Iranian Journal of Biosystem Engineering ISNN: 2423-7841
Homepage: http://ijswr.ut.ac.ir

Smart Control of a Microrobot for Navigation on Fluid Surface and Simulation of its Application in Microplastics Removal

Amar Salehi[1] | Soleiman Hosseinpour[2]✉ | Nasrollah Tabatabaei[3] | Mahmoud Soltani Firouz[4]

1. Department of Agricultural Machinery Engineering, Faculty of Agricultural, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran. E-mail: amar.salehi@ut.ac.ir
2. Corresponding Author, Department of Agricultural Machinery Engineering, Faculty of Agricultural, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran. E-mail: shosseinpour@ut.ac.ir
3. Department of Medical Nanotechnology, School of Advanced Technologies in Medicine, Tehran University of Medical Sciences, Tehran, Iran. E-mail: ntabatabaei@tums.ac.ir
4. Department of Agricultural Machinery Engineering, Faculty of Agricultural, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran. E-mail: mahsoltani@ut.ac.ir

Article Info

Article type: Research Article

Article history:

Received: Oct. 17, 2023

Revised: Jan. 2, 2024

Accepted: Jan. 7, 2024

Published online: Summer 2023

Keywords:
Control,
Deep reinforcement learning,
Microplastic,
Microrobot.

ABSTRACT

Microplastic contamination of food and beverages has become a global concern in recent years. As a novel approach, magnetic microrobots offer promising potential to address microplastic separation and degradation. However, achieving precise, intelligent, and automated navigation control for microrobots in such tasks remains a significant challenge. This level of control is typically achieved by modeling the complex dynamics of microrobots, the environment, and the actuation system. In this study, an alternative approach was presented using a model-free deep reinforcement learning algorithm (DRL) to navigate a magnetic microrobot on fluid surfaces. In order to simulate the process of reaching a microplastic particle on the fluid surface, the DRL system was implemented to train the microrobot to autonomously navigate from an initial position within the real-world environment to a specified target position. A magnetic actuation system based on two permanent magnets and one-axis Helmholtz coils was constructed to manipulate the position of the microrobot. During the training phase, the microrobot demonstrated high accuracy and speed in achieving the desired position. The evaluation results of the trained model also confirmed the microrobot's success in all episodes, with an average reward of 39.02 out of 40 and a standard deviation of 0.71. These findings indicate that the control system could effectively learn an optimal policy by employing DRL without any prior knowledge of environmental dynamics or the actuation system.

Cite this article: Salehi, A., Hosseinpour, S., Soltani Firouz, M., & Tabatabaei, N. (2023). Smart Control of a Microrobot for Navigation on Fluid Surface and Simulation of its Application in Microplastics Removal, *Iranian Journal of Biosystem Engineering*, 54 (3), 75-94. https://doi.org/10.22059/ijbse.2024.366451.665527
© The Author(s).
Publisher: The University of Tehran Press.
DOI: https://doi.org/10.22059/ijbse.2024.366451.665527

مجله مهندسی بیوسیستم ایران، دوره ۵۴، شماره ۳ شاپا: ۷۸۴۱-۲۴۲۳
Homepage: http://ijbse.ut.ac.ir

کنترل هوشمند میکروربات به منظور ناوبری روی سطح سیال و شبیه‌سازی کاربرد آن برای از بین بردن میکروپلاستیک‌ها

عمار صالحی [۱] | سلیمان حسین‌پور [۲]✉ | نصرالله طباطبائی [۳] | محمود سلطانی فیروز [۴]

۱. گروه مهندسی ماشین‌های کشاورزی، دانشکده کشاورزی، دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: amar.salehi@ut.ac.ir

۲. نویسنده مسئول، گروه مهندسی ماشین‌های کشاورزی، دانشکده کشاورزی، دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: shosseinpour@ut.ac.ir

۳. گروه نانوفناوری پزشکی، دانشکده فناوری‌های نوین پزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران. رایانامه: ntabatabaei@tums.ac.ir

۳. گروه مهندسی ماشین‌های کشاورزی، دانشکده کشاورزی، دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران. رایانامه: mahsoltani@ut.ac.ir

اطلاعات مقاله

نوع مقاله: مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۲/۷/۲۵
تاریخ بازنگری: ۱۴۰۲/۱۰/۱۲
تاریخ پذیرش: ۱۴۰۲/۱۰/۱۷
تاریخ انتشار: تابستان ۱۴۰۲

واژه‌های کلیدی:
کنترل،
میکروپلاستیک،
میکروربات،
یادگیری تقویتی عمیق.

چکیده

در سال‌های اخیر میکروپلاستیک‌های موجود در مواد غذایی و آشامیدنی به یک معضل جهانی تبدیل شده‌اند. میکرورбات‌های مغناطیسی به‌عنوان یک رویکرد نوین در حل این مشکل، پتانسیل خوبی برای جداسازی و از بین بردن میکروپلاستیک‌ها نشان داده‌اند. بااین‌حال، هدایت و ناوبری خودکار، هوشمند و دقیق میکروربات‌ها برای اجرای چنین وظایفی، همچنان یک چالش اصلی محسوب می‌شود. روش‌های مرسوم برای دستیابی به چنین سطحی از کنترل، اغلب به مدل‌سازی‌های پیچیده‌ای از دینامیک میکروربات، محیط و سیستم تحریک نیاز دارند. به‌عنوان یک رویکرد جایگزین، در این پژوهش یک سیستم کنترل مبتنی بر الگوریتم یادگیری تقویتی عمیق بدون مدل برای کنترل میکروربات مغناطیسی روی سطح سیال ارائه شد. هدف سیستم، آموزش میکروربات برای هدایت آن از یک نقطه در محیط واقعی به سمت موقعیت هدف بود تا فرآیند ناوبری به سوی موقعیت میکروپلاستیک شناور روی سطح سیال شبیه‌سازی شود. برای کنترل موقعیت میکروربات، یک سیستم تحریک مغناطیسی شامل دو آهنربای ثابت و یک سیم‌پیچ هلمهولتز تک‌محوره ساخته شد. نتایج آموزش میکروربات نشان داد که میکروربات توانست با دقت و سرعت بالایی به موقعیت هدف برسد. نتایج ارزیابی مدل آموزش‌یافته نیز حاکی از موقعیت میکروربات در رسیدن به نقطه هدف با میانگین پاداش ۳۹/۰۲ از ۴۰، و انحراف معیار ۰/۷۱ در تمام اپیزودها بود. این نتایج نشان می‌دهد که سیستم کنترل مبتنی بر الگوریتم یادگیری بدون داشتن هیچ‌گونه دانش قبلی از دینامیک محیط یا سیستم تحریک، یک سیاست بهینه را با استفاده از تعامل با محیط برای هدایت میکروربات کشف کرد.

استناد: صالحی، عمار؛ حسین‌پور، سلیمان؛ طباطبائی، نصرالله؛ و سلطانی فیروز، محمود (۱۴۰۲). کنترل هوشمند میکروربات به منظور ناوبری روی سطح سیال و شبیه‌سازی کاربرد آن برای از بین بردن میکروپلاستیک‌ها ، *مجله مهندسی بیوسیستم ایران*، ۵۴ (۳)، ۹۴-۷۵.
https://doi.org/10.22059/ijbse.2024.366451.665527

CC BY NC

ناشر: مؤسسه انتشارات دانشگاه تهران.
© نویسندگان.
DOI: https://doi.org/10.22059/ijbse.2024.366451.665527

مقدمه

در سال‌های اخیر، سیستم‌های میکروباتیک در زمینه‌های علمی متنوعی از پزشکی تا صنایع غذایی رشد قابل‌توجهی داشته و به دلیل قابلیت‌ها و کاربردهای وسیع آن‌ها در ابعاد میکرو، توجهات زیادی را در جامعه علمی به خود جلب کرده‌اند (Dan et al., 2022; Nauber et al., 2023). میکروربات‌ها با توجه به ساختار بسیار ریزشان این امکان را برای دانشمندان فراهم ساخته‌اند که به نقاطی بسیار ریز و غیرقابل‌دسترس، مخصوصاً در فرآیندهای پزشکی، دسترسی داشته باشند (Agrahari et al., 2020). در حوزه کشاورزی نیز برخی از پژوهشگران از میکروربات‌ها برای کاربردهایی نظیر استحصال قند و تولید سوختِ زیستی از درختان بدون نیاز به قطع کردن یا آسیب رساندن به آن‌ها، انتقال ژن به ذرت و غیره استفاده کرده‌اند. به عنوان نمونه، در یک پژوهش پیشگام، از میکروربات‌ها برای برداشت مواد شیمیایی با ارزش از سلول‌های گیاهی به صورت غیرمخرب استفاده شده است (Bae et al., 2021). در صنایع غذایی نیز میکروربات‌ها برای کاربردهای متنوعی هم‌چون گیرانداختن سلول‌های مخمر در داخل نوشیدنی‌ها (Villa et al., 2020)، جمع‌آوری میکروارگانیسم‌ها (مانند باکتری‌ها) در آب سیب (Campuzano et al., 2012)، خارج کردن باکتری‌های مضری چون استافیلوکوک اورئوس از شیر (Mayorga-Martinez et al., 2012) و غیره به کار گرفته شده‌اند. در این میان، اخیراً میکروربات‌ها پتانسیل هیجان‌انگیزی در پاک‌سازی محیطی و تصفیه آب از فلزات سنگین، آلودگی‌های میکروبی، روغن و غیره از خود نشان داده‌اند (Parmar et al., 2018; Urso et al., 2023; Vilela et al., 2017). با توجه به افزایش نگرانی‌ها درباره معضل جهانی میکروپلاستیک‌ها در آب‌های آشامیدنی، زنجیره‌های غذایی و محصولات آرایشی و بهداشتی که باعث ایجاد صدمات فیزیکی و بیماری‌های جبران‌ناپذیری در انسان و سایر جانداران می‌شوند (Leslie et al., 2022; Wright et al., 2013)، استفاده از میکروربات‌ها به‌عنوان یک رویکرد پایدار و مؤثر در جداسازی و از بین بردن میکروپلاستیک‌ها، افق‌های جدیدی را برای حل این معضل گشوده است (Zhou et al., 2021).

بااین‌حال، یکی از چالش‌هایی که این رویکرد برای حل معضل میکروپلاستیک‌ها با آن روبروست، کنترل و ناوبری خودکار، هوشمند و دقیق به سمت موقعیت هدف است (Medina & Schmidt, 2017; Yang et al., 2018) که نقش مهمی در به‌کارگیری عملی آن‌ها در چنین کاربردهایی دارد. هدف یک سیستم کنترل برای میکروربات، دست‌کاری شکل و اندازه میدان انرژی تحریک‌کننده به‌منظور حرکت دادن میکروربات برای دستیابی به یک رفتار دینامیکی خاص است. به این منظور، لازم است که میکروربات‌ها دارای قابلیت‌ها و ابزارهایی جهت حس کردن محیط اطراف، اتخاذ تصمیم و پیشرانش باشند. اگرچه این قابلیت‌ها و ابزارها در ربات‌های خودران در ابعاد ماکرو به‌طور گسترده از طریق ادغام و ترکیب ابزارهای نرم‌افزاری و سخت‌افزاری به‌خوبی در دسترس هستند (Lynch & Park, 2017; Salmani et al., 2018) اما در ابعاد میکرو به دلیل اندازه بسیار کوچک، امکان ادغام ابزارهای سخت‌افزاری نظیر حسگرها، منابع تأمین توان، ابزارهای ارتباطی و غیره وجود ندارد (Sitti, 2017).

میکروربات‌های مغناطیسی که به‌صورت از راه دور از طریق میدان مغناطیسی تحریک می‌شوند یکی از گزینه‌های مناسب و سازگار با محیط‌زیست برای حل مشکل پیشرانش در ابعاد میکرو محسوب می‌شوند. به دلیل امن بودن استفاده از این میکروربات‌ها مخصوصاً برای کاربردهای پزشکی، طیف گسترده‌ای از مکانیزم‌های ناوبری و شنا در محیط سیال در کنار عملکردهای مکملی همچون تحویل سلول، دارو و ژن (Choi et al., 2021)، بافت‌برداری (Sehyuk Yim et al., 2014) و تشخیص بیماری (Wang et al., 2020) را پوشش می‌دهد.

از سوی دیگر کنترل حرکت و ناوبری کارآمد و خودکار میکروربات‌های مغناطیسی در محیط سیال مستلزم استفاده از روش‌های هوشمند، دقیق و بهینه است. در ابعاد ماکرو، کنترل و ناوبری ربات‌ها اغلب به مدل‌سازی‌ها و شبیه‌سازی‌های بسیار تخصصی از ربات و محیطی که در آن عمل می‌کند نیازمند است (سلمانی زکریا و همکاران، ۱۳۹۷). به طور مشابه، رویکردهای مرسوم برای دستیابی به کنترل و ناوبری میکروربات‌ها نیز اغلب شامل روش‌های تقریبی و تجربی هستند که برای بررسی رفتار میکروربات در محیط‌های مختلف نیاز به مدل‌سازی دینامیکی پیچیده‌ای از کل سیستم، ازجمله دینامیک میکروربات، محیط و تحریک‌کننده دارند. این مدل‌ها اغلب فرضیات زیادی ازجمله مغناطش یکنواخت (Wang et al., 2022)، شکل ایده‌آل (Rahmer et al., 2017) و خطی بودن سیستم (Pawashe et al., 2012) را جهت ساده‌سازی سیستم اعمال می‌کنند که منجر به انحرافات رفتاری بین سیستم فیزیکی و سیستم مدل‌سازی‌شده می‌شوند. اگرچه ترکیب نظریه‌های کنترل مرسوم با الگوریتم‌های مسیریابی پیشرفته، راه‌حل امیدوارکننده‌ای است (Huang et al., 2017) اما این رویکردها نیز با بهره‌گیری از سیستم‌های بازخورد کلاسیک در مواجهه با محیط‌های پیچیده و غیرقابل‌پیش‌بینی با چالش‌های مهندسی بسیاری روبرو هستند (Medina & Schmidt, 2017; Xu et al., 2015).

