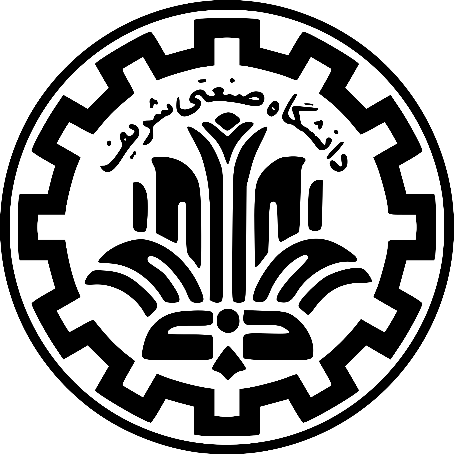
****

**دانشگاه صنعتی شریف**

**دانشکده مهندسی کامپیوتر**

**سمینار کارشناسی ارشد گرایش بیوانفورماتیک**

**عنوان:Optimizing Dynamic Treatment Regimes With Reinforcement Learning**

**نگارش:مصطفی توکلی**

**شهریور 1404**

**چکیده**

**رژیم‌های درمانی پویا (DTRs) ابزاری قدرتمند برای هدایت تصمیمات درمانی شخصی‌سازی‌شده در پزشکی هستند، به‌ویژه برای مدیریت بیماری‌های مزمن، که با در نظر گرفتن ویژگی‌های منحصربه‌فرد بیماران و تغییرات وضعیت آن‌ها در طول زمان، امکان تطبیق‌پذیری درمانی را فراهم می‌کنند. با این حال، تعیین DTRهای بهینه، به‌ویژه از داده‌های مشاهده‌ای پیچیده بالینی، با چالش‌های قابل توجهی همراه است. این گزارش به بررسی عمیق پتانسیل یادگیری تقویتی (RL) و رویکردهای یادگیری تقویتی عمیق (DRL) برای غلبه بر این چالش‌ها می‌پردازد. ما ضمن بسط مفاهیم بنیادین DTR ها و RL و اهمیت کاربرد آن‌ها در حوزه سلامت، به تحلیل جامع ادبیات موجود، شامل بررسی دقیق مقالات کلیدی در این زمینه می‌پردازیم. در انتها، چارچوب هیبریدی نوینی برای پیشبرد تحقیقات در این حوزه ارائه شده که بر افزایش تفسیر پذیری با ایجاد شفافیت در تصمیمات مدل و ایجاد قیود ایمنی برای محدود کردن ریسک هایی که بیمار را تهدید می کند، تاکید دارد.**

**1. مقدمه**

**پیشرفت‌های اخیر در هوش مصنوعی و یادگیری ماشین، به‌ویژه در حوزه یادگیری تقویتی، افق‌های جدیدی را برای شخصی‌سازی درمان در پزشکی گشوده است. در مواجهه با بیماری‌های مزمن و پیچیده، که نیازمند تصمیم‌گیری‌های متوالی و تطبیق‌پذیر در طول زمان هستند، "رژیم‌های درمانی پویا" (DTRs) به‌عنوان یک چارچوب قدرتمند مطرح شده‌اند DTR .ها مجموعه‌ای از قوانین تصمیم‌گیری فردی را تعریف می‌کنند که بر اساس تاریخچه پزشکی بیمار و ویژگی‌های در حال تغییر او، بهترین درمان را در هر مرحله از مداخله دیکته می‌کنند. هدف نهایی آن‌ها، تلاش برای رسیدن به بهترین وضعیت سلامتی و بیشترین منافع برای بیمار در بلندمدت است.**

**یادگیری تقویتی، به‌عنوان یک پارادایم یادگیری ماشین که در آن یک عامل (تصمیم‌گیرنده) از طریق تعامل با یک محیط (بیمار و بیماری) و دریافت پاداش (پیامدهای درمانی)، یاد می‌گیرد که چگونه یک "سیاست" بهینه را برای حداکثر کردن پاداش تجمعی در طول زمان اتخاذ کند، به‌طور طبیعی برای مدل‌سازی و حل مسائل DTRs مناسب است. این رویکرد می‌تواند عدم قطعیت‌ها را مدل‌سازی کرده، درمان‌ها را شخصی‌سازی کند و مداخله‌ها را در زمان واقعی تطبیق دهد.**

**با این حال، پیاده‌سازی RL در حوزه سلامت با چالش‌های قابل توجهی همراه است. این چالش‌ها شامل نیاز به داده‌های با کیفیت و در حجم بالا، مسائل مربوط به تفسیرپذیری مدل‌های پیچیده، و ملاحظات اخلاقی حیاتی در کاربردهای بالینی است. به‌علاوه، استفاده از داده‌های مشاهده‌ای، که برای مقیاس‌پذیری کاربردهای DTR-RL ضروری است، به ‌دلیل وجود متغیرهای مخدوش‌کننده مشاهده‌نشده، چالش‌های علّی مهمی را ایجاد می‌کند.**

**این گزارش با هدف ارائه یک درک جامع از DTRs و کاربرد RL در بهینه‌سازی آن‌ها تدوین شده است. ساختار گزارش به شرح زیر است: بخش 2 به اهمیت تحقیق می‌پردازد. بخش 3 یک مرور جامع و بسط‌یافته بر ادبیات موجود، شامل مفاهیم بنیادین DTRs، بررسی مقالات کلیدی و چالش‌های موجود ارائه می‌دهد. بخش 4 یک مدل پیشنهادی جامع برای پیشبرد تحقیقات در این زمینه ارائه می‌کند. در نهایت، واژه‌نامه و مراجع آورده شده‌اند.**

**2. اهمیت تحقیق**

**اهمیت تحقیق در زمینه بهینه‌سازی رژیم‌های درمانی پویا با استفاده از یادگیری تقویتی از چندین جنبه قابل بررسی است و نقش محوری در آینده پزشکی شخصی‌سازی‌شده ایفا می‌کند:**

* **پزشکی شخصی‌سازی‌شده و دقیق: DTRs در هسته خود، پزشکی شخصی‌سازی‌شده را برای مدیریت بیماری‌های مزمن مانند سرطان، دیابت، HIV محقق می‌سازند. با یادگیری از ویژگی‌های منحصربه‌فرد هر بیمار و واکنش او به درمان‌های قبلی، می‌توان درمان‌ها را به‌صورت پویا و متناسب با نیازهای در حال تغییر بیمار، تنظیم کرد. این رویکرد فراتر از مدل "یک اندازه برای همه" رفته و به پزشکان امکان می‌دهد تا با دقت بیشتری تصمیم‌گیری کنند، به‌ویژه در مواجهه با ناهمگونی (heterogeneity) بیماران و پاسخ‌های متغیر به درمان.**
* **مدیریت بهینه بیماری‌های مزمن: بسیاری از بیماری‌ها نیازمند مدیریت طولانی‌مدت و تصمیم‌گیری‌های متوالی هستند. DTRs چارچوبی را فراهم می‌کنند که در آن می‌توان اهداف درمانی بلندمدت را در نظر گرفت و بر اساس پیشرفت بیماری و واکنش بیمار، درمان را تطبیق داد. این امر می‌تواند منجر به بهبود پایدارتر در پیامدهای بیمار شود، زیرا سیستم به طور مداوم از نتایج درمان‌های قبلی درس می‌گیرد و سیاست‌های آتی را اصلاح می‌کند.**
* **پشتیبانی از تصمیم‌گیری مبتنی بر داده و هوش مصنوعی: با حجم فزاینده داده‌های سلامت (مانند پرونده‌های الکترونیکی سلامت، داده‌های حسگرهای پوشیدنی، نتایج آزمایشگاهی)، RL می‌تواند به ابزاری قدرتمند برای استخراج دانش از این داده‌ها تبدیل شود. این دانش می‌تواند به شکل توصیه‌های درمانی مبتنی بر شواهد به پزشکان ارائه شود و دقت و اثربخشی تصمیمات بالینی را افزایش دهد. این سیستم‌ها به پزشکان کمک می‌کنند تا با پیچیدگی داده‌ها و گزینه‌های درمانی مقابله کرده و تصمیماتی مبتنی بر شواهد قوی‌تر بگیرند.**
* **غلبه بر محدودیت‌های روش‌های سنتی: روش‌های آماری سنتی برای تخمین DTRs اغلب با چالش‌هایی مانند "نفرین ابعاد" (Curse of Dimensionality) (که با افزایش تعداد متغیرها و مراحل تصمیم‌گیری، پیچیدگی به طور تصاعدی افزایش می‌یابد) و "نفرین مدل‌سازی" (Curse of Modeling) (نیاز به مشخص کردن دقیق مدل‌های آماری پیش از تحلیل) مواجه هستند و ممکن است برای داده‌های پیچیده و با ابعاد بالا مناسب نباشند. RL، به‌ویژه با ادغام شبکه‌های عصبی عمیق (DRL)، می‌تواند این محدودیت‌ها را برطرف کرده و الگوهای پیچیده‌تری را در داده‌ها شناسایی کند، بدون اینکه نیاز به مشخص کردن صریح روابط بین متغیرها داشته باشد.**
* **بهبود پیامدهای بیمار و بهینه‌سازی منابع: هدف نهایی این تحقیقات، تلاش برای رسیدن به بهترین وضعیت سلامتی و بیشترین منافع برای بیماران است. با بهینه‌سازی رژیم‌های درمانی، می‌توان انتظار داشت که نرخ بقا افزایش یابد، عوارض جانبی کاهش یابد، کیفیت زندگی بیمار بهبود یابد و منابع درمانی (مانند داروها، زمان پزشک، بستری شدن در بیمارستان) به‌طور کارآمدتری مورد استفاده قرار گیرد. این بهینه‌سازی می‌تواند به کاهش هزینه‌های کلی مراقبت‌های بهداشتی نیز کمک کند.**

**3. ادبیات تحقیق**

**این بخش به بررسی مفاهیم بنیادین رژیم‌های درمانی پویا، چارچوب‌های محاسباتی برای بهینه‌سازی آن‌ها و مرور سه مقاله کلیدی در این زمینه می‌پردازد.**

