Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 10, Number 1, Spring 2021, pp. 72-92
Received: 02/18/2019, Revised: 03/24/2020, Accepted: 11/24/2020

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

Elnaz Kalhor[1], Amin Noori[2*], Sara Saboori Rad[3] and Mohammad ali sadrnia [4]
1- Faculty of Electrical and Biomedical Engineering, Sadjad University, Mashhad, Iran.
2*- Faculty of Electrical and Biomedical Engineering, Sadjad University, Mashhad, Iran.
3- Department of Dermatology, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
4- Faculty of Electrical and Robotics Engineering, Shahrood University of Technology, Shahrood, Iran.

[1]E.kalhor123@sadjad.ac.ir, [2*]Amin.noori@sadjad.ac.ir, [3]Sabourirads@mums.ac.ir, and [4]masadrnia@shahroodut.ac.ir

Corresponding author address: Amin Noori, Faculty of Electrical and Biomedical Engineering, Sadjad University of Technology, Mashhad, Iran, Post Code: 9188148848.

Abstract- This paper mainly aims to determine the optimal drug dosage for the purpose of reducing the population of cancer cells in melanoma patients. To do so, Reinforcement Learning method and the eligibility traces algorithm are employed, giving us the advantage of creating a compromise between the two algorithms of the reinforcement learning, being Monte-Carlo and Temporal Difference. Furthermore, it can be said that using this approach, there was no need to employ a mathematical model in the whole process. However, as its implementation on the real system was not possible, a delayed nonlinear mathematical model is used to investigate the performance of the proposed controller and simulate the behavior of the environment. It should be noted this mathematical model made use of no control method. This is the first time that population control of cancer cells is applied and tested on this model. To know of the optimal dosage of the drug, it should be mentioned that the drug is required to prevent the side effects on healthy/normal cells as much as possible. According to the obtained results, the eligibility traces algorithm is able to control and reduce the population of cancer cells through injecting the sub-optimal drug dose. This will increase the level of immunity in our body. Finally, to demonstrate the advantage of a selective method of increasing the rate of cancer cell death, this method is compared with the Q-learning algorithm and optimal control. By applying the fault to the sensor, the performance of the proposed controller to reduce cancer cells was investigated. The adaptability of the proposed method with the environment changes is checked afterwards. To this end, uncertainty in the system parameters and initial conditions are applied and the population of cancer cells are controlled in five melanoma patients. Moreover, having added noise to the system, it was shown that the eligibility traces algorithm is able to control the population of cancer cells and make it reach zero. Additionally, the convergence speed of both eligibility traces algorithm and Q learning algorithm in reducing the number of cancer cells for different learning rates was investigated.

Keywords- Side effects, Q-learning algorithm, Uncertainty, cancer cells population control, Melanoma, Reinforcement learning, Eligibility traces, Optimal control method, Convergence speed.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۱۰، شماره ۱، بهار ۱۴۰۰ ، صفحه ۷۲-۹۲
دریافت: ۱۳۹۷/۱۱/۲۹، بازنگری: ۱۳۹۹/۰۱/۰۵ پذیرش: ۱۳۹۹/۰۹/۰۴

تعیین دوز بهینه دارو برای کنترل جعیت سلول‌های سرطانی با لحاظ اثرات زیان‌بار دارو در بیمار مبتلا به ملانوما با استفاده از روش مسیرهای شایستگی

الناز کلهر[۱]، امین نوری*[۲]، سارا صبوری راد[۳]، محمدعلی صدرنیا [۴]
۱و۲- دانشکده برق و مهندسی پزشکی، دانشگاه سجاد، مشهد، ایران.
۳- دانشکده پوست، دانشگاه علوم پزشکی مشهد، مشهد، ایران.
۴- دانشکده مهندسی برق و رباتیک، دانشگاه صنعتی شاهرود، شاهرودؤ ایران.
[1]E.kalhor123@sadjad.ac.ir, [2*]Amin.noori@sadjad.ac.ir, [3]Sabourirads@mums.ac.ir, [4]masadrnia@shahroodut.ac.ir

* نشانی نویسنده مسئول: امین نوری، مشهد، بلوار جلال آل احمد،جلال آل احمد۶۴، دانشگاه صنعتی سجاد، دانشکده برق و مهندسی پزشکی، کد پستی: ۹۱۸۸۱۴۸۸۴۸

چکیده- هدف اصلی در این مقاله، تعیین میزان بهینه دوز دارو برای کاهش جمعیت سلول‌های سرطانی در بیماران مبتلا به سرطان ملانوما می‌باشد. برای این کار از روش مسیرهای شایستگی که یکی از روش‌های حل مسئله یادگیری تقویتی می‌باشد، استفاده شده است. این روش مزایای دو روش مرسوم یادگیری تقویتی شامل یادگیری تفاوت گذرا و مونت کارلو را دارا می‌باشد. از دیگر مزایای این روش می‌توان به بی‌نیاز بودن آن به مدل ریاضی اشاره کرد ولی چون امکان پیاده‌سازی بر روی سیستم واقعی امکان پذیر نبوده است، برای بررسی عملکرد کنترلر پیشنهادی از مدل ریاضی غیرخطی تاخیردار جهت شبیه‌سازی رفتار محیط استفاده گردیده است. با توجه به بررسی‌هایی که تاکنون انجام شده است،لازم به ذکر می‌باشد که بر روی این مدل ریاضی هیچ نوع روش کنترلی پیاده‌سازی نشده است و این اولین باری می‌باشد که کنترل جمعیت سلول‌های سرطانی برای این مدل انجام گرفته است. در کنترل بهینه دوز دارو، میزان دارو می‌بایست به گونه‌ای باشد تا از اثرات زیان‌بار دارو بر روی سلول‌های سالم تا حد امکان جلوگیری شود. با توجه به نتایج حاصل از شبیه‌سازی، مشاهده می‌شود که روش انتخابی توانسته است با تزریق زیر بهینه میزان دوز دارو، جمعیت سلول‌های سرطانی را کنترل کرده، کاهش داده و به صفر برساند که این امر، در کنار افزایش سلول‌های ایمنی بدن رخ داده است. در انتها برای نشان دادن مزیت روش انتخابی در افزایش سرعت برای کاهش سلول‌های سرطانی، این روش با روش الگوریتم یادگیری Q که یکی دیگر از روش‌های حل مسئله یادگیری تقویتی می‌باشد و روش کنترل بهینه مقایسه شده است. با اعمال عیب به سنسور سیستم نیز، عملکرد کنترلر پیشنهادی برای کاهش سلول‌های سرطانی در حضور عیب مورد بررسی قرار گرفت. برای بررسی یکی از مزایای روش یادگیری تقویتی که تطبیق‌پذیری آن با محیط می‌باشد، با لحاظ عدم قطعیت در پارامترهای سیستم و شرایط اولیه، کنترل جمعیت سلول‌های سرطانی در پنج بیمار مبتلا به سرطان ملانوما انجام شده است. در نهایت نیز، با افزودن نویز به سیستم نشان داده شده است که روش مسیرهای شایستگی باز هم قادر به کنترل جمعیت سلول‌های سرطانی و رساندن آن‌ها به صفر بوده است. همچنین سرعت همگرایی هر دو روش مسیرهای شایستگی و الگوریتم یادگیری Q در کاهش سلول‌های سرطانی به ازای نرخ‌های آموزش مختلف مورد بررسی قرار گرفته است.

واژه‌های کلیدی: اثرات زیان‌بار دارو، الگوریتم یادگیری Q، عدم قطعیت،کنترل جمعیت سلول‌های سرطانی، ملانوما، یادگیری تقویتی، مسیرهای شایستگی، کنترل بهینه، سرعت همگرایی.حداکثر ده کلمه به‌عنوان کلمات کلیدی انتخاب شود. این کلمات باید بیانگر موضوعات اصلی و فرعی مقاله باشند.

۱- مقدمه

سلول‌های پوست به طور طبیعی رشد می‌کنند و به شکل سلول‌های جدید در می‌آیند. هر روز سلول‌های پوست پیر می‌شوند و سلول‌های جدید جایگزین آن‌ها می‌شوند. گاهی اوقات این روند دچار اختلال می‌شود و سلول‌های جدید در جایی که پوست به آن‌ها احتیاجی ندارد، به وجود می‌آیند و سلول‌های پیر هم در زمانی که باید، از بین نمی‌روند. این سلول‌های اضافه ایجاد توده‌ای از بافت می‌کنند که باعث ایجاد سرطان پوست می‌شوند. یکی از متداول‌ترین و خطرناک‌ترین انواع سرطان پوست، ملانوما می‌باشد که با رشد بیش از اندازه در سلول‌های رنگ‌ساز پوست ایجاد می‌شود. این سرطان در دو نوع بدخیم و خوش‌خیم می‌باشد که نوع بدخیم آن، قابلیت انتشار به تمام ارگان‌های داخلی بدن را دارد و حتی منجر به مرگ می‌شود. اما در صورت تشخیص در مراحل اولیه، قابل درمان خواهد بود. از داروهای مورد استفاده برای کنترل و درمان این نوع سرطان می‌توان به داروی اینترفرون با نیمه عمر ده ساعت اشاره کرد [۱]. سعی و خطا در تعیین دوز بهینه دارو جزء جدایی‌ناپذیر برای کنترل سلول‌های سرطانی در بیماران مبتلا به سرطان، توسط پزشکان می‌باشد. این امر نه تنها زمان‌بر بوده بلکه اثرات زیان‌بار دارو برای بیمار را نیز به همراه خواهد داشت. همچنین ممکن است در بعضی از مواقع باعث مرگ وی شود. در استفاده از روش‌های هوشمند، این سعی و خطا در ابتدا بر روی مدل ریاضی نامی بیمار صورت می‌گیرد و در نهایت بهینه‌ترین میزان دوز دارو برای کنترل جمعیت سلول‌های سرطانی پیشنهاد می‌شود. سپس، با سعی و خطای بسیار محدودتر تعیین بهترین میزان دوز دارو برای بیمار واقعی انجام می‌شود. در زمینه کنترل و درمان سرطان ملانوما با استفاده از روش‌های هوشمند می‌توان به موارد محدود زیر اشاره کرد: مزدیاسنا و همکارانش[۱] [۲]، با استفاده از الگوریتم ژنتیک، بهترین میزان دوز دارو را در یک مدل غیرخطی از بیمار مبتلا به سرطان ملانومای B16F10 بدست آورده‌اند. این نوع سرطان در واقع یک رده سلولی تومور موش است، که در تحقیقات، به عنوان الگویی برای سرطان پوست مورد استفاده قرار می‌گیرد. تومورهای B16 مدل‌های مفید برای مطالعه متاستاز و تشکیل تومورهای جامد هستند [۳]. این الگوریتم در زمان یادگیری بسیار زمان‌بر می‌باشد و همچنین قابلیت تعیین میزان دوز دارو در حالت برخط را ندارد. این موارد از معایب روش الگوریتم ژنتیک محسوب می‌شود. وانگ و همکارانش[۲] [۴]، از تاثیر بازدارنده‌ها برای کاهش سلول‌های سرطانی استفاده کرده‌اند و نشان داده‌اند که بازدارنده mTOR[۳] باعث کاهش سلول‌های سرطانی شده است. این نوع بازدارنده همانطور که از اسم آن مشخص می‌باشد، باعث مهار فعالیت mTOR می‌شود. مسیر mTOR یک مسیر آنزیمی مربوط به پروتئین کیناز است که باعث تولید فاکتورهای رشد سلول می‌شود. پس این نوع بازدارنده از رشد سلول‌ها جلوگیری می‌کند. همچنین برای درمان برخی از سرطان‌ها نیز از mTOR به همراه شیمی‌درمانی استفاده می‌شود [۵]. سیرین و همکارانش[۴] [۶]، کنترل جمعیت سلول‌های سرطانی را با استفاده از روش یادگیری تقویتی مجموعه‌ای[۵] و با تنظیم مستقیم ژن‌ها انجام داده‌اند. در این مقاله تزریق دارو به‌گونه‌ای انجام گرفته است که باعث تغییر در ساختار ژن‌ها می‌شود. پاداش‌دهی نیز بر اساس نحوه قرارگیری ژن‌ها تعریف شده است. کلهر و همکارش [۷]، کنترل سلول‌های سرطانی را بیماران مبتلا به سرطان ملانوما با استفاده از الگوریتم ژنتیک انجام داده‌اند. در این مقاله، تابع برازندگی به صورت مجموع دوز داروی تزریقی و غلظت سلول‌های سرطانی تعریف شده است. دلیل این انتخاب کاهش هم‌زمان سلول‌های سرطانی و دوز داروی تزریقی برای کاهش اثرات زیان‌بار دارو می‌باشد.

اما از روش‌های هوشمند برای کنترل انواع سرطان‌ها و بیماری‌ها استفاده شده است. یکی از این روش‌ها الگوریتم یادگیری- Q[۶] بوده که از روش‌های حل مسئله یادگیری تقویتی می‌باشد این روش دارای خاصیت تطبیق‌پذیری با محیط در برابر نویز و اغتشاش می‌باشد. اما به‌روزرسانی ارزش جفت حالت و عمل در لحظه فعلی فقط با تاثیرپذیری از ارزش جفت حالت و عمل لحظه بعد انجام می‌گیرد. این در حالی می‌باشد که روش مسیرهای شایستگی علاوه بر مورد فوق، با برخورداری از مزیت نگاه به عقب، کل زنجیره حالت و عمل‌های لحظات قبل را در بروزرسانی ارزش جفت حالت و عمل لحظه فعلی در نظر می‌گیرد که این امر باعث بهبود سرعت یادگیری و افزایش دقت خواهد شد [۸]. از جمله مقالاتی که به کنترل جمعیت سلول‌های سرطانی و دیگر بیماری‌ها با استفاده از روش یادگیری تقویتی پرداخته‌اند، می‌توان به موارد زیر اشاره کرد.

نوری و همکارش [۹]، کنترل میزان گلوکز خون در بیماران مبتلا به دیابت توسط روش یادگیری تقویتی انجام داده‌اند. دی پولا و همکارانش[۷] [۱۰]، با ترکیب روش یادگیری تقویتی و توزیع گوسی، میزان گلوکز خون در سیستم دارای عدم قطعیت را کنترل کرده‌اند. در [۱۱،۱۲،۱۳]، با استفاده از الگوریتم یادگیری Q، سلول‌های سرطانی را در بیمار مبتلا به سرطان کنترل کرده‌اند. در [۱۲]، مقدار باقی‌مانده دارو نیز در هر لحظه در بدن بیمار محاسبه شده است. در کاری که پاداماناب هن و همکارانش[۸] [۱۳] انجام داده‌اند، نتایج شبیه‌سازی نشان داده است که با کاهش سلول‌های

سرطانی و حذف آن‌ها میزان دوز داروی تزریقی به بیمار نیز کاهش پیدا کرده است. این امر کاهش اثرات زیانبار دارو بر روی سلول‌های سالم را به همراه خواهد داشت. یکی از مزایای یادگیری تقویتی که به آن اشاره شد، تطبیق‌پذیری آن با محیط می‌باشد. در این مقاله برای نشان دادن این امر، عدم قطعیت در پارامترهای سیستم و شرایط اولیه لحاظ شده است و الگوریتم پیشنهادی برای سه بیمار دیگر توانسته است سلول‌های سرطانی را کنترل کرده و به صفر برساند. در این سه بیمار نیز میزان دوز داروی تزریقی با کاهش سلول‌های سرطانی کاهش پیدا کرده است. نوری و همکارانش [۱۴]، با استفاده از یادگیری تقویتی تعیین میزان بهینه دوز دارو در بیماران مبتلا به هپاتیت را انجام داده‌اند. در این مقاله پاداش‌دهی بر اساس میزان کاهش ویروس‌های آزاد انجام گرفته است. پیترسن و همکارانش[۹] [۱۵]، درمان بیماری سپسین[۱۰] توسط روش‌های هوشمند انجام داده‌اند. در این بیماری سیستم دفاعی بدن دچار اختلال شده و به بافت‌های بدن آسیب می‌زند. در این مقاله، از روش یادگیری عمیق برای افزایش ایمنی بدن در مقابله با این بیماری استفاده شده است.