با پیشرفت‌های خیره‌کننده فناوری‌های هوش مصنوعی و با الهام از رفتار موجودات زنده در یادگیری و سازگاری با محیط اطراف از

طریق تجربیات گذشته، الگوریتم‌های یادگیری تقویتی به‌عنوان یک راه‌حل جایگزین برای چالش کنترل حرکت میکروربات معرفی شده است (Tsang et al., 2020). در رویکرد یادگیری تقویتی به‌عنوان یکی از زیرمجموعه‌های هوش مصنوعی، امکان یادگیری یک عامل[1] از طریق تعامل با محیط واقعی در سیستم‌های مهندسی مصنوعی فراهم می‌شود (Ding et al., 2020). همین ویژگی، وجه تمایز یادگیری تقویتی با سایر الگوریتم‌های یادگیری ماشینِ نظارت‌شده و بدون نظارت است؛ زیرا تمرکز اصلی در این الگوریتم‌ها، مبتنی بر کشف الگوهای پنهان در داده‌های برچسب‌گذاری شده یا بدون برچسب است درحالی‌که هدف اصلی یادگیری تقویتی به‌جای کشف ساختارهای پنهان، بیشینه کردن پاداش یا/و کمینه کردن جریمهٔ سیستم است (Sutton & Barto, 2018). در این رویکرد، پاداش و جریمه نوعی بازخورد از محیط در نظر گرفته می‌شود که به اقدامات و حالاتی که عامل در مراحل زمانی مختلف دارد بستگی خواهد داشت (Ding et al., 2020). انعطاف‌پذیری و سازگاری این رویکرد در مواجهه با سناریوهای کنترلی متنوع بدون نیاز به دانش قبلی از دینامیک محیط، ویژگی ممتازی است که مخصوصاً در محیط‌های پیچیده و غیرقابل‌پیش‌بینی، بسیار راهگشا خواهد بود. با توجه به پیچیدگی ذاتی محیط‌هایی که میکروربات‌ها در آن عمل می‌کنند و هم‌چنین در پرتو نتایج امیدوارکننده و موفقیت‌آمیز الگوریتم‌های یادگیری تقویتی در طیف وسیعی از مسائل کنترل رباتیک (Kober et al., 2013; Singh et al., 2022)، پژوهش‌های بسیاری در سال‌های اخیر با تمرکز بر استفاده از الگوریتم‌های یادگیری تقویتی در کنترل و ناوبری میکروربات‌ها صورت گرفته است (Amoudruz & Koumoutsakos, 2022; Colabrese et al., 2017; Qiu et al., 2020; Yang et al., 2022). بااین‌حال، اغلب پژوهش‌های فعلی در این حوزه عمدتاً مبتنی بر مدل‌سازی‌های محاسباتی و شبیه‌سازی در محیط‌های ساده است؛ زیرا با توجه به اینکه الگوریتم‌های یادگیری تقویتی استراتژی‌های کنترلی را با استفاده از روش آزمون‌وخطا یاد می‌گیرند، به‌کارگیری آن‌ها در دنیای واقعی معمولاً نه‌تنها پرهزینه و زمان‌بر است بلکه بعضاً نگرانی‌هایی درباره ایمنی آن‌ها مطرح شده است (Dulac-Arnold et al., 2021).

در حال حاضر سیستم‌های میکروروباتیک در مراحل ابتدائی توسعه قرار دارند و یک زمینه تحقیقاتی نوظهور با پتانسیل‌های بسیار زیاد محسوب می‌شود. به عنوان یک چشم‌انداز کلی، این ابزارها بیشتر در زمینه پزشکی به کار گرفته شده‌اند اما حتی در این زمینه علمی نیز علی‌رغم پیشرفت‌های خیره کننده، هم‌چنان راه بسیار زیادی برای کاربرد آن‌ها در شرایط in vivo وجود دارد. انتظار می‌رود که در آینده نزدیک، میکروروباتیک راه خود را برای درمان پایدار بسیاری از بیماری‌ها و ناتوانی‌ها، مانند سرطان باز کند. با این توصیفات، می‌توان خاطر نشان ساخت که در حال حاضر اغلب پژوهش‌های حوزه میکروروباتیک به صورت اثبات مفهوم (proof of concept) اجرا می‌گردند؛ زیرا در عمل در محیط واقعی یا حتی آزمایشگاهی با توجه به چالش‌های بسیار زیاد در میکروروباتیک، که در مقاله سانچز و اشمیت (۲۰۱۷) به تفصیل تشریح شده است (Medina & Schmidt, 2017)، امکان اجرا نیافته‌اند. پژوهش حاضر نیز از این قاعده مستثنی نیست و برای اثبات مفهوم اجرا گردیده است؛ در وهله اول به منظور توسعه سیستم کنترل مبتنی بر هوش مصنوعی و در وهله دوم برای شبیه‌سازی فرآیند از بین بردن میکروپلاستیک‌ها در محیط سیال.

در این پژوهش، برای کنترل یک میکروربات دیسکی مغناطیسی در سیستم واقعی در شرایط آزمایشگاهی، یک سیستم تحریک مغناطیسی مبتنی بر میدان‌های مغناطیسی ثابت ساخته شد. پس‌ازآن از یک الگوریتم یادگیری تقویتی عمیق به نام عملگر–منتقد نرم (SAC)[2] برای کنترل میکروربات به‌صورت دوبعدی روی سطح آب استفاده گردید. الگوریتم عملگر–منتقد نرم (Haarnoja et al., 2018) می‌تواند باوجود داده‌های جمع‌آوری‌شدهٔ محدود، استراتژی کنترل را به‌طور خودکار و با سرعت و کارایی بیش‌تری در مقایسه با سایر الگوریتم‌های یادگیری تقویتی یاد بگیرد (Haarnoja et al., 2018) که مشکل پرهزینه بودن و زمان‌بر بودن فرآیند یادگیری میکروربات را حل می‌کند. وظیفه میکروربات رسیدن به یک نقطه هدف برای شبیه‌سازی فرآیند حرکت به‌سوی میکروپلاستیک‌های شناور روی سطح سیال می‌باشد. در این پژوهش، تجزیه‌وتحلیلی برای مدل‌سازی میدان مغناطیسی تولیدشده توسط سیستم تحریک مغناطیسی یا مدل‌سازی دینامیکی میکروربات در محیط سیال انجام نگردید. فرضیه این پژوهش بر توسعه یک سیستم کنترل با کارایی بالا با استفاده از یادگیری تقویتی استوار است که نیازی به توسعه یک مدل تخصصی از سیستم نخواهد داشت.

پیشینهٔ پژوهش

با پیشرفت و توسعه روزافزون الگوریتم‌های یادگیری تقویتی در سال‌های اخیر، پژوهش‌های بسیاری برای کنترل میکروربات‌ها با استفاده

1. Agent
2. Soft Actor-Critic

از این رویکرد انجام شده است. الگوریتم‌های یادگیری تقویتی عمدتاً به دودسته بدون مدل و مدل‌محور تقسیم می‌شوند (Sutton & Barto, 2018). همان‌گونه که از نام این دودسته مشخص است الگوریتم‌های بدون مدل نیازی به یک مدل از پیش طراحی‌شده برای پیش‌بینی وضعیت‌های مختلف محیط ندارند. این مزیت باعث می‌شود که اجرای آن‌ها در مقایسه با الگوریتم‌های مدل‌محور ساده‌تر باشد. الگوریتم‌های یادگیری تقویتی بدون مدل به سه زیرگروه دسته‌بندی می‌شود: روش‌های ارزش‌محور، روش‌های سیاست‌محور و روش‌های عملگر-منتقد. در چارچوب الگوریتم یادگیری تقویتی، مفاهیمی چون ارزش و سیاست به ترتیب به پاداش بلندمدت مورد انتظار تنزیل‌شده[1] و استراتژی اتخاذشده توسط عامل تصمیم‌گیرنده برای اقدام بعدی در پاسخ به وضعیت فعلی اطلاق می‌شود.

یکی از مرسوم‌ترین الگوریتم‌های ارزش‌محور، الگوریتم موسوم به Q-learning است (Watkins, 1989). این الگوریتم، تابع ارزش (Q) را با در نظر گرفتن پاداش‌های فعلی و با بیشینه کردن مقادیر وضعیت آینده به‌روزرسانی می‌کند. پژوهشگران زیادی از این الگوریتم برای کنترل و ناوبری انواع میکروربات‌ها استفاده کرده‌اند (Colabrese et al., 2017; Gustavsson et al., 2017). پژوهشگران از الگوریتم Q-learning برای یادگیری استراتژی‌های کارآمد برای حرکت در جریان ثابت (Qiu et al., 2020)، یادگیری سیاست‌های پیشرانش بهینه توسط میکروشناگر موسوم به سه‌کره‌ای (Tsang et al., 2020) و برنامه‌ریزی مسیر برای ناوبری میکروربات در زمان واقعی (Jiang et al., 2023) استفاده کرده‌اند. بااین‌حال، یکی از معایب این الگوریتم، محدود بودن تعداد وضعیت‌های سیستم است که عملاً تنها در سیستم‌های با فضای عمل گسسته می‌تواند کاربرد داشته باشد (Jiang et al., 2022). این موضوع سبب می‌شود که در سناریوهای دنیای واقعی با فضای عمل پیوسته کارایی لازم را نداشته باشد. برای حل این مشکل، یک روش کنترل یادگیری تقویتی عمیق مبتنی بر Q-learning عمیق پیشنهاد شده است که اخیراً به‌منظور کنترل حرکت میکروربات‌ها در محیط‌های ناشناخته‌ای چون محیط‌های با موانع ثابت و متحرک زیاد (Yang et al., 2020a) و یا رگ‌های خونی شبیه‌سازی‌شده (Yang et al., 2022) مورداستفاده قرار گرفته است.

بااین‌حال، یکی از مهم‌ترین چالش‌های الگوریتم‌های ارزش‌محور در کنترل میکروربات‌ها، تعیین مسیر بهینه سراسری در محیط است که باعث می‌شود عامل‌ها در یک مسیر بهینه محلی گیر کنند (Sutton & Barto, 2018). این چالش، تعمیم‌پذیری این روش برای وضعیت‌ها و محیط‌های مختلف را دچار مشکل می‌کند. از طرف دیگر، این روش در مواجهه با فضاهای وضعیت و عمل بزرگ و پیچیده، به منابع محاسباتی و ذخیره اطلاعات قابل‌توجهی نیاز دارد (Van Hasselt et al., 2016). برای حل این مشکل، روش سیاست‌محور ارائه شده است که از همگرایی و کارایی بیش‌تری در مواجهه با فضاهای عمل پیوسته برخوردار است (Zhang & Yu, 2020). اما بااین‌حال برای دستیابی به عملکردی قابل‌قبول به تعداد زیادی نمونه در تعامل با محیط نیاز دارد. این موضوع مخصوصاً در سیستم‌های واقعی حائز اهمیت است؛ زیرا این روش برای کسب نتیجه و عملکرد بهینه عملاً باید گام‌های یادگیری بسیار زیادی (حتی تا یک‌میلیون گام) به کار گیرد (Mnih et al., 2015; Zhu et al., 2020)، درحالی‌که انجام چنین کاری در سیستم‌های واقعی بسیار چالش‌برانگیز، پرهزینه و زمان‌بر است.

روش عملگر-منتقد، ترکیبی از روش ارزش‌محور و سیاست‌محور است که از مزایای هر دو روش برای حل مشکلات ذکرشده استفاده می‌کند. در این روش، دو فرآیند تعامل عملگر و منتقد وجود دارد که هدف عملگر بهبود سیاست فعلی و هدف منتقد، ارزیابی سیاست فعلی است. پژوهش‌های متعدد کارآمدی این روش در کنترل و ناوبری میکروربات در محیط‌های پیچیده را اثبات کرده‌اند. این عملکرد مناسب عمدتاً به دلیل توانایی آن در مدیریت سیاست‌های کنترل غیرخطی است. تاکنون از این روش برای یادگیری الگوهای حرکتی توسط میکروربات‌ها به‌منظور شنا در یک جهت خاص (Zou et al., 2022)، هدایت میکروربات به سمت ناحیه هدف در محیط‌ها و شرایط مختلف (Yang et al., 2020b) و یادگیری نحوه بهره‌برداری از نیروهای هیدرودینامیکی توسط میکروربات برای حرکت در جهت خاص (Borra et al., 2022) استفاده شده است.