**3.1. رژیم‌های درمانی پویا (DTRs): مفهوم و نمادگذاری**

**رژیم‌های درمانی پویا (DTRs) توالی‌ای از قوانین تصمیم‌گیری فردی را تعریف می‌کنند که در هر مرحله از مداخله، چگونگی تعیین درمان بیماران را بر اساس تاریخچه درمانی و متغیرهای کمکی در حال تغییر آن‌ها دیکته می‌کنند.**

**مفهوم و نمادگذاری: فرض کنید i=1,dots,N به بیماران، t=1,dots,T به مراحل تصمیم‌گیری و O\_t به بردار ویژگی‌های بیمار که در طول دوره درمانی t انباشته شده است، اشاره دارد A\_t . یک متغیر نشانگر درمان چنددسته‌ای با مقدار مشاهده‌شده a\_tinmathcalA∗t=1,dots,K\_t است، که در آن K\_t(K\_tgeq2) تعداد گزینه‌های درمانی در مرحله tاُم است. O∗T+1 کل تاریخچه مشاهده بالینی بیمار تا پایان مرحله T را نشان می‌دهد. ما Y را به‌عنوان پاسخ انتخاب‌شده به درمان در نظر می‌گیریم که مقادیر بالاتر Y مطلوب‌تر هستند. بنابراین، داده‌های مشاهده‌شده بر روی بیماران $H^T = {(A^1\\_i, \\dots, A^T\\_i, O^\\top\\_{T+1,i})}\\_{i=1}^N$ هستند که تاریخچه کامل بیمار را در طول زمان T توصیف می‌کنند.**

**مفهوم گردآوری تاریخچه بیمار در طول زمان را می‌توان به‌عنوان یک فرآیند ترتیبی (sequential process) تصور کرد. در ابتدا، در خط مبنا، O\_1 جمع‌آوری می‌شود، مانند ویژگی‌های جمعیت‌شناختی، بالینی و سبک زندگی بیمار. سپس، اولین تصمیم درمانی A\_1 تعیین می‌شود. بنابراین، تاریخچه بیمار در مرحله 1 به سادگی با H\_1=O\_1 نمایش داده می‌شود. پس از آن، پاسخ بیمار به اولین درمان به‌عنوان O\_2 و متغیر پاسخ Y\_1 ارزیابی می‌شود. در نتیجه، تاریخچه در مرحله 2 به H\_2=O\_1,A\_1,Y\_1,X\_2 تبدیل می‌شود. این فرآیند تا آخرین مرحله ادامه می‌یابد.**

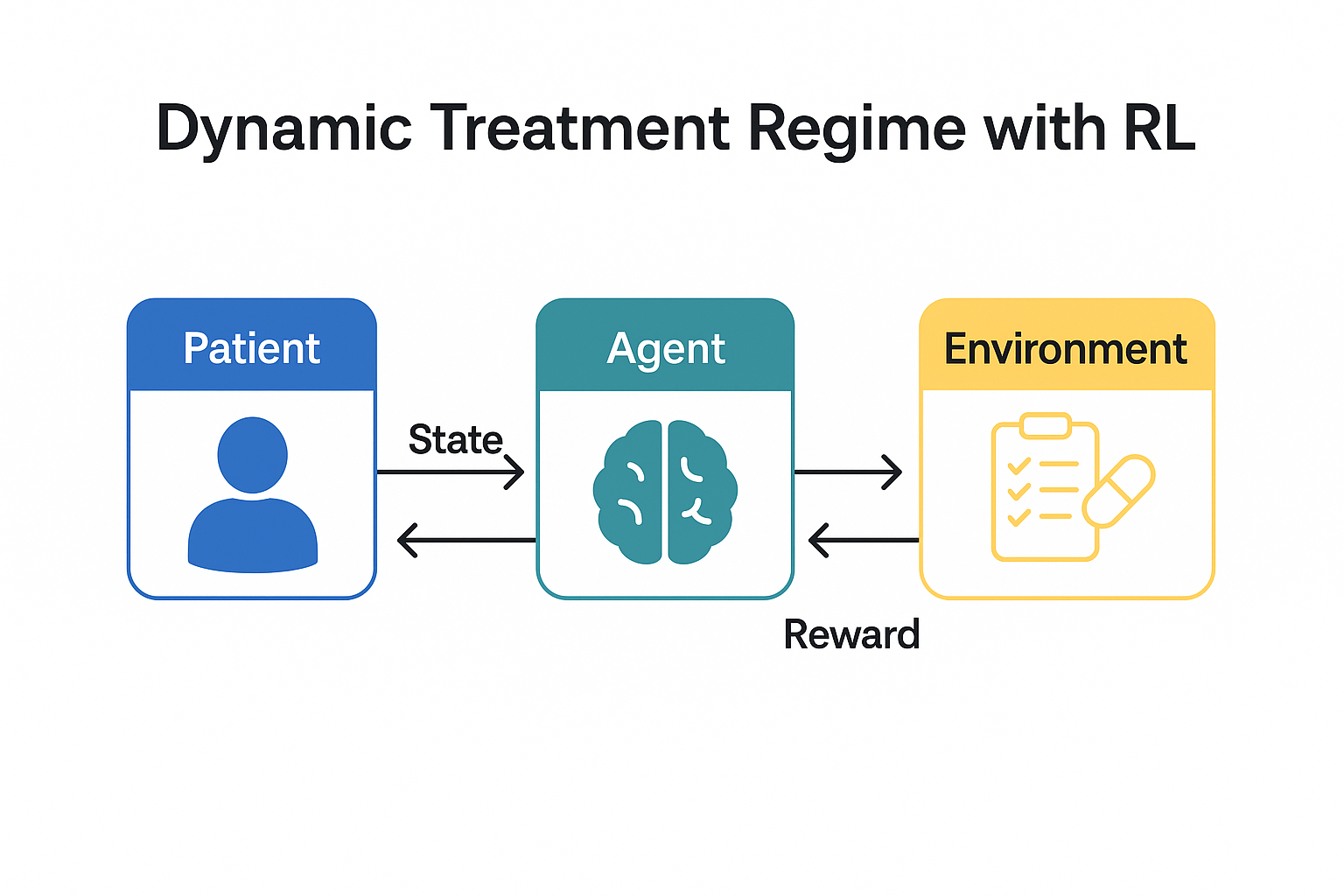
**یک مدل رژیم درمانی پویا (DTR) توالی‌ای از قوانین تصمیم‌گیری درمانی فردی, r=(r\_1,dots,r\_T) را تعریف می‌کند که در آن r\_t نگاشتی از تاریخچه بیمار در مرحله t به دامنه تخصیص درمان mathcalA\_t است. یکی از روش‌های تعریف و شناسایی DTR بهینه، در نظر گرفتن یک چارچوب پیامدسازی متقابل (Counterfactual Framework) برای استنتاج علّی [1] و شروع از آخرین مرحله درمان به ترتیب معکوس است. چارچوب پیامدسازی متقابل استفاده‌شده شامل مقایسه آنچه اتفاق افتاده (پیامد مشاهده‌شده) با آنچه اگر درمان دیگری اعمال می‌شد، اتفاق می‌افتاد (پیامد پیامدسازی متقابل یا مشاهده‌نشده) است. در این چارچوب، ایده این است که سناریوهای پیامدسازی متقابل را بسازیم که نشان‌دهنده آنچه در شرایط مختلف رخ می‌داد، هستند.**

**برای توضیح مفهوم پیامدسازی متقابل، یک مثال عملی را در زمینه ارزیابی یک داروی جدید ضد فشار خون طراحی شده برای کاهش فشار خون در بیماران مبتلا به فشار خون بالا در نظر بگیرید. در یک کارآزمایی بالینی تصادفی‌شده ایده‌آل، شرکت‌کنندگان به طور تصادفی برای دریافت داروی جدید یا دارونما (یک ماده غیرفعال) تخصیص داده می‌شوند، که تضمین می‌کند مقایسه بین درمان‌ها دقیق و بدون سوگیری است. در سناریوی واقعی، بیمار A برای دریافت داروی جدید تخصیص داده می‌شود و کاهش قابل توجهی در فشار خون را تجربه می‌کند. این پیامد مشاهده‌شده، اندازه‌گیری مستقیمی از تأثیر دارو بر فشار خون بیمار تحت شرایط درمان را فراهم می‌کند. سناریوی پیامدسازی متقابل، با این حال، شامل تخمین آنچه برای فشار خون بیمار A اتفاق می‌افتاد، اگر به جای دریافت داروی جدید به گروه دارونما تخصیص داده می‌شد، است. این پیامد فرضی، که به عنوان پیامد پیامدسازی متقابل شناخته می‌شود، سطح فشار خونی را نشان می‌دهد که بیمار A اگر دارونما دریافت می‌کرد، تجربه می‌کرد. با مقایسه پیامد واقعی (کاهش فشار خون مشاهده‌شده با داروی جدید) با این پیامد پیامدسازی متقابل (سطح فشار خون فرضی با دارونما)، می‌توانیم اثر علّی داروی جدید را تخمین بزنیم. این مقایسه با برجسته‌سازی تفاوت بین اثر مشاهده‌شده و اثر بالقوه در صورت دریافت درمان جایگزین توسط بیمار، بینش‌هایی را در مورد اثربخشی دارو ارائه می‌دهد.**

**به‌طور معمول، پیامدهای پیامدسازی متقابل به‌طور مستقیم قابل مشاهده نیستند، اما به‌عنوان یک نقطه مرجع برای ارزیابی تأثیر علّی یک درمان عمل می‌کنند. در مرحله نهایی T، فرض کنید Y∗(A\_1,dots,A\_T−1,a\_T)، یا به‌اختصار Y∗(a\_T)، نشان‌دهنده پیامد پیامدسازی متقابل یک بیمار تحت درمان با a\_TinmathcalA∗T به شرط درمان‌های قبلی (A\_1,dots,A∗T−1) باشد، و $Y^\\*(r\\_T)$ را به‌عنوان پیامد پیامدسازی متقابل تحت قانون رژیم r\_T و تاریخچه بیمار HT تعریف کنید. یعنی:**