از دیگر روش‌هایی که برای کنترل سرطان مورد استفاده قرار می‌گیرد، روش‌های کنترل کلاسیک (کنترل بهینه) می‌باشد. از معایب این روش‌ها می‌توان به وابستگی آن‌ها به مدل ریاضی و همچنین خطی‌سازی سیستم اشاره کرد. در خطی‌سازی سیستم، تقریبی از مدل ریاضی بدست می‌آید که در خطی‌سازی سیستم‌های پزشکی رفتار مدل خطی‌شده کاملا مشابه با رفتار سیستم اصلی نخواهد بود. این امر باعث عدم تطبیق‌پذیر بودن رفتار مدل با دینامیک بدن بیمار واقعی خواهد بود. در [۲۰-۱۶] کنترل سلول‌های سرطانی با استفاده از روش‌های کنترل کلاسیک (کنترل بهینه) انجام گرفته است. غلمان و همکارانش[۱۱] [۱۶]، از مدل ریاضی دارای تاخیر از بیمار مبتلا به سرطان استفاده کرده‌اند. نتایج شبیه‌سازی این مقاله نشان داده است که روش پیشنهادی توانسته است سلول‌های سرطانی را کاهش داده و از بین ببرد. مالینزی و همکارانش[۱۲] [۱۷] نیز، برای ارتقاء شیمی‌درمانی از یکی از روش‌های کنترل بهینه به نام"اصل حداکثری پونتریاگین" استفاده کرده‌اند. این روش سلول‌های سرطانی را کاهش داده است، اما قادر به حذف کامل آن‌ها نبوده است. هلن مور[۱۳] [۱۸]، کنترل سلول‌های سرطانی در بیماران مبتلا به سرطان خون را انجام داده است. این روش با روش تزریق دوز داروی ثابت مقایسه شده است. با توجه به نتایج شبیه‌سازی، روش کنترل بهینه با تزریق دوز داروی کمتر توانسته است سلول‌های سرطانی را کاهش دهد، اما هر دو روش قادر به حذف کامل آن‌ها نبوده‌اند. روچا و همکارانش[۱۴] [۱۹]، برای کاهش هم‌زمان سلول‌های سرطانی و میزان دوز داروی تزریقی، از روش کنترل بهینه چند هدفه استفاده کرده‌اند. تزریق دوز دارو تا زمانی که سلول‌های سرطانی از بین نرفته‌اند انجام می‌گیرد و با حذف کامل آن‌ها تزریق دارو متوقف شده است. خالوزاده و همکارانش [۲۰]، در بیماران مبتلا به سرطان پستان که در مرحله سوم این بیماری می‌باشند، رژیم دارویی بهینه را به‌گونه‌ای تنظیم کرده‌اند که اندازه تومور به اندازه مطلوب برای انجام عمل جراحی و برداشتن آن برسد.

در تمام مقالات ذکر شده تابع هزینه بر اساس میزان دوز داروی تزریقی و جمعیت سلول‌های سرطانی بیان شده است .از معایب این روش‌ها همان‌طور که بیان شد، می‌توان به وابستگی آن‌ها به مدل ریاضی اشاره کرد و تمامی این روش‌ها بر مبنای خطی‌سازی می‌باشند که در بعضی مواقع رفتار سیستم در این حالت با رفتار سیستم اصلی تفاوت خواهد داشت. بنابراین دوز داروی پیشنهاد شده توسط چنین روش‌هایی از قابلیت اطمینان بالا برخوردار نمی‌باشد. در جدول(۱) دسته بندی مقالات و خلاصه‌ای از کار هر مقاله ذکر شده است.

در کار گذشته [۷]، از روش الگوریتم ژنتیک برای کاهش جمعیت سلول‌های سرطانی استفاده شد. این روش برخط نبوده و زمان‌بر می‌باشد و همچنین قابلیت کنترل سلول‌های سرطانی در حضور عدم قطعیت و نویز را نیز ندارد. مدل ریاضی که در این مقاله برای بیان دینامیک بدن بیمار مبتلا به ملانوما استفاده شد، ناقص بوده و به‌طور کامل دینامیک بدن بیمار مبتلا به ملانوما را توصیف نمی‌کند. با توجه به موارد ذکر شده، در این مقاله با راهنمایی پزشک مربوطه از مدل ریاضی کامل‌تر که به طور دقیق رفتار تمام پارامترهای دخیل در سیستم دفاعی بدن بیمار مبتلا به ملانوما را توصیف می‌کند استفاده شده است. لازم به ذکر می‌باشد که تا به حال کنترل جمعیت سلول‌های سرطانی توسط چنین مدلی انجام نگرفته است. برای بهبود روش‌های ذکر شده در کاهش و کنترل جمعیت سلول‌های سرطانی و تعیین میزان بهینه دوز دارو نیز از روش مسیرهای شایستگی [۱۵] که یکی از روش‌های حل مسئله یادگیری تقویتی [۱۶] می‌باشد استفاده شده است. از مزایای این روش نسبت به دیگر روش‌های یادگیری تقویتی می‌توان به نگاه همزمان رو به جلو و رو به عقب این الگوریتم در به‌روزرسانی ارزش جفت حالت و عمل لحظه فعلی اشاره کرد. این روش مزایای دو روش مرسوم در یادگیری تقویتی به نام‌های یادگیری تفاوت گذرا و مونت کارلو را دارا می‌باشد در این روش برای بروزرسانی ارزش جفت حالت و عمل در لحظه فعلی، نه تنها ارزش حالت و عمل در

لحظه بعد تاثیرگذار است، بلکه ارزش کل زنجیره حالت و عمل‌های لحظات قبل نیز در بروزرسانی آن تاثیرگذار خواهد بود.

جدول ۱: مرور و دسته‌بندی مقالات

ردیف	نام نویسندگان	روش کنترلی	سال انتشار
۱	مزدیاسنا و همکارانش[۲]	استفاده از الگوریتم ژنتیک برای کنترل سرطان ملانوما	۲۰۱۵
۲	وانگ و همکارانش[۴]	استفاده از بازدارنده mTOR برای کاهش سلول‌های سرطانی در بیمار مبتلا به ملانوما	۲۰۱۴
۳	سیرین و همکارانش[۶]	کنترل جمعیت سلول‌های سرطانی با استفاده از روش یادگیری تقویتی مجموعه‌ای و با تنظیم مستقیم ژن‌ها	۲۰۱۳
۴	کلهر و همکارش[۷]	کنترل سلول‌های سرطانی و دوز دارو در بیماران مبتلا به ملانوما با استفاده از الگوریتم ژنتیک	۲۰۱۷
۵	نوری و همکارش[۸]	کنترل میزان گلوکز خون در بیماران مبتلا به دیابت توسط روش یادگیری تقویتی	۲۰۱۷
۶	دی پولا و همکارانش[۹]	استفاده از ترکیب روش‌های یادگیری تقویتی و توزیع گوسی برای کنترل میزان گلوکز خون در سیستم دارای عدم قطعیت	۲۰۱۵
۷	سی زی بولا و همکارانش [۱۰]	استفاده از الگوریتم یادگیری Q برای کنترل سلول‌های سرطانی در بیمار مبتلا به سرطان	۲۰۱۳
۸	جکوبس [۱۱]	استفاده از الگوریتم یادگیری Q برای کنترل سلول‌های سرطانی در بیمار مبتلا به سرطان و مقدار باقی‌مانده دارو در هر لحظه در بدن بیمار	۲۰۱۴
۹	پاداماناب هن و همکارانش [۱۲]	کنترل سلول‌های سرطانی با استفاده از الگوریتم یادگیری Q و اعمال عدم قطعیت	۲۰۱۷
۱۰	نوری و همکارانش [۱۳]	تعیین میزان بهینه دوز دارو در بیماران مبتلا به هپاتیت با استفاده از روش یادگیری تقویتی	۲۰۱۱
۱۱	پیترسن و همکارانش [۱۴]	درمان بیماری سپسین توسط روش‌های هوشمند	۲۰۱۸
۱۲	غلمان و همکارانش [۱۵]	استفاده از روش کنترل بهینه برای کنترل جمعیت سلول‌های سرطانی در مدل ریاضی دارای تاخیر از بیمار مبتلا به سرطان	۲۰۱۸
۱۳	مالینزی و همکارانش [۱۶]	استفاده از روش "اصل حداکثری پونتریاگین" برای ارتقاء شیمی درمانی و کاهش سلول‌های سرطانی	۲۰۱۸
۱۴	هلن مور[۱۷]	کنترل سرطان خون با استفاده از روش کنترل بهینه	۲۰۱۸
۱۵	روچا و همکارانش[۱۸]	استفاده از روش کنترل بهینه چند هدفه برای کاهش هم‌زمان سلول‌های سرطانی و میزان دوز داروی تزریقی	۲۰۱۸
۱۶	خالوزاده و همکارانش [۱۹]	تنظیم رژیم دارویی بهینه در بیماران مبتلا به سرطان پستان جهت رسیدن تومور به اندازه مطلوب برای انجام عمل جراحی	۲۰۰۸

این روش با دارا بودن خاصیت تطبیق‌پذیری، در حضور عدم قطعیت و نویز قابلیت کنترل جمعیت سلول‌های سرطانی را دارا می‌باشد. با وجود اینکه در این مقاله از مدل ریاضی دارای تاخیر از بیمار مبتلا به سرطان ملانوما استفاده شده است، اما روش یادگیری تقویتی بر خلاف روش‌های کنترل کلاسیک، بی‌نیاز به مدل ریاضی می‌باشد. اما لازم به ذکر است که در این مقاله به دلیل عدم دسترسی به بیمار واقعی با راهنمایی پزشک مربوطه از مدل ریاضی غیرخطی دارای تاخیر از بیمار مبتلا به ملانوما استفاده شده است. با توجه به بررسی‌هایی که تاکنون گرفته است، لازم به ذکر می‌باشد که کنترل جمعیت سلول‌های سرطانی توسط این مدل با هیچ کدام از روش‌های هوشمند توسط گذشتگان انجام نگرفته است.

بنابراین نوآوری‌های اصلی این مقاله، در نظر گرفتن مدل ریاضی دارای تاخیر از بیمار مبتلا به سرطان ملانوما می‌باشد. لازم به ذکر می‌باشد که این مدل ریاضی با راهنمایی پزشک مربوطه انتخاب شده است. این مدل ریاضی نسبت به دیگر مدل‌های ریاضی ملانوما کامل‌تر می‌باشد و به طور دقیق رفتار تمام پارامترهای دخیل در سیستم دفاعی بدن بیمار مبتلا به ملانوما را توصیف می‌کند. برای کنترل جمعیت سلول‌های سرطانی از روش مسیرهای شایستگی استفاده شده است. این روش یکی از بهترین و کارامدترین روش‌های حل مسئله در یادگیری تقویتی می‌باشد. در این روش همانطور که بیان شد، بروز رسانی ارزش جفت حالت و عمل در هر لحظه علاوه بر اینکه به ارزش حالت و عمل لحظه بعد وابسته می‌باشد، ارزش کل حالت و عمل‌های لحظات قبل نیز در بروز رسانی آن تاثیرگذار خواهد بود. این امر باعث افزایش سرعت یادگیری خواهد شد. برای نشان دادن مزیت و برتری این روش نسبت به دیگر روش‌هایی که جهت کنترل جمعیت سلول‌های سرطانی در بیمار مبتلا به ملانوما استفاده شده است، این روش با یکی دیگر از روش‌های یادگیری تقویتی به نام الگوریتم یادگیری Q و روش کنترل بهینه مقایسه شده است. روش تزریق دوز داروی ثابت نیز از دیگر روش‌هایی می‌باشد که جهت مقایسه با روش پیشنهادی استفاده شده است.

از کاربردهای مهم روش ذکر شده می‌توان به استفاده از آن در تعیین میزان دوز دارو برای هر بیمار به طور مجزا اشاره کرد. این امر مفهوم drug personalization می‌باشد و اخیرا در مسائل پزشکی بسیار مورد توجه قرار گرفته است. این امر توسط روش یادگیری تقویتی به این صورت انجام می‌گیرد که با استفاده از جدول Q بدست امده یادگیری برای هر بیمار مجدد انجام می‌گیرد

و با سعی و خطای کمتر دوز بهینه دارو برای هر یک تعیین می‌گردد.

بخش‌های بعدی مقاله به شرح زیر می‌باشد. در بخش دوم مقاله، مدل ریاضی بیمار مبتلا به سرطان ملانوما بیان شده است. در بخش سوم، در رابطه با مسئله یادگیری تقویتی و مسیرهای شایستگی به عنوان یک روش برای کنترل سلول‌های سرطانی توضیحاتی مطرح شده است و در بخش چهارم نیز شبیه‌سازی و نتایج حاصل از آن ارایه شده است. در این بخش برای نشان دادن مزیت روش انتخابی، این روش با روش‌های کنترل بهینه و الگوریتم یادگیری Q مقایسه شده است. در این بخش همچنین با اعمال عیب به سنسور سیستم، عملکرد کنترلر پیشنهادی در کاهش سلول‌های سرطانی در حضور عیب مورد بررسی قرار گرفته است. همچنین برای نشان دادن یکی از مزایای روش یادگیری تقویتی که تطبیق‌پذیری آن با محیط می‌باشد، عدم قطعیت در پارامترهای سیستم و شرایط اولیه اعمال گردیده است و کنترل جمعیت سلول‌های سرطانی در پنج بیمار مبتلا به سرطان ملانوما انجام گرفته است. در این بخش با افزودن نویز به متغیرهای سیستم نیز، نشان داده شده است که روش انتخابی باز هم قادر به کنترل سلول‌های سرطانی در بیماران مبتلا به سرطان ملانوما بوده است. از دیگر مواردی که در این بخش مورد بررسی قرار گرفت، مقایسه سرعت همگرایی دو روش مسیرهای شایستگی و الگوریتم یادگیری Q می‌باشد.

۲- مدل ریاضی

مدل‌های ریاضی زیادی برای بیان دینامیک بیماری سرطان ملانوما وجود دارد که در مراجع [۲۶-۲۱] توضیحات کامل مربوط به هر یک از آن‌ها آورده شده است. مدل ریاضی که در این مقاله استفاده شده است، توسط مارزیو پنیسی[۱۷] ارایه شده است [۲۷]. یکی از مزایای این مدل ریاضی در نظرگرفتن یک معادله دیفرانسیل جدا برای نمایش رفتار سلول‌های سرطانی می‌باشد. از دیگر مزایای این مدل می‌توان به درنظر گرفتن تعداد متغیرهای بیشتر اشاره کرد. این امر باعث نزدیکی هر چه بیشتر مدل ریاضی به دینامیک بدن بیمار واقعی خواهد شد. این مزیت در دیگر مدل‌های ریاضی از بیمار مبتلا به سرطان ملانوما دیده نشده است. متغیرهای درنظر گرفته شده در این مدل از موارد مهمی می‌باشند که در سیستم دفاعی بدن بیمار مبتلا به سرطان ملانوما بسیار تاثیرگذار هستند. در این مدل ریاضی برای توصیف دینامیک بدن بیمار از دو محفظه استفاده شده است که نام آن‌ها: محفظه نقطه تزریق[۱۸] و محفظه‌ی پوست[۱۹] می‌باشد.

معادلات بیان‌کننده دینامیک بیمار مبتلا به ملانوما در روابط(۱) تا(۷) بیان شده‌اند:

$$\frac{dE}{dt} = K_{in}(t,p) - \alpha_{11}E - \alpha_8 E \quad (۱)$$

$$\frac{dAb}{d} = K_{in}(t,q) - \alpha_{11}Ab - \alpha_{10}Ab \quad (۲)$$

$$\frac{dE_s}{dt} = \alpha_7\left[\frac{A_s}{A_s + K_1}\right]E_s + \alpha_{11}E(t-\tau) + \alpha_6 NA + \alpha_8 E_s \quad (۳)$$

$$\frac{dC}{dt} = \left(\alpha_1 - \alpha_2 Ln(C)\right).C - \alpha_3\left[\frac{A_s + K_2}{A_s + K_3}\right]E_s C \quad (۴)$$

$$\frac{dA}{dt} = \alpha_4\left[\alpha_3\left[\frac{A_s + K_2}{A_s + K_3}\right]E_s C\right] - \alpha_5 A - \alpha_6 NA \quad (۵)$$

$$\frac{dN}{dt} = h(M - N) - \alpha_6 NA \quad (۶)$$

$$\frac{dA_s}{dt} = \alpha_{11}Ab(t-\tau) - \alpha_9 A_s E_s - \alpha_{10}A_s \quad (۷)$$

معرفی هر یک از متغیرهای حالت در جدول(۲) بیان شده است.

جدول۲: معرفی متغیرهای حالت سیستم[۲۷]

نام متغیرحالت	معرفی متغیر	واحد
E	میزان CTL (OT1 Cytotoxic T Lymphocytes) های فعال در محفظه‌ی نقطه‌ی تزریق	$Cell/_{Kg}$
Ab	میزان آنتی بادی CD137 تزریق شده در محفظه‌ی نقطه‌ی تزریق	$mg/_{ml}$
E_s	میزان OT1CTLهای فعال در محفظه پوست	$Cell/_{Kg}$
C	جمعیت سلول‌های ملانوما	mm^2
A	میزان آنتی‌ژن‌ها	$\mu g\ of\ protein$
N	میزان CTLهای ساده	$cell/_{days}$
A_s	میزان آنتی‌بادی‌ها در محفظه‌ی پوست	$mg/_{ml}$

بر اساس جدول فوق، CTLها سلول‌های ایمنی بدن بوده و به عنوان عامل کشنده برای از بین بردن سلول‌های سرطانی می‌باشند. CTLهای ساده در بدن هر فردی وجود دارند و به محض ورود سلول‌های سرطانی فعال شده و خاصیت کشندگی پیدا می‌کنند و تبدیل به CTLهای فعال می‌شوند. در روابط (۱) تا (۷)، معادله‌ی (۱) بیانگر OT1CTLهای فعال در محفظه‌ی نقطه‌ی تزریق می‌باشد و معادله‌ی (۲) نیز نمایشگر تغییرات آنتی‌بادی‌ها در این محفظه می‌باشد. تغییرات OT1CTLهای فعال در محفظه پوست نیز با معادله‌ی(۳) نمایش داده شده است. در این رابطه، تغییرات CTLها در بیمار مبتلا به سرطان ملانوما در هر لحظه، به تغییراتOT1CTLها در چند لحظه قبل و در محفظه نقطه تزریق وابسته می‌باشد، که بخش دوم این معادله بیان‌کننده‌ی همین موضوع می‌باشد [۲۷]. رابطه‌ی (۴) تغییرات جمعیت سلول‌های سرطانی ملانوما را نشان می‌دهد. در معادله‌ی(۵)، روند تغییرات

آنتی‌ژن‌ها نشان داده شده است. معادله‌ی(۶) نمایشگر تغییرات CTLهای ساده بوده و معادله‌ی(۷) نیز تغییرات آنتی‌بادی‌ها در محفظه‌ی پوست را نشان می‌دهد. در این معادله نیز میزان آنتی‌بادی‌ها در لحظه‌ی فعلی به تغییرات آن‌ها در چند لحظه‌ی قبل وابسته می‌باشد [۲۷]. در این مدل ریاضی، برای کاهش جمعیت سلول‌های سرطانی دو نوع دارو در نظر گرفته شده است. نوع اول آن که در معادله (۱) نشان داده شده است، تزریق CTL می‌باشد. نوع دوم آن تزریق یک نوع آنتی‌بادی به نام CD-137 می‌باشد که در معادله (۳) نشان داده شده است. مقدار پارامترهای ثابت سیستم در جدول (۳) آورده شده‌اند.