این پژوهش‌ها عمدتاً بر اساس مدل‌سازی عددی میکروربات‌ها و محیط پیرامون آن‌ها صورت گرفته‌اند و درنتیجه با توجه به ماهیت ساده‌سازی این نوع مدل‌سازی‌ها، در محیط‌های واقعی و با شرایط غیرقابل‌پیش‌بینی با چالش‌های بسیاری روبرو خواهند بود؛ بنابراین با توجه به چالش‌های ذکرشده، الگوریتمی که بتواند عملکرد مناسبی در کنترل حرکت و ناوبری میکروربات‌ها در محیط واقعی داشته باشد ضروری است. الگوریتم عملگر-منتقد نرم (Haarnoja et al., 2018) که یکی از محبوب‌ترین الگوریتم‌های روش عملگر-منتقد است،

1. Discounted expected long-term reward

اخیراً نتایج امیدوارکننده‌ای در کنترل میکروربات‌ها در محیط واقعی از خود نشان داده است. به‌عنوان‌مثال، بهرنس[1] و رودر[2] (۲۰۲۲) از این الگوریتم برای کنترل حرکت یک میکروربات مغناطیسی مارپیچ هوشمند استفاده کردند (Behrens & Ruder, 2022). این میکروربات از طریق یک سیستم الکترومغناطیسی چرخان غیرخطی تحریک می‌شد و می‌توانست رفتارهای شنای بهینه را به‌طور مستقل در محیط سیال بدون نیاز به دانش قبلی از محیط یاد بگیرد که درنهایت منجر به کاهش زمان و منابع لازم جهت توسعه سیستم می‌شد. هدف عامل یادگیری تقویتی، یادگیری یک سیاست عملی بود که پاداش کلی را برای هدایت میکروربات در یک محیط کاری دایره‌ای به سمت موقعیت هدف به حداکثر برساند. این پژوهش نشان داد که الگوریتم عملگر-منتقد نرم به‌خوبی توانسته است سیاستِ تقریباً بهینه برای حرکت دایره‌ای میکروربات حول یک نقطه ثابت در محیط سیال را یاد بگیرد (Behrens & Ruder, 2022). بااین‌حال یکی از محدودیت‌های این پژوهش، بروز رفتارها و اقدامات غیر بهینه (حرکت منفی به سمت عقب) توسط میکروربات است که عملکرد کلی سیستم را تحت‌الشعاع قرار می‌دهد. در پژوهش دیگری که توسط کای[3] و همکاران (۲۰۲۲) انجام شده است از الگوریتم عملگر-منتقد نرم برای کنترل میکروربات مغناطیسی نرم به‌منظور رسیدن به نقطه هدف (به‌منظور شبیه‌سازی فرآیند تحویل دارو) و شناور شدن درون لوله محتوی سیال (به‌منظور شبیه‌سازی فرآیند آزادسازی دارو) استفاده کردند (Cai et al., 2022). در این پژوهش، تاریخچهٔ وضعیت-عمل میکروربات و نرخ جریان تخمینی به‌عنوان بخشی از ورودی برای آموزش در نظر گرفته شد. در این پژوهش، ابتدا میکروربات در یک محیط شبیه‌سازی‌شده آموزش یافت و سپس بدون هیچ فرآیند آموزشی اضافی و تنظیم پارامتر در محیط واقعی به کار گرفته شد. نتایج ارزیابی الگوریتم نشان داد که میکروربات‌ها توانستند وظیفه از پیش تعیین‌شده را با موفقیت تکمیل کنند (Cai et al., 2022). بااین‌حال، این پژوهش همچنان به سطحی از دانش تخصصی از دینامیک میکروربات و محیط نیاز دارد. ضمن این‌که همچنان فرآیند آموزش میکروربات در محیط شبیه‌سازی‌شده و با پیش‌فرض‌های زیادی جهت ساده‌سازی انجام شده است.

باید توجه داشت که پژوهش‌های انجام شده با تمرکز بر کاربرد الگوریتم‌های یادگیری تقویتی در محیط واقعی میکروروباتیک بسیار محدود هستند. با بررسی منابع علمی در این زمینه در میان بیش از ۲۳ پژوهش شناسایی‌شده که از الگوریتم‌های یادگیری تقویتی برای کنترل و ناوبری میکروربات استفاده کرده‌اند، تنها در ۳ پژوهش، الگوریتم‌های یادگیری تقویتی در محیط واقعی آزمایشگاهی اجرا شده است (Behrens & Ruder, 2022; Cai et al., 2022; Jiang et al., 2023). از طرف دیگر تمام این پژوهش‌ها از سیستم الکترومغناطیسی برای تحریک میکروربات استفاده کرده‌اند که با توجه به محدودیت‌های الکتریکی آن، اغلب تحت تأثیر میدان مغناطیسی شدید و محدودیت گرادیان میدان قرار دارد (Lee & Ha, 2017). سیستم‌های تحریک الکترومغناطیسی همچنین ازنظر ابعاد دارای محدودیت هستند که در محیط‌های کاری بزرگ یک چالش مهم محسوب می‌شود. همچنین استفاده از آن‌ها، مخصوصاً برای آموزش میکروربات‌ها که مستلزم گام‌های یادگیری نسبتاً زیادی است با مصرف انرژی الکتریکی و تولید گرمای قابل‌توجهی همراه هست.

استفاده از میکروربات‌ها به‌منظور از بین‌بردن و یا جداسازی میکروپلاستیک‌ها در محیط سیال در پژوهش‌های مختلفی گزارش شده است. به‌عنوان‌مثال، وانگ[4] و همکاران (۲۰۱۹) از میکروربات‌های حساس به نور برای جداسازی میکروپلاستیک‌های استخراج شده از محصولات آرایشی و همچنین نمونه‌های آب دریای بالتیک استفاده کردند (Wang et al., 2019). همچنین میکروربات‌های نوری و مغناطیسی برای حذف و تجزیه میکروپلاستیک‌های موجود در یک فضای پیچیده با کانال پرپیچ‌وخم مورد استفاده قرار گرفتند (Beladi-Mousavi et al., 2021). در پژوهش‌های دیگری از میکروربات‌ها برای جداسازی و حتی تجزیه میکروپلاستیک‌ها استفاده شده است (Sun et al., 2020; Zhou et al., 2021). علاوه بر این‌ها، اگرچه پژوهش‌های دیگری با انواع مختلفی از میکروربات‌ها برای تجزیه یا جداسازی میکروپلاستیک‌ها گزارش شده‌اند، اما فرآیند کنترل در اغلب آن‌ها یا از طریق یک اپراتور متخصص و یا با استفاده از سیستم‌های کنترل کلاسیک انجام شده است (Urso et al., 2021; Zhou et al., 2021). با بررسی منابع علمی، پژوهش مشابهی که از سیستم تحریک مغناطیسی ثابت و الگوریتم‌های یادگیری تقویتی برای کنترل میکروربات جهت رسیدن به نقطه هدف به‌منظور شبیه‌سازی فرآیند حرکت به‌سوی میکروپلاستیک‌های شناور روی سطح سیال در محیط واقعی آزمایشگاهی (و یا حتی شبیه‌سازی‌شده) استفاده کرده باشد یافت نگردید.

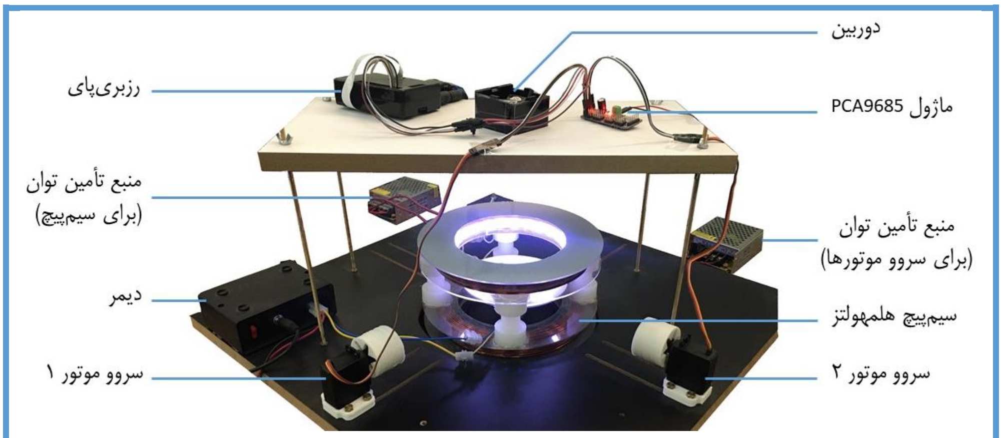
1. Behrens
2. Ruder
3. Cai
4. Wang

روش‌شناسی پژوهش

میکروربات و سیستم تحریک مغناطیسی

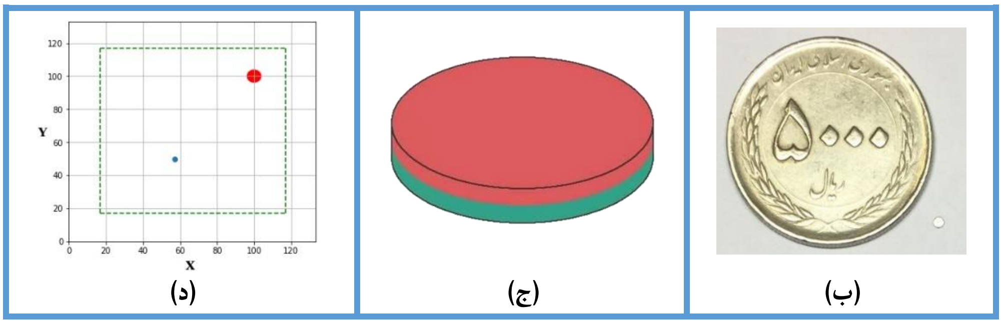
در این پژوهش به‌منظور آزمایش کارایی سیستم یادگیری تقویتی برای کنترل میکروربات مغناطیسی، یک سیستم تحریک مغناطیسی ثابت با الهام از پژوهش‌های (Khalesi et al., 2022; Yousefi & Nejat Pishkenari, 2021) ساخته شد (شکل ۱–الف). میکروربات استفاده شده در این پژوهش از جنس آلیاژ نئودیمیوم (NdFeB) گرید N45 با سه لایه محافظتی از جنس نیکل و کبالت (ساخت شرکت Webcraft آلمان) و به شکل یک دیسک با شعاع ۷۵۰ میکرومتر و ضخامت ۵۰۰ میکرومتر است (شکل ۱–ب) که مغناطش آن به‌صورت محوری (عمود بر سطح دیسک) است (شکل ۱–ج). سیستم تحریک مغناطیسی شامل دو آهنربای دائمی چرخان و یک جفت سیم‌پیچ موسوم به هلمهولتز[1] تک‌محوره است. آهنرباهای چرخان به شکل مکعب و با اندازه $20 \times 20 \times 20\ mm^3$ هستند که از آلیاژ نئودیمیوم (NdFeB) گرید N42 با سه لایه محافظتی از جنس نیکل و کبالت ساخته شده‌اند. از دو سروو موتور MG996R با محدوده زاویهٔ کاری ۱۸۰ درجه و دقت $0.19\ s/60°$ در ولتاژ کاری 4.8 V و $0.15\ s/60°$ در ولتاژ کاری 6 V ساخت شرکت tower pro چین برای چرخاندن آهنرباها استفاده شد (TowerPro, 2023). اتصال آهنرباها به سروو موتورها از طریق یک بستر مخصوص انجام گردید که با پرینتر سه‌بعدی مدل Ender-3 Pro و فیلامنت PLA ساخته شد. ازآنجاکه آهنرباها دارای دو قطب مثبت و منفی در جهت محوری و چگالی شار میدان مغناطیسی معادل ۱/۳۲ تسلا هستند (Magnetics, 2023)، تغییر زاویه آن‌ها باعث تغییر میدان و گرادیان در فضای کاری و درنتیجه تغییر مکان میکروربات شناور روی سطح سیال در صفحه xy با دو درجه آزادی می‌شود. نظر به اینکه هر آهنربا یک درجه آزادی دورانی دارد، استفاده از دو آهنربا به‌صورت عمود بر هم به‌طوری‌که خط واصل بین مکان هر آهنربا و مرکز فضای کاری، محور دوران هرکدام از آن‌ها باشد، باعث ایجاد یک سیستم کنترل کاملاً تحریک‌شده[2] خواهد شد. سروو موتورها و آهنرباهای متصل به آن‌ها در فاصله ۱۷ سانتی‌متری از مرکز فضای کاری قرار داده شدند.

از طرف دیگر، برای چرخاندن سروو موتورها از یک بورد رزبری‌پای مدل 3B V1.2 استفاده شد. سروو موتورها از طریق یک ماژول درایور ۱۶ کاناله ۱۲ بیتی PWM با تراشه PCA9685 به رزبری‌پای متصل شدند. ماژول درایور از یک طرف به یک منبع تأمین توان DC با ولتاژ ۵ ولت و جریان ۵ آمپر و از طرف دیگر از طریق کانکتور GPIO و پروتکل ارتباطی I2C به بورد رزبری‌پای متصل شد. ازآنجاکه میکروربات‌ها تنها به‌صورت دوبعدی و در صفحه xy حرکت می‌کنند از یک سیم‌پیچ هلمهولتز تک محوره برای ایجاد یک میدان مغناطیسی یکنواخت قوی برای هم‌تراز کردن میکروربات و درنتیجه عمود کردن گشتاور مغناطیسی میکروربات در جهت عمود بر فضای کاری (در جهت z) استفاده شد.



(الف)

1. Helmholtz
2. Fully-actuated



شکل ۱. سیستم فعال‌سازی مغناطیسی و میکروربات: (الف) سیستم فعال‌سازی مغناطیسی شامل یک جفت سیم‌پیچ هلمهولتز و دو آهنربای دائمی، (ب) اندازه میکروربات در مقایسه با سکه ۵۰۰۰ ریالی، (ج) جهت محور مغناطش میکروربات. رنگ قرمز و سبز به ترتیب نشان‌دهندهٔ قطب N و S میکروربات است، (د) فضای کار میکروربات در یک صفحهٔ شبکه‌ای؛ جسم قرمز نشان‌دهنده موقعیت میکروپلاستیک (هدف)، جسم آبی نشان‌دهنده موقعیت میکروربات و خطچین سبزرنگ نشان‌دهنده حاشیه فضای کاری است.

سیم‌پیچ هلمهولتز آرایش خاصی از دو سیم‌پیچ دایره‌ای یکسان است که به موازات یکدیگر قرار گرفته‌اند و با فاصله‌ای برابر با شعاع آنها از هم جدا شده‌اند (شکل ۱-الف). این پیکربندی یک میدان مغناطیسی یکنواخت و قوی بین سیم‌پیچ‌ها ایجاد می‌کند. در سیم‌پیچ‌های هلمهولتز، برای آن‌که جریان عبوری در سیم‌پیچ‌ها برابر باشد باید آن‌ها را به‌صورت سری به یکدیگر متصل کرد. به این منظور از یک منبع تأمین توان DC با حداکثر ولتاژ ۱۲ ولت و حداکثر جریان ۵ آمپر استفاده شد و سیم‌پیچ‌ها نیز به‌صورت سری به یکدیگر متصل شدند. در این پژوهش، قطر متوسط سیم‌پیچ‌ها ۱۵ سانتی‌متر انتخاب شد و هر یک از آن‌ها از ۱۶۶ دور سیم نازک مسی با قطر ۱/۳ میلی‌متر تشکیل شده‌اند که به دور یک بوبین شفاف از جنس پلکسی‌گلاس سیم‌پیچی شده‌اند. طبق رابطه بیوساوار اندازه‌ی میدان مغناطیسی در سیم‌پیچ هلمهولتز تابع تعداد دور، جریان سیم‌پیچ و شعاع سیم‌پیچ است (رابطهٔ ۱):

$$B_Z(Z)\Big|_{|Z|\le\frac{a}{2}} \approx \frac{8\mu_0 NI}{5\sqrt{5}a}$$ رابطهٔ ۱)

که در آن a فاصله‌ی دو سیم‌پیچ از یکدیگر بر حسب متر، I جریان دو سیم‌پیچ بر حسب آمپر، N تعداد دور سیم‌پیچ، Z فاصله بین نقطه واصل مرکز دو سیم‌پیچ و $\mu 0$ ضریب گذردهی خلا برحسب تسلا در متر بر آمپر است.