**Y∗(rT​)=aT​=1∑KT​​Y∗(aT​)I{rT​(HT)=aT​}(1)**

**عملکرد قانون درمان r\_T با میانگین پیامد پیامدسازی متقابل EY∗(r\_T) اندازه‌گیری می‌شود، زمانی که همه بیماران از r\_T پیروی می‌کنند، و قانون درمان بهینه ropt\_T شرط EYopt(ropt\_T)geqEY∗(r\_T) را برای تمام کلاس‌های بالقوه رژیم‌های درمانی R\_T، که r\_TinR\_T هستند، برآورده می‌کند.**



**برای ایجاد ارتباط بین پیامدهای پیامدسازی متقابل و داده‌های مشاهده‌شده، ما به سه فرض اثبات‌شده در ادبیات تکیه می‌کنیم [2, 3]:**

* **سازگاری (Consistency): پیامد واقعی بیمار با پیامدی که اگر بیمار درمان واقعی دریافت کرده بود، مطابقت دارد.**
* **ارزش درمان واحد پایدار (Stable Unit Treatment Value - SUTVA): پیامد یک فرد تحت تأثیر درمان‌های اعمال‌شده بر سایر بیماران قرار نمی‌گیرد.**
* **تبادل‌پذیری ترتیبی (Sequential Exchangeability): تخصیص درمان در هر نقطه زمانی، مستقل از پیامدهای بالقوه آینده است، به شرط تاریخچه درمان گذشته و تاریخچه متغیرهای کمکی.**

**این فرض‌ها بیان می‌کنند که پیامدهای مشاهده‌شده بازتاب خوبی از پیامدهای بالقوه هستند، پیامدهای هر بیمار تحت تأثیر درمان‌های دریافت‌شده توسط سایر بیماران نیست، تخصیص درمان از پیش تعیین‌شده نیست، و با فرض عدم مخدوش‌کنندگی توسط عوامل اندازه‌گیری‌نشده، تخصیص درمان در نقطه زمانی فعلی مستقل از پیامدهای بالقوه آینده است، به شرط تاریخچه کامل درمان گذشته و تاریخچه پزشکی بیمار تا آن نقطه. از روش نمره گرایش (Propensity Score) برای تقویت این فرض (تبادل‌پذیری ترتیبی) در تحلیل داده‌های طولی استفاده می‌شود.**

**تحت این فرض‌ها، قانون بهینه در مرحله نهایی T را می‌توان به صورت زیر نوشت:**

**rTopt​=argrT​∈RT​max​E[aT​=1∑KT​​E(Y∣AT​=aT​,HT)I{rT​(HT)=aT​}](2)**

**که در آن انتظار بیرونی نسبت به توزیع مشترک داده‌های مشاهده‌شده HT گرفته می‌شود. به طور مشابه، تحت فرض‌های فوق، قانون بهینه ropt\_t در مرحله t را می‌توان به صورت زیر تعریف کرد:**

**rtopt​=argrt​∈Rt​max​E[at​=1∑Kt​​E(Y^t​∣At​=at​,Ht​)I{rt​(Ht​)=at​}](3)**

**که در آن R\_t مجموعه تمام قوانین بالقوه در مرحله t است. hatY\_T=Y در مرحله T، و در t, hatY\_t را می‌توان به صورت بازگشتی با استفاده از بهینگی بلمن (Bellman’s Optimality) تعریف کرد:**

**Y^t​=E[Y^t+1​∣At+1​=rt+1opt​(Ht+1​),Ht+1​],t=1,…,T−1**

**یعنی، پیامد مورد انتظار با فرض رعایت قوانین بهینه در تمام مراحل آینده.**

**3.2. رویکردهای یادگیری ماشینی و یادگیری تقویتی برای DTRs**

**توانایی پیش‌بینی چگونگی واکنش بیمار به دارو، تصمیمات درمانی را از روش آزمون و خطا دور کرده و بارهای بهداشتی و مالی مرتبط با بیماری را کاهش می‌دهد. در این راستا، رویکردهای یادگیری ماشینی اعمال شده بر روی مجموعه‌داده‌های مشاهده‌ای بالینی، نوید بزرگی را برای ارائه پزشکی شخصی‌سازی‌شده می‌دهند [4]. در ادامه، الگوریتم‌های یادگیری تقویتی مبتنی بر درخت و سایر مدل‌های مرتبط که برای مطالعه تجربی DTRs انتخاب شده‌اند، مورد بحث و بسط قرار خواهند گرفت.**

**3.2.1. یادگیری تقویتی مبتنی بر درخت (T-RL) الگوریتم یادگیری تقویتی مبتنی بر درخت (T-RL) به طور مستقیم رژیم‌های درمانی پویا (DTRs) بهینه را در یک محیط چندمرحله‌ای و چنددرمانی [5, 6] از داده‌های بالینی مشاهده‌ای برای تعیین درمان بهینه تخمین می‌زند. این الگوریتم از یک ساختار درخت تصمیم استفاده می‌کند، که در آن یک گره، نقطه‌ای در درخت تصمیم را نشان می‌دهد که در آن یک تصمیم بر اساس مقدار یک ویژگی خاص گرفته می‌شود، با تقسیم مکرر یک گره والد به دو گره فرزند، که با گره ریشه که شامل کل نمونه یادگیری است، شروع می‌شود [7]. این رویکرد به دلیل ساختار سلسله مراتبی و بصری درخت، قابلیت تفسیرپذیری بالاتری نسبت به مدل‌های جعبه سیاه ارائه می‌دهد.**

**تخمین رژیم درمانی پویا (DTR) چالش مهمی را ایجاد می‌کند. این چالش بر شناسایی درمان بهینه برای بیمار با استفاده از داده‌های بالینی متمرکز است که از طریق مشاهده مستقیم قابل دسترسی نیستند. درمان بهینه در مرحله t, ropt\_t, تنها می‌تواند به طور غیرمستقیم از طریق درمان‌ها و پیامدهای مشاهده‌شده استنتاج شود. این کار با تخمین میانگین پیامد پیامدسازی متقابل (Counterfactual Mean Outcomes) با توجه به تمام درمان‌های ممکن با استفاده از چارچوب علّی و سه فرض بیان شده در بخش 3.1، قابل دستیابی است. انشعاب انتخاب‌شده در هر گره باید میانگین پیامد پیامدسازی متقابل را افزایش دهد، که می‌تواند به‌عنوان معیاری برای خلوص در درخت‌های DTR عمل کند، با هدف کلی تلاش برای رسیدن به بهترین وضعیت سلامتی و بیشترین منافع برای بیمار در بلندمدت در کل جمعیت مورد نظر. از این رو، رویکرد T-RL درخت‌های تصمیم را در هر مرحله می‌سازد و فرآیند بهینه‌سازی شامل مقایسه‌های درمانی متعدد را از طریق یک معیار خلوص مدیریت می‌کند که با استفاده از برآوردگرهای وزن‌دهی احتمال معکوس افزوده (Augmented Inverse Probability Weighted Estimators - AIPW) همانطور که در کار [8, 9] بیان شده، ساخته شده است. این فرآیند به‌طور بازگشتی با استفاده از استقراء پس‌رو (Backward Induction) استفاده می‌شود که می‌تواند مراحل تصمیم‌گیری متعدد را به‌طور مؤثر مدیریت کند.**

**برای تلاش برای رسیدن به بهترین وضعیت سلامتی و بیشترین منافع برای بیمار در بلندمدت برای کل جمعیت، هر گره در یک درخت DTR باید به گونه‌ای تقسیم شود که پیامد جمعیت را بهبود بخشد، و از این می‌توان برای اندازه‌گیری خلوص تقسیم گره همانطور که در بالا ذکر شد، استفاده کرد. برای بیان آن در نمادگذاری، فرض کنید eta\_a(H)=E(hatY∣A=a,H) پیامد پیامدسازی متقابل و pi(H) نمره گرایش تخمین‌زده شده باشد [2]. برآوردگر AIPW برای میانگین پیامد پیامدسازی متقابل تحت یک درمان داده‌شده a به صورت زیر محاسبه می‌شود:**

**P[πa​(H)I(A=a)​Y+(1−πa​(H)I(A=a)​)ηa​(H)](4)**

**که در آن P عملگر امید ریاضی تجربی است. برای مرحله T, تخمین برای پیامد پیامدسازی متقابل تحت یک قانون درمان r\_T ($E(Y^\\*\\_T(r\\_T))$) به صورت زیر تعریف می‌شود:**

**P[πT,AT​​(HT)I(AT​=rT​(HT))​Y+(1−πT,AT​​(HT)I(AT​=rT​(HT)​)ηT,rT​​(HT)](5)**

**که در آن pi\_T,A\_T(HT) نمره گرایش و eta\_T,r\_T(HT) میانگین شرطی تخمین‌زده شده است. بنابراین، T-RL میانگین پیامد پیامدسازی متقابل را از طریق هر یک از گره‌ها با بهینه‌سازی معادله (5) به منظور تلاش برای رسیدن به بهترین وضعیت سلامتی و بیشترین منافع برای بیمار در بلندمدت حداکثر می‌کند.**

**3.2.2. یادگیری تقویتی مبتنی بر درخت تصادفی (ST-RL) در یک محیط چندمرحله‌ای و چنددرمانی، روش جستجوی درخت تصادفی برای تخمین رژیم‌های درمانی پویا بهینه (ST-RL) که توسط [10] معرفی شده است، بر روی داده‌های آزمایش‌های تصادفی‌شده یا مطالعات مشاهده‌ای اعمال می‌شود ST-RL . یک رویکرد مبتنی بر درخت تصادفی را برای تخمینDTR ها اتخاذ می‌کند. این بدان معناست که یک درخت تصمیم در هر مرحله از درمان می‌سازد، که در آن هر گره در درخت یک تصمیم درمانی ممکن را نشان می‌دهد. برخلاف T-RL که به صورت قطعی (deterministically) درخت را می‌سازد، ST-RL از جستجوی تصادفی بهره می‌برد. برای ساخت یک درخت تصمیم، ST-RL ابتدا میانگین پیامدهای پیامدسازی متقابل را از طریق مدل‌های رگرسیون غیرپارامتریک مدل‌سازی می‌کند. سپس، از یک الگوریتم زنجیره مارکوف مونت کارلو [11, 12] برای جستجوی تصادفی قانون تصمیم‌گیری ساختاردرخت بهینه استفاده می‌کند. این بدان معناست که به طور تصادفی یک تصمیم درمانی را در هر مرحله انتخاب می‌کند و سپس پیامدهای این قانون تصمیم‌گیری را ارزیابی می‌کند. قانون تصمیم‌گیری با بهترین پیامدها سپس به‌عنوان DTR بهینه انتخاب می‌شود. این رویکرد به ST-RL اجازه می‌دهد تا فضاهای جستجوی بزرگتر را کاوش کند و از گیر افتادن در بهینه‌های محلی جلوگیری کند.**