۳- کنترل جمعیت سلول‌های سرطانی با استفاده از یادگیری تقویتی

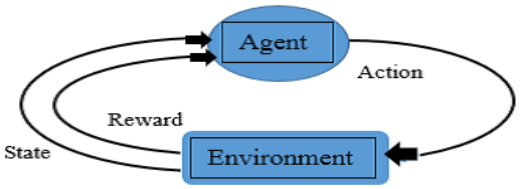
روش استفاده شده در این مقاله برای از بین بردن جمعیت سلول‌های سرطانی، یکی از کارآمدترین روش‌های حل مسئله یادگیری تقویتی (RL)، به نام روش مسیرهای شایستگی می‌باشد.

جدول۳: مقادیر در نظر گرفته‌شده برای پارامترهای ثابت سیستم[۲۷]

پارامتر	معرفی پارامتر	مقدار
α_1	نرخ رشد CTLها	0.0065
α_2	نرخ کاهش CTLها	0.00038
α_3	ماکزیمم نرخ مرگ به وسیله OT1CTLهای فعال	1×10^{-8}
α_4	نرخ حذف آنتی‌بادی‌ها در محفظه‌ی پوست به وسیله‌ی CTLها	0.0063
α_5	نرخ از هم پاشیده شدن آنتی‌بادی‌ها	0.0023
α_6	نرخ تغییر CTLها به OT1CTLهای فعال در محفظه‌ی پوست	0.0030
α_7	حداکثر نرخ تکرار OT1CTLهای فعال	0.0028
α_8	نرخ مرگ طبیعی OT1CTLهای فعال در محفظه‌ی پوست و نقطه تزریق	0.0014
α_9	نرخ مرگ آنتی‌بادی‌ها در محفظه‌ی تزریق	3.03×10^{-8}
α_{10}	سرعت از بین‌رفتن آنتی‌بادی در دو محفظه	0.001
α_{11}	سرعت حرکت از محفظه‌ی تزریق به محفظه‌ی پوست	0.00027
h	نرخ تزریق دوباره از CTLها به وسیله غده تیموس	0.0003
K_1	آستانه تکرار آنتی‌بادی‌ها	10
K_2	مینیمم آستانه نرخ نابودی سلول‌های سرطانی	1
K_3	ماکزیمم آستانه نرخ نابودی سلول‌های سرطانی	50
M	تعدادCTLها در موقعیت‌های بی‌خطر	196000
C_0	مقدار اولیه سلول‌های سرطانی	180000
P	مقدار داروی تزریق شده بهOT1CTL های فعال در محفظه‌ی تزریق	760000
Q	مقدار داروی تزریق شده به آنتی‌بادی در محفظه‌ی تزریق	$1\times10^{+6}$
τ	تاخیر	1

۳-۱- یادگیری تقویتی (RL)

در حل مسئله با استفاده از روش یادگیری تقویتی عامل و محیط نقش بسیار مهمی را ایفا می‌کنند. عامل با جستجوی فراوان در محیط اطلاعات موجود را در هر لحظه دریافت می‌کند و بر اساس آن‌ها اطلاعات خود را بروز می‌کند. این روش بین روش با ناظر و بدون ناظر قرار دارد. بدین معنا که عامل با سعی و خطای فراوان در محیط و دریافت پاداش و جریمه از آن طی تکرار بالا سعی می‌کند بهترین عمل را در هر لحظه انتخاب کند و به سیاست بهینه برسد [۸]. در هر لحظه حالت‌های سیستم (s_t) بدست می‌آید و عامل بر اساس آن‌ها بهترین عمل (a_t) را انتخاب می‌کند تا به خروجی مطلوب برسد. عمل انتخابی توسط عامل بر حالت‌های سیستم تاثیر گذاشته و حالت‌های بعدی (s_{t+1}) آن بدست می‌آید. بر اساس این حالت‌ها عامل از محیط پاداش (r_{t+1}) را دریافت می‌کند [۲۷]. عمکرد بین عامل و محیط در شکل (۱) نشان داده شده است.



شکل۱: عملکرد بین عامل و محیط

رابطه‌ی (۸)، تخمین تابع ارزش تحت سیاست π را نشان می‌دهدکه با V^{π} نمایش داده می‌شود [۸].

$$V^{\pi}(s)=E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+1}|s_t=s\right\} \qquad (۸)$$

در رابطه‌ی فوق، γ نرخ فراموشی و بین صفر تا یک انتخاب می‌شود. k گام زمانی است. در این رابطه، ارزش هر حالت تحت سیاست π، بر اساس محاسبه امید ریاضی برای مجموع حاصل ضرب نرخ فراموشی و مقدار پاداش‌های دریافتی از لحظه فعلی (s_t) تا انتهای مسیر می‌باشد که با توجه به رابطه (۸)، تاثیر پاداش‌های دریافتی در ارزش حالت فعلی (s_t) در هر گام زمانی با ضرب شدن در γ^k خواهد بود. به این صورت خواهد بود که در ابتدا k برابر یک می‌باشد و میزان پاداش دریافتی در لحظه $k+1$ (لحظه بعد)، در γ ضرب خواهد شد. سپس k برابر با دو شده و میزان پاداش دریافتی در لحظه بعد نیز در γ^2 ضرب می‌شود. تاثیر پاداش دریافتی در این لحظه کمتر از لحظه قبل می‌باشد. این روند تا انتهای مسیر و تاثیر دادن تمام پاداش‌های دریافتی ادامه پیدا می‌کند و در هر لحظه تاثیر پاداش دریافتی بر روی ارزش حالت و عمل فعلی کمتر خواهد شد. رابطه (۹) بروزرسانی ارزش حالت و عمل تحت سیاست π را در هر گام زمانی بیان می‌کند.

$$Q^{\pi}(s,a) = E_{\pi}\{R_t|s_t = s, a_t = a\} = E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s, a_t = a\right\} \quad (۹)$$

همان‌طور که برای رابطه (۸) بیان شد، در رابطه فوق نیز γ نرخ فراموشی و k گام زمانی می‌باشد. اما در این رابطه ارزش جفت حالت و عمل تحت سیاست π محاسبه می‌گردد. بدین صورت که ابتدا مجموع حاصل ضرب نرخ فراموشی و مقدار پاداش‌های دریافتی از لحظه فعلی (s_t) تا انتهای مسیر بدست می‌آید و سپس امید ریاضی برای آن محاسبه می‌گردد. در این رابطه پاداش‌های دریافتی تاثیر خود را در ارزش جفت حالت و عمل لحظه فعلی (s_t, a_t) در هر گام زمانی با ضرب شدن در γ^k می‌گذارند. . این روند نیز همانند تعریف بالا می‌باشد و در هر لحظه تاثیر پاداش دریافتی بر روی ارزش حالت و عمل فعلی کمتر خواهد شد. رابطه (۱۰)، بیان‌کننده معادله بلمن در بروزرسانی ارزش حالت در هر لحظه می‌باشد [۸].

$$V^{\pi}(s) = \sum_{a}\pi(s,a)\sum_{s'}P^{a}_{ss'}\left[R^{a}_{ss'} + \gamma V^{\pi}(s')\right] \quad (۱۰)$$

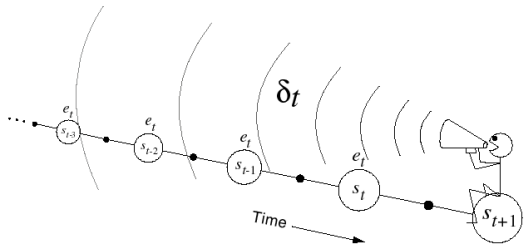
در رابطه فوق، $\pi(s,a)$، سیاست مورد نظر برای انتخاب عمل a است. $P^{a}_{ss'}$، احتمال رفتن از حالت فعلی (s) به حالت بعدی (s') با انجام عمل a می‌باشد. $R^{a}_{ss'}$، کل پاداش‌های دریافتی در زمان رفتن از حالت فعلی (s) به حالت بعدی (s') با انجام عمل a است. γ نرخ فراموشی بوده و بین صفر تا یک انتخاب می‌شود. $V^{\pi}(s')$، ارزش حالت بعدی (s') تحت سیاست π می‌باشد. برای بروزرسانی ارزش جفت حالت و عمل توسط معادله بلمن از رابطه (۱۱) استفاده می‌شود. در این رابطه بهترین ارزش برای جفت حالت و عمل (s,a) بدست خواهد آمد [۸].

$$Q^{*}(s,a) = \sum_{s'}P^{a}_{ss'}\left[R^{a}_{ss'} + \gamma \max_{a'}Q^{*}(s',a')\right] \quad (۱۱)$$

در رابطه فوق، $\max_{a'}Q^{*}(s',a')$ در نظر گرفتن بهترین ارزش برای جفت حالت و عمل لحظه بعد (s',a') می‌باشد.

۳-۲- مسیرهای شایستگی

روش مسیرهای شایستگی تعاملی بین دو روش دیگر حل مسئله‌ی یادگیری تقویتی به نام‌های مونت کارلو[۲۰] (MC) و یادگیری تفاوت‌گذرا[۲۱] (TD) می‌باشد. در این روش، بروزرسانی ارزش جفت حالت و عمل در هر گام انجام می‌گیرد و بدین صورت خواهد بود که برای به روزرسانی ارزش حالت و عمل لحظه فعلی (s_t) علاوه بر نگاه رو به جلو، نگاه رو به عقب (نگاه دوسویه) نیز وجود خواهد داشت. نگاه دوسویه بدین معنا می‌باشد که برای بروزرسانی ارزش جفت حالت و عمل در لحظه فعلی، نه تنها ارزش حالت و عمل در لحظه بعد تاثیرگذار است، بلکه ارزش کل زنجیره حالت و عمل‌های لحظات قبلی نیز در بروزرسانی آن تاثیرگذار خواهد بود. این امر موجب افزایش سرعت در کاهش جمعیت سلول‌های سرطانی با تزریق بهینه‌ترین میزان دوز دارو خواهد شد که کاهش اثرات زیان‌بار دارو را به همراه دارد [۸]. نگاه به عقب در روش مسیرهای شایستگی در شکل (۲) قابل نمایش می‌باشد.



شکل۲: نگاه به عقب در روش مسیرهای شایستگی [۲۷]

شایستگی زوج حالت و عمل توسط رابطه‌ی (۱۲) بدست می‌آید.

$$e_t(s,a) = e_t(s,a)+1 \quad (۱۲)$$

برای بروزرسانی ارزش جفت حالت و عمل از رابطه‌ی (۱۳) استفاده می‌شود [۸].

$$Q(s,a) = Q(s,a) + \alpha\delta e(s,a) \quad (۱۳)$$

در رابطه فوق، α نرخ یادگیری بوده و مقداری بین صفر و یک انتخاب می‌شود. δ نیز طبق رابطه‌ی (۱۴) محاسبه می‌شود.

$$\delta = r + \gamma Q(s',a') - Q(s,a) \quad (۱۴)$$

در رابطه‌ی (۱۴)، r میزان پاداش دریافتی و $Q(s',a')$ ارزش جفت حالت و عمل در لحظه بعدی $(t+1)$ می‌باشد. و α نیز نرخ یادگیری می‌باشد. بروزرسانی میزان شایستگی نیز، توسط رابطه (۱۵) محاسبه می‌شود [۸].

$$e_t(s,a) = \gamma\lambda \times e_t(s,a) \quad (۱۵)$$

در رابطه فوق، γ نرخ فراموشی می‌باشد. λ، پاداش‌ها را از گام فعلی تا انتهای مسیر وزن‌دهی می‌کند و مقدار آن ثابت و بین صفر و یک انتخاب می‌شود. طبق این رابطه، بروزرسانی میزان شایستگی زوج حالت و عمل برای حالت‌های قبلی و حالت فعلی، با بودن در هر حالت (s_t) در مقدار $\gamma\lambda$ ضرب خواهد شد [۸].

لازم به ذکر می‌باشد که تمامی پارامترها در روابط فوق با سعی و خطا بدست آمده‌اند و سعی شده است که بهترین مقدار برای آن‌ها انتخاب شود. فرایند یادگیری در RL بسیار زمان‌بر می‌باشد و ابعاد مسئله بزرگ می‌باشد. در نظر گرفتن فرایند بهینه‌سازی در هر گام زمانی برای پیدا کردن این پارامترها بسیار وقت‌گیر خواهد بود و حجم محاسباتی بالایی دارد. اما برای تعیین نرخ یادگیری (α) در بخش‌های بعدی مقاله صحبت شده است.

۳-۳- تطبیق روش مسیرهای شایستگی با درمان سرطان ملانوما

در این مقاله هدف تعیین میزان بهینه دوز دارو برای کاهش جمعیت سلول‌های سرطانی می‌باشد، به‌گونه‌ای که از اثرات زیان‌بار آن بر روی سلول‌های سالم جلوگیری شود. در این مسئله، محیط، مدل غیرخطی دارای تاخیر از یک بیمار مبتلا به سرطان ملانوما می‌باشد. عامل با جستجو در محیط و دریافت پاداش و جریمه یاد می‌گیرد، بهترین عمل را انتخاب کرده که بهترین میزان دوز دارو را برای کاهش جمعیت سلول‌های سرطانی و کاهش اثرات زیان‌بار دارو می‌باشد. برای این منظور از روش مسیرهای شایستگی در مسئله یادگیری تقویتی استفاده شده است.

۳-۳-۱- تعریف پارامترهای یادگیری تقویتی

حالت در نظر گرفته شده در این مقاله، جمعیت سلول‌های سرطانی می‌باشد که به صورت گسسته در ۲۰۰ بازه تعریف شده‌اند. بازه‌بندی از ۵۰۰۰ تا ۱۰۰۰۰۰۰ در نظر گرفته شده است. میزان داروهای تزریقی $(K_{in}(t,q), K_{in}(t,p))$ نیز به عنوان عمل در مسئله‌ی یادگیری تقویتی می‌باشند و هر دو بین بازه ۵۷۶۰۰۰ تا ۱۰۰۰۰۰۰ تقسیم‌بندی شده‌اند. عامل یاد می‌گیرد در هر لحظه بهترین میزان دوز دارو را تعیین کند تا علاوه بر کاهش جمعیت سلول‌های سرطانی، از تاثیرات زیانبار دارو بر روی سلول‌های ایمنی بدن نیز جلوگیری شود. تعداد اعمال انتخابی توسط عامل ۱۶۱ عدد می‌باشد. سیاست در نظر گرفته شده برای انتخاب اعمال $\varepsilon - greedy$ می‌باشد که بر اساس رابطه (۱۶) محاسبه می‌شود [۸].

$$a_t = \begin{cases} a_t^* & with\ probability\ 1-\varepsilon \\ random\ action\ with\ probability\ \varepsilon \end{cases} \quad (۱۶)$$

طبق رابطه بالا، بهترین عمل a_t^* با احتمال $\varepsilon - ۱$ و بقیه اعمال با احتمال یکسان ε انتخاب می‌شوند. تابع پاداش در مسئله یادگیری تقویتی با توجه به هدف مورد نظر در هر مسئله تعیین می‌گردد و برخلاف روش‌های کنترل بهینه از هیچ رابطه ریاضی خاصی تبعیت نمی‌کند. در این مقاله با توجه به اینکه هدف کاهش جمعیت سلول‌های سرطانی و افزایش ایمنی بدن با تزریق بهترین میزان دوز دارو، می‌باشد، تابع پاداش بر اساس تغییرات میزان سلول‌های سرطانی بدن و همچنین تغییرات هر دو نوع دارو در نظر گرفته شده است. با این تعریف برای تابع پاداش، از اثرات زیان‌بار دارو بر روی سلول‌های سالم نیز جلوگیری خواهد شد. تابع پاداش به صورت رابطه (۱۷) تعریف شده است [۲۸].