با این پیکربندی و با توجه به رابطهٔ ۱، این سیم‌پیچ هلمهولتز تک‌محوری قادر به تأمین میدان مغناطیسی حداکثر ۹/۶ میلی‌تسلا خواهد بود (Khalesi et al., 2022; Yousefi & Nejat Pishkenari, 2021).

فضای کاری سیستم شامل یک ظرف پتری دیش با قطر ۶ سانتی‌متر است که در مرکز سیم‌پیچ و بین آن‌ها قرار گرفته است. با توجه به روند طولانی فرآیند یادگیری میکروربات و احتمال تبخیر آب، از آب دیونیزه برای جلوگیری از تشکیل حباب‌های هوا در پتری دیش استفاده شد. به‌این‌ترتیب سطح مشترک آب-هوا به‌عنوان یک محدودیت فیزیکی برای اثبات مفهوم در نظر گرفته شد. سطح مشترک آب-هوا باعث می‌شود که نیروی کشش سطحی، جهت‌گیری مغناطیسی میکروربات را به‌صورت عمودی محدود کند و نیروها را در جهت z متعادل نماید و درنتیجه با توجه به وزن بسیار ناچیز میکروربات (۰/۰۰۶۷ گرم)، آن را روی سطح آب شناور نگه دارد (Dong & Sitti, 2020).

پردازش تصویر جهت تشخیص میکروربات

برای تشخیص میکروربات و استخراج موقعیت آن در صفحه xy، فرآیند پردازش تصویر صورت گرفت. به این منظور از یک دوربین با نرخ ۳۰ فریم بر ثانیه و رزولوشن ۱۰۸۰×۱۹۲۰ در بالای فضای کاری استفاده شد. دوربین به یک لپ‌تاپ Acer Aspire 3 A315-57G-77K6 با پردازنده‌ی مرکزی i7 اینتل مدل 1065G7 و ۸ گیگابایت رم متصل بود. ازآنجاکه نوردهی مناسب یکی الزامات پردازش تصویر برای عملکرد مناسب تشخیص اشیاء محسوب می‌شود، از یک LED نواری ۱۵ وات و یک دیمر برای تنظیم شدت نور استفاده شد.

فرآیند پردازش تصویر با استفاده از کتابخانه OpenCV و با زبان برنامه‌نویسی پایتون نسخه 3.10.9 در نرم‌افزار Spyder نسخه 5.3.3 انجام گرفت. در مرحله اول پس از دریافت مداوم فریم‌ها از دوربین، آخرین فریم دریافتی برش داده شد و سپس وارد مرحله پردازش تصویر رنگی گردید. در این مرحله از تکنیک تبدیل تصویر به مقیاس خاکستری برای سهولت کار با تصویر استفاده شد؛ زیرا تصاویر در مقیاس خاکستری به محاسبات کمتری نیاز دارند. سپس تصویر خاکستری با استفاده از روش آستانه تطبیقی به یک تصویر باینری که در آن مقادیر پیکسل صفر (سیاه) یا ۲۵۵ (سفید) است، تبدیل شد. روش آستانه تطبیقی برای کنترل تغییرات در شرایط نوری استفاده می‌شود

که هدف از آن جداسازی میکروربات از پس‌زمینه و نویز است. در مرحله بعد، از دو عملیات پایه مورفولوژی ریاضی به نام‌های سایش و گسترش برای تصحیح و تمیز کردن تصویر باینری استفاده شد. عملگر سایش، نواحی سفید در تصویر باینری را کوچک می‌کند که منجر به حذف نویزهای ریز می‌شود. همچنین عملگر گسترش، با گسترش نواحی سفید و صاف کردن مرزهای میکروربات، منجر به تشخیص دقیق‌تر آن می‌شود. این فرآیند با پیدا کردن کانتورها[1] که نشان‌دهنده مرزهای اجزای متصل در تصویر باینری هستند ادامه یافت. کانتورها برای شناسایی شکل و موقعیت جسم در تصویر بسیار مهم هستند. بزرگ‌ترین کانتور که برجسته‌ترین شیء در تصویر است برای تشخیص دقیق میکروربات به کار گرفته شد. مختصات مرکز دایره محصور، اطلاعات موردنیاز در مورد موقعیت میکروربات را فراهم می‌کند. این داده برای رسم دایره در اطراف میکروربات و محاسبه موقعیت آن ضروری است. درنهایت مختصات پیکسل مرکز جسم میکروربات با توجه به تنظیم موقعیت مبدأ مختصات، محاسبه گردید.

الگوریتم یادگیری تقویتی

یادگیری تقویتی زیر مجموعه‌ای از هوش مصنوعی و یادگیری ماشین است که در آن یک عامل یاد می‌گیرد که از طریق تعامل با محیط اقدام به تصمیم‌گیری کند. در این حالت عامل، بازخوردی را در قالب پاداش یا تنبیه دریافت می‌کند و به او امکان می‌دهد رفتار یا سیاست بهینه را در طول زمان بیاموزد. هدف از یادگیری تقویتی این است که عامل یاد بگیرد اقداماتی را انجام دهد که پاداش انباشته دریافتی را به حداکثر برساند. به‌طورکلی، رویکرد یادگیری تقویتی نیازمند یک چارچوب ریاضی اساسی و پایه برای درک و حل مسائل مربوط به آن است. همان‌گونه که پیش‌تر هم بیان شد سازوکار یادگیری تقویتی، تعامل با محیط در طی یک سری گام‌های زمانی به‌منظور یادگیری سیاست بهینه برای بیشینه کردن پاداش تجمعی است. فرایند تصمیم‌گیری مارکوف[2] روشی ساختاریافته برای نمایش چنین مسائل تصمیم‌گیری ارائه می‌دهد. یک فرآیند تصمیم‌گیری مارکوف می‌تواند به‌صورت $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ تعریف شود که در آن $\mathcal{S}$ فضای وضعیت، $\mathcal{A}$ فضای عمل، $\mathcal{P}(s_{t+1}|s_t, a_t)$ توزیع احتمال انتقالی، $\mathcal{R}(s_t, a_t)$ تابع پاداش و γ ضریب تنزیل است (Bellman, 1957). یک عامل در وضعیت اولیه خود s_1 از توزیع ثابت $\mathcal{P}(s_1)$ آغاز می‌کند و در هر گام زمانی t، یک عمل $a_t \in \mathcal{A}$ را از وضعی $s_t \in \mathcal{S}$ انتخاب می‌کند و به وضعیت بعدی $s_{t+1} \sim \mathcal{P}(\cdot\,|s_t, a_t)$ منتقل می‌شود. بعد از هر عمل، عامل یک پاداش $r_t = \mathcal{R}(s_t, a_t)$ دریافت می‌کند. وظیفهٔ عامل در یک الگوریتم یادگیری استاندارد، یادگیری سیاست بهینه $\pi(a_t|s_t)$ برای بیشینه کردن پاداش تجمعی (رابطهٔ ۲) است (Bellman, 1957):

$$\sum_{t=1}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r_t] \qquad \text{(رابطهٔ ۲)}$$

که در آن، T تعداد کل گام‌های زمانی و ρ_π بازدیدهای تنزیل‌یافته وضعیت-عمل است.

الگوریتم‌های یادگیری تقویتی کلاسیک زیادی (همانند Q-learning و SARSA) در تطابق با این ساختار توسعه یافته‌اند، اما اغلب این الگوریتم‌ها تنها در سیستم‌های ساده‌ای که امکان کاوش همه وضعیت‌ها به راحتی وجود دارد کاربرد دارند؛ زیرا در این سیستم‌ها امکان نگاشت جدول عمل-وضعیت و در نتیجه تخمین عمل بعدی به راحتی وجود دارد. اما در سیستم‌های پیچیده با فضای عمل و فضای وضعیت پیوسته مخصوصاً در دنیای واقعی عملاً امکان استفاده از رویکردهای کلاسیک یادگیری تقویتی وجود ندارد؛ زیرا برآورد عمل‌ها در این حالت به تعداد بسیار زیادی نمونه یا تعامل با محیط (حتی تا یک میلیون یا بیشتر) برای یادگیری سیاست بهینه نیاز خواهد داشت. در این حالت باید حتما یک الگوریتم اضافی برای تخمین ارزش بر اساس برخی از عملکردهای وضعیت وجود داشته باشد. به این منظور عمدتاً از شبکه‌های عصبی مصنوعی (برای فضای وضعیت مبتنی بر موقعیت) یا شبکه‌های عصبی پیچشی (برای فضای وضعیت مبتنی بر تصویر) برای تقریب توابع استفاده می‌شود که می‌تواند پاداش‌های کنونی را با حالت‌هایی که قبلاً تجربه کرده تخمین بزند. این رویکرد منجر به معرفی الگوریتم‌های یادگیری تقویتی عمیق شد.

در این پژوهش از الگوریتم یادگیری تقویتی عمیق SAC که زیرمجموعه روش عملگر-منتقد محسوب می‌شود برای کنترل میکروربات استفاده شد. این الگوریتم از رویکرد Off-policy استفاده می‌کند که در آن با بهره‌گیری از یک چارچوب تنظیم آنتروپی[3]، میان

---

1. Contours
2. Markov Decision Process (MDP)
3. Entropy regularization

سیاست جمع‌آوری داده و سیاست عامل در حین یادگیری تفاوت قائل می‌شود (Haarnoja et al., 2018). آنتروپی مؤید این نکته است که یک متغیر چقدر می‌تواند غیرقابل‌پیش‌بینی باشد. این الگوریتم با ذخیره تجربیات قبلی در بافری به نام بافر بازپخش[1] و استفاده مجدد از آن‌ها، نیاز عامل به جمع‌آوری نمونه‌های جدید برای هر تکرار را کاهش می‌دهد و از داده‌های جمع‌آوری‌شده به بهترین شکل استفاده می‌کند. وضعیت‌ها، اقدامات، پاداش‌ها و وضعیت‌های جدیدی که یک عامل در اپیزودهای قبلی تجربه کرده است در بافر حافظه ذخیره و به‌عنوان «تجربه» علامت‌گذاری می‌شوند. ازآن‌جهت SAC، «نرم» نامیده می‌شود که یک تابع ارزش نرم و یک سیاست نرم را معرفی می‌کند. نرم بودن در اینجا به استفاده از یک رویکرد احتمالی برای تابع ارزش و سیاست اطلاق می‌شود. SAC در هر تکرار مراحل ارزیابی و بهبود سیاست نرم را انجام می‌دهد. الگوریتم *SAC* از سه شبکه عصبی مصنوعی برای محاسبهٔ عمل تخمینی برای وضعیت کنونی و تولید سیگنال خطای اختلاف‌زمانی برای هر گام زمانی استفاده می‌کند. این سه شبکه عبارت‌اند از: یک شبکه عصبی برای تقریب سیاست در شبکه عملگر، یک شبکه عصبی برای تقریب مقدار سیاست در شبکه ارزش و درنهایت یک شبکه عصبی برای تقریب *Q*–ارزش در شبکه منتقد (de Jesus et al., 2021).

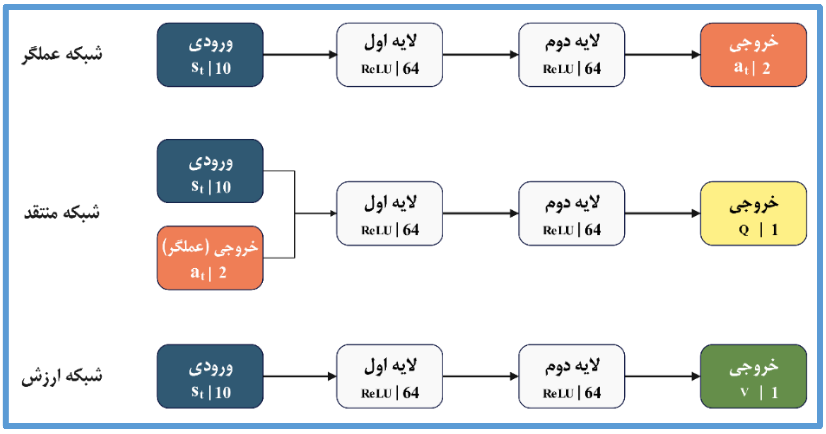
هدف سیستم کنترل میکروربات، تغییر زاویه آهنرباهای دائم به‌منظور حرکت دادن میکروربات برای دستیابی به رفتار دینامیکی موردنظر بود. برای پیاده‌سازی الگوریتم SAC به‌منظور کنترل موقعیت میکروربات مغناطیسی شناور روی سطح آب از زبان برنامه‌نویسی پایتون نسخه 3.10.9 در نرم‌افزار Spyder نسخه 5.3.3 و کتابخانه Stable-Baselines3 در PyTorch استفاده شد. ازآنجاکه کتابخانه Stable-Baselines3 به یک محیط برای اجرای الگوریتم‌های یادگیری تقویتی نیاز دارد، با استفاده از API استاندارد Gymnasium شرکت OpenAI، یک محیط سفارشی‌سازی‌شده طراحی شد. ازآنجاکه مسئله کنترل برای سیستم میکروربات این پژوهش به‌عنوان یک مسئله اپیزودیک با فضای عمل پیوسته و فضای وضعیت پیوسته فرمول‌بندی شده است، در این محیط، فضای وضعیت ، فضای عمل و موقعیت هدف (موقعیت میکروپلاستیک) به ترتیب با موقعیت میکروربات (x و y)، زوایای آهنرباها (angle1 و angle2) و موقعیت یک نقطهٔ مشخص در فضای کاری (x_t و y_t) تعریف شدند. فضای کاری در این محیط به‌صورت یک فضای شبکه‌ای با سایز ۱۳۴×۱۳۴ پیکسل با ۱۷ پیکسل حاشیه تعریف شد (شکل ۱–د).

