintenance therapy chemotherapeutic dosing for childhood a cutely mphoblastic leukemia," *Journal of Theoretical Biology*, vol. 264, no. 3, pp. 990-1002, 2010.

[9] M. Engelhart, D. Lebiedz, and S. Sager, "Optimal control for selected cancer chemotherapy ODE models: a view on the potential of optimal schedules and choice of objective function," *Mathematical Biosciences*, vol. 229, no. 1, pp. 123-134, 2011.

[10] A. Ghaffari, M. Nazari, and F. Arab, "Suboptimal mixed vaccine and chemotherapy in finite duration cancer treatment: state-dependent Riccati equation control," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 37, no. 1, pp. 45-56, 2015.

[11] N. Babaei and M. Salamci, "Controller design for personalized drug administration in cancer therapy: Successive approximation approach," *Optimal Control Applications and Methods*, pp. 1-38, 201.۷

[12] N. Babaei and M. U. Salamci, "Mixed therapy in cancer treatment for personalized drug administration using model reference adaptive control," *European Journal of Control*, vol. In press, 2019.

[13] K.C. Tan, E.F. Khor, J. Cai, C. Heng, and T. H.Lee" ,Automating the drug scheduling of cancer chemotherapy via evolutionary computation," *Artifficial Intelligence in Medicine*, vol. 25, no. 2, pp. 169-185, 2002.

[14] S.-M.Tse, Y.Liang, K.-S.Leung, K.-H.Lee, and T.S.-K.Mok, "A memetic algorithm for multiple-drug cancer chemotherapy schedule optimization " *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 37, no. 1, pp. 84-91, 2007.

[15] D.Vrabie, K.G.Vamvoudakis, and F.L.Lewis, *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principle*. Lomdon, UK: Institution of Engineering andTechnology, 2013.

[16] M. Sedighizadeh and A. Rezazadeh, "Adaptive PID controller based on reinforcement learning for wind turbine control," *World Academy of Science, Engineering and Technology*, vol. 13, pp. 1-23, 2008.

[17] P. Abbeel, A. Coates, M. Quigley, and A. Y. Ng, "An application of reinforcement learning to aerobatic helicopter flight," *Advances in Neural Information Processing Systems*, vol. 19, pp. 1-8, 2007.

[18] I. Carlucho, M. De Paula, and G. G. Acosta, "An adaptive deep reinforcement learning approach for MIMO PID control of mobile robots," *ISA Transactions*, vol. In press, 2020.

[19] C. Pi, K. Hu, S. Cheng, and I. Wu, "Low-level autonomous control and tracking of quadrotor using reinforcement learning," *Control Engineering Practice*, vol. 95, 2020.

[20] W. Koch, R. Mancuso, R. West, and A. Bestavros, "Reinforcement Learning for UAV Attitude Control," *ACM Transactions on Cyber-Physical Systems*, vol. 22, 2019.

[211] R. Padmanabhan ,N. Meskina, and W. M. Haddad, "Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment," *Mathematical Biosciences*, vol. 293, pp. 11-20, 2017.

[22] J. Martin-Guerrero, F. Gomez, E.Soria-Olivas, J. Schmidhuber, M.Climente–Marti, and N.Jemenez-Torres, "A reinforcement learning approach for individualizing erythropoiet in dosages in

change the dynamics of a system:Application to finite duration cancer treatment," presented at the 22nd Annual Conference of Mechanical Engineering, Ahvaz, Iran, 2014 .

[35] M. Nazari, A. Ghaffari, and F. Arab, "Finite duration treatment of cancer by using vaccine therapy and optimal chemotherapy: state-dependent riccati equation control and extended kalman filter," *Journal of Biological Systems*, vol. 23, no. 1, 2015.

[36] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA,: MIT Press, 1998.

[37] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, no. 1, pp. 137-285, 1996.

## پیوست آ: یادگیری تقویتی

یادگیری تقویتی یکی از شاخه‌های یادگیری ماشین است که امکان یادگیری از تجربیات را برای عاملی که با محیط تعامل دارد، فراهم می‌سازد. در حین کاوش استراتژی بهینه با یادگیری تقویتی، تجربه‌های جدید با سعی و خطا به وجود می‌آیند. در حین تعامل با سیستم، عامل یادگیر تلاش می‌کند اعمالی را یاد بگیرد که پاداش تجمعی کسب شده را بیشینه می‌کنند. به طور کلی، مسائل یادگیری تقویتی با فرآیندهای تصمیم‌گیری مارکوف (MDP) از طریق زنجیره مونت کارلو و برنامه‌ریزی پویا مدل و به صورت تکراری حل می‌شوند [۳۶]. یکی از رویکردهای یادگیری تقویتی که به مدل دقیق سیستم نیازی ندارد و می‌تواند در حین تغییرات سیستم مورد استفاده قرار گیرد روش یادگیری Q پیشنهاد شده توسط واتکین [۲۷] می‌باشد. چهار دنباله مجموعه متناهی حالات S، مجموعه متناهی اعمال A تعریف شده برای هر حالت $s \in S$، تابع پاداش R برای هدایت عامل مطابق با عمل $a \in A$، و ماتریس احتمال انتقال P، در چارچوب فرایند تصمیم‌گیری مارکوف متناهی برای فهم رفتار سیستم استفاده می‌شود.