**برای برازش $E(Y^\\* | A\\_t = a\\_t, H\\_t)$ ST-RL, از درخت‌های رگرسیون افزایشی بیزی (Bayesian Additive Regression Trees - BART) استفاده کرده و پیامدهای پیامدسازی متقابل را برای هر بیمار پیش‌بینی می‌کند. در پیاده‌سازی استقراء پس‌رو ST-RL در مرحله نهایی T از متغیر پاسخ نهایی برای تخمین DTR استفاده می‌شود. در هر مرحله میانی $1 \< t \< T$ پیامد پیامدسازی متقابل به رژیم‌های درمانی بهینه در تمام مراحل آینده بستگی دارد و نیاز به پیش‌بینی دارد.**

**برای کاهش انباشت سوگیری از استقراء پس‌رو چندمرحله‌ای، پیامد میانی (که "شبه‌پیامد" (Pseudo-Outcome - PO) نیز نامیده می‌شود) هم مقدار متغیر پاسخ میانی مشاهده‌شده واقعی در مرحله t و هم زیان پیش‌بینی‌شده آینده ناشی از درمان‌های زیربهینه را در بر می‌گیرد(. شبه‌پیامد (PO) به صورت زیر خواهد بود:**

**PO=Y+t′=t+1∑T​{E[Y∗∣rt′opt​(Ht′​),Ht′​]−E[Y∣At​=at​,Ht​]}**

**که در آن شبه‌پیامد (PO) به‌عنوان پیامد برای مرحله t−1 در فرآیند استقراء پس‌رو استفاده می‌شود. این فرمول به مدل کمک می‌کند تا عدم قطعیت‌ها و وابستگی‌های زمانی را در تصمیم‌گیری‌های متوالی بهتر مدیریت کند.**

**3.2.3. روش مبتنی بر درخت علّی (Causal Tree-Based Method) روش‌های مبتنی بر درخت که برای استنتاج توصیه‌های درمانی پیشنهاد شده‌اند، به‌طور سنتی با استفاده از درخت‌های طبقه‌بندی و رگرسیون (CART) یا جنگل‌های تصادفی ساخته شده‌اند، بنابراین بدون مدل‌سازی اثرات علّی. با این حال، برای درک تأثیر مداخلات درمانی مختلف بر روی یک بیمار، به ویژه در چارچوب DTR ها که هدف آن توصیه‌های شخصی‌سازی‌شده است، بررسی پیامدهای علّی این درمان‌ها بسیار مهم است. از این رو، [13] یک رویکرد درخت علّی را معرفی کرد که برای ارزیابی مستقیم اثرات درمان طراحی شده و در عین حال یک تفسیر علّی را ارائه می‌دهد. این امر از طریق پیاده‌سازی یک معیار انشعاب سفارشی برای درخت‌های تصمیم حاصل می‌شود.**

**یادگیری درخت علّی، که برای تخمین اثرات درمان طراحی شده است، بهینه‌سازی معیارهای متمایزی را در مقایسه با یادگیری درخت تصمیم، که صرفاً بر پیش‌بینی و طبقه‌بندی تمرکز دارد، به کار می‌گیرد. در حالی که درخت‌های تصمیم از معیارهای انشعاب متمرکز بر دقت در پیش‌بینی یک متغیر هدف استفاده می‌کنند، یادگیری درخت علّی نیاز به ادغام دو اکتشافی (Heuristics) ضروری دارد. اکتشافی اول مربوط به دستیابی به تعادل بین افراد تحت درمان و بدون درمان در هر برگ برای یک درمان داده‌شده است. این تعادل برای تخمین دقیق تفاوت‌های پیامد و سپس محاسبه تخمین اثر درمان بدون سوگیری حیاتی است. اکتشافی دوم شامل تقسیم برگ‌ها به گروه‌های متمایز با پیامدهای متفاوت است. این تقسیم تضمین می‌کند که دقت اثرات درمان تخمین‌زده شده به خطر نمی‌افتد و درخت قادر به شناسایی گروه‌هایی از بیماران با پاسخ‌های متفاوت به درمان است.**

**الگوریتم یادگیری درخت علّی با استفاده از یک معیار انشعاب اصلاح‌شده به‌عنوان اکتشافی برای تعیین انشعاب‌های بهینه، این چالش‌ها را برطرف می‌کند. به جای تکیه بر معیارهای انشعاب درخت تصمیم معمولی مانند ناخالصی Gini (Gini Impurity) و بهره اطلاعاتی (Information Gain)، یادگیری درخت علّی از خطای میانگین مربعات مورد انتظار برای اثرات درمان (Expected Mean Squared Error for Treatment Effects - EMSE) استفاده می‌کند. این معیار به طور خاص برای تخمین اثرات درمانی ناهمگن همانطور که توسط [14] معرفی شده، طراحی شده است.**

**با توجه به نمره گرایش (Propensity Score) و تاریخچه پزشکی بیمار، در مرحله,t EMSE را می‌توان به صورت زیر فرموله کرد:**

**EMSE=t=1∑T​i=1∑K​πt​(Ht​)(Y(at​)−Y∗(rt​))2+τ2(6)**

**که در آن tau اثر علّی درمان بر پیامد است. این تفاوت در پیامد مورد انتظار بین بیمارانی است که درمان دریافت می‌کنند و بیمارانی که درمان دریافت نمی‌کنند. مقدار tau به تفاوت مربعات بین پیامد واقعی تحت قانون تخصیص درمان بهینه و پیامد تحت قانون تخصیص درمان واقعی اضافه می‌شود. این بدان معناست که EMSE به میزان اثر علّی درمان بر پیامد افزایش می‌یابد. دلیل این امر آن است که قانون تخصیص درمان بهینه، قانونی است که تفاوت در پیامد مورد انتظار بین بیمارانی که درمان دریافت می‌کنند و بیمارانی که درمان دریافت نمی‌کنند را به منظور تلاش برای رسیدن به بهترین وضعیت سلامتی و بیشترین منافع برای بیمار در بلندمدت حداکثر می‌کند.**

**برای به دست آوردن میانگین پیامد پیامدسازی متقابل مورد انتظار، ابتدا خطای مربعات مورد انتظار را برای هر قانون درمان ممکن با توجه به تاریخچه بیمار محاسبه می‌کنیم:**

**Expected squared error=i=1∑K​πt​(Ht​)(Yat​​−Y∗(rt​))2**

**ما قانون درمان را با کمترین خطای مربعات مورد انتظار انتخاب می‌کنیم و میانگین پیامد پیامدسازی متقابل مورد انتظار را برای تخصیص درمان انتخاب‌شده به صورت زیر محاسبه می‌کنیم:**

**Expected counterfactual mean=i=1∑K​πat​​Y∗(rt​)**

**که در آن $Y^\\*(r\\_t)$ پیامد پیامدسازی متقابل تحت درمان بهینه است.**

**فرمول EMSE (خطای میانگین مربعات مورد انتظار)، هنگام تقسیم گره‌های درخت علّی برای شناسایی رژیم‌های درمانی پویا، تخصیص درمانی را انتخاب می‌کند که خطای مربعات پیش‌بینی‌شده را به حداقل می‌رساند. این انتخاب با در نظر گرفتن وضعیت فعلی داده‌ها، اثر علّی تخمین‌زده‌شده درمان بر پیامد، و احتمالات تخصیص هر درمان صورت می‌گیرد. بنابراین، تخصیص درمان بهینه، خطای مربعات مورد انتظار را به حداقل می‌رساند.**

**3.3. بررسی مقالات کلیدی**

**در این بخش، سه مقاله کلیدی در زمینه DTRs و RL مورد بررسی دقیق قرار می‌گیرند تا رویکردهای نوین و چالش‌های حل‌شده در این حوزه را روشن سازند.**

**3.3.1. مقاله 1: Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data (Liu et al., 2018)**

* **روش‌شناسی: این مقاله یک چارچوب یادگیری تقویتی عمیق (DRL) دو مرحله‌ای را برای تخمین DTRs از داده‌های مشاهده‌ای ثبت پزشکی معرفی می‌کند [15]. مرحله اول شامل یک گام یادگیری نظارت‌شده (Supervised Learning) است که با استفاده از شبکه‌های عصبی عمیق، به پیش‌بینی اقدامات متخصصان بالینی می‌پردازد و دانش موجود را در مدل تعبیه می‌کند. مرحله دوم شامل یک گام DRL است که با بهره‌گیری از اصول Q-learning ، تابع ارزش بلندمدت DTRs را تخمین می‌زند. این چارچوب از خودکدگذارها (Auto-encoders) برای کاهش ابعاد بالای فضای حالت و عمل استفاده می‌کند که در داده‌های بالینی رایج است. تابع پاداش آن‌ها یک پاداش اکتشافی بر اساس زمان بقای بدون بازگشت، تعدیل‌شده با کیفیت زندگی بیمار است که تلاش می‌کند تا اهداف بالینی را در تابع پاداش منعکس کند.**
* **مشارکت‌ها: این مقاله یکی از اولین چارچوب‌های DRL را برای DTRs بر روی داده‌های ثبت پزشکی مشاهده‌ای ارائه می‌دهد که قادر به مدیریت فضاهای حالت و عمل با ابعاد بالا است، چالشی که در پزشکی بسیار رایج است. نتایج تجربی آن‌ها بر روی داده‌های واقعی پیوند سلول‌های بنیادی خونساز برای بیماری GVHD، دقت امیدوارکننده‌ای در پیش‌بینی تصمیمات متخصصان انسانی و تابع پاداش مورد انتظار بالایی در DTR های مبتنی بر DRL نشان می‌دهد که کارایی این رویکرد را در شرایط واقعی نشان می‌دهد.**
* **نقاط قوت:**
  + **قابلیت مدیریت داده‌های با ابعاد بالا و فضاهای عمل پیچیده، که یک مزیت بزرگ برای داده‌های سلامت است.**
  + **استفاده از داده‌های ثبت پزشکی واقعی، که اعتبار کاربرد مدل در محیط بالینی را افزایش می‌دهد.**
  + **ترکیب یادگیری نظارت‌شده با یادگیری تقویتی برای بهره‌گیری از دانش متخصصان و تسریع فرآیند یادگیری.**
* **محدودیت‌ها:**
  + **استفاده از تابع پاداش اکتشافی که ممکن است کاملاً منعکس‌کننده اهداف بالینی پیچیده و چندوجهی نباشد و نیاز به تنظیمات دقیق دارد.**
  + **چالش‌های مرتبط با داده‌های سانسورشده (مانند زمان بقا) و نیاز به تنظیمات دقیق برای عدم تصادفی بودن در داده‌های مشاهده‌ای، که می‌تواند بر اعتبار نتایج تأثیر بگذارد.**
  + **تمرکز بر یادگیری آفلاین (Offline Learning) که به معنای یادگیری از داده‌های از پیش جمع‌آوری شده است و نیاز به بررسی سیاست‌های اکتشافی در محیط‌های آنلاین برای تطبیق‌پذیری در زمان واقعی را مطرح می‌کند.**
  + **احتمال عدم دقت در خارج از نمونه آموزشی (Out-of-training-sample accuracy) که می‌تواند تعمیم‌پذیری مدل را در سناریوهای جدید محدود کند.**