شبکه عصبی مورداستفاده دارای ۱۰ ورودی است. چهار ورودی به‌عنوان چهار مشاهده اخیر از موقعیت x میکروربات، چهار ورودی به‌عنوان چهار مشاهده اخیر از موقعیت y میکروربات و دو ورودی از موقعیت x_t و y_t میکروپلاستیک. ازآنجاکه میان عمل و مشاهده تأخیر زمانی وجود دارد، تاریخچه چهار مشاهده آخر به‌عنوان ورودی به شبکه عصبی داده شد. زاویهٔ آهنربای اول و زاویهٔ آهنربای دوم نیز به‌عنوان خروجی‌های شبکه عصبی برای سروو موتورها ارسال می‌شوند. تمام ورودی‌ها و خروجی‌ها برای مقایسه‌پذیر بودن داده‌ها، نرمال‌سازی شدند. وضعیت کنونی میکروربات در محیط به‌عنوان ورودی به شبکه عملگر داده شد. این ورودی توسط دو لایه شبکه عصبی کاملاً متصل با ۶۴ نرون به لایه خروجی متصل می‌شود. لایه خروجی نیز زوایای آهنرباها را تولید می‌کند. ازآنجاکه از یک واحد خطی اصلاح‌شده (ReLU) به‌عنوان تابع فعال‌سازی استفاده می‌شود مقدار عمل بین [-1,1] در نظر گرفته شده است. این مقدار درنهایت پس قرارگیری در رابطهٔ ۳ به‌عنوان زاویه آهنربا (بین ۰ تا ۱۸۰ درجه) برای سروو موتور ارسال می‌شود.

$$angle = (a_t \times 90) + 90 \quad \text{(۳ رابطهٔ)}$$

از طرف دیگر، وضعیت کنونی و عمل انجام شده توسط عامل، به‌عنوان ورودی به شبکه منتقد داده می‌شود. خروجی شبکه منتقد، Q–ارزش است. درنهایت نیز در شبکه ارزش، مقدار وضعیت کنونی برآورد می‌شود. این دو شبکه نیز از ۲ لایه شبکه عصبی کاملاً متصل برای پردازش ورودی‌های وضعیت استفاده می‌کنند. Q–ارزش و مقدار وضعیت کنونی نیز از طریق یک تابع فعال‌سازی خطی، فعال می‌شوند. ساختار شبکهٔ عصبی مورداستفاده در الگوریتم SAC در شکل ۲ نمایش داده شده است. در جدول ۱ نیز فراپارامترهای مورداستفاده در الگوریتم SAC نمایش داده شده است. در ابتدای هر اپیزود، میکروربات در یک موقعیت مشخص قرار داده می‌شد. هر اپیزود به‌گونه‌ای تعریف شد که با رسیدن میکروربات به موقعیت هدف (موقعیت میکروپلاستیک) و یا با حداکثر ۱۰۰ گام زمانی به پایان می‌رسید و اپیزود بعدی آغاز می‌شد.

1. Replay buffer



شکل ۲. ساختار شبکه عصبی مورداستفاده در الگوریتم SAC.

پس از انتخاب هر عمل توسط عامل، موقعیت میکروربات از طریق دوربین تعیین می‌شد و با توجه به فاصله اقلیدسی (d) بین میکروربات و موقعیت هدف (رابطهٔ ۴)، برای آن پاداش یا جریمه (*reward*) طبق رابطهٔ ۵ محاسبه می‌شد.

$$d = \sqrt{(y-y_t)^2 + (x-x_t)^2}$$ رابطهٔ ۴)

$$reward = \begin{cases} +40 \;, & d < 2 \\ -1 \;, & microrobot\ hits\ the\ border \\ r_c \times d \;, & elsewhere \end{cases}$$ رابطهٔ ۵)

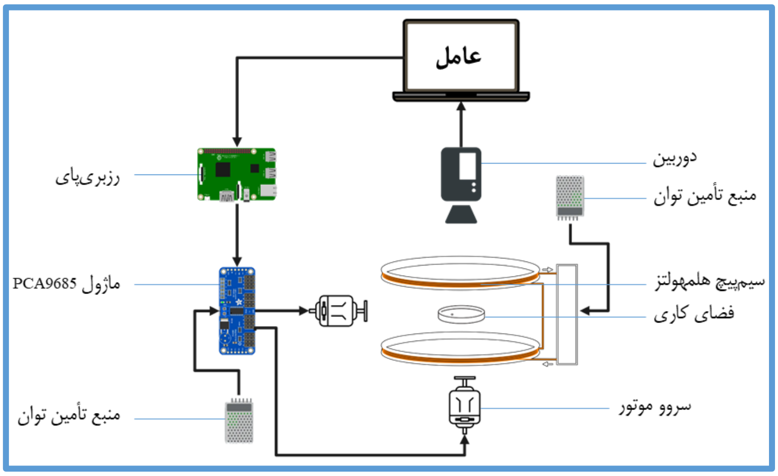
که در آن، r_c=-0.007 ثابت پاداش است که از آن برای نرمال‌سازی جریمه استفاده می‌شود.

همان‌گونه که در رابطهٔ ۵ مشخص است، اگر فاصله اقلیدسی میکروربات تا موقعیت هدف کمتر از ۲ پیکسل باشد، یک پاداش بزرگ ۴۰+ دریافت می‌کند و اپیزود پایان می‌یابد. این پاداش برای تشویق عامل برای رسیدن به هدف در سریع‌ترین زمان ممکن طراحی گردید. همین‌طور برای جلوگیری از برخورد میکروربات با حاشیهٔ فضای کاری، از یک جریمه ثابت ۱− استفاده شده است. درنهایت با توجه به فاصله میکروربات از موقعیت هدف، یک جریمه پویا بین [0,1] برای میکروربات تعریف شده است. بنابراین محدودهٔ پاداش اکتسابی برای هر اپیزود بین [-100,+40] خواهد بود که در آن بهترین سناریو عبارت است از رسیدن به هدف با یک گام زمانی (۴۰+) و بدترین سناریو، خارج شدن از حاشیهٔ فضای کاری در تمام گام‌های زمانی است.

جدول ۱. فراپارامترهای مورداستفاده در این پژوهش.

فراپارامتر	مقدار
سیاست	**MlpPolicy**
ساختار شبکه عصبی مصنوعی	[۶۴ ,۶۴]
نرخ یادگیری	۰/۰۰۰۳
اندازه بافر	۱۰۰۰۰۰
اندازه دسته‌ای	۲۵۶
ضریب به‌روزرسانی نرم	۰/۰۰۵
گاما	۰/۹۹
شروع یادگیری	۱۰۰۰

الگوریتم SAC از طریق لپ‌تاپ اجرا شد که موقعیت میکروربات را از دوربین و سپس از قسمت پردازش تصویر دریافت کرده و زوایای آهنرباها را با استفاده از پروتکل ارتباطی TCP/IP به رزبری‌پای ارسال می‌کرد. در این فرآیند، پس از برقراری ارتباط بین رزبری‌پای و لپ‌تاپ از طریق یک شبکه محلی وای‌فای، زوایای آهنرباها با استفاده از برنامه‌نویسی سوکت و کتابخانهٔ socket در پایتون، به‌صورت رشته برای رزبری‌پای ارسال شد. رزبری‌پای نیز پس از دریافت داده‌ها و تبدیل آن‌ها به اعداد صحیح، آن‌ها را از طریق پروتکل ارتباطی I2C برای ماژول درایور PCA9685 ارسال می‌کرد و درنهایت این ماژول زوایای دریافتی را برای سروو موتورها ارسال می‌نمود (شکل ۳).



شکل ۳. دیاگرام کلی فرآیند دریافت ورودی‌ها و ارسال خروجی‌های الگوریتم.

نتایج فرآیند یادگیری در حین و پس از آموزش میکروربات با استفاده از ابزار مصورسازی *TensorboardX* نظارت می‌شد. الگوریتم از طریق معیارهایی نظیر میانگین پاداش اپیزودها، میانگین طول اپیزودها، ضریب آنتروپی، زیان ضریب آنتروپی، زیان عملگر و زیان منتقد مورد ارزیابی قرار گرفت. لازم به توضیح است که معیار دقت میکروربات، نرخ موفقیت آن در رسیدن به هدف در نظر گرفته شد.

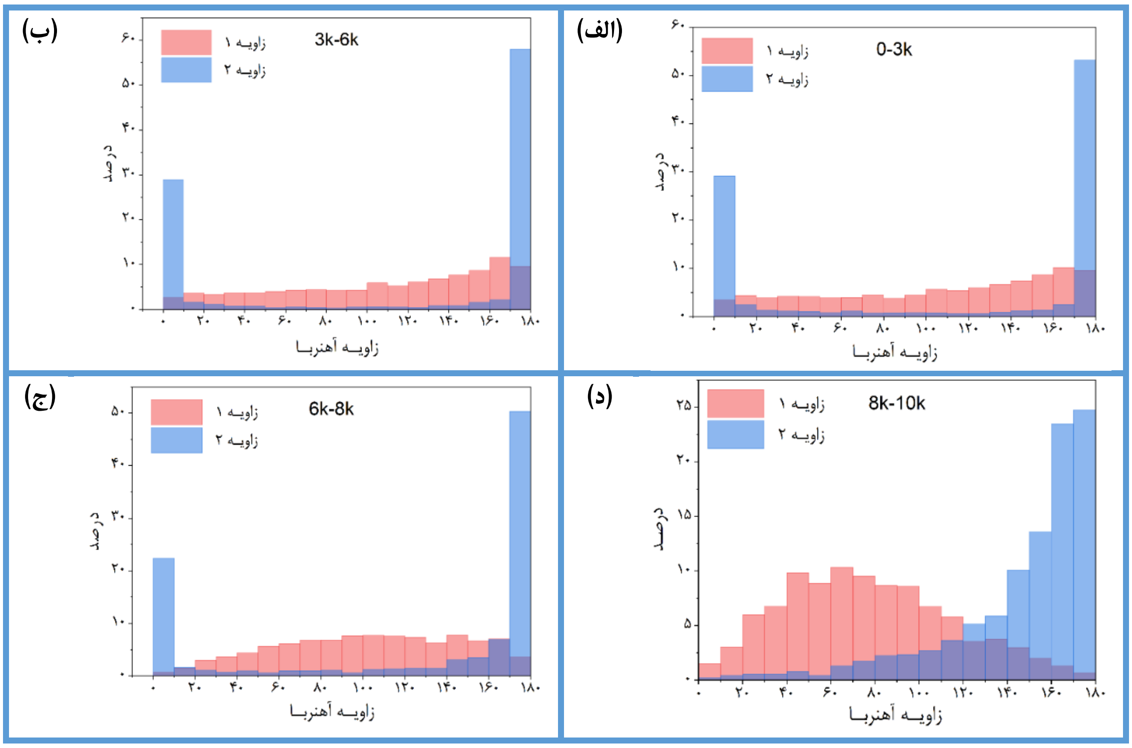
همان‌گونه که پیشتر هم عنوان شد یکی از نقاط قوت الگوریتم بدون مدل استفاده شده در این پژوهش، عدم نیاز آن به محاسبه، مدلسازی و شبیه‌سازی پارامترهای مربوط به رفتار میکروربات، محیط و سیستم فعال‌سازی است؛ زیرا همگی این پارامترها به عنوان بخشی از تجربه سیستم از محیط (موسوم به وضعیت‌ها)، در عمل‌های شبکه نمود می‌یابند و اساساً الگوریتم یادگیری تقویتی با توجه به همه‌ی پارامترهای دخیل در سیستم، اقدام به برآورد یک عمل مشخص برای حالت بعدی می‌کند و در نتیجه ماهیتاً نیازی به تعریف این پارامترها ندارد. از طرف دیگر، با توجه به حجم بسیار زیاد اغتشاشات خارجیِ غیرقابل پیش‌بینی در زمان به کارگیری میکروربات‌ها در شرایط واقعی، الگوریتم‌های یادگیری تقویتی معمولاً به گونه‌ای توسعه می‌یابند که نسبت به تمام تغییرات محیطی (چه آن‌هایی که از قبل قابل پیش‌بینی بوده‌اند و چه آن‌هایی که امکان پیش‌بینی‌شان فراهم نبوده است) بسیار مقاوم هستند. به عنوان مثال، در یک پژوهش یک ربات به منظور یادگیری راه رفتن روی سطح صاف آموزش داده شد. با وجود آن‌که این ربات تنها در شرایط عادی و روی سطح صاف آموزش دیده بود اما پس از تکمیل مدل به راحتی قادر به حرکت در شیب‌ها و موانع از پیش تعریف‌نشده و حتی اغتشاشات خارجی از طرف کاربر بود (Haarnoja et al., 2018). در پژوهش حاضر نیز با توجه به استفاده از الگوریتم یادگیری تقویتی بدون مدل، عملاً نیازی به تعریف پارامترهای مربوط به اغتشاشات خارجی وجود نداشت.