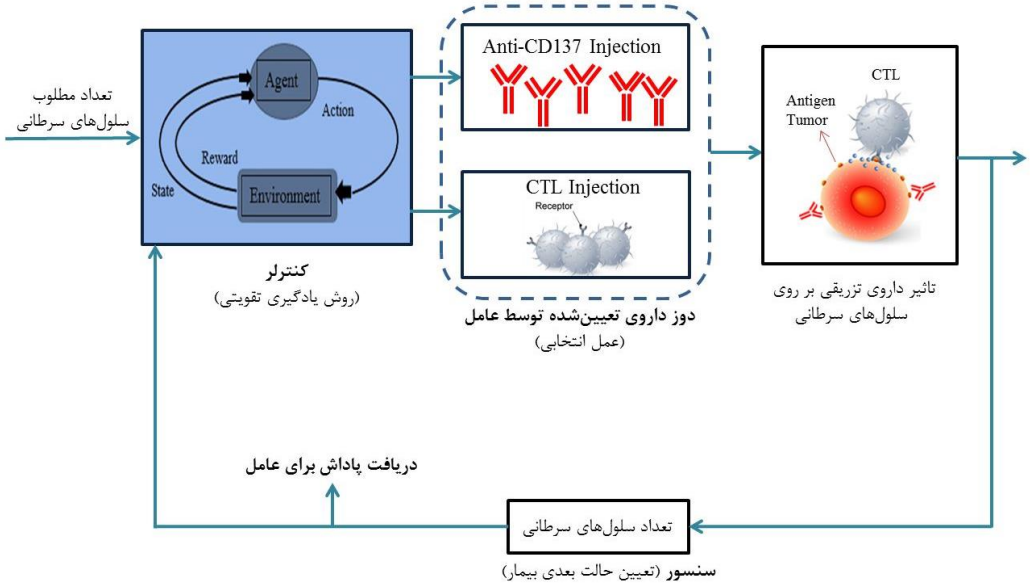
$$Reward = -\log\left(\frac{C_{new}}{C_{old}}\right) - (dose1_{new} - dose1_{old}) - (dose2_{new} - dose2_{old}) \quad (۱۷)$$

در رابطه (۱۷)، C_{old} جمعیت سلول‌های سرطانی بدن در لحظه‌ی قبل، C_{new} جمعیت سلول‌های سرطانی بدن در لحظه‌ی فعلی، $(dose1_{new} - dose1_{old})$، بیان‌کننده میزان اختلاف داروی تزریقی نوع اول در لحظه فعلی و لحظه قبل می‌باشد. $(dose2_{new} - dose2_{old})$ بیان‌کننده میزان اختلاف داروی تزریقی نوع دوم در لحظه فعلی و لحظه قبل می‌باشد. بدین ترتیب پاداش دریافتی مثبت خواهد بود. روند عملکرد پارامترهای یادگیری تقویتی در کنترل جمعیت سلول‌های سرطانی در شکل(۳) قابل نمایش می‌باشد.

با توجه به شکل (۳) و همان‌طور که ذکر شد، پارامترهای یادگیری تقویتی حالت، عمل و پاداش دریافتی می‌باشند. حالت در این مقاله، حالت‌های مختلف بیمار برای مقادیر مختلف سلول‌های سرطانی در هر لحظه می‌باشد. عملی که عامل در هر لحظه انتخاب می‌کند، میزان دوز داروی تزریقی می‌باشد و همان‌طور که بیان شد، در این مقاله دو نوع داروی تزریقی برای کاهش جمعیت سلول‌های سرطانی در نظر گرفته شده است. روند کاهش سلول‌های سرطانی با توجه به آنچه که در شکل(۳) نشان داده شده است، بدین صورت می‌باشد که ابتدا در مرحله یادگیری، جدولی تحت عنوان "جدول ارزش‌گذاری بر جفت حالت و عمل(جدولQ)" تعریف می‌گردد. این جدول با مقدار صفر مقداردهی اولیه می‌شود. مقادیر این جدول در روند اجرای الگوریتم توسط رابطه(۱۱) بروزرسانی می‌گردند. عامل در هر لحظه با تاثیرپذیری از این جدول عمل مورد نظر که میزان دوز داروی تزریقی می‌باشد را انتخاب می‌کند. این دارو به مدل غیرخطی از بیمار مبتلا به سرطان ملانوما اعمال می‌گردد. سپس حالت بعدی بیمار(جمعیت سلول‌های سرطانی) محاسبه می‌گردد. در این گام عامل به خاطر عمل انتخابی که چه میزان باعث کاهش یا افزایش سلول‌های سرطانی شده است، از محیط پاداش دریافت می‌کند. در این مقاله همانطور که بیان شد و طبق رابطه(۱۷) پاداش دریافتی بر اساس تغییرات سلول‌های سرطانی و میزان دوز داروی تزریقی در نظر گرفته شده است. این انتخاب به این دلیل می‌باشد که علاوه بر کاهش سلول‌های سرطانی، میزان دوز داروی تزریقی نیز کاهش پیدا کند تا از اثرات زیان‌بار دارو بر روی سلول‌های سالم جلوگیری گردد. این اعمال در چندین گام زمانی تکرار می‌گردند. در هر گام بروزرسانی جدول Q بر عمل انتخابی توسط عامل تاثیر گذاشته و باعث بهبود آن می‌گردد.

۴- شبیه‌سازی

در این بخش برای اینکه عملکرد روش پیشنهادی در افزایش سرعت کاهش سلول‌های سرطانی مورد بررسی قرار گیرد، از یکی از روش‌های کنترل کلاسیک(کنترل بهینه) و یکی از الگوریتم‌های روش یادگیری تقویتی به نام یادگیری Q جهت مقایسه با روش



شکل ۳: روند عملکرد پارامترهای یادگیری تقویتی در کنترل جمعیت سلول‌های سرطان

انتخابی استفاده شده است.

روش الگوریتم یادگیری Q یکی از الگوریتم‌های مرسوم در روش یادگیری تفاوت گذرا می‌باشد که در آن به‌روزرسانی ارزش جفت حالت و عمل در همان لحظه انجام می‌شود و نیازی به اتمام فرایند جهت ارزش‌گذاری زوج‌های حالت و عمل نمی‌باشد. به‌روزرسانی ارزش جفت حالت و عمل در هر لحظه (s_t, a_t) طبق معادله (۱۸) محاسبه می‌شود [۸].

$$Q(s_t,a_t) = Q(s_t,a_t) + \alpha[r_{t+1} + \gamma maxQ(s_{t+1},a_t) - Q(s_t,a_t)] \quad (۱۸)$$

در رابطه (۱۸)، α نرخ یادگیری و γ که نرخ فراموشی می‌باشد. لازم به ذکر می‌باشد تابع پاداش، نحوه انتخاب اعمال و پارامترهای مربوط به نرخ‌های یادگیری و فراموشی در هر دو روش مسیرهای شایستگی و الگوریتم یادگیری Q یکسان درنظر گرفته شده‌اند.

در بخش کنترل بهینه از روش "رگولاتور درجه خطی[۲۲]" با نام اختصاری LQR برای مقایسه با روش انتخابی استفاده شده است. این روش برای سیستم‌های خطی مورد استفاده قرار می‌گیرد. هدف اصلی در این روش یافتن ورودی کنترلی u^* است که با اعمال آن به سیستم خطی تعریف شده در رابطه (۱۹)، علاوه بر اینکه سیستم پایدار می‌شود و قیود تعریف شده برای آن برقرار می‌گردد، تابع هزینه تعریف شده نیز می‌بایست مینیمم گردد. در این روش متغیرهای حالت نیز با کمترین تلاش کنترلی به سمت صفر همگرا می‌شوند[۲۹].

$$\dot{x}(t) = A(x)x(t) + B(x)u(t) \quad (۱۹)$$

در رابطه فوق، $\dot{x}(t)$ بردار حالت، $u(t)$ بردار ورودی و A, B ماتریس‌های فضای حالت می‌باشند.

در این روش تابع هزینه بر اساس رابطه(۲۰) تعریف می‌گردد[۲۹].

$$J(x,u,t) = \frac{1}{2}x^T(t_f)Hx(t_f) + \int_{t_0}^{t_f}\left\{\frac{1}{2}x^T(t)Qx(t) + \frac{1}{2}u^T(t)Ru(t)\right\}dt \quad (۲۰)$$

در رابطه فوق، H و Q ماتریس‌های وزنی متقارن مثبت می‌باشند و R نیز ماتریس وزنی اکیدا مثبت می‌باشد. R و Q پارامترهای کنترلی می‌باشند. t_0 زمان اولیه و t_f زمان نهایی برای اجرای الگوریتم و رسیدن به مقدار مطلوب می‌باشد. رابطه(۲۱) فرم ورودی کنترلی را نشان می‌دهد[۲۹].

$$u^* = -R^{-1}B^TKx(t) \quad (۲۱)$$

در رابطه(۲۱)، K ضریب لاگرانژ می‌باشد و از حل یک معادله ریکاتی بدست می‌آید. این معادله در رابطه (۲۲) نشان داده شده است،[۲۹].

$$k = -KA - A^TK - Q + KBR^{-1}B^TK \quad (۲۲)$$

برای سیستم مورد استفاده در این مقاله که یک مدل غیر خطی از بیمار مبتلا به سرطان ملانوما می‌باشد، ابتدا خطی‌سازی انجام شده است و سپس کنترلر مربوطه با استفاده از روش LQR طراحی شده است. برای شبیه‌سازی از نرم‌افزار متلب استفاده شده است. شرایط اولیه برای یک بیمار مبتلا به ملانوما در جدول (۴) آورده شده است.

جدول۴: شرایط اولیه برای یک بیمار مبتلا به ملانوما [۲۷]

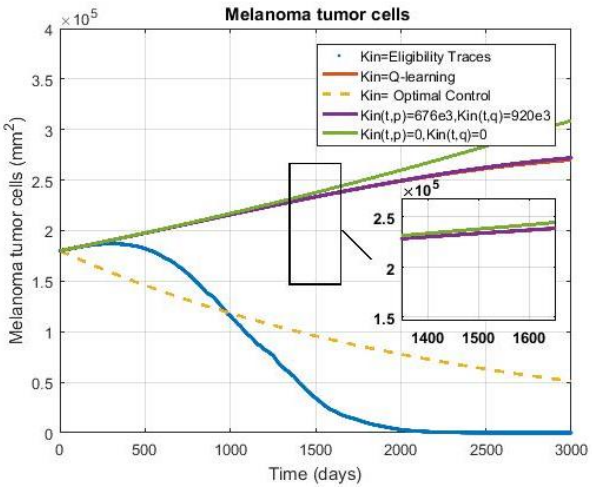
متغیر	مقدار اولیه	متغیر	مقدار اولیه
$0\left({}^{mg}/_{ml}\right)$	$Ab_s(0)$	$18\times10^4(mm^2)$	$M(0)$
$0(\mu g\ of\ protein)$	$Ta(0)$	$0\left({}^{Cell}/_{Kg}\right)$	$Ac(0)$
$196\times10^3\left({}^{cell}/_{days}\right)$	$Na(0)$	$0\left({}^{mg}/_{ml}\right)$	$Ac_s(0)$
		$0\left({}^{mg}/_{ml}\right)$	$Ab(0)$

برای اینکه عملکرد روش مسیرهای شایستگی بهتر نشان داده شود، رفتار سیستم در حالت تعیین دارو با استفاده از روش مسیرهای شایستگی علاوه بر روش یادگیری Q و کنترل بهینه با تزریق دوز داروی صفر و مقدار ثابت در هر لحظه نیز، مقایسه می‌شود. مقدار ثابت برای هر دو دارو، مقادیر $K_{in}\ (t,p)=676\times10^3\left({}^{Cell}/_{Kg}\right)$، برای OT1CTLها و $K_{in}\ (t,q)=920\times10^3\left({}^{mg}/_{ml}\right)$ برای آنتی‌بادی‌ها در نظر گرفته شده است. شکل (۴) جمعیت سلول‌های سرطانی در حالت تعیین دوز دارو با استفاده روش‌های مسیرهای شایستگی، یادگیری Q، کنترل بهینه، دوز داروی ثابت و دوز صفر را نشان می‌دهد.

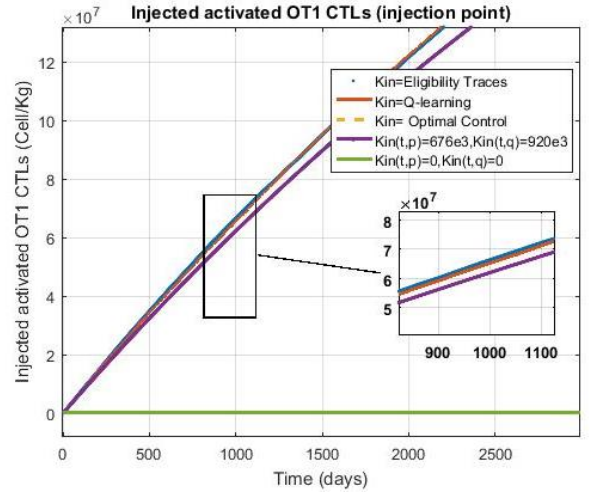
همانطور که در شکل (۴) نشان داده شده است، جمعیت سلول‌های سرطانی در حالت تعیین دوز دارو با استفاده از روش مسیرهای شایستگی در مدت زمان درمان در نظر گرفته شده کاهش سریع‌تری نسبت به دیگر روش‌ها داشته است. تزریق دوز دارو در روش انتخابی نیز کمتر بوده که این امر کاهش اثرات زیانبار آن را به همراه خواهد داشت. این در حالی می‌باشد که دیگر روش‌ها قادر به از بین بردن سلول‌های سرطانی به‌طور کامل در این مدت زمان نبوده‌اند.

شکل (۵) رفتار OT1CTLهای فعال در محفظه‌ی نقطه‌ی تزریق در حالت تعیین دوز دارو با استفاده از پنج حالت ذکر شده را نشان می‌دهد.

OT1CTLها به عنوان سلول‌های ایمنی بدن و عامل کشنده برای سلول‌های سرطانی می‌باشند. بنابراین تا زمانی که سلول‌های سرطانی در بدن از بین نرفته‌اند، مقدار آن‌ها افزایش پیدا کرده و در نهایت با حذف سلول‌های سرطانی رشد آنها متوقف شده و به مقدار ثابت می‌رسند. همان‌طور که در شکل(۵) نشان داده شده است، این امر در حالت استفاده از روش‌های مسیرهای شایستگی و الگوریتم یادگیری Q با سرعت بیشتری نسبت به تزریق دوز داروی ثابت اتفاق افتاده است. در حالت استفاده از روش کنترل بهینه نیز

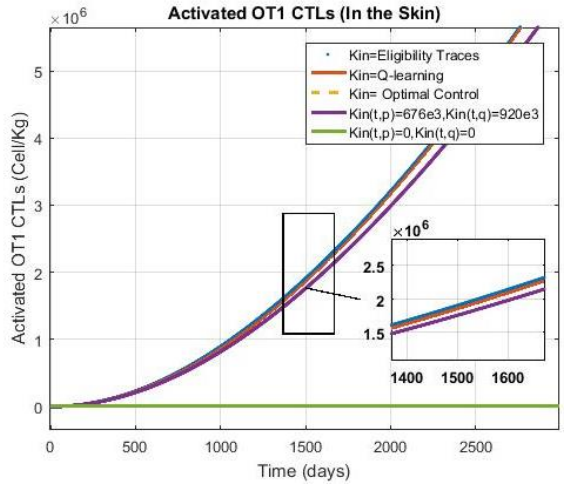


شکل۴: رفتار جمعیت سلول‌های سرطانی در حالت استفاده از روش مسیرهای شایستگی، الگوریتم یادگیری Q، کنترل بهینه، مقدار ثابت و دوز داروی صفر



شکل۵: رفتار OT1CTLهای فعال تزریق‌شده به نقطه تزریق در حالت استفاده از روش مسیرهای شایستگی، الگوریتم یادگیری Q، کنترل بهینه، مقدار ثابت و دوز داروی صفر

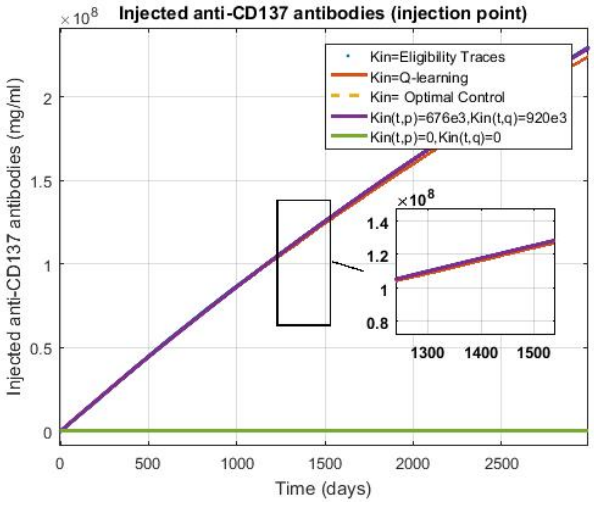
سرعت افزایش OT1CTLها در مدت زمان درنظر گرفته شده بسیار پایین بوده و نزدیک به صفر می‌باشد. شکل (۶) رفتار OT1CTLهای فعال در محفظه پوست در حالت تعیین میزان دوز دارو با استفاده روش‌های مسیرهای شایستگی، یادگیری Q، کنترل بهینه و دوز داروی ثابت و دوز صفر را نشان می‌دهد. در این شکل نیز، OT1CTLهای که در محفظه پوست می‌باشند، برای ازبین بردن سلول‌های سرطانی در این ناحیه افزایش می‌یابند. در حالت استفاده از روش‌های مسیرهای شایستگی و الگوریتم یادگیری Q این متغیر نیز با سرعت بیشتری نسبت به تزریق دوز داروی ثابت افزایش یافته است. بدون تزریق دارو نیز تغییراتی در این متغیر دیده نمی‌شود. در حالت استفاده از روش کنترل بهینه نیز سرعت افزایش OT1CTLها در محفظه پوست و در مدت زمان درنظر



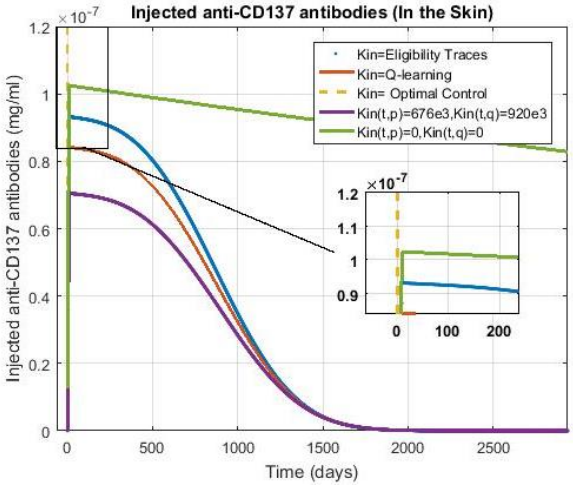
شکل۶: رفتار OT1CTLهای فعال در محفظه پوست، در حالت استفاده از روش مسیرهای شایستگی، الگوریتم یادگیری Q، کنترل بهینه مقدار ثابت و دوز داروی صفر

گرفته شده بسیار پایین بوده و نزدیک به صفر می‌باشد. شکل (۷) تغییرات میزان آنتی بادی CD137 تزریق شده در محفظه‌ی نقطه‌ی تزریق در کل مدت درمان و با استفاده از هر پنج روش ذکر شده را نشان می‌دهد. همان‌طور که درشکل (۷) نشان داده شده است، میزان آنتی بادی‌های ترشح شده در محفظه نقطه تزریق نیز برای کاهش سلول‌های سرطانی افزایش پیدا می‌کنند. سرعت افزایش آن‌ها در حالت استفاده از روش مسیرهای شایستگی بیشتر از روش الگوریتم یادگیری Q می‌باشد. در حالت استفاده از روش کنترل بهینه نیز سرعت تغییرات بسیار پایین بوده و نزدیک به صفر می‌باشد.