**3.3.2. مقاله 2: Learning Optimal Dynamic Treatment Regimens Subject to Stagewise Risk Controls (Zhao et al., 2024)**

* **روش‌شناسی: این مقاله یک چارچوب رژیم‌های درمانی پویا مبتنی بر منفعت-ریسک (Benefit-Risk DTRs - BR-DTRs) را معرفی می‌کند که کنترل ریسک را در هر مرحله از درمان در نظر می‌گیرد، چالشی حیاتی در تصمیم‌گیری‌های بالینی [18]. آن‌ها یک رویه یادگیری پس‌رو (Backward Learning) توسعه می‌دهند که از آخرین مرحله درمان شروع کرده و به عقب بازمی‌گردد. در هر مرحله، یک مسئله ماشین بردار پشتیبان (SVM) وزن‌دار با قیود نرم اصلاح‌شده (modified smooth constraint) حل می‌شود تا سیاست‌های بهینه که هم منافع را حداکثر و هم ریسک‌ها را مدیریت می‌کنند، تخمین زده شود. این روش بر روی بهینه‌سازی سیاست با در نظر گرفتن مصالحه (trade-off) بین منفعت و ریسک تمرکز دارد و از توابع جایگزین (Surrogate Functions) مانند زیان هینچ (Hinge Loss) و زیان شیب‌دار (Ramp Loss) برای تقریب اهداف پیچیده استفاده می‌کند.**
* **مشارکت‌ها: این مقاله چارچوبی کلی و نوآورانه برای DTR ها تحت قیود ریسک مرحله‌ای ارائه می‌دهد که به طور مستقیم به نگرانی‌های ایمنی بیمار پاسخ می‌دهد. این رویکرد دارای بنیان‌های نظری قوی است و ضمانت‌های نظری (مانند سازگاری فیشر و کران‌های خطای غیرمجاورتی) را برای برآوردگرهای پیشنهادی خود اثبات می‌کند که اعتبار ریاضی روش را تضمین می‌کند.**
* **نقاط قوت:**
  + **توجه صریح و مدل‌سازی ریسک در هر مرحله از درمان، که اهمیت بالایی در تصمیم‌گیری‌های پزشکی دارد.**
  + **دارای بنیان‌های نظری قوی و ضمانت‌های ریاضی، که اعتماد به نتایج مدل را افزایش می‌دهد.**
  + **رویکردی انعطاف‌پذیر که می‌تواند برای مدیریت ریسک در شرایط مختلف بالینی و با انواع داده‌ها به کار رود.**
* **محدودیت‌ها:**
  + **عملکرد ممکن است با تعداد بسیار زیادی از مراحل درمان کاهش یابد، زیرا پیچیدگی محاسباتی با افزایش مراحل افزایش می‌یابد.**
  + **وابستگی به "فرض ریسک حاد" که ممکن است همیشه در سناریوهای بالینی پیچیده صادق نباشد و نیاز به ارزیابی دقیق فرض‌ها در کاربردهای واقعی دارد.**
  + **تمرکز اصلی بر درمان‌های دوتایی و قیود ریسک منفرد، که ممکن است برای محیط‌های چنددرمانی یا چندریسکی نیاز به تعمیم و بسط بیشتری داشته باشد.**

**3.3.3. مقاله 3: Near-Optimal Reinforcement Learning in Dynamic Treatment Regimes**

**(Zhang & Bareinboim, 2020)**

* **روش‌شناسی: این مقاله به چالش مهم یادگیری RL از داده‌های مشاهده‌ای که ذاتاً مخدوش (confounded) هستند، می‌پردازد [19]. آن‌ها الگوریتم‌های جدیدی مانند UC-DTR خوش‌بینی تحت عدم قطعیت برای RL آنلاین و UCc-DTR (که کران‌های علّی جدیدی را از داده‌های مشاهده‌ای برای تسریع یادگیری در بر می‌گیرد) را معرفی می‌کنند. این الگوریتم‌ها از معادلات بلمن و برنامه‌ریزی خطی برای برنامه‌ریزی خوش‌بینانه استفاده می‌کنند و به دنبال دستیابی به "پشیمانی تقریباً بهینه" (Near-optimal Regret) هستند. پشیمانی معیاری برای سنجش فاصله عملکرد الگوریتم از عملکرد بهینه است.**
* **مشارکت‌ها: این مطالعه اولین الگوریتم تطبیقی را با پشیمانی تقریباً بهینه در محیط DTR خالص ارائه می‌دهد که یک پیشرفت نظری مهم محسوب می‌شود. همچنین کران‌های علّی قابل اثبات دقیقی را از داده‌های مشاهده‌ای ارائه می‌کند و الگوریتم کارآمدی برای بهره‌گیری از مشاهدات مخدوش‌شده معرفی می‌کند که به مقابله با سوگیری‌های ناشی از مخدوش‌کنندگی کمک می‌کند.**
* **نقاط قوت:**
  + **توانایی یادگیری آنلاین (Online Learning) که برای محیط‌های پویا و در حال تغییر مناسب است و امکان به‌روزرسانی سیاست‌ها در زمان واقعی را فراهم می‌کند.**
  + **دارای تضمین‌های نظری قوی در مورد عملکرد الگوریتم (پشیمانی تقریباً بهینه)، که اعتماد به قابلیت‌های مدل را افزایش می‌دهد.**
  + **پرداختن به چالش مهم استفاده از داده‌های مشاهده‌ای با وجود مخدوش‌کنندگی، که در داده‌های بالینی واقعی بسیار رایج است.**
* **محدودیت‌ها:**
  + **وابستگی به دانش حوزه (Domain Knowledge) برای برخی از اجزای الگوریتم، که ممکن است در کاربردهای عمومی چالش‌برانگیز باشد.**
  + **پیچیدگی محاسباتی برنامه‌ریزی خوش‌بینانه که ممکن است در مقیاس‌های بسیار بزرگ داده‌ها یا تعداد زیاد مراحل درمان چالش‌برانگیز باشد و نیاز به منابع محاسباتی قابل توجهی دارد.**
  + **تأثیر ابعاد بالا بر عملکرد پشیمانی الگوریتم، به این معنی که با افزایش ابعاد فضای حالت و عمل، رسیدن به عملکرد بهینه دشوارتر می‌شود.**

**3.4. چالش‌ها و پیشرفت‌های کلی در RL برای DTRs**

**چالش‌ها:**

* **کمبود و کیفیت داده‌ها: داده‌های بالینی اغلب پراکنده، ناقص، نویزدار، ناهمگن و با حجم کم برای برخی شرایط نادر هستند. علاوه بر این، جمع‌آوری داده‌های مناسب برای آموزش RL (شامل توالی اقدامات، مشاهدات و پاداش‌ها) می‌تواند پیچیده باشد.**
* **تفسیرپذیری (Interpretability): مدل‌های پیچیده RL، به‌ویژه DRL ، اغلب به عنوان "جعبه سیاه" عمل می‌کنند. این بدان معناست که درک منطق پشت توصیه‌های آن‌ها برای پزشکان دشوار است، که مانع اصلی در پذیرش بالینی آن‌ها است.**
* **تعمیم‌پذیری (Generalizability): سیاست‌های آموخته‌شده ممکن است به خوبی به جمعیت‌های جدید بیمار یا محیط‌های بالینی مختلف تعمیم نیابند، به دلیل تفاوت در پروتکل‌های درمانی، جمعیت‌شناسی بیمار یا کیفیت داده‌ها.**
* **ایمنی و ملاحظات اخلاقی: ماهیت اکتشافی RL و نیاز به آزمون و خطا در محیط‌های بالینی که ایمنی بیمار حیاتی است، چالش‌های اخلاقی جدی ایجاد می‌کند. نمی‌توان یک عامل RL را به سادگی بر روی بیماران واقعی آموزش داد.**
* **طراحی تابع پاداش: تعریف یک تابع پاداش مناسب که به‌طور دقیق اهداف بالینی پیچیده و بلندمدت را منعکس کند، دشوار است. پاداش باید هم پیامدهای مثبت (مانند بهبود سلامت) و هم منفی (مانند عوارض جانبی یا هزینه‌ها) را در بر گیرد.**
* **مخدوش‌کنندگی و استنتاج علّی: داده‌های مشاهده‌ای به‌طور ذاتی مستعد مخدوش‌کنندگی (Confounding) هستند، به‌ویژه از نوع مشاهده‌نشده، که می‌تواند باعث شود مدل روابط علّی نادرست را یاد بگیرد و سیاست‌های بهینه‌ای را ارائه ندهد.**
* **پیچیدگی مدل و محاسبات: مدل‌های RL پیشرفته می‌توانند از نظر محاسباتی بسیار سنگین باشند و نیاز به قدرت پردازشی قابل توجهی برای آموزش و استنتاج دارند.**