آموزش عامل، یک فرایند پیوسته است که در طی آن تعاملات با محیط در قدم‌های زمانی گسسته رخ می‌دهند. در یک قدم t عامل، حالت کنونی $s_t$ را مشاهده نموده و یک عمل $a_t$ را برای تعامل با محیط انتخاب می‌کند. محیط به عمل پاسخ داده و به حالت جدید $s_{t+1}$ منتقل می‌شود. ماتریس انتقال حالت $P_{a_t}(s_t, s_{t+1})$ احتمال انتقال از حالت $s_t \in S$ به حالت $s_{t+1}$ در نتیجه عمل $a_t \in A$ را مشخص می‌کند. یک تابع $\pi$ که عمل $a_t \in A$ را بر مبنای حالت $s_t \in S$ تولید می‌کند، یک سیاست نام دارد [۳۷]. به بیان دیگر $\pi: S \to A \Rightarrow a_t = \pi(s_t)$. عامل می‌تواند یک پاداش متناسب با انتقال $(s_t, a_t, s_{t+1})$ دریافت نماید. پاداش متناسب $r_{t+1} \in R$ نشان‌دهنده میزان مطلوب بودن عمل انتخاب شده $a_t$ می‌باشد. یک مسئله یادگیری تقویتی، عملا یافتن سیاستی است که مجموع پاداش‌های R(s, a) کسب شده در طول زمان را بیشینه نماید. در یادگیری Q، عامل بعد از هر گام، تابع عمل-مقدار Q(s, a) را که نشان دهنده کیفیت هر عمل در هر حالت است به روز می‌کند. یادگیری Q، یک الگوریتم یادگیری تقویتی مبتنی بر مقدار است که برای یافتن یک سیاست بهینه انتخاب عمل توسط تابع Q استفاده می‌شود.

hemodialysis patients," *Expert Systems with Applications*, vol. 36, pp. 9737-9742, 2009.

[23] B.L. Moore, L.D. Pyeatt, V. Kulkarni, P. Panousis, Kevin, and A.G.Doufas, "Reinforcement learning for closed-loop propofol anesthesia : a study in human volunteers," *Journal of Machine Learning Research*, vol. 15, pp. 655-696, 2014.

[24] P. Yazdjerdi, N. Meskin, M. Al-Naemi, A. Al Moustafa, and L. Kovács, "Reinforcement learning-based control of tumor growth under anti-angiogenic therapy," *Computer Methods and Programs in Biomedicine*, vol. 173, pp. 15-26, 2019.

[25] R. Padmanabhana, N. Meskin, and W. M. Haddad, "Optimal adaptive control of drug dosing using integral reinforcement learning," *Mathematical Biosciences*, vol. 309, pp. 131-142, 2019.

[26] M. Tejedor, A. Z. Woldaregay, and F. Godtliebsen, "Reinforcement learning application in diabetes blood glucose control: A systematic review," *Artificial Intelligence in Medicine*, vol. 104, pp. 101-183, 2020.

[27] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279-292, 1992.

[28] L. G. De Pillis and A. E. Radunskaya, "The dynamics of an optimally controlled tumor model: a case study," *Mathematical and Computer Modeling*, vol. 37, pp. 1221–1244, 2003.

[29] A. Talkington, C. Dantoin, and R. Durrett, "Ordinary Differential Equation Models for Adoptive Immunotherapy," *bulletin of Mathematical Biology*, vol. 80, no. 5, pp. 1059-1083.۲۰۱۸ ,

[30] L. G. De Pillis, W. Gu, and A. E. Radunskaya, "Mixed immunotherapy and chemotherapy of tumors: modeling, applications and biological interpretations," *Journal of Theoretical Biology*, vol. 238, pp. 841–862, 2006.

[31] Y. Batmani and H. Khaloozadeh, "Optimal chemotherapy in cancer treatment: state dependent Riccati equation control and extended Kalman filter," *Optimal Control Applications and Methods*, vol. 34, pp. 562-577, 2012.