در شکل (۸) تغییرات میزان آنتی‌بادی‌ها در محفظه‌ی پوست در حالت تعیین دوز دارو با استفاده از هر پنج روش ذکر شده قابل



شکل۷: تغییرات آنتی‌بادیCD137 تزریق شده به نقطه تزریق در حالت استفاده از روش مسیرهای شایستگی، الگوریتم یادگیری Q، کنترل بهینه، مقدار ثابت و دوز داروی صفر

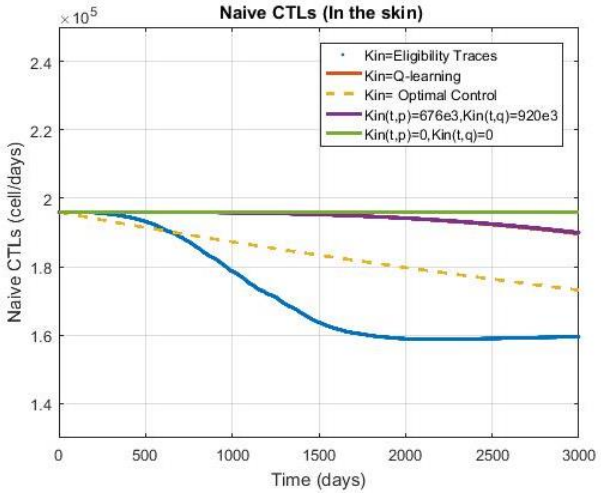


شکل۸: تغییرات آنتی‌بادیCD137 در محفظه پوست، در حالت استفاده از روش مسیرهای شایستگی، الگوریتم یادگیری Q، کنترل، مقدار ثابت و دوز داروی صفر

نمایش می‌باشد.

با توجه به شکل (۸)، آنتی‌بادی‌ها در محفظه پوست در ابتدا که جمعیت سلول‌های سرطانی بیشتر می‌باشند افزایش می‌یابند و با کاهش سلول‌های سرطانی میزان آن‌ها کاهش پیدا کرده و از بین می‌روند. همانطور که در این شکل نشان داده شده است، آنتی‌بادی‌ها در محفظه پوست در حالت استفاده از روش کنترل بهینه افزایش بسیار زیادی داشته است. شکل (۹)، رفتار CTLهای ساده[۲۳] با استفاده از هر پنج روش ذکر شده را نشان می‌دهد.

CTLهای ساده در بدن هر فرد به مقدارهای متفاوتی وجود دارند. با وارد شدن سلول‌های سرطانی به بدن، تبدیل به CTL فعال و کشنده می‌شوند. بنابراین CTLهای ساده با حذف سلول‌های سرطانی کاهش پیدا کرده و سپس افزایش می‌یابند. همان‌طور که در شکل (۹) نشان داده شده است، در حالت استفاده از روش

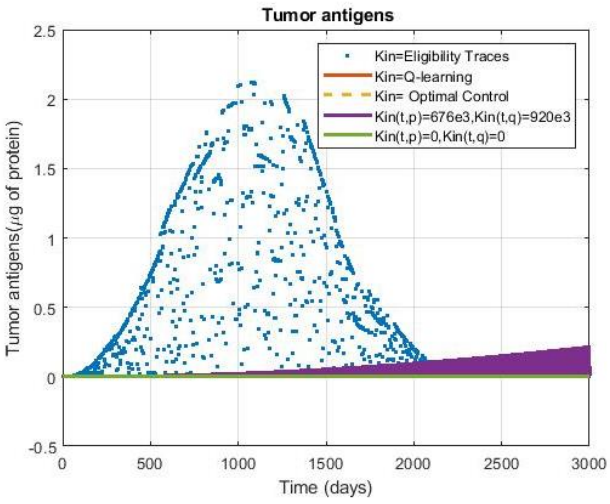


شکل۹: رفتار CTLهای ساده، در حالت استفاده از روش مسیرهای شایستگی، الگوریتم یادگیری Q، کنترل بهینه، مقدار ثابت و دوز داروی صفر

مسیرهای شایستگی طبق روند ذکر شده، ابتدا CTLهای ساده کاهش پیدا کرده‌اند و با حذف سلول‌های سرطانی مقدار آنها ثابت می‌شود. این در حالی می‌باشد که در مدت زمان در نظرگرفته شده در حالت استفاده از روش تزریق دوز داروی ثابت، الگوریتم یادگیری Q و کنترل بهینه، CTLهای ساده فقط کاهش پیدا کرده‌اند. کاهش CTLهای ساده به معنای تبدیل آن‌ها به CTLهای فعال می‌باشد. که کاهش سریع آن‌ها باعث تولید CTLهای فعال بیشتر می‌شود. در حالت بدون تزریق دارو نیز مقدار این آن‌ها تغییری نکرده است. بنابراین روش مسیرهای شایستگی بهتر توانسته است میزان CTLها را کنترل کند. شکل (۱۰)، تغییرات میزان آنتی‌ژن تومور را در حالت استفاده از روش مسیرهای شایستگی، الگوریتم یادگیری Q، کنترل بهینه، دوز داروی ثابت و بدون تزریق دارو نشان می‌دهد.

همان‌طور که در شکل فوق نشان داده شده است، میزان آنتی‌ژن‌های تومور در حالت استفاده از روش مسیرهای شایستگی با سرعت بیشتری نسبت به دیگر روش‌ها کاهش پیدا کرده‌اند. دلیل این امر، استفاده از سلول‌های سرطانی به صورت رابطه $\left(-log\left(\frac{y_{new}}{y_{old}}\right)\right)$ در تابع پاداش می‌باشد. که با تاثیر مستقیم سلول‌های سرطانی در تغییرات تومور آنتی‌ژن‌ها کاهش آن‌ها با سرعت بالا اتفاق افتاده است. همان‌طور که ملاحظه می‌گردد، این امر در حالت استفاده از الگوریتم یادگیری Q و کنترل بهینه اتفاق نیفتاده است. لازم به ذکر می‌باشد که به دلیل اینکه روند تغییرات آنتی‌ژن تومور بهتر نشان داده شود، این شکل با تعداد گام زمانی بیشتر رسم شده است.

میزان دوز داروی تزریقی توسط هر چهار روش در جدول (۵) آورده شده است. بر اساس این جدول مجموع دوز داروی تزریقی با استفاده از روش مسیرهای شایستگی کمتر از روش‌های الگوریتم یادگیری Q ، کنترل بهینه و روش تزریق دوز داروی ثابت می‌باشد که باعث کاهش اثرات زیان‌بار دارو خواهد شد. همانطور که در شکل (۴) نشان داده شد، سلول‌های سرطانی در حالت استفاده از روش مسیرهای شایستگی با سرعت بیشتر و در تعداد گام کمتر نسبت به دیگر روش‌ها از بین رفته‌اند. علاوه بر این همانطور که در جدول (۵) نشان داده شده است، مجموع دوز داروی تزریقی به بیمار در کل مدت درمان در حالت استفاده از این روش کمتر می‌باشد که این امر باعث کاهش اثرات زیان‌بار دارو و کاهش آسیب رسیدن به سلول‌های سالم خواهد شد.



شکل۱۰: تغییرات آنتی‌ژن تومور، در حالت استفاده از روش مسیرهای شایستگی، الگوریتم یادگیری Q، کنترل بهینه، مقدار ثابت و دوز داروی صفر

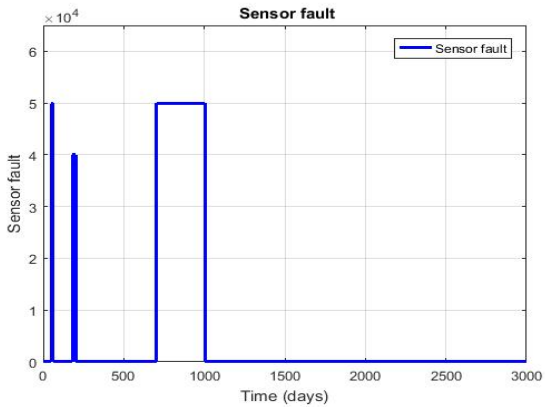
جدول۵: مقایسه مجموع داروهای تزریقی به بیمار در سه حالت ثابت، الگوریتم یادگیری Q، مسیرهای شایستگی و با استفاده از روش کنترل بهینه

روش تزریق دارو	مجموع داروی تزریقی به بیمارتا زمان از بین رفتن سلول‌های سرطانی
تزریق داروی ثابت	$K_{in}(t,p)$=9.464× 10^9 $Cell/Kg$
	$K_{in}(t,q)$=1.288× 10^{10} mg/ml
تزریق دارو با استفاده از روش الگوریتم یادگیریQ	$K_{in}(t,p)$=1.4318× 10^{10} $Cell/Kg$
	$K_{in}(t,q)$=1.8269× 10^{10} mg/ml
تزریق دارو با استفاده از روش مسیر شایستگی	$K_{in}(t,p)$=1.7958× 10^9 $Cell/Kg$
	$K_{in}(t,q)$=2.3017× 10^9 mg/ml
کنترل بهینه	$K_{in}(t,p)$=۲,۲۳۴۴× 10^9 $Cell/Kg$
	$K_{in}(t,q)$=1.0847× 10^{13} mg/ml

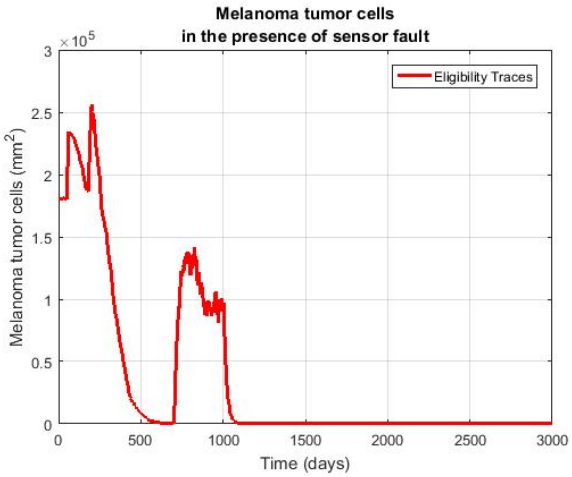
۴-۱- کنترل جمعیت سلول‌های سرطانی در حضور عیب در سنسور با استفاده از روش مسیرهای شایستگی

سیستم‌های پزشکی، سیستم‌های بسیار حساسی می‌باشند. بنابراین عملکرد صحیح آنها نیز از اهمیت ویژه‌ای برخوردار است. بروز عیب در این سیستم‌ها باعث خسارات جانی جبران‌ناپذیری خواهد شد. بنابراین تشخیص سریع، دقیق و به‌موقع عیب در سیستم‌های پزشکی از مواردی می‌باشد که در چند ساله اخیر بسیار مورد توجه قرار گرفته است. در این بخش عملکرد کنترلر پیشنهادی (روش مسیرهای شایستگی) در حضور عیب در سنسور سیستم مورد بررسی قرار گرفته است. این امر منطبق بر بخش یادگیری تقویتی مقاله [۳۰] انجام گرفته است. عیب وارد به سنسور سیستم در شکل(۱۱) نشان داده شده است.

عملکرد کنترلر پیشنهادی (روش مسیرهای شایستگی) در کاهش جمعیت سلول‌های سرطانی در حضور عیب در سنسور سیستم در شکل (۱۲) نشان داده شده است.



شکل۱۱ :عیب وارد شده به سنسور سیستم



شکل۱۲: رفتار جمعیت سلول‌های سرطانی در حالت استفاده از روش مسیرهای شایستگی

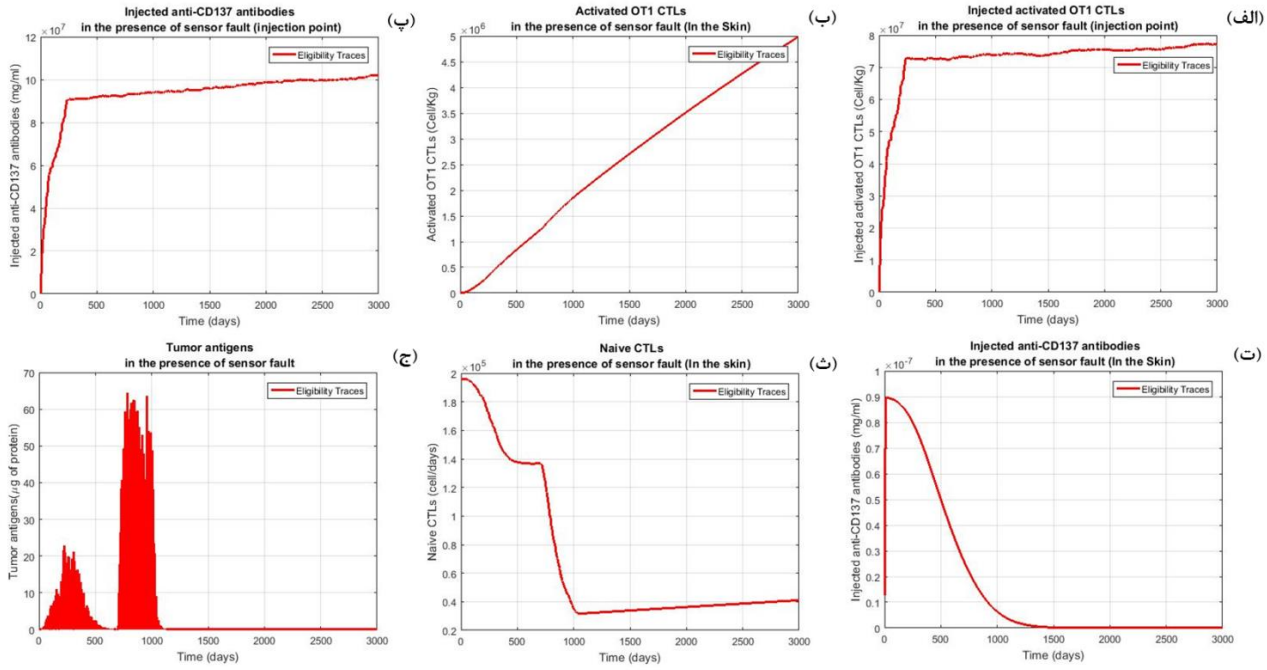
همانطور که در شکل(۱۲) نشان داده شده است، با رخ دادن عیب در سنسور سیستم باز هم کنترلر پیشنهادی توانسته است جمعیت سلول‌های سرطانی را کنترل کرده و آن‌ها را از بین ببرد. رفتار دیگر متغیرها نیز در حضور عیب در سنسور سیستم کنترل شده می‌باشد و در شکل (۱۳) نمایش داده شده‌اند. همانطور که در این شکل نشان داده شده است، در نمودار (الف)، CTLهای تزریق شده به محفظه نقطه تزریق تا زمان از بین نرفتن کامل سلول‌های سرطانی افزایش یافته است سپس به یک مقدار ثابت رسیده است. در نمودار (ب) نیزCTL های محفظه پوست برای کاهش سلول‌های سرطانی افزایش پیدا کرده‌اند. در نمودار (پ)، آنتی‌بادی نوع CD-137 تزریق شده به نقطه تزریق تا زمان کاهش سلول‌های سرطانی به بیمار تزریق شده و در نهایت به مقدار ثابت رسیده است. در نمودار (ت) نیز، آنتی‌بادی نوع CD-137 تزریق شده به محفظه پوست ابتدا افزایش پیدا کرده و با کاهش سلول‌های سرطانی از بین رفته است. CTLهای ساده در نمودار (ث) نیز با تبدیل شدن به CTL کاهش پیدا کرده و به یک مقدار ثابت می‌رسند. در نمودار (ج) آنتی‌ژن توموری نیز در هنگام اعمال عیب افزایش پیدا کرده است. مجموع میزان دوز داروی تزریقی در حالت اعمال عیب به سنسور سیستم بدین صورت بدست آمده است:

برای OT1CTLها مقدار $Kin(t,p) = 1.3767 \times 10^9 \frac{Cell}{Kg}$ و برای آنتی‌بادی‌ها مقدار $Kin(t,q) = 1.355 \times 10^9 \frac{mg}{ml}$ حاصل گردید.