**پیشرفت‌ها و راه‌حل‌های پیشنهادی:**

* **ادغام دانش پزشکی: استفاده از دانش متخصصان (مانند پزشکان) برای شکل‌دهی به فضاهای حالت/عمل، تعریف پاداش‌ها یا تنظیم ساختار مدل، می‌تواند به هدایت فرآیند یادگیری کمک کند.**
* **یادگیری از داده‌های ناقص/نویزدار: توسعه روش‌های قوی برای مقابله با داده‌های نامتعادل، سانسورشده و نویزدار، که در سوابق پزشکی واقعی رایج هستند.**
* **افزایش تفسیرپذیری: استفاده از تکنیک‌های هوش مصنوعی قابل توضیح (XAI) مانند SHAP/LIME یا مدل‌های ذاتاً قابل تفسیر مانند درختان تصمیم علّی، برای شفاف‌سازی تصمیمات مدل.**
* **یادگیری تقویتی علّی (Causal RL): ترکیب RL با روش‌های استنتاج علّی برای تخمین اثرات واقعی درمان و مقابله با مخدوش‌کنندگی، به منظور تضمین اینکه مدل روابط علّی را یاد می‌گیرد.**
* **یادگیری تقویتی معکوس (Inverse RL): یادگیری تابع پاداش از مشاهدات متخصصان به جای تعریف دستی آن، که می‌تواند به تعریف پاداش‌های واقع‌بینانه‌تر و مرتبط‌تر با بالین کمک کند.**
* **یادگیری تقویتی چندهدفه (Multi-objective RL): بهینه‌سازی همزمان چندین پیامد (مثلاً بقا و کیفیت زندگی) با در نظر گرفتن مبادلات بین آن‌ها.**
* **استفاده از محیط‌های شبیه‌سازی‌شده: بهره‌گیری از شبیه‌سازهای پزشکی مانند EpiCare [17] برای آموزش اولیه (pre-training) و ارزیابی سیاست‌ها در محیط‌های امن و کنترل‌شده، قبل از اعمال در محیط‌های واقعی.**

**4. مدل پیشنهادی**

**با توجه به محدودیت‌های شناسایی‌شده در ادبیات و چالش‌های کلی پیاده‌سازی RL در حوزه سلامت، مدل پیشنهادی ما بر روی سه محور اصلی برای بهبود و پیشبرد تحقیقات در زمینه بهینه‌سازی DTRs با RL تمرکز دارد: افزایش تفسیرپذیری و ایمنی، یادگیری مقاوم از داده‌های مشاهده‌ای، و طراحی توابع پاداش پیچیده‌تر و بالینی‌محور. این رویکرد جامع، به دنبال رفع موانع کلیدی برای پذیرش گسترده DTRs مبتنی بر RL در عمل بالینی است.**

**4.1. افزایش تفسیرپذیری و ایمنی در DRL برای DTRs مدل‌های DRL فعلی اغلب به دلیل ماهیت "جعبه سیاه" خود و عدم شفافیت در فرآیند تصمیم‌گیری، با عدم پذیرش در محیط‌های بالینی مواجه هستند. این امر می‌تواند منجر به بی‌اعتمادی پزشکان و بیماران به توصیه‌های خودکار شود. پیشنهاد می‌شود یک چارچوب DRL هیبریدی توسعه یابد که قدرت DRL در مدیریت ابعاد بالا و شناسایی الگوهای پیچیده را با ویژگی‌های تفسیرپذیری و تضمین‌های ایمنی ترکیب کند.**

* **رویکرد:**
  + **هوش مصنوعی قابل توضیح :(XAI) می‌توان از تکنیک‌های XAI مانند SHAP (SHapley Additive exPlanations) یا LIME (Local Interpretable Model-agnostic Explanations) برای تبیین تصمیمات سیاست RL استفاده کرد. این تکنیک‌ها قادرند "اهمیت ویژگی" (Feature Importance) را برای هر تصمیم مشخص کرده و توضیح دهند که کدام ویژگی‌های بیمار در هر مرحله (مانند سن، نتایج آزمایشگاهی، تاریخچه بیماری) بیشترین تأثیر را بر توصیه درمانی داشته‌اند. این توضیحات می‌تواند به صورت بصری یا متنی به پزشک ارائه شود.**
  + **مدل‌های ذاتاً قابل تفسیر: به جای تکیه صرف بر DRLهای پیچیده، می‌توان از مدل‌های ذاتاً قابل تفسیر مانند درختان علّی (Causal Trees) یا جنگل‌های علّی (Causal Forests) در لایه‌های خاصی از شبکه یا به عنوان بخشی از یک استراتژی تصمیم‌گیری چندمرحله‌ای استفاده کرد. این درختان به طور مستقیم اثرات علّی درمان‌ها را بر اساس ویژگی‌های بیمار مدل‌سازی می‌کنند و می‌توانند به عنوان قوانین تصمیم‌گیری قابل فهم عمل کنند.**
  + **تضمین‌های ایمنی: می‌توان از تکنیک‌های یادگیری تقویتی با محدودیت (Constrained RL) استفاده کرد. در این رویکرد، علاوه بر حداکثر کردن پاداش، قیود ایمنی خاصی (مانند عدم تجویز دوزهای سمی، یا جلوگیری از افت شدید فشار خون) در فرآیند بهینه‌سازی مدل گنجانده می‌شود. این قیود می‌توانند به عنوان توابع جریمه (Penalty Functions) در تابع هدف RL یا به عنوان محدودیت‌های سخت در فرآیند بهینه‌سازی عمل کنند.**
* **جزئیات رویکرد ما:**

**هدف روش پیشنهادی:**  
ترکیب یادگیری تقویتی عمیق با ابزارهای تفسیرپذیری و تضمین ایمنی به‌گونه‌ای که:

* **شفافیت تصمیم‌گیری** (Explainability)
* **تضمین عدم نقض قیدهای ایمنی** (Safety Constraints)
* **حفظ توانایی DRL در حل مسائل با فضای حالت بزرگ**

**۱. معماری کلی عامل  
در هستهٔ این چارچوب، از یک شبکه‌ی Dueling DQN استفاده می‌شود که دو جریان موازی برای یادگیری ارزش کلی حالت (V(s)) و مزیت نسبی هر عمل (A(s,a)) دارد. تقسیم‌بندی معماری به این دو مسیر باعث می‌شود عامل بتواند به شکلی جداگانه فاکتور «مهم بودن وضعیت» و «مزیت انجام عمل» را بیاموزد و سپس با ترکیب آن‌ها به‌صورت**

**Q(s,a) = V(s) + \Bigl(A(s,a) - \tfrac{1}{|\mathcal A|}\sum\_{a'} A(s,a')\Bigr)**

**مقادیر Q ثابتی تولید کند. برای جلوگیری از بیش‌برازش و بهبود تثبیت یادگیری، در هر لایهٔ مخفی از LayerNorm و Dropout استفاده می‌گردد. در نهایت یک زیرشبکه‌ی مجزا—«لایهٔ ایمنی»—خروجی‌های Q(s,a) را بر اساس احتمال «مجاز بودن» اقدام در آن وضعیت تعدیل می‌کند. این لایه عددی در بازهٔ [0,1] تولید می‌کند و آن را ضرب در Q-value می‌کند؛ نتیجتاً هرگاه اقدام پیشنهادی قیدهای بالینی را نقض کند، Q-value آن به صفر یا مقدار بسیار کوچک نزدیک شده و عملاً غیرقابل انتخاب می‌شود.**

**همچنین تابع پاداش به صورت چندمعیاره تعریف شده است که پنج محور بالینی را همزمان بهینه می‌کند:**

* **تثبیت فشار خون (stabilization): پاداش مثبت اگر فشار خون به محدودهٔ هدف برسد.**
* **رعایت قیدهای ایمنی (safety): جریمهٔ سنگین برای اقدامات مضر مانند تجویز وازوپرسور در فشار خون بالا.**
* **کارایی درمان (efficiency): پاداش بر اساس تغییر مثبت واقعی در فشار خون.**
* **جلوگیری از درمان بیش از حد (overtreatment): جریمه برای تجویز غیرضروری دارو.**
* **جلوگیری از درمان کمتر از حد (undertreatment): جریمه برای عدم تجویز وقتی شواهد بالینی نیاز را نشان می‌دهد.**

**هر مؤلفه با ضریب وزن‌دهی دلخواه بالینی (wi​) ترکیب و سپس به بازهٔ محدود برای پایدارسازی یادگیری نگاشت می‌شود.**

**۲. تضمین ایمنی (Constrained RL)  
پیش از هر تصمیم، قیدهای کلیدی بالینی )حداقل و حداکثر MAP، آستانهٔ لاکتات، حداکثر دوز و (…بررسی می‌شوند. تابع is\_unsafe(s,a) در صورتی‌که اقدام نقض‌کنندهٔ هر یک از این آستانه‌ها باشد، علامت می‌دهد که آن عمل ناایمن است. در نتیجه، Q-value مربوطه به طور خودکار کاهش چشمگیری می‌یابد یا به −∞ نزدیک می‌شود تا هرگز از آن استفاده نشود. این مکانیزم تضمین می‌کند عامل هیچ‌گاه اقدامی خارج از چارچوب ایمنی بالینی انجام ندهد.**

**۳. تفسیرپذیری (XAI)  
برای جلب اعتماد پزشکان و فراهم آوردن شفافیت، دو ابزار متقارن به کار گرفته می‌شود:**

1. **LIME**
   * **توضیح محلی تصمیمات با ساخت یک مدل خطی ساده پیرامون نقطهٔ کنونی.**
   * **خروجی نمودار میله‌ای از «اهمیت ویژگی‌ها» که نشان می‌دهد کدام متغیر)مثلاً MAPیا لاکتات( بیشترین تأثیر را در انتخاب عمل داشته است.**
2. **درخت علّی (Causal Decision Tree)**
   * **از یک مدل درخت تصمیم با عمق محدود برای استخراج قواعد «اگر (ویژگی ≤ آستانه) آنگاه تصمیم» استفاده می‌شود.**
   * **هر گره حاوی اطلاعات impurity، تعداد نمونه‌ها و توزیع کلاس‌هاست و رنگ‌آمیزی بر اساس غالب‌ترین کلاس (تجویز/عدم تجویز) به درک سریع منطق مدل کمک می‌کند.**
   * **نمایش گرافیکی با کتابخانه‌هایی مانند dtreeviz مختصات بصری واضحی از مسیرهای تصمیم به پزشک ارائه می‌دهد.**