[32] A. Ghaffari, M. Nazari, M. Khazaee, and B. Bahmaei, "Changing the dynamics of a system by using finite duration inputs: Application to cancer modeling and treatment," *Journal of Solid and Fluid Mechanics*, vol. 4, no. 1, pp. 79-91, 2014.

[33] M. Nazari and A. Ghaffari, "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, no. 3, pp. 1-19, 2015.

[34] A. Ghaffari, M. Nazari, B. Bahmaie, and B. Ghaffari, "How finite duration inputs are able to

روش مبتنی بر یادگیری تقویتی، با یک سیاست اولیه دلخواه شروع به کار می‌کند و این سیاست در تعامل با سیستم به روز می‌شود [۳۶]. با دریافت اطلاعات بیشتر در قالب حالت، عمل و پاداش، سیاست به سمت سیاست بهینه پیش می‌رود. در روش یادگیری $Q$ هر مدخل جدول $Q$ با استفاده از اطلاعات دریافت شده از حالت، عمل و پاداش، به روز می‌شود. در این جدول سطرها حالت‌های بالقوه را نشان می‌دهند و ستون‌ها عمل‌ها را نشان می‌دهند. مدخل $Q(s,a)$ در جدول $Q$ میزان مطلوب بودن عمل $a$ را نسبت به حالت $s$ نشان می‌دهد. برای مدلسازی مطلوب بودن، مدخل $Q(s,a)$ مجموع پاداش مورد انتظار از انجام عمل $a$ در حالت $s$ را نشان می‌دهد. بنابراین الگوریتم یادگیری $Q$ تلاش می‌کند تا مقدار مورد انتظار پاداش تنزیل یافته را بیشینه نماید:

$$Q^{\pi}(s_t,a_t)=E\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}\right] \quad (۱۳)$$

پارامتر $0\le\gamma\le 1$ فاکتور تنزیل است که میزان اهمیت پاداش‌های کنونی و آینده را نشان می‌دهد. وقتی $\gamma=0$ عامل تنها پاداش کنونی را در نظر می‌گیرد و وقتی $\gamma$ به سمت ۱ می‌رود، پاداش‌های آینده نیز اهمیت پیدا می‌کنند. در هر صورت، به دلیل اینکه اهمیت پاداش‌های آینده به صورت نمایی کاهش پیدا می‌کند، الگوریتم به دنبال دریافت سریع‌تر پاداش‌های دقیق و مثبت خواهد بود. جدول $Q$ با استفاده از پاداش $r_{t+1}$، انتقال حالت $s_t\rightarrow s_{t+1}$ و عمل $a_t$ به روز می‌شود.

$$\begin{aligned}&Q(s_t,a_t)\\&\leftarrow Q(s_t,a_t)\\&+\eta\times\left[r_{t+1}+\gamma\max_{a_{t+1}}Q(s_{t+1},a_{t+1})\text{-}Q(s_t,a_t)\right]\end{aligned} \quad (۱۴)$$

نرخ یادگیری $\eta\in[0,1]$ میزان تصحیح مدخل جدول بعد از هر تکرار را تغییر می‌دهد. هر دو پارامتر $\eta$ و $\gamma$ می‌توانند در طول زمان کاهش داده شوند. بعد از هر مشاهده از محیط، جدول $Q$ به روز می‌شود. در حالت نهایی، جدول $Q$ دیگر به روز نمی‌شود و مقدار مدخل جدول به پاداش دریافتی مقداردهی می‌شود. تکرار الگوریتم تا زمانی که به همگرایی برسیم ادامه دارد. پارامتر آستانه $\delta$ در شرط $\Delta Q_{t+1}\triangleq|Q_{t+1}\text{-}Q_t|\le\delta$ برای بررسی همگرایی الگوریتم به کار می‌رود [۲۷, ۳۶]. عامل یا همان کنترلر در هر گام زمانی $t$ و حالت $s_t$، عمل $a_t$ را مطابق سیاست زیر انتخاب می‌کند:

$$a_t=\underset{a_k}{\operatorname{argmax}}\,Q_t(s_t,a_k) \quad (۱۵)$$

عامل در هر حالت عمل‌های تصادفی را نیز با احتمال $\epsilon$ انتخاب می‌کند که $\epsilon$ یک مقدار کوچک مثبت است. این امر سیاست حریصانه $\epsilon$ نام دارد. در نهایت با ادامه کاوش وقتی $t\rightarrow\infty$ جدول بهینه $Q$ استخراج خواهد شد. در بسیاری موارد، همگرایی الگوریتم در قدم‌های متناهی رخ خواهد داد.