۴-۲- اعمال عدم قطعیت در پارامترهای سیستم و شرایط اولیه

در این بخش پایداری روش یادگیری در حضور اغتشاش و عدم قطعیت‌ها مورد بحث قرار گرفته است. به طور کلی برای اثبات پایداری مقاوم روش‌های هوش مصنوعی از جمله یادگیری تقویتی در حضور اغتشاش و نویز وجود ندارد. این امر تنها در زمانی اتفاق می‌افتد که یک سری شرایط خاص بر روی سیستم مورد نظر و محدودسازی بر روی پارامترهای این روش در نظر گرفته شود. در زیر مقالاتی که با لحاظ شرایط خاص و محدودسازی یک پارامتر در این روش، اثبات پایداری مقاوم را انجام داده‌اند، به اختصار بیان شده است:

در [۳۱]، با فرض اینکه سیستم خطی و نامتغیر با زمان می‌باشد، اثبات پایداری روش یادگیری تقویتی توسط روش لیاپانوف انجام شده است. اشکال عمده‌ای که در این مقاله وجود دارد، استفاده از فرم خطی شده سیستم برای اثبات روش می‌باشد که این امر دینامیک سیستم را به‌خصوص در سیستم‌های پزشکی بسیار تحت تاثیر قرار می‌دهد. این در حالی می‌باشد که مزیت استفاده از روش یادگیری تقویتی توانایی عملکرد این روش در استفاده از فرم غیرخطی سیستم می‌باشد و استفاده از حالت خطی سیستم برای این روش معنایی ندارد. لازم به ذکر می‌باشد که اگر سیستم واقعی در دسترس باشد، استفاده از مدل غیرخطی هم نیازی نیست. در [۳۲]، با در نظر گرفتن اینکه اغتشاش به عمل انتخابی توسط عامل اعمال می‌گردد، پایداری مقاوم روش یادگیری تقویتی اثبات می‌گردد. در [۳۳] با لحاظ شرایط لیپشیتز[۲۴]، اثبات پایداری انجام گرفته است. در این مقاله همچنین محدویت بر روی عمل انتخابی اعمال شده است. در [۳۴] سیستم در نظر گرفته شده شبه خطی می‌باشد و پایداری روش یادگیری تقویتی در حضور اعمال عدم قطعیت در تابع پاداش اثبات شده است. اما در این مقاله از روش یادگیری تقویتی به طور عام استفاده شده است و شرط به خصوصی برای الگوریتم یادگیری و سیستم در نظر گرفته نشده است. به عنوان مثال همانند بسیاری از مقالاتی که جهت بررسی پایداری مقاوم، سیستم را خطی کرده‌اند، در این مقاله سیستم خطی نشده است. به دلیل اینکه سیستم‌های پزشکی با اعمال خطی‌سازی عملکردشان با سیستم واقعی تفاوت زیادی خواهد داشت، از فرم غیرخطی سیستم و گسسته‌سازی جهت کنترل آن توسط روش یادگیری تقویتی استفاده شده است. در این مقاله



شکل۱۳ : رفتار متغیرهای سیستم در حضور عیب سنسور

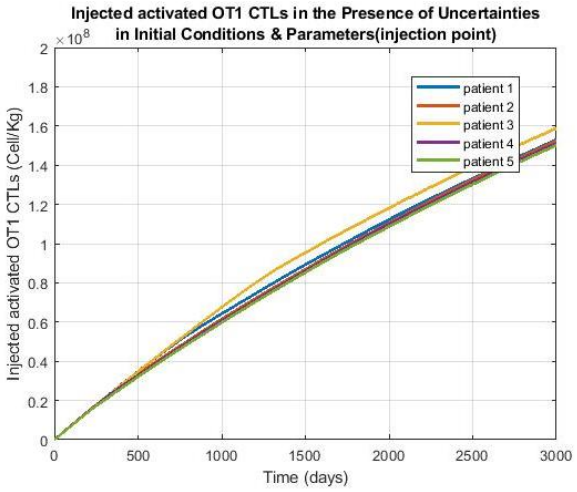
هدف اثبات پایداری روش ذکر شده نمی‌باشد و فقط به منظور بررسی عملکرد کنترلر پیشنهادی در کاهش سلول‌های سرطانی در دیگر بیماران مبتلا به ملانوما، عدم قطعیت در پارامترهای مدل اعمال شده است. پارامترهای ثابت در مدل ریاضی برای تمام بیماران مبتلا به ملانوما یکسان نیست و از بیماری به بیمار دیگر دارای تفاوت اندکی می‌باشد. در این بخش برای اینکه عملکرد کنترلر طراحی شده در کاهش سلول‌های سرطانی برای دیگر بیماران مبتلا به ملانوما نشان داده شود، عدم قطعیت در پارامترهای سیستم و شرایط اولیه اعمال شده است. بنابراین کنترل سلول‌های سرطانی با بهترین میزان تزریق دوز دارو در ۵ بیمار مبتلا به سرطان ملانوما بررسی شده است. برای این کار، ابتدا عدم قطعیت در پارامترها و شرایط اولیه به ترتیب با تغییرات ۵ و ۲۵ درصد، با توزیع یکنواخت لحاظ شده است. این تغییرات به صورت روابط (۲۳) و (۲۴) قابل نمایش می‌باشد.

$$P-(5\%)P<P<P+(5\%)P \qquad (۲۳)$$

$$I-(25\%)I<I<I+(25\%)I \qquad (۲۴)$$

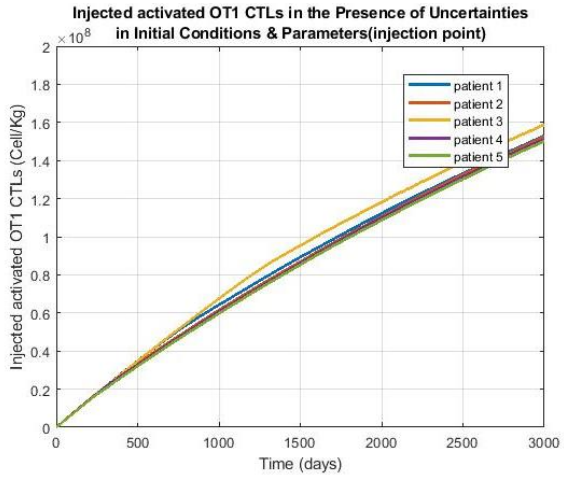
در رابطه(۲۳)، P معرف پارامترهای ثابت مدل ریاضی می‌باشد. این رابطه بیان می‌دارد که مقادیر انتخابی برای هر پارامتر به صورت تصادفی و در بازه ۵٪ کمتر و بیشتر آن انتخاب می‌شود. رابطه(۲۴) نیز بیانگر تغییرات برای شرایط اولیه می‌باشد. I بیان‌گر شرایط اولیه می‌باشد و مقادیر انتخابی برای هر کدام از شرایط اولیه به صورت تصادفی و در بازه ۲۵٪ کمتر و بیشتر آن انتخاب می‌شود.

شکل (۱۴) رفتار OT1CTLهای فعال تزریق‌شده به نقطه تزریق با لحاظ کردن عدم قطعیت را نمایش می‌دهد.



شکل۱۴: رفتار OT1CTLهای فعال تزریق‌شده به نقطه تزریق، در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی

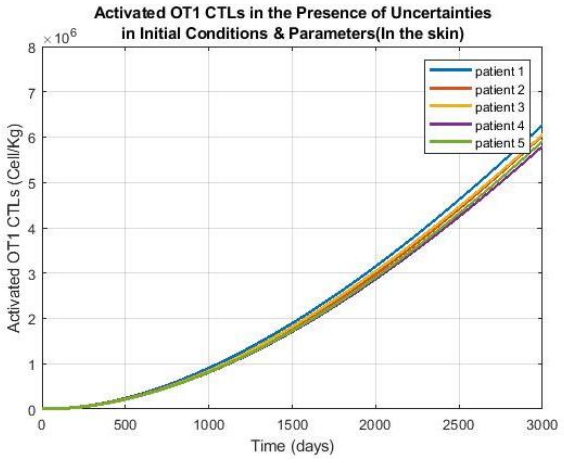
همانطور که در شکل (۱۴) مشاهده می‌گردد، یادگیری که در سیستم اصلی با استفاده از روش مسیرهای شایستگی انجام شده است، به خوبی توانسته است میزان OT1CTLهای فعال در دیگر بیماران مبتلا به سرطان ملانوما را افزایش دهد، که این امر نشان‌دهنده خاصیت تطبیق‌پذیری روش مسئله یادگیری تقویتی می‌باشد. شکل (۱۵)، تغییرات میزان آنتی‌بادی نوع CD-137 تزریق شده به نقطه تزریق را در پنج بیمار مبتلا به سرطان ملانوما نشان می‌دهد.



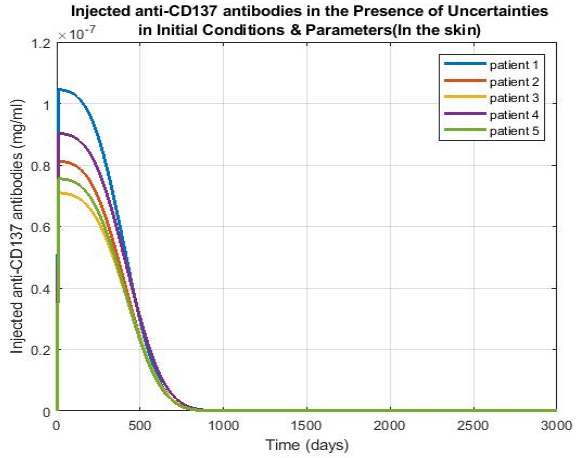
شکل۱۵: رفتار آنتی‌بادی نوع CD-137 تزریق شده به نقطه تزریق در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی

همان‌طور که در شکل (۱۵) نشان داده شده است، آنتی‌بادی‌های تزریق شده به نقطه تزریق در پنج بیمار افزایش یافته است. شکل (۱۶)، تغییرات رفتار OT1CTLهای فعال در محفظه پوست را برای پنج بیمار مبتلا به ملانوما نشان می‌دهد. با توجه به شکل (۱۶)، OT1CTLهای فعال در پنج بیمار مبتلا به ملانوما همانند تغییرات این متغیر در مدل اصلی، افزایش یافته است. شکل (۱۷)، تغییرات میزان آنتی‌بادی نوع CD-137 تزریق شده به پوست را برای پنج بیمار مبتلا به سرطان ملانوما نشان می‌دهد.

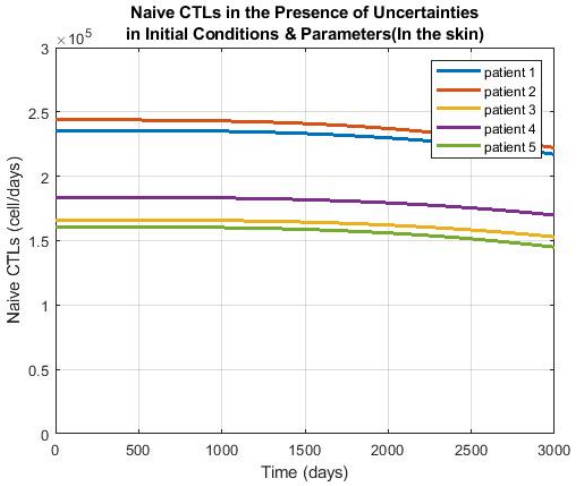
همان‌طور که در شکل (۱۷) نشان داده شده است، میزان آنتی‌بادی نوع CD-137 تزریق شده به پوست در تمام بیماران در ابتدا که جمعیت سلول‌های سرطانی بیشتر می‌باشد، افزایش یافته است و سپس کاهش یافته و صفر شده است. شکل (۱۸)، رفتار CTLهای ساده در پنج بیمار مبتلا به ملانوما را نشان می‌دهد.



شکل۱۶: رفتار OT1CTLهای فعال بدن، در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی



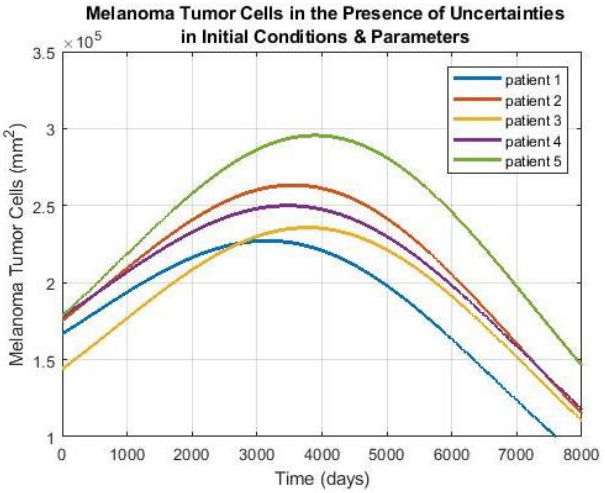
شکل۱۷: رفتار آنتی‌بادی نوع CD-137 تزریق شده به پوست در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی



شکل۱۸: رفتار CTLهای ساده در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی

همانطور که در شکل (۱۸) نشان داده شده است، CTLهای ساده در ابتدا که سلول‌های وارد بدن بیمار نشده‌اند، مقدار بیشتری داشته و با ورود سلول‌های سرطانی تبدیل به CTL فعال شده و خودشان کاهش پیدا کرده و در نهایت به مقداری ثابت می‌رسند. شکل (۱۹)، تغییرات رفتار جمعیت سلول‌های سرطانی در پنج بیمار مبتلا به سرطان ملانوما را نشان می‌دهد.

همانطور که در شکل (۱۹) نشان داده شده است، روش یادگیری تقویتی با لحاظ عدم قطعیت در سیستم باز هم توانسته است جمعیت سلول‌های سرطانی را در دیگر بیماران مبتلا به سرطان ملانوما به خوبی کنترل کرده و کاهش دهد. بیمار تا انتهای عمر می‌بایست دارو مصرف کند که این امر از نظر پزشکی قابل قبول می‌باشد. این امر به دلیل خاصیت تطبیق‌پذیری روش یادگیری تقویتی با محیط می‌باشد. شکل (۲۰)، تغییرات میزان آنتی‌ژن تومور را در پنج بیمار مبتلا به سرطان ملانوما نشان می‌دهد.



شکل۱۹: جمعیت سلول‌های سرطانی در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی

باتوجه به شکل (۲۰)، میزان آنتی‌ژن تومور در تمام بیماران مبتلا به ملانوما ابتدا زیاد شده و سپس کاهش پیدا کرده و به صفر رسیده است.

۴-۳- افزودن نویز به سیستم در حضور عدم قطعیت در پارامترها و شرایط اولیه

پارامترهای بدن یک بیمار در طول روز ثابت نیست و به دلیل فعالیت‌های روزانه مدام در حال تغییر می‌باشد. در این بخش تاثیر نویز و اغتشاش در کنترل سلول‌های سرطانی در بیماران مبتلا به سرطان ملانوما بررسی شده است. نویز وارده به سیستم، نویز سفید گوسی می‌باشد و با استفاده از دستور wgn[۲۵] به معادلات سیستم اضافه شده است.