**۴. تجربهٔ اولویت‌دار بهبود یافته (Optimized PER)  
برای پایدارسازی و تسریع یادگیری:**

* **ذخیره در آرایه‌های پیش‌تخصیص‌یافته )بدون overhead حافظه(**
* **وزن‌دهی تجربیات با ضریب αو به‌روزرسانی دینامیک β در طول آموزش**
* **clipping دینامیک TD-error به آستانهٔ مشخص برای جلوگیری از اولویت‌دهی نابجا و نوسانات شدید**

**۵. تنظیم خودکار پارامترها و رویکردهای مقاوم**

* **:Adaptive LR Scheduler نرخ یادگیری به‌محض آنکه متوسط TD-error بالاتر از آستانه شود، کاهش خودکار می‌یابد.**
* **:Partial Network Reset: با احتمال اندکی در فواصل منظم، چند لایه یا وزن تصادفی مجدداً مقداردهی می‌شوند تا از گیر کردن در مینیمم‌های محلی جلوگیری گردد.**
* **:Reward Scaling & Clipping جلوگیری از انفجار یا کم‌رنگ شدن مقادیر پاداش با نرمال‌سازی و قطع‌سازی در یک بازهٔ معین.**

**4.2. یادگیری مقاوم از داده‌های مشاهده‌ای و مقابله با مخدوش‌کنندگی: یکی از بزرگترین چالش‌ها در DTR-RL، وجود متغیرهای مخدوش‌کننده مشاهده‌نشده در داده‌های مشاهده‌ای بالینی است که می‌تواند منجر به یادگیری سیاست‌های سوگیرانه و غیرقابل اعتماد شود. این مشکل ناشی از عدم تخصیص تصادفی درمان در محیط‌های بالینی واقعی است.**

* **رویکرد:**
  + **تکنیک‌های استنتاج علّی پیشرفته: می‌توان از تکنیک‌های پیشرفته استنتاج علّی (مانند استفاده از متغیرهای ابزاری (Instrumental Variables)، مدل‌های ناهمگونی مشاهده‌نشده (Unobserved Heterogeneity Models)، یا رویکردهای یادگیری نمایش علّی (Causal Representation Learning) [20]) در فاز پیش‌پردازش داده‌ها یا حتی به طور مستقیم در معماری شبکه RL استفاده کرد. این روش‌ها تلاش می‌کنند تا اثرات مخدوش‌کننده‌های مشاهده‌نشده را خنثی کنند.**
  + **برآوردگرهای مقاوم : برای مقابله با چالش‌های علّی در داده‌های مشاهده‌ای، مدل پیشنهادی می‌تواند از برآوردگرهای وزن‌دهی احتمال معکوس افزوده (AIPW) برای تخمین پیامدهای پیامدسازی متقابل به صورت مقاوم استفاده کند. فرمول‌های (4) و (5) که در بخش 3.2.1 برای محاسبه میانگین پیامد پیامدسازی متقابل تحت یک درمان یا یک قانون درمانی مشخص استفاده می‌شوند، می‌توانند هسته ارزیابی سیاست‌ها را در حضور سوگیری‌های ناشی از داده‌های مشاهده‌ای تشکیل دهند. این برآوردگرها با اصلاح برای نمرات گرایش، تلاش می‌کنند تا عدم تصادفی بودن تخصیص درمان در داده‌های مشاهده‌ای را کاهش داده و تخمین‌های بدون سوگیری ارائه دهند.**
  + **ادغام با شبیه‌سازها: می‌توان از محیط‌های شبیه‌سازی‌شده دقیق مانند [17] EpiCare برای آموزش اولیه (pre-training) مدل بر روی داده‌های کاملاً کنترل‌شده و بدون مخدوش‌کنندگی استفاده کرد. سپس، مدل می‌تواند با داده‌های واقعی و محدود تنظیم دقیق (fine-tuning) شود. این رویکرد ترکیبی به مدل کمک می‌کند تا اصول کلی را در محیطی امن بیاموزد و سپس آن‌ها را با پیچیدگی‌های داده‌های واقعی تطبیق دهد و مقاومت مدل در برابر نویز و سوگیری را افزایش دهد.**

**4.3. بهینه‌سازی به شکل DTR چندهدفه: پیامدهای بیمار اغلب چندوجهی هستند (مثلاً افزایش بقا، بهبود کیفیت زندگی، کاهش عوارض جانبی، بهینه‌سازی هزینه درمان) و نیاز به تعادل میان آن‌ها وجود دارد. همچنین، تعریف یک تابع پاداش عددی واحد و دقیق برای RL در پزشکی که به‌طور کامل اهداف بالینی را منعکس کند، دشوار است.**

* **رویکرد:**
  + **یادگیری تقویتی چندهدفه (Multi-objective RL): می‌توان رویکردهای یادگیری تقویتی چندهدفه را پیاده‌سازی کرد که به‌طور همزمان چندین پیامد را بهینه‌سازی کنند. این ممکن است شامل استفاده از تکنیک‌های بهینه‌سازی پارتو (Pareto Optimization) برای یافتن مجموعه‌ای از سیاست‌های بهینه باشد که مبادلات مختلف بین اهداف را نشان می‌دهند. این مجموعه سیاست‌ها می‌تواند به پزشکان گزینه‌های متنوعی را ارائه دهد تا بر اساس اولویت‌های بیمار و شرایط بالینی، بهترین انتخاب را انجام دهند.**
  + **یادگیری تقویتی معکوس (Inverse RL): برای تعریف تابع پاداش، می‌توان از یادگیری تقویتی معکوس (Inverse RL) استفاده کرد. این روش به جای تعریف دستی تابع پاداش، آن را از مشاهدات رفتار متخصصان بالینی یا از داده‌های ترجیحات بیمار استنتاج می‌کند. این رویکرد می‌تواند منجر به پاداش‌هایی شود که واقع‌بینانه‌تر و مرتبط‌تر با بالین هستند، زیرا منعکس‌کننده اهداف و ارزش‌های واقعی متخصصان و بیماران است.**
  + **توابع پاداش ترکیبی و دینامیک: می‌توان توابع پاداش ترکیبی را طراحی کرد که ترکیبی از شاخص‌های بقا، کیفیت زندگی، رویدادهای نامطلوب و حتی هزینه‌ها باشند. وزن‌دهی این مولفه‌ها می‌تواند به صورت دینامیک و بر اساس تغییر وضعیت بیمار یا ترجیحات او در طول زمان تنظیم شود.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| روش / مقاله | سال / مرجع | نوع روش | ایدهٔ کلی | مزایا اصلی | محدودیت‌ها / ریسک‌ها | متریک‌ها / ارزیابی‌ها |
| Deep RL دو مرحله‌ای (Liu et al.) | 2018 | DRL هیبریدی | ابتدا رفتار بالینی را با یادگیری نظارت‌شده پیش‌آموزش می‌دهد، سپس با DRL policy را بهینه می‌کند؛ از autoencoder برای کاهش بُعد استفاده شده | قابلیت مدیریت فضای حالت/عمل بزرگ؛ استفاده از دانش بالینی برای جهت‌دهی آموزش | نیاز به تابع پاداش هویوریستیک؛ آفلاین بودن و تعمیم‌پذیری محدود | دقت در پیش‌بینی اقدامات، مقدار انتظار تابع پاداش (surrogate reward) |
| T-RL (Tree-based RL) | — | روش درختی (قابل تفسیر) | ایجاد درخت تصمیم برای هر مرحله و استفاده از برآوردگرهای AIPW برای ارزیابی پیامدها | تفسیرپذیری بالا (قواعد If-Then)؛ مناسب برای توضیح بالینی | محدودیت در فضاهای حالت/عمل بسیار بزرگ | برآورد میانگین پیامد متقابل، معیارهای خلوص تقسیم |
| ST-RL (Stochastic Tree Search) | ~2020 | جستجوی درخت تصادفی | جستجوی تصادفی در فضای درخت‌ها با استفاده از مدل‌های غیرپارامتریک مانند BART برای برآورد پیامدها | کاوش بهتر فضای قوانین؛ جلوگیری از بهینه‌های محلی | هزینه محاسباتی بالا؛ نیاز به نمونه‌برداری گسترده | عملکرد DTR، زمان اجرا، تنوع قوانین پیدا شده |
| روش مبتنی بر درخت علّی (Causal Tree) | — | درخت علّی | تقسیم درخت با معیارهای مخصوص جهت تخمین اثرات درمان (EMSE)؛ تمرکز بر برآورد ناهمگونی اثر | برآورد اثرات علّی ناهمگن؛ تفسیرپذیری قوی | حساسیت به داده‌های کم یا سانسورشده؛ نیاز به پیش‌پردازش | EMSE، ثبات برآوردها، خطاهای برآورد |
| BR-DTR (Benefit-Risk stagewise) | Zhao et al., 2024 | کنترل منفعت-ریسک مرحله‌ای | بهینه‌سازی منفعت-ریسک در هر مرحله با قیود نرم؛ الگوریتم‌های شبیه SVM برای مرحله‌ها | ایمنی صریح در هر مرحله؛ تضمین‌های نظری محلی | مقیاس‌پذیری محدود با افزایش مراحل؛ پارامترسازی حساس | معیار منفعت-ریسک، کران‌های خطا |
| UC-DTR (Zhang & Bareinboim) | 2020 | RL خوش‌بینانه با کران علّی | الگوریتم خوش‌بینانه برای یادگیری آنلاین از داده‌های مشاهده‌ای با کران‌های نظری | تضمین‌های نظری (regret)؛ مقاومت نسبی در برابر مخدوشی | پیچیدگی محاسباتی بالا؛ نیاز به فرضیات خاص | پشیمانی (regret)، کران‌های نظری، عملکرد آنلاین |
| چارچوب پیشنهادی | — | DRL هیبریدی + XAI + constrained-RL+Causal Decision Tree | معماری ترکیبی: Dueling-DQN با لایهٔ ایمنی، چندمعیاره کردن پاداش، AIPW برای مقابله با مخدوشی، ابزارهای XAI (LIME, causal trees) | ترکیب کارایی و تفسیرپذیری؛ توجه به ایمنی بالینی | پیچیدگی معماری؛ نیاز به تنظیم وزن‌های پاداش و ارزیابی بالینی | معیارهای بالینی (ایمنی، ثبات)، معیارهای یادگیری (TD-error، OPE)، شاخص‌های تفسیرپذیری |