یافته‌های پژوهش و بحث

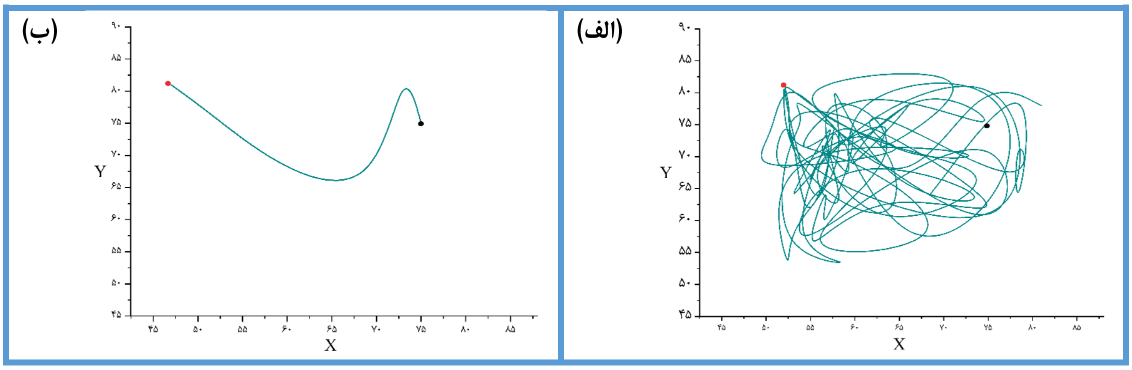
در این پژوهش، فرآیند آموزش میکروربات در دو تکرار انجام شد (آموزش اول و آموزش دوم). پارامترهای شبکه عصبی عملگر و منتقد در ابتدای هر فرآیند یادگیری به‌طور تصادفی مقداردهی اولیه شدند. عامل تا حداکثر ۱۰ هزار گام زمانی آموزش داده شد. در شکل ۴ نمودار توزیع فراوانی خروجی‌های الگوریتم SAC که شامل زاویه ۱ (زاویه آهنربای اول که توسط سروو موتور شماره ۱ چرخانده می‌شود) و زاویه ۲ (زاویه آهنربای دوم که توسط سروو موتور شماره ۲ چرخانده می‌شود) به درصد هستند در چهار بخش از گام زمانی ۰ تا ۳ هزار (شکل ۴–الف)، از گام زمانی ۳ هزار تا ۶ هزار (شکل ۴–ب)، از گام زمانی ۶ هزار تا ۸ هزار (شکل ۴–ج) و از گام زمانی ۸ هزار تا ۱۰ هزار (شکل ۴–د) نمایش داده شده است. نمودار توزیع فراوانی زاویه ۱ و ۲ ویژگی مهمی از رفتار الگوریتم را نشان می‌دهد. با توجه به شکل، به نظر می‌رسد که این الگوریتم محدودۀ خاصی از مقادیر زاویه را برای هر دو سروو موتور ترجیح می‌دهد. برای زاویه ۱، خروجی الگوریتم برای آهنربای اول (زاویه ۱) تا ۶ هزار گام زمانی، از توزیع نسبتاً یکنواختی برخوردار است (شکل ۴–الف و ب). این یکنواختی نشان می‌دهد که الگوریتم زوایای مختلفی را برای اولین سروو موتور به کار گرفته است تا با بررسی تمام عمل‌های ممکن در محیط، بهترین سیاست ممکن را برای کنترل میکروربات کشف کند. بااین‌حال، قابل‌توجه‌ترین مشاهدات از توزیع فراوانی زاویه ۲ مربوط به زوایای ۱۸۰–۱۷۰ درجه است که بخش قابل‌توجهی از مقادیر را تا ۸ هزار گام زمانی به خود اختصاص داده است. این نشان می‌دهد که الگوریتم سوگیری قابل‌ملاحظه‌ای به مقادیر زاویه ۲ به‌ویژه در محدوده ۱۷۰–۱۸۰ درجه از خود نشان داده است و به‌شدت از زوایای بالا برای سروو موتور دوم استفاده کرده

است. بااین‌حال، شایان‌ذکر است که در همین بازه گام‌های زمانی، مقادیر زاویه ۲ در محدوده ۰–۱۰ درجه نیز وزن قابل‌توجهی دارند.

در ادامه فرآیند آموزش از گام زمانی ۸ هزار تا ۱۰ هزار، عامل نحوه انتخاب زوایای خود را برای رسیدن به بهترین سیاست کنترلی بهبود می‌بخشد. همان‌گونه که در شکل ۴–د نمایش داده شده است، عامل برای زاویه ۱ به‌طور محسوسی از زوایای میانی (۳۰ تا درجه۱۲۰) و برای زاویه ۲ از زوایای انتهایی (۱۲۰ تا درجه۱۸۰) استفاده می‌کند که منجر به حرکت سریع‌تر و دقیق‌تر میکروربات به سمت موقعیت میکروپلاستیک می‌شود. برخلاف گام‌های زمانی بازه ۰ تا ۸ هزار که عامل برای زاویه ۲ از زوایای اولیه (۰ تا ۲۰ درجه) نیز استفاده می‌کرد، در گام‌های پایانی، عامل سیاست انتخاب این زوایا را کاملاً تغییر داده تا به یک استراتژی بهینه برای کنترل میکروربات دست یابد. نتیجهٔ اتخاذ این سیاست بهینه برای میکروربات، پیمودن مسیری کوتاه و سریع به سوی موقعیت میکروپلاستیک بود. شکل ۵ نشان‌دهنده اولین و آخرین اپیزود فرآیند یادگیری است. همان‌گونه که در شکل ۵–الف مشخص است در اپیزود اول، میکروربات با پیمودن تمام وضعیت‌های ممکن در حداکثر ۱۰۰ گام زمانی تعریف شده برای هر اپیزود، سعی در کشف محیط دارد، اما در اپیزود نهایی با توجه به یافتن یک سیاست بهینه جهت ناوبری میکروربات، در حداکثر ۵ گام‌های زمانی از موقعیت اولیه خود (رنگ قرمز) به موقعیت میکروپلاستیک (رنگ مشکلی) می‌رسید (شکل ۵–ب). این موضوع نشان می‌دهد که الگوریتم SAC به‌خوبی توانست بدون نیاز به هیچ دانش و مدل‌سازی قبلی از دینامیک محیط، میکروربات و سیستم تحریک، با موفقیت به یک استراتژی کنترلی بهینه برای ناوبری میکروربات نائل آید. هم‌چنین نرخ موفقیت میکروربات پس از پایان مرحله آموزش و طی تمام ۱۰ هزار گام زمانی، ۸۹ درصد مشاهده شد. در حالی‌که نرخ موفقیت میکروربات در گام‌های زمانی اولیه تا ۲ هزار گام زمانی، ۴۸ درصد بود. این نتیجه کاملاً مورد انتظار بود؛ چرا که در مراحل ابتدایی، الگوریتم در حال اجرای آزمون و خطا برای یافتن سیاست بهینه و بهره‌گیری از سیاست اکتشاف بیشتر (exploration) به جای بهره‌برداری (exploitation) از یک سیاست تقریباً موفق بود.

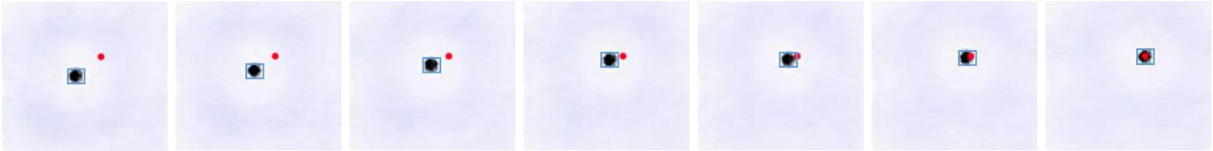


شکل ۴. نمودار توزیع فراوانی خروجی الگوریتم یادگیری تقویتی شامل زوایای سروو موتورها در گام‌های زمانی مختلف. (الف) از ۰ تا ۳ هزار گام زمانی، (ب) از ۳ تا ۶ هزار گام زمانی، (ج) از ۶ تا ۸ هزار گام زمانی، (د) از ۸ تا ۱۰ هزار گام زمانی.



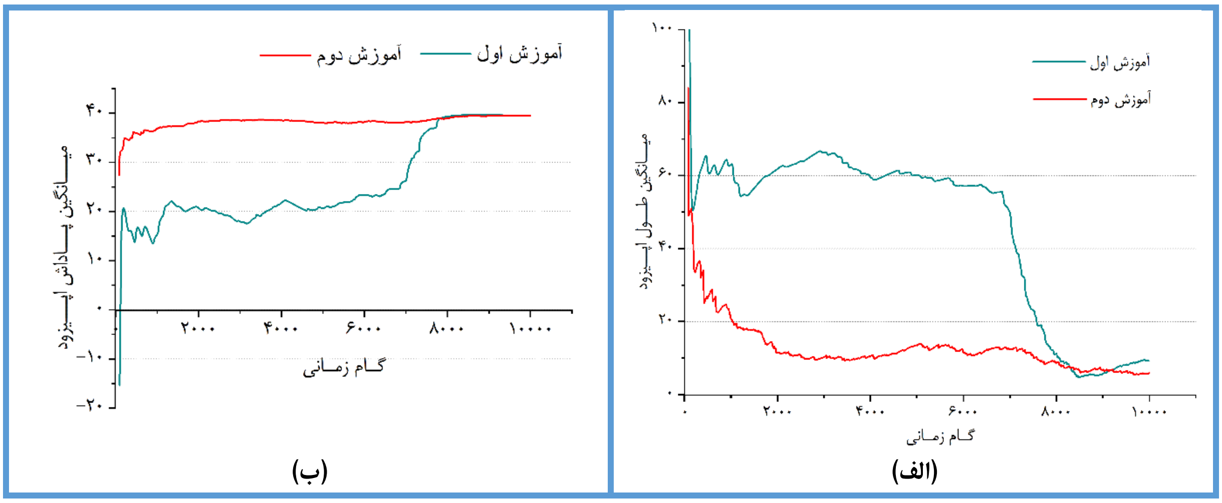
شکل ۵. مقایسه مسیر پیموده شده توسط میکروربات در (الف) اپیزود اول با ۱۰۰ گام زمانی و (ب) اپیزود آخر با ۵ گام زمانی

شکل ۶ نیز نشان‌دهنده هدایت میکروربات به سمت موقعیت هدف (رنگ قرمز) توسط الگوریتم یادگیری تقویتی در طول فرآیند ارزیابی میکروربات است.



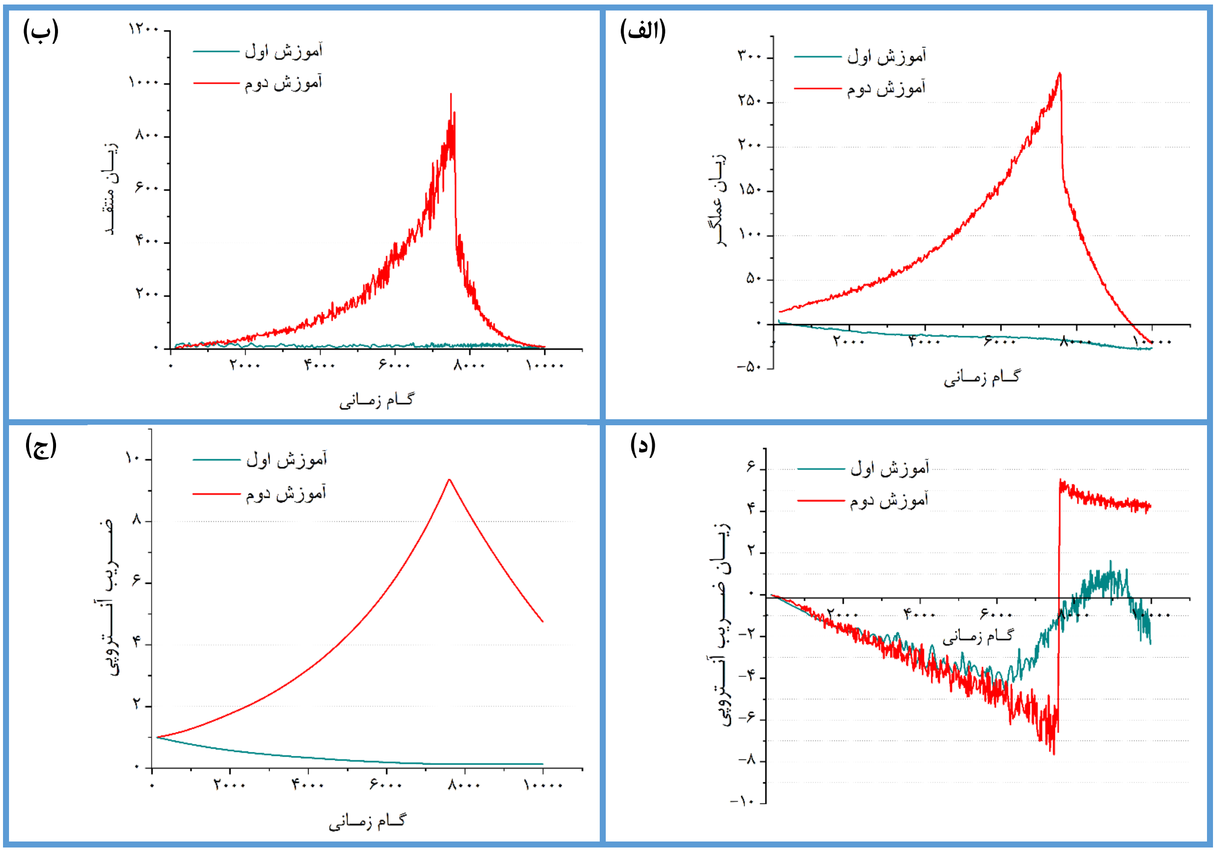
شکل ۶. تصویری از هدایت میکروربات (رنگ سیاه با حاشیه آبی) به سمت موقعیت هدف (رنگ قرمز) در یکی از اپیزودهای فرآیند ارزیابی میکروربات.