برای مدل‌سازی اکثر فرایندهای تصادفی در طبیعت از نویز سفید گوسی استفاده می‌شود. زیرا بسیاری از فرایندهای طبیعی از نویز گوسی تبعیت می‌کنند. همچنین آنالیز طیف زیادی از نویزها به آنالیز نویز گوسی با واسطه و یا با تقریب مرتبط می‌باشند. برای افزودن نویز به دینامیک سیستم، از پارامترهای w_1 تا w_7 استفاده است. که در هر گام زمانی مقداری متفاوت و احتمالاتی به خود می‌گیرند. معادلات جدید به فرم روابط (۲۵) تا (۳۱) تغییر خواهند کرد. بازه تغییرات نویز وارد شده به هر یک از متغیرها بر اساس میزان تغییرات آن‌ها با اعمال عدم قطعیت به سیستم می‌باشد. این نویز خطای اندازه‌گیری تغییرات را نشان می‌دهد.

$$\frac{dE}{dt} = K_{in}(t,p) - \alpha_{11}E - \alpha_8 E + w_1 \quad (۲۵)$$

$$\frac{dAb}{dt} = K_{in}(t,q) - \alpha_{11}Ab - \alpha_{10}Ab + w_2 \quad (۲۶)$$

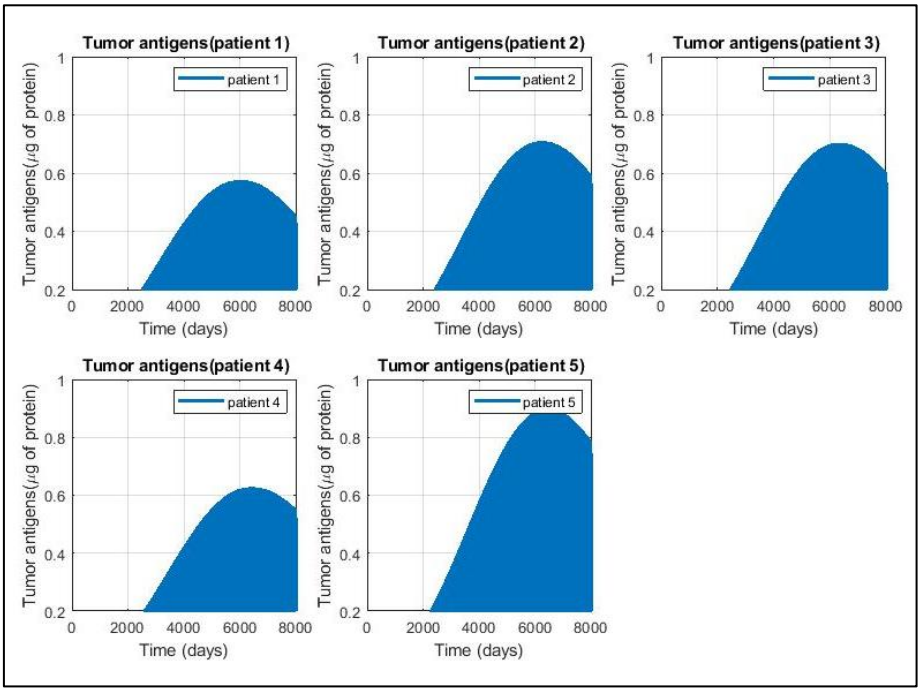
$$\frac{dE_s}{dt} = \alpha_7\left[\frac{A_s}{A_s+K_1}\right]E_s + \alpha_{11}E(t-\tau) + \alpha_6 NA + \alpha_8 E_s + w_3 \quad (۲۷)$$

$$\frac{dC}{dt} = \left(\alpha_1 - \alpha_2 Ln(C)\right).C - \alpha_3\left[\frac{A_s+K_2}{A_s+K_3}\right]E_s C + w_4 \quad (۲۸)$$

$$\frac{dA}{dt} = \alpha_4\left[\alpha_3\left[\frac{A_s+K_2}{A_s+K_3}\right]E_s C\right] - \alpha_5 A - \alpha_6 NA + w_5 \quad (۲۹)$$

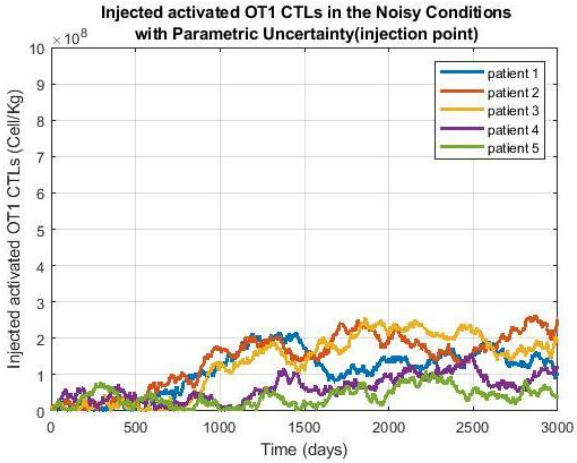
$$\frac{dN}{dt} = h(M-N) - \alpha_6 NA + w_6 \quad (۳۰)$$

$$\frac{dA_s}{dt} = \alpha_{11}Ab(t-\tau) - \alpha_9 A_s E_s - \alpha_{10}A_s + w_7 \quad (۳۱)$$



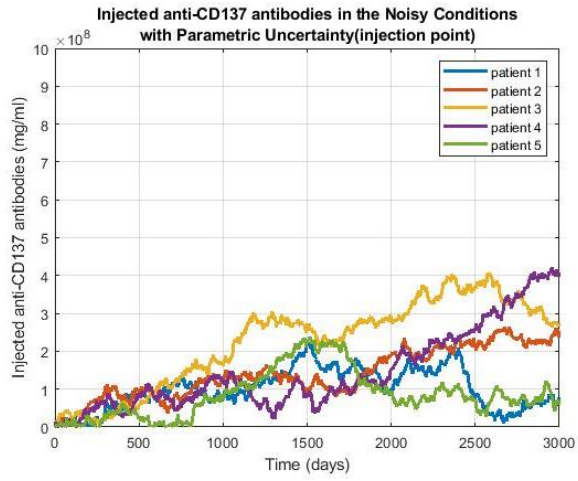
شکل۲۰: تغییرات آنتی‌ژن تومور در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی

با توجه به معادلات فوق، رفتار OT1CTLهای فعال تزریق‌شده به نقطه تزریق برای پنج بیمار مختلف در حضور نویز در محیط در شکل (۲۱) نمایش داده شده است. همان‌طور که در این شکل نشان داده شده است، OT1CTLهای فعال تزریق‌شده به نقطه تزریق در پنج بیمار مبتلا به ملانوما در حضور نویز با وجود تلرانس زیاد، باز هم افزایش یافته‌اند.

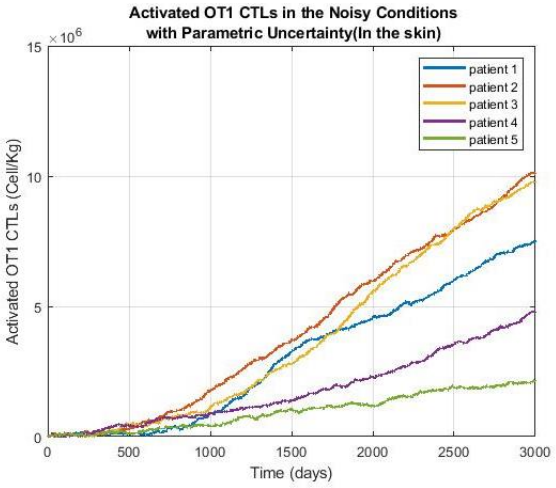


شکل۲۱: رفتار OT1CTLهای فعال تزریق‌شده به نقطه تزریق، در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی در حضور نویز

شکل (۲۲)، تغییرات میزان آنتی‌بادی نوع CD-137 تزریق شده به نقطه تزریق را در پنج بیمار مبتلا به سرطان ملانوما در حضور نویز در سیستم نشان می‌دهد. همانطور که در شکل (۲۲) نشان داده شده است، آنتی‌بادی نوع CD-137 تزریق شده به نقطه تزریق در پنج بیمار مبتلا به سرطان ملانوما در حضور نیز در سیستم افزایش یافته‌اند. شکل (۲۳)، تغییرات رفتار OT1CTLهای فعال در محفظه پوست را برای پنج بیمار مبتلا به سرطان ملانوما در حضور نویز و با استفاده از روش مسیرهای شایستگی را نشان می‌دهد.

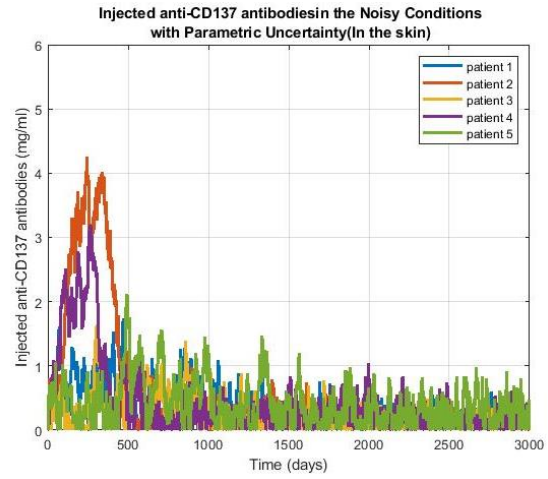


شکل۲۲: رفتار آنتی‌بادی نوع CD-137 تزریق شده به نقطه تزریق در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی در حضور نویز



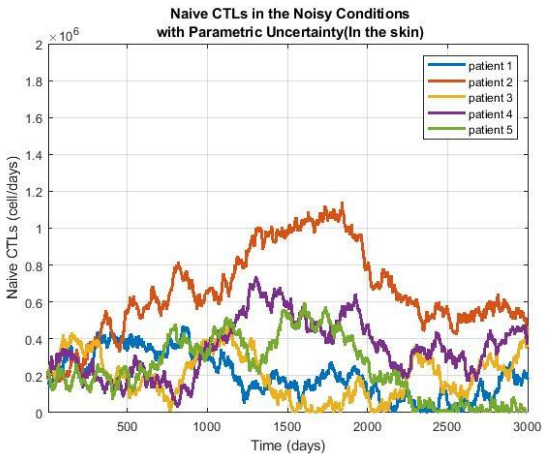
شکل۲۳: رفتار OT1 CTLهای فعال بدن در محفظه پوست، در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی در حضور نویز

همانطور که در شکل فوق نشان داده شده است، OT1CTLهای فعال در محفظه پوست در پنج بیمار مبتلا به سرطان ملانوما نیز با استفاده از روش‌مسیرهای شایستگی و در حضور نویز افزایش پیدا کرده‌اند. شکل (۲۴)، تغییرات میزان آنتی‌بادی نوع CD-137 تزریق شده به محفظه پوست را در پنج بیمار مبتلا به ملانوما در حضور نویز در سیستم نشان می‌دهد.

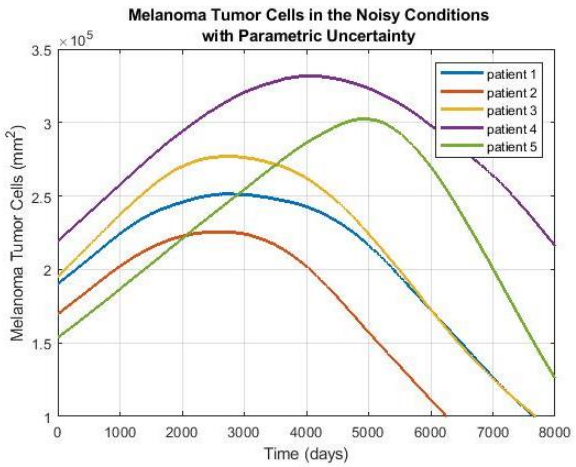


شکل۲۴: رفتار آنتی‌بادی نوع CD-137 تزریق شده به پوست در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی در حضور نویز

همان‌طور که در شکل فوق نشان داده شده است، آنتی‌بادی نوع CD-137 تزریق شده به پوست در حضور نویز، در ابتدا با توجه به افزایش سلول‌های سرطانی، افزایش یافته‌اند و سپس کم شده و در نزدیکی صفر تلرانس خیلی کم داشته‌اند. شکل (۲۵)، بیان‌گر رفتار CTLهای ساده در پنج بیمار مبتلا به سرطان ملانوما در حضور نویز می‌باشد.



شکل۲۵: رفتار CTLهای ساده در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی در حضور نویز



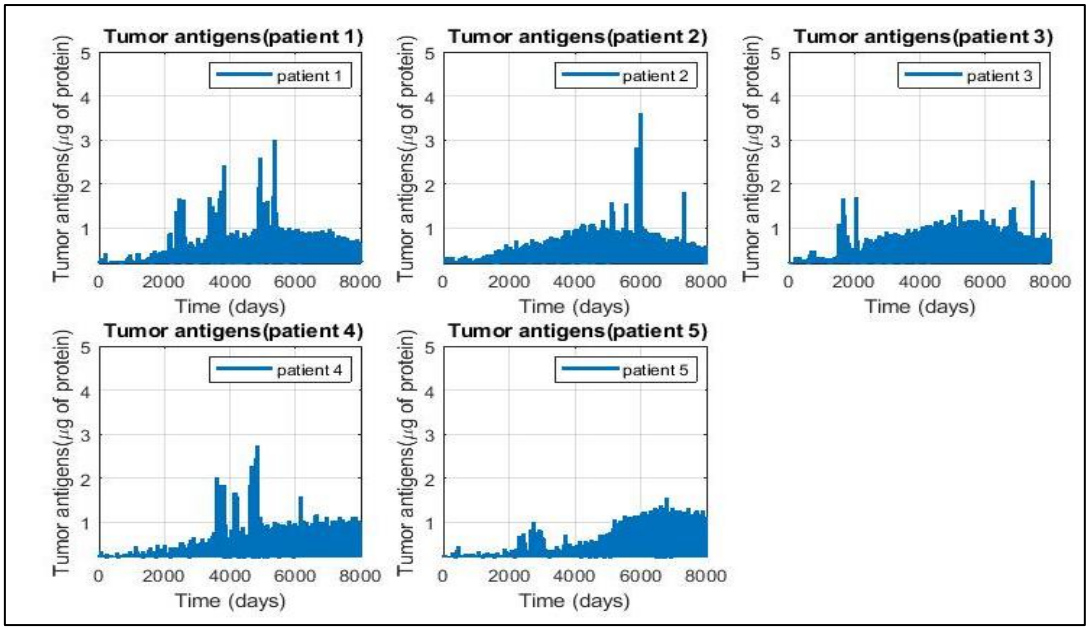
شکل۲۶: جمعیت سلول‌های سرطانی در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی در حضور نویز

همانطور که در شکل فوق نشان داده شده است، در حالت وارد شدن نویز به سیستم، CTLهای ساده با ورود سلول‌های سرطانی تبدیل به CTL فعال شده و در بدن افزایش می‌یابند در نهایت با از بین رفتن سلول‌های سرطانی به مقداری ثابت می‌رسند. شکل (۲۶)، تغییرات رفتار جمعیت سلول‌های سرطانی در پنج بیمار مبتلا به ملانوما در حضور نویز را نشان می‌دهد.

همانطور که در شکل (۲۶) نشان داده شده است، روش یادگیری تقویتی با اضافه شدن نویز به سیستم باز هم توانسته است جمعیت سلول‌های سرطانی را در دیگر بیماران مبتلا به سرطان ملانوما کنترل کرده و به صفر برساند. این امر به دلیل خاصیت تطبیق‌پذیری روش یادگیری تقویتی با محیط می‌باشد. شکل (۲۷)، تغییرات میزان آنتی‌ژن تومور را در پنج بیمار مبتلا به سرطان ملانوما نشان می‌دهد. با توجه به این شکل، میزان آنتی‌ژن تومور در تمام بیماران مبتلا به ملانوما ابتدا زیاد شده و سپس کاهش پیدا کرده و به صفر رسیده است.

۴-۴- مقایسه روش‌های مسیرهای شایستگی و یادگیری Q از نظر سرعت همگرایی

در این گونه روش‌ها پارامتر α نرخ یادگیری می‌باشد. تعیین مقدار درست برای آن بر سرعت همگرایی و زمان رسیدن به سیاست بهینه تاثیرگذار خواهد بود. نرخ یادگیری مقداری بین صفر و یک می‌باشد. مقدار صفر باعث می‌شود که عامل چیزی یاد نگیرد و در نظر گرفتن مقدار یک باعث می‌شود که عامل فقط اطلاعات جدید را ملاک قرار دهد. با برقراری دو شرط بیان شده در روابط (۳۲) و (۳۳) تضمین می‌شود که در بهترین ارزش جفت حالت و عمل همگرایی اتفاق می‌افتد [۳۵].



شکل۲۷: تغییرات آنتی‌ژن تومور در پنج بیمار مبتلا به سرطان ملانوما با استفاده از روش مسیرهای شایستگی در حضور نویز

$$\sum_{t=1}^{\infty} \alpha_t(x,\alpha) = \infty \quad (۳۲)$$

$$\sum_{t=1}^{\infty} {\alpha_t}^2(x,\alpha) < \infty \quad (۳۳)$$

روابط فوق بیان می‌دارند که با در نظر گرفتن مقادیر متفاوت برای نرخ همگرایی در هر لحظه، مجموع آن‌ها برای همگرایی کل الگوریتم لازم می‌باشد. نرخ همگرایی برای جفت حالت و عمل در هر لحظه طبق رابطه(۳۴) محاسبه می‌گردد [۳۵].

$$\alpha_t(x,\alpha) = \begin{cases} \frac{1}{N_t(s,a)} & if\ (x,\alpha) = (x_t,a) \\ 0 & other\ wise \end{cases} \quad (۳۴)$$

تعیین نرخ یادگیری نیز بر اساس رابطه فوق انجام شده است. در این رابطه، x حالت و a عمل انجام شده در هر لحظه می‌باشد. $N_t(s,a)$ تعداد جفت حالت و عمل‌هایی که تا لحظه t اتفاق افتاده است. بر اساس شرایط ذکر شده ، روابط (۳۵) و (۳۶) برقرار خواهند. بود [۳۵].

$$|Q_t(x,a) - Q^*(x,a)| \le \frac{B}{t^{R(1-\gamma)}} \quad (۳۵)$$

$$|Q_t(x,a) - Q^*(x,a)| \le B\sqrt{\frac{\log\log t}{t}} \quad (۳۶)$$

γ نرخ فراموشی می‌باشد. در روابط فوق، مقدار پارامتر B ثابت و بزرگتر از صفر می‌باشد. پارامتر R بر اساس رابطه (۳۷) بدست می‌آید [۳۵].

$$R = \frac{P_{min}}{P_{max}} \quad (۳۷)$$

در رابطه فوق، P_{min} مینیمم احتمال وقوع جفت حالت و عمل و P_{max} ماکزیمم احتمال وقوع جفت حالت و عمل می‌باشد. در رابطه فوق، اگر $\gamma \ge 1 - \frac{P_{max}}{2P_{min}}$ باشد، رابطه (۳۶) کندتر عمل خواهد کرد و اگر $\gamma < 1 - \frac{P_{max}}{2P_{min}}$ باشد، رابطه (۳۷) به کندی عمل خواهد کرد.