**5. واژه‌نامه**

* **رژیم درمانی پویا :(DTR - Dynamic Treatment Regime) مجموعه‌ای از قوانین تصمیم‌گیری فردی که درمان بهینه را در مراحل مختلف بیماری بر اساس تاریخچه بیمار و وضعیت فعلی او دیکته می‌کند.**
* **یادگیری تقویتی :(RL - Reinforcement Learning) یک پارادایم یادگیری ماشین که در آن یک عامل با تعامل با محیط و دریافت پاداش، یک سیاست بهینه برای تلاش برای رسیدن به بهترین وضعیت و بیشترین پاداش در بلندمدت را می‌آموزد.**
* **مدل یادگیری تقویتی عمیق(DRL - Deep Reinforcement Learning) : ترکیبی از یادگیری تقویتی و شبکه‌های عصبی عمیق برای حل مسائل پیچیده‌تر با فضاهای حالت و عمل بزرگ.**
* **فرآیند تصمیم مارکوف : (MDP - Markov Decision Process) چارچوب ریاضی برای مدل‌سازی فرآیندهای تصمیم‌گیری ترتیبی که در آن نتایج فعلی فقط به وضعیت و عمل فعلی بستگی دارند (خاصیت مارکوف).**
* **Q -یادگیری: (Q-learning) یک الگوریتم یادگیری تقویتی بدون مدل (Model-Free) که یک تابع ارزش-عمل (Q-value) را برای یافتن سیاست بهینه یاد می‌گیرد. این تابع نشان‌دهنده ارزش مورد انتظار انجام یک عمل خاص در یک وضعیت خاص و سپس پیروی از سیاست بهینه است.**
* **سیاست :(Policy) یک استراتژی یا مجموعه قوانین که به عامل می‌گوید در هر وضعیت (State) چه عملی (Action) را انجام دهد.**
* **تابع ارزش :(Value Function) انتظار از مجموع پاداش‌های آتی که از یک وضعیت خاص یا انجام یک عمل خاص در آن وضعیت به دست می‌آید. این تابع به عامل کمک می‌کند تا بهترین مسیر را انتخاب کند.**
* **پاداش : (Reward) یک سیگنال عددی (مثبت یا منفی) که کیفیت یک عمل انجام‌شده در یک وضعیت خاص را نشان می‌دهد و راهنمای یادگیری عامل است.**
* **حالت :(State) توصیف کامل و کافی از وضعیت فعلی محیط (مثلاً مجموعه ویژگی‌های بالینی بیمار در یک لحظه خاص).**
* **عمل :(Action) عملی که عامل می‌تواند در یک وضعیت خاص انجام دهد (مثلاً تجویز یک دوز خاص از دارو یا انجام یک عمل جراحی).**
* **داده‌های مشاهده‌ای :(Observational Data) داده‌هایی که از طریق مشاهده رویدادها در محیط‌های واقعی (مانند بالین بیمارستان) جمع‌آوری می‌شوند و نه از طریق آزمایش‌های کنترل‌شده تصادفی‌شده . این داده‌ها معمولاً مستعد سوگیری هستند.**
* **مخدوش‌کنندگی :(Confounding) وجود متغیرهایی در داده‌های مشاهده‌ای که هم بر تخصیص درمان و هم بر پیامد بیمار تأثیر می‌گذارند و می‌توانند منجر به ارتباطات کاذب (غیرعلّی) شوند.**
* **ارزیابی خارج از ‌سیاست :(OPE - Off-policy Evaluation) فرآیند تخمین ارزش یک سیاست بدون نیاز به اجرای آن سیاست در محیط، با استفاده از داده‌های از پیش جمع‌آوری شده تحت یک سیاست دیگر (سیاست رفتاری).**
* **آزمایش تصادفی چندگانه ترتیبی (SMART - Sequential Multiple Assignment Randomized :Trial) نوعی طراحی کارآزمایی بالینی که به محققان امکان می‌دهد DTR های مختلف را به طور سیستماتیک در مراحل مختلف درمان مقایسه و ارزیابی کنند.**
* **تفسیرپذیری :(Interpretability) قابلیت درک و توضیح چگونگی رسیدن یک مدل هوش مصنوعی به یک تصمیم یا پیش‌بینی خاص، که برای اعتماد و پذیرش مدل در کاربردهای حساس مانند پزشکی حیاتی است.**
* **تعمیم‌پذیری :(Generalizability) توانایی یک مدل آموزش‌دیده برای حفظ عملکرد خود بر روی داده‌ها یا جمعیت‌های جدید و ندیده‌شده که با داده‌های آموزشی متفاوت هستند.**
* **سانسور :(Censoring) وضعیتی در داده‌های بقا که در آن زمان دقیق رویداد (مثلاً مرگ یا عود بیماری) برای برخی از افراد به دلیل از دست دادن پیگیری یا پایان مطالعه به‌طور کامل مشاهده نمی‌شود.**
* **پاداش اکتشافی :(Heuristic Reward) یک تابع پاداش که بر اساس قواعد تجربی، شهود متخصصان یا دانش دامنه تعریف می‌شود، نه لزوماً بر اساس یک تعریف ریاضی دقیق از پاداش‌های ذاتی محیط.**
* **فرض ریسک حاد :(Acute Risk Assumption) یک فرض در برخی مدل‌ها که ریسک‌ها به‌صورت فوری یا در یک بازه زمانی کوتاه پس از درمان ظاهر می‌شوند و نه به صورت بلندمدت یا تجمعی.**
* **یادگیری تقویتی معکوس: (IRL - Inverse Reinforcement Learning) فرآیندی که در آن تابع پاداش یک عامل از مشاهدات رفتار بهینه (مثلاً تصمیمات متخصصان بالینی یا ترجیحات بیمار) استنتاج می‌شود، به جای تعریف صریح آن.**
* **یادگیری فدرال :(Federated Learning) یک رویکرد یادگیری ماشین غیرمتمرکز که در آن مدل‌ها در چندین دستگاه یا سرور محلی آموزش می‌بینند و تنها به‌روزرسانی‌های مدل به یک سرور مرکزی ارسال می‌شوند، که به حفظ حریم خصوصی داده‌ها کمک می‌کند.**
* **خودکدگذارهای واریانسی انتقالی :(tVAEs - Transitional Variational Autoencoders) نوعی شبکه عصبی عمیق که برای مدل‌سازی توالی‌های داده‌ای و تولید داده‌های جدید استفاده می‌شود، به‌ویژه در زمینه شبیه‌سازی داده‌های پزشکی و درک پویایی تغییر وضعیت بیمار.**

**6. مراجع**

1. **Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9), 1393-1512.**
2. **Hernán, M. A., & Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7), 578-586.**
3. **Hernán, M. A., & Robins, J. M. (2010). Causal inference. *Chapman and Hall/CRC*.**
4. **Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13), 1216-1219.**
5. **Laber, E. B., & Zhao, Y. (2015). Tree-based methods for individualized treatment regimes. *Biometrics*, 71(4), 856-865.**
6. **Daniel Rodriguez Duque, Erica E.M. Moodie, and David A. Stephens. (2023). Bayesian inference for optimal dynamic treatment regimes in practice. *ArXiv***
7. **Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1984). *Classification and regression trees*. CRC press.**
8. **Zhang B, Tsiatis A, Laber E, et al. (2013)Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika. 2013 doi: 10.1093/biomet/ast014.**
9. **Ying-Qi-Zhao., & Eric B Lander. (2014). Estimation of Optimal Dynamic Treatment Regimes. *PMC*,** [**PMC4247353**](https://pmc.ncbi.nlm.nih.gov/articles/PMC4247353/)**.**
10. [**n**](https://www.tandfonline.com/author/Sun%2C+Yilun)**Yilun Sun & Lu Wang. (2020). Stochastic tree search for estimating optimal dynamic treatment regimes.  *Journal of the American Statistical Association*,** [**doi.org/10.1080/01621459.2020.1819294**](https://doi.org/10.1080/01621459.2020.1819294)
11. **Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. CRC press.**
12. **Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97-109.**
13. **Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using causal forests. *Journal of the American Statistical Association*, 113(523), 1229-1242.**
14. **Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353-7360.**
15. **Ning Liu, Ying Liu, Brent Logan, Zhiyuan Xu, Jian Tang, Yanzhi Wang. (2018). Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data. *PMC*.** [**PMC5856473**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5856473/)
16. **Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, Michael R Kosorok. (2020). Estimating Dynamic Treatment Regimes in Mobile Health Using V-learning. *PMC*.** [**PMC7500510**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7500510/)
17. **Mason Hargrave, Alex Spaeth, Logan Grosenick. (2024). EpiCare: A Reinforcement Learning Benchmark for Dynamic Treatment Regimes. *PMC*.** [**PMC12124763**](https://www.google.com/search?q=https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12124763/)
18. **Mochuan Liu, Yuanjia Wang, Haoda Fu,** **Donglin Zeng. (2024). Learning Optimal Dynamic Treatment Regimes Subject to Stagewise Risk Controls. *Journal of Machine Learning Research*.**  [**Link**](https://www.jmlr.org/papers/volume25/23-0072/23-0072.pdf)
19. **Zhang, S., & Bareinboim, E. (2020). Near-Optimal Reinforcement Learning in Dynamic Treatment Regimes. *33rd Conference on Neural Information Processing Systems*.**  [**Link**](https://causalai.net/r48.pdf)
20. **Sophia Yazzourh, Nicolas Savy, Philippe Saint-Pierre,** **Michael R. Kosorok. (2024). Medical Knowledge Integration Into Reinforcement Learning Algorithms for Dynamic Treatment Regimes. *International Statistical Review*.**  [**Link**](https://arxiv.org/pdf/2407.00364)