بااین‌حال، برای ارزیابی رفتار سیستم به معیارهای مربوط به الگوریتم SAC نیاز است. دو معیار مهم در ارزیابی سیستم کنترلی مبتنی بر الگوریتم SAC، میانگین طول اپیزود و میانگین پاداش اپیزود در طول گام‌های زمانی است. پیکربندی الگوریتم در این پژوهش به‌گونه‌ای تعریف شده بود که دوره‌های کوتاه‌تر و پاداش بیش‌تر، نشان‌دهندۀ بهبود کارایی یادگیری است. همان‌گونه که در شکل ۷ نشان داده شده است در اپیزودهای ابتدایی (در حدود ۱۰۰–۳۰۰ گام زمانی)، طول اپیزودها در هر دو تکرار نسبتاً زیاد و پاداش میانگین کم هستند. این موضوع در فرآیند یادگیری طبیعی است؛ زیرا عامل در حال یادگیری و کاوش محیط و تلاش برای رسیدن به هدف با عمل‌های تصادفی است. در ادامه با پیشرفت آموزش، روند واضح کاهش طول اپیزودها و افزایش پاداش میانگین مخصوصاً در آموزش دوم قابل‌مشاهده است. این نشان می‌دهد که عامل، عملکرد خود را بهبود می‌بخشد و در رسیدن به هدف کارآمدتر می‌شود. این روند به‌صورت پیوسته تا ۸ هزار گام زمانی ادامه یافت و حتی طول اپیزودها تا ۵ گام‌زمانی کاهش نشان داد و همین‌طور پاداش میانگین نیز تا ۳۹ افزایش یافت. این نتایج به این معنی است که میکروربات با سرعت بیشتر و در طی گام‌های زمانی بسیار کمتری به موقعیت هدف یا میکروپلاستیک می‌رسیده است. این علامت بسیار مثبتی است؛ زیرا اثربخشی رویکرد یادگیری تقویتی را نشان می‌دهد. در گام‌های زمانی پایانی (از ۸ هزار تا ۱۰ هزار)، طول اپیزودها و میانگین پاداش در هر دو آموزش تثبیت می‌شود و به مقدار نسبتاً ثابت و قابل قبولی می‌رسد که نشان می‌دهد عامل به یک سیاست تقریباً بهینه همگرا شده است و آموزش بیشتر ممکن است عملکرد را به‌طور قابل‌توجهی بهبود نبخشد؛ زیرا حداکثر پاداش قابل دریافت در این پیکربندی، معادل ۴۰+ واحد است و همان‌گونه که در شکل ۷–ب مشاهده می‌شود عامل در هر دو آموزش در اپیزودهای پایانی به‌طور میانگین تا ۳۹/۵ واحد پاداش دریافت کرده است.



شکل ۷. نمودار ارزیابی عملکرد الگوریتم SAC: (الف) میانگین طول اپیزودها در طول گام‌های زمانی، (ب) میانگین پاداش اپیزودها در طول گام‌های زمانی.

علاوه بر این دو معیار، از زیان عملگر و زیان منتقد نیز برای ارزیابی آموزش عامل استفاده شد. این دو تابع زیان برای بهبود عمل‌ها و Q-ارزش در محیط به کار می‌رود. شکل ۸-الف زیان عملگر برای هر دو آموزش را در طول دوره یادگیری نشان می‌دهد که در آن گام‌های زمانی در محور x و زیان عملگر در محور y به نمایش در آمده‌اند. زیان عملگر اطلاعات ارزشمندی در مورد فرآیند یادگیری سیاست ارائه می‌دهد. شکل ۸-الف روند کاهش نسبی زیان عملگر در آموزش اول و افزایش آن در آموزش دوم در طول گام‌های زمانی را نشان می‌دهد. در ابتدا، زیان عامل در آموزش دوم، افزایش سریعی را نشان می‌دهد که می‌توان آن را به رفتار اکتشافی عامل نسبت داد؛ زیرا میکروربات یاد می‌گیرد که در محیط حرکت کند. بااین‌حال در آموزش دوم شیب کاهش زیان عملگر بسیار ملایم است که نمایانگر پیشرفت قابل‌توجه الگوریتم در بهینه‌سازی سیاست و سرعت زیاد آن برای انطباق و یادگیری شرایط کنترلی و محیط است. پس‌ازآن، با کاهش زیان عملگر در هر دو آموزش، میکروربات حرکات کارآمدتری را نشان داد و در گام‌های کمتری به موقعیت میکروپلاستیک می‌رسید. در شکل ۸-ب زیان منتقد نشان داده شده است. زیان منتقد نشان می‌دهد که شبکه منتقد تا چه میزان عملکرد ارزش را تقریب می‌کند و هم‌چنین بازخوردی را به شبکه عملگر ارائه می‌دهد. زیان منتقد کمتر نشان‌دهنده تقریب بهتر تابع ارزش است. همان‌گونه که در شکل قابل‌ملاحظه است، در گام‌های زمانی ابتدائی، زیان منتقد در آموزش اول دارای ثبات نسبی و در آموزش دوم علاوه برداشتن روند افزایشی، دارای نوسان نیز هست. این نوسانات ممکن است به دلیل توازن بین اکتشاف[1] و بهره‌برداری[2] ذاتی در یادگیری تقویتی باشد. در این حالت، عامل، عمل‌های جدیدی را بررسی می‌کند که منجر به نوساناتی در پیش‌بینی منتقد می‌شود. علی‌رغم نوسانات ذکرشده، با افزایش گام‌های زمانی و پیشرفت یادگیری، روند کلی زیان منتقد مخصوصاً در آموزش دوم به‌صورت کاهشی است که نشان می‌دهد شبکه منتقد در حال یادگیری تقریب بهتر تابع ارزش و ارائه بازخورد دقیق‌تر به عملگر است.



شکل ۸. نمودار ارزیابی عملکرد شبکه‌های عصبی عملگر و منتقد: (الف) زیان عملگر در طول گام‌های زمانی، (ب) زیان منتقد در طول گام‌های زمانی، (الف) ضریب آنتروپی در طول گام‌های زمانی، (ب) زیان ضریب آنتروپی در طول گام‌های زمانی.

از طرف دیگر ضریب آنتروپی معیار دیگری است که نقش مهمی در ارزیابی عملکرد الگوریتم SAC برای اکتشاف و بهره‌برداری از

1. Exploration
2. Exploitation

سیاست عامل ایفا می‌کند. وظیفه این ضریب، متعادل کردن اکتشاف و بهره‌برداری با سیاست تصادفی است. نتایج نشان می‌دهد که الگوریتم در آموزش اول به‌طورکلی سطح متوسطی از آنتروپی را برای ارتقای اکتشاف در فضای عمل حفظ می‌کند. آنتروپی در آموزش دوم، اکتشاف بیشتر را در گام‌های ابتدائی تحمیل می‌کند، اما به‌تدریج میزان آنتروپی کاهش می‌یابد و عامل به سیاست خود اطمینان بیشتری جهت بهره‌برداری پیدا می‌کند (شکل ۸–ج). بااین‌حال برای نشان دادن تعادل بین اکتشاف و بهره‌برداری در الگوریتم SAC به معیار دیگری به نام زیان ضریب آنتروپی نیاز است که سطح آنتروپی را در توزیع عمل تنظیم می‌کند. ضریب بالاتر منجر به سیاست اکتشافی بیشتر می‌شود درحالی‌که ضریب پایین‌تر، بهره‌برداری را در اولویت قرار می‌دهد. ماهیت پویای زیان ضریب آنتروپی منعکس‌کننده سازگاری الگوریتم است. همان‌گونه که در شکل ۸–د نمایش داد شده است، زیان ضریب آنتروپی برای هر دو تکرار در ابتدا روندی کاهشی دارد که نشان‌دهندهٔ تمرکز عامل بر اکتشاف سیاست برای بهبود یادگیری است. پس‌ازآن، افزایش قابل‌توجهی در زیان ضریب آنتروپی در گام‌های زمانی خاصی رخ می‌دهد که ممکن است مربوط به زمانی باشد که الگوریتم برای مقابله با رکود سیاست، اکتشاف را اولویت‌بندی می‌کند. پس از نوسانات اولیه، زیان ضریب آنتروپی متمایل به تثبیت و تعادل بین اکتشاف و بهره‌وری در سیاست آموخته‌شده می‌شود. الگوریتم SAC در مرحله یادگیری، ۰/۳ ثانیه تأخیر در پاسخ را به سیستم تحمیل می‌کرد؛ یعنی پس از دریافت ورودی‌ها، ۰/۳ ثانیه طول می‌کشید تا عمل مربوطه که همان زوایای آهنرباها بودند تولید شوند. از طرف دیگر ۰/۰۶ ثانیه طول می‌کشید تا موقعیت میکروربات به عنوان ورودی به شبکه وارد شود. بنابراین کل سرعت پاسخ سیستم ۰/۳۶ ثانیه در هر گام زمانی بود. در نهایت پس از طی مراحل آموزش، مدل آموزش‌دیده از طریق اجرای ۴۰ اپیزود مجزا مورد ارزیابی قرار گرفت. برای آن‌که کارایی مدل در موقعیت‌هایی که در آن آموزش ندیده است مورد ارزیابی قرار گیرد، موقعیت اولیه میکروربات و موقعیت میکروپلاستیک در فرآیند ارزیابی، متفاوت با فرآیند آموزش در نظر گرفته شد. نتایج ارزیابی نشان داد که میکروربات به‌خوبی توانست با کسب میانگین پاداش ۳۹/۰۲، و انحراف معیار ۰/۷۱ در هر ۴۰ اپیزود با موفقیت و با دقت ۱۰۰ درصد به موقعیت میکروپلاستیک برسد. هم‌چنین میکروربات به‌طور میانگین در هر اپیزود ۴/۵۹ گام‌زمانی طی کرد. این نتایج نشان‌دهندهٔ دقت و سرعت بالای میکروربات در رسیدن به موقعیت میکروپلاستیک بدون نیاز به هرگونه دانش قبلی از سیستم تحریک و دینامیک میکروربات و محیط است.

نتیجه‌گیری و پیشنهادها

در این پژوهش یک سیستم کنترل حلقه بسته مبتنی بر یک الگوریتم یادگیری تقویتی عمیق به نام SAC برای کنترل حرکت و ناوبری یک میکروربات مغناطیسی شناور روی سطح سیال در محیط واقعی توسعه داده شد. آزمایش‌ها با تمرکز بر کنترل حرکت میکروربات از یک نقطه در محیط به سمت موقعیت هدف اجرا شد تا فرآیند ناوبری به سوی موقعیت میکروپلاستیک شناور روی سطح مایعات بدون نیاز به مدل‌سازی دینامیکی سیستم تحریک یا میکروربات و محیط شبیه‌سازی شود. میکروربات‌ها با استفاده از یک دستگاه تحریک مغناطیسی شامل آهنرباهای ثابت و سیم‌پیچ هلمهولتز تک‌محوره به حرکت درمی‌آمدند. الگوریتم موردنظر دارای ۱۰ ورودی ازجمله چهار مشاهده اخیر از موقعیت x میکروربات، چهار مشاهده اخیر از موقعیت y میکروربات و موقعیت x_t و موقعیت y_t میکروپلاستیک بود. در مقابل، زاویهٔ آهنربای اول و زاویهٔ آهنربای دوم به‌عنوان خروجی‌های الگوریتم برای کنترل موقعیت میکروربات‌ها در نظر گرفته شد. یک تابع پاداش به‌عنوان بازخورد اصلی برای سیستم کنترل تعریف گردید.

نتایج مراحل آموزش و ارزیابی نشان داد که میکروربات مغناطیسی می‌تواند با استفاده از الگوریتم SAC و با گام‌های زمانی کم و با دقت قابل‌قبول مسیر بهینه برای رسیدن به موقعیت هدف (موقعیت میکروپلاستیک) در محیط سطح سیال را یاد بگیرد. میکروربات پژوهش حاضر به‌خوبی توانست سیاست‌های تحریک مغناطیسی را از ورودی وضعیت‌های سیستم یاد بگیرد بدون آن‌که نیازی به دانش و مدل‌سازی قبلی از دینامیک میکروربات، سیستم تحریک مغناطیسی و یا محیط اطراف داشته باشد. این ویژگی یکی از نقاط قوت پژوهش در مقایسه با سایر سیستم‌های کنترل میکروربات‌های مغناطیسی (Xu et al., 2015) محسوب می‌شود؛ زیرا با توجه به کارایی قابل‌قبول آن در محیط‌های واقعی، می‌تواند بدون نیاز به مدل‌سازی‌های پیچیده و ساده‌سازی‌های مربوط به آن، وظایف میکروروباتیک را با دقت و سرعت بالا به انجام برساند و درنتیجه راه را برای توسعه سیستم‌های میکروروباتیک هموار نماید. قابلیت میکروربات حاضر برای یادگیری سیاست‌های کنترلی بدون نیاز به مدل، هم‌چنین می‌تواند زمان و منابع موردنیاز برای توسعه سیستم‌های میکروروباتیک با کارایی بالا را به میزان قابل‌توجهی کاهش دهد. این ویژگی، پتانسیل سیستم توسعه داده شده را جهت به‌کارگیری در کاربردهایی نظیر جداسازی و از بین بردن میکروپلاستیک‌های شناور روی سطح مایعات نشان می‌دهد. به‌عنوان‌مثال، در پژوهش‌های آینده می‌توان با بارگذاری لیپاز بر روی سطح

میکروربات موردمطالعه در این پژوهش، باعث تجزیه آنزیمی میکروپلاستک‌ها شد. هم‌چنین پژوهش‌های آینده می‌تواند متمرکز بر آموزش میکروربات در محیط‌های پیچیده با جریان آشفته یا مملو از موانع ثابت و متحرک باشد.

منابع

سلمانی، یوسف؛ موسی زاده، حسین؛ علیمردانی، رضا؛ جعفربیگلو، حمید؛ عمرانی، الهام؛ مخصوص، اشکان؛ و کیاپی، علی (۱۳۹۷). ارزیابی الگوریتم ناوبری قایق ربات خودران و مقایسه آن با نتایج شبیه‌سازی. *مهندسی بیوسیستم ایران*، ۴۹(۳)، ۳۶۶-۳۵۳.

REFERENCES

Agrahari, V., Agrahari, V., Chou, M.-L., Chew, C. H., Noll, J., & Burnouf, T. (2020). Intelligent micro-/nanorobots as drug and cell carrier devices for biomedical therapeutic advancement: Promising development opportunities and translational challenges. *Biomaterials*, *260*, 120163.

Amoudruz, L., & Koumoutsakos, P. (2022). Independent Control and Path Planning of Microswimmers with a Uniform Magnetic Field. *Advanced Intelligent Systems*, *4*(3), 2100183.

Bae, H., Paludan, M., Knoblauch, J., & Jensen, K. H. (2021). Neural networks and robotic microneedles enable autonomous extraction of plant metabolites. *Plant Physiology*, *186*(3), 1435–1441.

Behrens, M. R., & Ruder, W. C. (2022). Smart Magnetic Microrobots Learn to Swim with Deep Reinforcement Learning. *Advanced Intelligent Systems*, *4*(10), 2270049.

Beladi-Mousavi, S. M., Hermanová, S., Ying, Y., Plutnar, J., & Pumera, M. (2021). A Maze in Plastic Wastes: Autonomous Motile Photocatalytic Microrobots against Microplastics. *ACS Applied Materials & Interfaces*, *13*(21), 25102–25110.

Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, 679–684.

Borra, F., Biferale, L., Cencini, M., & Celani, A. (2022). Reinforcement learning for pursuit and evasion of microswimmers at low Reynolds number. *Physical Review Fluids*, *7*(2), 023103.

Cai, M., Wang, Q., Qi, Z., Jin, D., Wu, X., Xu, T., & Zhang, L. (2022). Deep Reinforcement Learning Framework-Based Flow Rate Rejection Control of Soft Magnetic Miniature Robots. *IEEE Transactions on Cybernetics*, 1–13.

Campuzano, S., Orozco, J., Kagan, D., Guix, M., Gao, W., Sattayasamitsathit, S., Claussen, J. C., Merkoçi, A., & Wang, J. (2012). Bacterial Isolation by Lectin-Modified Microengines. *Nano Letters*, *12*(1), 396–401.

Choi, J., Hwang, J., Kim, J. young, & Choi, H. (2021). Recent Progress in Magnetically Actuated Microrobots for Targeted Delivery of Therapeutic Agents. *Advanced Healthcare Materials*, *10*(6), 1–24.

Colabrese, S., Gustavsson, K., Celani, A., & Biferale, L. (2017). Flow Navigation by Smart Microswimmers via Reinforcement Learning. *Physical Review Letters*, *118*(15), 158004.

Dan, J., Shi, S., Sun, H., Su, Z., Liang, Y., Wang, J., & Zhang, W. (2022). Micro/nanomotor technology: the new era for food safety control. *Critical Reviews in Food Science and Nutrition*, 1–21.

de Jesus, J. C., Kich, V. A., Kolling, A. H., Grando, R. B., Cuadros, M. A. de S. L., & Gamarra, D. F. T. (2021). Soft Actor-Critic for Navigation of Mobile Robots. *Journal of Intelligent & Robotic Systems*, *102*(2), 31.

Ding, Z., Huang, Y., Yuan, H., & Dong, H. (2020). Introduction to Reinforcement Learning. In *Deep Reinforcement Learning* (pp. 47–123). Springer Singapore.

Dong, X., & Sitti, M. (2020). Controlling two-dimensional collective formation and cooperative behavior of magnetic microrobot swarms. *The International Journal of Robotics Research*, *39*(5), 617–638.

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., & Hester, T. (2021). Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, *110*(9), 2419–2468.

Gustavsson, K., Biferale, L., Celani, A., & Colabrese, S. (2017). Finding efficient swimming strategies in a three-dimensional chaotic flow by reinforcement learning. *The European Physical Journal E*, *40*(12), 110.

Haarnoja, T., Ha, S., Zhou, A., Tan, J., Tucker, G., & Levine, S. (2018). Learning to walk via deep reinforcement learning. ArXiv Preprint ArXiv:1812.11103.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 1861–1870). PMLR.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., & Abbeel,

P. (2018). Soft actor-critic algorithms and applications. *ArXiv Preprint ArXiv:1812.05905*.

Huang, L., Rogowski, L., Kim, M. J., & Becker, A. T. (2017). Path planning and aggregation for a microrobot swarm in vascular networks using a global input. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 414–420.

Jiang, J., Yang, L., & Zhang, L. (2023). DQN-based on-line Path Planning Method for Automatic Navigation of Miniature Robots. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 5407–5413.

Jiang, J., Yang, Z., Ferreira, A., & Zhang, L. (2022). Control and Autonomy of Microrobots: Recent Progress and Perspective. *Advanced Intelligent Systems*, *4*(5), 2100279.

Khalesi, R., Yousefi, M., Nejat Pishkenari, H., & Vossoughi, G. (2022). Robust independent and simultaneous position control of multiple magnetic microrobots by sliding mode controller. *Mechatronics*, *84*, 102776.

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, *32*(11), 1238–1274.

Lee, J., & Ha, J.-I. (2017). Direction Priority Control Method for Magnetic Manipulation System in Current and Voltage Limits. *IEEE Transactions on Industrial Electronics*, *64*(4), 2914–2923.

Leslie, H. A., van Velzen, M. J. M., Brandsma, S. H., Vethaak, A. D., Garcia-Vallejo, J. J., & Lamoree, M. H. (2022). Discovery and quantification of plastic particle pollution in human blood. *Environment International*, *163*, 107199.

Lynch, K. M., & Park, F. C. (2017). *Modern robotics*. Cambridge University Press.

Magnetics, Magnetop. (2023). Neodymium Cube - 20mm x 20mm x 20mm - N42. https://www.magnetop.com.au/neodymium-cube-20mm-x-20mm-x-20mm-n42

Mayorga-Martinez, C. C., Castoralova, M., Zelenka, J., Ruml, T., & Pumera, M. (2023). Swarming Magnetic Microrobots for Pathogen Isolation from Milk. *Small*, *19*(6), 2205047.

Medina-Sánchez, M., & Schmidt, O. G. (2017). Medical microbots need better imaging and control. *Nature*, *545*(7655), 406–408.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Nauber, R., Goudu, S. R., Goeckenjan, M., Bornhäuser, M., Ribeiro, C., & Medina-Sánchez, M. (2023). Medical microrobots in reproductive medicine from the bench to the clinic. *Nature Communications*, *14*(1), 728.

Parmar, J., Vilela, D., Villa, K., Wang, J., & Sánchez, S. (2018). Micro- and Nanomotors as Active Environmental Microcleaners and Sensors. *Journal of the American Chemical Society*, *140*(30), 9317–9331.

Pawashe, C., Floyd, S., Diller, E., & Sitti, M. (2012). Two-Dimensional Autonomous Microparticle Manipulation Strategies for Magnetic Microrobots in Fluidic Environments. *IEEE Transactions on Robotics*, *28*(2), 467–477.

Qiu, J., Huang, W., Xu, C., & Zhao, L. (2020). Swimming strategy of settling elongated micro-swimmers by reinforcement learning. *Science China Physics, Mechanics & Astronomy*, *63*(8), 284711.

Rahmer, J., Stehning, C., & Gleich, B. (2017). Spatially selective remote magnetic actuation of identical helical micromachines. *Science Robotics*, *2*(3).

Salmani, Y., Mousazadeh, H., Alimardani, R., Jafarbiglu, H., Omrani, E., Makhsoos, A., & Kiapey, A. (2018). Evaluation of a Navigation Algorithm for Robot Boat and Comparison to Simulation Results. *Iranian Journal of Biosystem Engineering*, *49*(3), 353–366. (In Persian)

Sehyuk Yim, Gultepe, E., Gracias, D. H., & Sitti, M. (2014). Biopsy using a Magnetic Capsule Endoscope Carrying, Releasing, and Retrieving Untethered Microgrippers. *IEEE Transactions on Biomedical Engineering*, *61*(2), 513–521.

Singh, B., Kumar, R., & Singh, V. P. (2022). Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review*, *55*(2), 945–990.

Sitti, M. (2017). *Mobile Microrobotics* (R. Arkin (ed.)). The MIT Press.

Sun, M., Chen, W., Fan, X., Tian, C., Sun, L., & Xie, H. (2020). Cooperative recyclable magnetic microsubmarines for oil and microplastics removal from water. Applied Materials Today, 20, 100682.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

TowerPro. (2023). MG996R. https://www.towerpro.com.tw/product/mg996r/

Tsang, Alan C. H., Demir, E., Ding, Y., & Pak, O. S. (2020). Roads to Smart Artificial Microswimmers.

Advanced Intelligent Systems, *2*(8), 1900137.

Tsang, Alan Cheng Hou, Tong, P. W., Nallan, S., & Pak, O. S. (2020). Self-learning how to swim at low Reynolds number. *Physical Review Fluids*, *5*(7), 074101.

Urso, M., Ussia, M., & Pumera, M. (2021). Breaking Polymer Chains with Self-Propelled Light-Controlled Navigable Hematite Microrobots. *Advanced Functional Materials*, *31*(28).

Urso, M., Ussia, M., & Pumera, M. (2023). Smart micro- and nanorobots for water purification. *Nature Reviews Bioengineering*.

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep Reinforcement Learning with Double Q-Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1).

Vilela, D., Stanton, M. M., Parmar, J., & Sánchez, S. (2017). Microbots Decorated with Silver Nanoparticles Kill Bacteria in Aqueous Media. *ACS Applied Materials & Interfaces*, *9*(27), 22093–22100.

Villa, K., Vyskočil, J., Ying, Y., Zelenka, J., & Pumera, M. (2020). Microrobots in Brewery: Dual Magnetic/Light-Powered Hybrid Microrobots for Preventing Microbial Contamination in Beer. *Chemistry – A European Journal*, *26*(14), 3039–3043.

Wang, L., Kaeppler, A., Fischer, D., & Simmchen, J. (2019). Photocatalytic TiO 2 Micromotors for Removal of Microplastics and Suspended Matter. *ACS Applied Materials & Interfaces*, *11*(36), 32937–32944.

Wang, Q., Li, T., Fang, D., Li, X., Fang, L., Wang, X., Mao, C., Wang, F., & Wan, M. (2020). Micromotor for removal/detection of blood copper ion. *Microchemical Journal*, *158*, 105125.

Wang, X., Hu, C., Pane, S., & Nelson, B. J. (2022). Dynamic Modeling of Magnetic Helical Microrobots. *IEEE Robotics and Automation Letters*, *7*(2), 1682–1688.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*.

Wright, S. L., Thompson, R. C., & Galloway, T. S. (2013). The physical impacts of microplastics on marine organisms: A review. *Environmental Pollution*, *178*, 483–492.

Xu, T., Yu, J., Yan, X., Choi, H., & Zhang, L. (2015). Magnetic Actuation Based Motion Control for Microrobots: An Overview. *Micromachines*, *6*(9).

Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., Nelson, B. J., Scassellati, B., Taddeo, M., Taylor, R., Veloso, M., Wang, Z. L., & Wood, R. (2018). The grand challenges of Science Robotics. *Science Robotics*, *3*(14).

Yang, Y., Bevan, M. A., & Li, B. (2020a). Efficient Navigation of Colloidal Robots in an Unknown Environment via Deep Reinforcement Learning. *Advanced Intelligent Systems*, *2*(1), 1900106.

Yang, Y., Bevan, M. A., & Li, B. (2020b). Micro/Nano Motor Navigation and Localization via Deep Reinforcement Learning. *Advanced Theory and Simulations*, *3*(6), 2000034.

Yang, Y., Bevan, M. A., & Li, B. (2022). Hierarchical Planning with Deep Reinforcement Learning for 3D Navigation of Microrobots in Blood Vessels. *Advanced Intelligent Systems*, *4*(11), 2200168.

Yousefi, M., & Nejat Pishkenari, H. (2021). Independent position control of two identical magnetic microrobots in a plane using rotating permanent magnets. *Journal of Micro-Bio Robotics*, *17*(1), 59–67.

Zhang, H., & Yu, T. (2020). Taxonomy of Reinforcement Learning Algorithms. In *Deep Reinforcement Learning* (pp. 125–133). Springer Singapore.

Zhou, H., Mayorga-Martinez, C. C., & Pumera, M. (2021). Microplastic Removal and Degradation by Mussel-Inspired Adhesive Magnetic/Enzymatic Microrobots. *Small Methods*, *5*(9), 2100230.

Zhu, H., Yu, J., Gupta, A., Shah, D., Hartikainen, K., Singh, A., Kumar, V., & Levine, S. (2020). The ingredients of real-world robotic reinforcement learning.

Zou, Z., Liu, Y., Young, Y.-N., Pak, O. S., & Tsang, A. C. H. (2022). Gait switching and targeted navigation of microswimmers via deep reinforcement learning. *Communications Physics*, *5*(1), 158.

Smart Control of a Microrobot for Navigation on Fluid Surface and Simulation of its Application in Microplastics Removal

EXTENDED ABSTRACT

Introduction

Over the past few decades the presence of microplastics in foods and beverages has caused irreversible damage and disease to humans and other organisms. To address this issue, it is necessary to develop a sustainable and efficient method of separation and degradation of microplastics. There has been considerable interest in using microrobots in order to achieve this objective. Nevertheless, this approach faces a critical challenge in terms of automating, intelligent, and precise control and navigation. As traditional methods are typically based on approximate and empirical approaches, they require complex dynamic modeling of the environment and actuation systems in order to analyze microrobot behavior. As a result of remarkable advancements in artificial intelligence technology, reinforcement learning algorithms (RL) have been introduced as a potential alternative method of addressing the challenge of microrobot navigation control. The RL makes it possible to train agents by enabling them to interact with real-world environments. The objective of the microrobot was to reach the target point, simulating the process of approaching microplastic particles floating on the surface of a fluid. In this study, the hypothesis is to create a high-performance control system using RL, eliminating the need to develop a specialized modeling of the magnetic field or fluid dynamics.

Material and methods

In this study, a magnetic actuation system was constructed to control a disk-shaped magnetic microrobot in a real-world environment. Changing the angle of the magnets affects the magnetic field and gradient within the workspace, which in turn, affects the position of the microrobot floating on the fluid surface in the xy plane. In order to align the microrobot in the Z-direction, a Helmholtz coil was used to generate a uniform magnetic field. In order to detect and determine the microrobot's position in the xy plane, an image processing procedure was employed. In this study, a Soft Actor-Critic algorithm (SAC) was utilized for microrobot control. Due to its sample-efficient capability, SAC is considered as a suitable choice, particularly in real-world scenarios. The training process was conducted through two repetitions. After observing the microrobot's state, the agent took action, and then the control system provided a feedback in the form of a reward or penalty primarily based on the microrobot's distance from the target position.

Results and discussion

Through 10,000 training steps, the SAC algorithm evaluated all available actions within the environment to determine the optimal policy for microrobot control. As a result, the agent enhances its actions in order to achieve an optimal control strategy. During microrobot training, there was a noticeable trend of shorter episode lengths and higher average rewards. These results show that implementing this optimal policy for the microrobot resulted in following a shorter and quicker path to reach the microplastic position.

Conclusions

It was demonstrated that the SAC algorithm could effectively achieve an optimal control strategy for microrobot navigation without requiring any prior knowledge of environmental dynamics, microrobot behavior, or actuation systems.