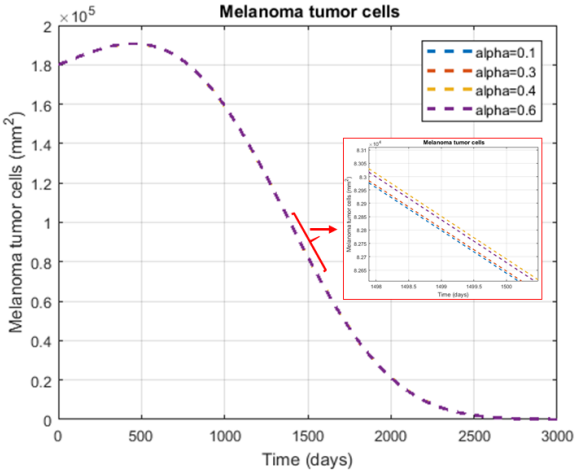
همان‌طور که مشاهده می‌شود، در نهایت جدولQ به مقدار بهینه خود Q^* نزدیک می‌گردد. اثبات این مسئله و همگرایی در روش یادگیری تقویتی در [۳۵-۴۰] به طور کامل آورده شده است.

شکل‌های (۲۸) و (۲۹) تاثیر αهای مختلف در کاهش سلول‌های سرطانی توسط روش‌های مسیرهای شایستگی و الگوریتم یادگیری Q را نشان می‌دهد.

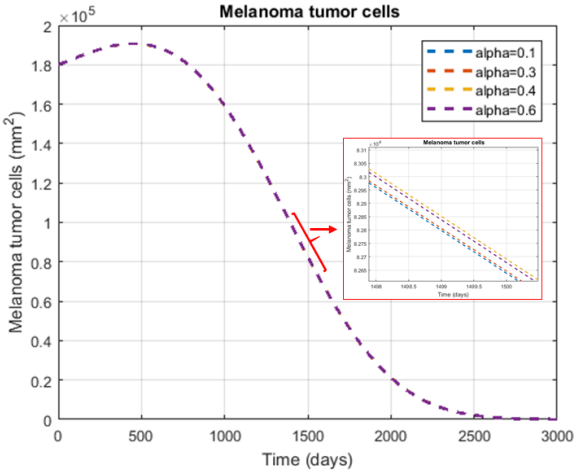
با توجه به آنچه در شکل‌های فوق نشان داده شده است، هرچه میزان نرخ یادگیری افزایش یافته و به یک نزدیک می‌شود، سرعت کاهش سلول‌های سرطانی کاهش می‌یابد.

۵- نتیجه‌گیری

در این مقاله تعیین بهینه دوز دارو در یک مدل ریاضی دارای تاخیر زمانی از بیمار مبتلا به سرطان ملانوما با استفاده از روش



شکل۲۸ : تاثیر αهای مختلف در کاهش سلول‌های سرطانی توسط روش مسیرهای شایستگی



شکل۲۹ : تاثیر αهای مختلف در کاهش سلول‌های سرطانی توسط روش الگوریتم یادگیریQ

مسیرهای شایستگی صورت گرفت. روش مسیرهای شایستگی یکی از روش‌های حل مساله یادگیری تقویتی می‌باشد. این نسبت به روش‌های مرسوم حل مساله یادگیری تقویتی همانند یادگیری تفاوت‌گذرا و مونت کارلو، سرعت و دقت بالاتری دارد. در اکثر روش‌های مبتنی بر سیگنال پاداش و در عین حال تعامل با محیط، ابتدا دوز داروی بهینه با استفاده از یک مدل ریاضی از بیمار تعیین می‌شود و پس از این، برنامه دارویی در عمل با سعی و خطای جزئی برای بیمار واقعی تعیین می‌گردد. یکی از مواردی مهمی که در تعیین دوز دارو باید مورد توجه قرار گیرد، تعیین میزان بهینه آن می‌باشد. این امر به گونه‌ای انجام می‌گیرد که علاوه بر کنترل جمعیت سلول‌های سرطانی، اثرات زیان‌بار دارو نیز کاسته شود. مدل ریاضی مورد استفاده در این مقاله، مدل ریاضی دارای تاخیر زمانی از بیمار مبتلا به سرطان ملانوما بوده است. که دینامیک بدن بیمار مبتلا به سرطان ملانوما را با توجه به دخیل کردن تاخیر زمانی به خوبی نشان می‌دهد. برای نشان دادن عملکرد بهتر

روش مسیرهای شایستگی در افزایش سرعت کاهش سلول‌های سرطانی، این روش با روش الگوریتم یادگیری Q که یکی از روش‌های حل مسئله یادگیری تقویتی می‌باشد، و همچنین روش‌های کنترل بهینه و تزریق دوز داروی ثابت در هر گام زمانی مقایسه شده است.

در روش مسیرهای شایستگی در مقایسه با دیگر روش‌های استفاده شده، علاوه بر آن‌که جمعیت سلول‌های سرطانی بسیار سریع‌تر به صفر همگرا شد، میزان دوز داروی تزریقی تا زمان حذف کامل سلول‌های سرطانی نیز به میزان چشم‌گیری کاهش یافت. شایان ذکر است که در این مقاله با توجه به بررسی‌های انجام شده، برای اولین بار کنترل جمعیت سلول‌های سرطانی بر روی این مدل ریاضی انجام گرفته است. با وجود اینکه در این مقاله از مدل ریاضی دارای تاخیر از بیمار مبتلا به سرطان ملانوما استفاده شد، اما روش یادگیری تقویتی بر خلاف روش‌های کنترل بهینه، کاملا بی‌نیاز به مدل ریاضی می‌باشد و صرفا جهت شبیه‌سازی رفتار محیط و عدم دسترسی به بیمار واقعی از آن استفاده شده است که این مورد یکی از مزایای عمده روش یادگیری تقویتی می‌باشد. مدل ریاضی استفاده شده نیز با توجه به نظر پزشک انتخاب شده است و نسبت به دیگر مدل‌های ریاضی از سرطان ملانوما کامل‌تر می‌باشد. روش‌های کنترل بهینه بر اساس خطی‌سازی می‌باشند. اما مدل ریاضی استفاده شده در این مقاله غیرخطی می‌باشد که روش انتخابی باز هم عملکرد خیلی خوبی را دارا بوده است. با توجه به کار انجام شده در خصوص کنترل سلول‌های سرطانی در بیماران مبتلا به ملانوما با استفاده از الگوریتم ژنتیک [۷]، این روش برخط نبوده و بسیار زمان‌بر می‌باشد. همچنین در برابر نویز و عدم قطعیت رفتار مناسبی را از خود نشان نمی‌دهد. این در حالی می‌باشد که روش مسیرهای شایستگی دارای خاصیت تطبیق‌پذیری با محیط بوده و در هر لحظه دارای خاصیت یادگیری می‌باشد. برای این کار، با لحاظ عدم قطعیت در پارامترهای سیستم و شرایط اولیه کنترل جمعیت سلول‌های سرطانی در پنج بیمار مبتلا به سرطان ملانوما انجام شد. در نهایت نیز با افزودن نویز به سیستم و عیب در سنسورآن نشان داده شد که روش مسیرهای شایستگی باز هم قادر به کنترل جمعیت سلول‌های سرطانی و رساندن آن‌ها به صفر بوده است. بنابراین این روش نسبت به دیگر روش‌های هوشمند دارای وضعیت بهتری می‌باشد.

بر روی مدل ریاضی استفاده شده در این مقاله آنالیز حساسیت انجام شده است و اعتبارسنجی بر روی آن صورت گرفته است. همچنین روش یادگیری تقویتی یک سیاست دارویی ارایه می‌دهد که در کنار کار پزشکان به تصمیم‌گیری هر چه بهتر آن‌ها برای درمان بیمار و کاهش سعی و خطا در تعیین دوز بهینه دارو کمک خواهد کرد. در این مقاله نیز از این روش برای تزریق دارو به بیمار مبتلا به ملانوما استفاده شد. اما می‌توان برای هر بیمار دیگری نیز از این روش استفاده نمود. همچنین با توجه به سیاست بدست‌آمده می‌توان دستگاهی هوشمند با قابلیت تطبیق‌پذیری ساخت تا با بررسی حالت‌های بیمار در هر لحظه میزان دوز داروی مورد نیاز به بدن وی تزریق گردد.

مراجع

[1] M. Suryapraba, G. Rajanarayanee and P. Kumari, "Analysis of Skin Cancer Classification Using GLCM Based On Feature Extraction in Artificial Neural Network," International Journal of Emerging Technology in Computer Science & Electronics, Vol. 13, 2015.

[2] S. Mazdeyasna, A. H. Jafari, J. Hadjati, A. Allahverdy, and M.Alavi-Moghaddam. "Modeling the Effect of Chemotherapy on Melanoma B16F10 in Mice Using Cellular Automata and Genetic Algorithm in Tapered Dosage of FBS and Cisplatin." Frontiers in Biomedical Technologies 2.2, Vol. 2, No. 2, pp. 103-108, 2015.

[۳] جواد بهار آرا، زهرا طیرانی نجاران، الهه امینی، فرزانه سالک عبداللهی. "اثر مهاری کروسین بر ملانوژنز در سلول‌های رده ملانومای موشی B16F10." ماهنامه علمی پژوهشی دانشگاه علوم پزشکی شهید صدوقی یزد، دوره ۲۴، شماره ۶، ۴۹۰-۴۷۹، شهریور ۱۳۹۵.

[4] X. Wang, S. Lu and J. Guo. "Treatment algorithm of metastatic mucosal melanoma." Chinese clinical oncology 3.3, Vol. 3, No. 3, 2014.

[5] Y. Zheng and Y. Jiang, "mTOR inhibitors at a glance", Molecular and cellular pharmacology, Vol. 7, No. 2, 2015.

[6] U. Sirin, F. Polat and R. Alhajj. "Employing batch reinforcement learning to control gene regulation without explicitly constructing gene regulatory networks." Proceedings of the 23rd International Joint Conference on Artificial Intelligence, 2013.

[۷] الناز کلهر، امین نوری."کنترل سلول‌های سرطانی در بیماران مبتلا به ملانوما با استفاده از الگوریتم ژنتیک و لحاظ اثرات زیان‌بار دارو"، بیست و چهارمین کنفرانس ملی و دومین کنفرانس بین‌المللی مهندسی زیست پزشکی ایران، ۱۰-۸ آذر ۱۳۹۶.

[8] R. Sutton and A. Barto. Reinforcement learning: An introduction, MIT Press, 2011.

[9] A. Noori and M. A. Sadrnia. "Glucose level control using Temporal Difference methods." In Electrical Engineering (ICEE), 2017 Iranian Conference on, pp. 895-900, 2017.

[10] M. De Paula, L. O. Ávila and E. C. Martínez. "Controlling blood glucose variability under uncertainty using reinforcement learning and Gaussian processes." Applied Soft Computing 35, Vol. 35, pp. 310-332, 2015.

[11] G. Czibula, I. M. Bocicor and I. Czibula. "Temporal ordering of cancer microarray data through a reinforcement learning based approach." PloS one 8, Vol. 8, No. 4, 2013.

[12] M. Jacobs. "Personalized Anticoagulant Management Using Reinforcement Learning." Ph. D. dissertation, Dep. of Bioengineering, University of Louisville, 2014.

[13] Padmanabhan, R., Meskin, N., & Haddad, W. M. "Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment." Mathematical biosciences 293, Vol. 293, pp. 11-20, 2017.

[14] A. Noori, M. B. Naghibi Sistani and N. Pariz. "Hepatitis B virus infection control using reinforcement learning", ICEEE 2011.

[15] B. K. Petersen, J. Yang, W. S. Grathwohl, C.Cockrell, C. Santiago, G. An and D. M.Faissol. "Precision medicine as a control problem: Using simulation and deep reinforcement learning to discover

adaptive, personalized multi-cytokine therapy for sepsis:. arXiv preprint arXiv:1802.10440. 2018.

[16] L. Göllmann and H. Maurer.” Optimal control problems with time delays: Two case studies in biomedicine”. Mathematical Biosciences & Engineering, Vol. 15, No. 5, pp. 1137-1154, 2018.

[17] J. Malinzi, R. Ouifki, A. Eladdadi, D. F. Torres and K. A. White. “Enhancement of chemotherapy using oncolytic virotherapy: Mathematical and optimal control analysis”. arXiv preprint arXiv:1807.04329, 2018.

[18] H. Moore. “How to mathematically optimize drug regimens using optimal control”. Journal of pharmacokinetics and pharmacodynamics, Vol. 45. No.1, pp. 127-137, 2018.

[19] A. M. A. Rocha, M. F. P. Costa and E. M. Fernandes. “On a multiobjective optimal control of a tumor growth model with immune response and drug therapies”. International Transactions in Operational Research,Vol. 25, No. 1, pp. 269-294, 2018.

[20] H. Khaloozadeh, P. Yazdanbakhsh and F. Homaei-Shandiz. “The Optimal Dose of Drug in Neoadjuvant Chemotherapy before Surgery for the Patients Suffering from Breast Cancer Stage III”. Iranian Journal of Biomedical Engineering,Vol. 1, No.4, pp. 319-334, 2008.

[21] S. Eikenberry, T. Craig and K. Yang. “Tumor-immune interaction, surgical treatment, and cancer recurrence in a mathematical model of melanoma.”PLoSComputBiol, Vol. 5, No. 4, 2009.

[22] Y. Kogan, A. Zvia and E. Moran. “A mathematical model for the immunotherapeutic control of the Th1/Th2 imbalance in melanoma.” Discrete and Continuous Dynamical Systems Series B, Vol. 18, No.4, pp. 1017-1030, 2013.

[23] L. G. DePillisZ and A. Radunskaya. “A model of dendritic cell therapy for melanoma.”Frontiers in oncology, Vol. 3, 2013.

[24] X. Sun, J.Bao and Y. Shao. “Mathematical modeling of therapy-induced cancer drug resistance: connecting cancer mechanisms to population survival rates.” Scientific reports 6, Vol. 6, No. 22498, 2016.

[25] A. Kłusek, W. Dzwinel and V. Vasilyev. “Supermodeling in simulation of melanoma progression.” Procedia Computer Science, Vol. 80, pp. 999-1010, 2016.

[26] A. Isabel. “On the geometric modulation of skin lesion growth: a mathematical model for melanoma.” Research on Biomedical Engineering AHEAD, Vol. 32, No. 1, pp. 2446-4740, 2016.

[27] M. Pennisi. “A mathematical model of immune-system-melanoma competition.” Computational and mathematical methods in medicine, Vol. 2012, No. 850754, 2012.

[28] H. Gholizade-Narm and A. Noori. “Control the population of free viruses in nonlinear uncertain HIV system using Q-learning.” International Journal of Machine Learning and Cybernetics, Vol. 9, No. 7, pp. 1169-1179, 2017.

[29] N. A. Alias, Linear Quadratic Regulator (LQR) controller design for Inverted Pendulum, Ph. D. dissertation, Universiti Tun Hussein Onn Malaysia, 2013.

[30] F. Farivar, M. N. Ahmadabadi. "Continuous reinforcement learning to robust fault tolerant control for a class of unknown nonlinear systems." Applied Soft Computing. Vol. 37, pp. 702-714, 2015.

[31] M. Jin and J. Lavaei, "Stability-certified reinforcement learning: A control-theoretic perspective", arXiv preprint arXiv:1810.11505, 2018.

[32] C. Tessler, Y. Efroni, Y. and S. Mannor, "Action Robust Reinforcement Learning and Applications in Continuous Control", arXiv preprint arXiv:1901.09184, 2019.

[33] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees". In Advances in neural information processing systems, pp. 908-918, 2017.

[34] J. Morimoto and K. Doya, "Robust reinforcement learning". Neural computation, Vol. 17, No. 2, pp. 335-359, 2005.

[35] C.Szepesvári, "The asymptotic convergence-rate of Q-learning". In Advances in Neural Information Processing Systems. pp. 1064-1070, 1998.

[36] B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu and L. Song, "SBEED: Convergent reinforcement learning with nonlinear function approximation", arXiv preprint arXiv:1712.10285. 2017.

[37] A. Geramifard, M. Bowling, M. Zinkevich and R. S. Sutton, "iLSTD: Eligibility traces and convergence analysis", In Advances in Neural Information Processing Systems, pp. 441-448, 2007.
H. Yu, "On convergence of emphatic temporal-difference learning", In Conference on Learning Theory, pp. 1724-1751, 2015.

پاورقی‌ها:

[1] Mazdeyasna et al
[2] Wang et al
[3] mechanism Target of Rapamycin
[4] Sirin et al
[5] Batch Reinforcement Learning
[6] Q-learning
[7] De Paula et al
[8] Padmanabhan et al
[9] Petersen et al
[10] sepsis
[11] Göllmann et al
[12] Malinzi et al
[13] Helen Moore
[14] Rocha et al
[15] Eligibility Traces
[16] Reinforcement Learning
[17] Marzio Pennisi
[18] Injection point compartment
[19] Skin compartment
[20] Monte Carlo
[21] Temporary Differences Learning
[22] Linear Quadratic Regulator
[23] Naïve CTL
[24] Lipschitz continues
[25] white Gaussian noise