همگرایی هوش مصنوعی و زیست‌فناوری: چشم‌انداز نوین برای تحول علوم زیستی

آرین نوحی، شکیبا درویش علیپور آستانه
دانشکده زیست فناوری، پردیس علوم و فناوری نوین، دانشگاه سمنان، سمنان، ایران
Darvishalipour@semnan.ac.ir
نامه علوم پایه شماره ۱۵، بهار و تابستان ۱۴۰۴

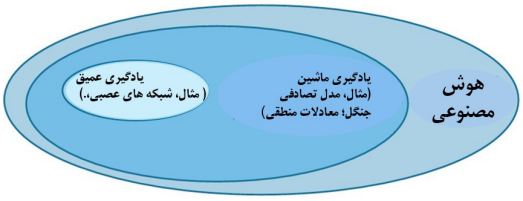
چکیده

پیشرفت فناوری‌های پرسرعت در علوم‌زیستی، مانند توالی‌یابی نسل جدید (NGS) و طیف‌سنجی جرمی، حجم زیادی از داده‌ها تولید می‌کند. هدف اصلی، کمی‌سازی داده‌های دی‌ان‌ای، آران‌ای، پروتئین‌ها و متابولیت‌ها است؛ مولکول‌هایی که در تعیین ساختار، عملکرد و پویایی سامانه‌های زنده مانند سلول و بافت نقش دارند. تحلیل این داده‌ها مستلزم به‌کارگیری ابزارهای پیشرفته است، زیرا مدیریت، ذخیره‌سازی و استخراج مفاهیم زیستی از این حجم گسترده اطلاعات، چالش‌برانگیز است. با ظهور هوش مصنوعی و ترکیب آن با زیست‌فناوری، چشم‌انداز پژوهش‌های زیستی دگرگون شده است. کاربردهای یادگیری ماشین در تحلیل داده‌های حجیم، امکان درک دقیق‌تر عملکرد اجزای سامانه‌های زیستی و تعاملات میان آن‌ها را فراهم می‌سازد. هم‌افزایی میان هوش مصنوعی و زیست‌فناوری، نقشی مهم در تسریع کشف دارو، پزشکی فردمحور، مهندسی زیستی و ویرایش ژن ایفا می‌کند و در حوزه‌هایی مانند ژنومیکس و پروتئومیکس نیز کاربرد گسترده‌ای دارد. افزون بر این، این همگرایی با ارائه راهکارهای سازگار با محیط زیست، مانند بیوپلاستیک‌ها و بیودیزل، می‌تواند در رفع چالش‌های جهانی مؤثر باشد.

کلیدواژگان: زیست‌فناوری، علوم زیستی، فناوری پرسرعت، هوش مصنوعی، یادگیری عمیق، یادگیری ماشین

مقدمه

جان مک‌کارتی[1] در سال ۱۹۵۶، در کنفرانس دارتموث، برای نخستین بار اصطلاح «هوش مصنوعی[2]» (AI) را به کار برد تا به حوزه‌ای اشاره کند که هدف آن شبیه‌سازی فرایندهای هوشمند در ماشین‌ها بود. هوش مصنوعی به توانایی سامانه‌ها برای انجام وظایفی مانند یادگیری از تجربه، استدلال، و واکنش مناسب به محیط اطراف اشاره دارد. یادگیری ماشین[3] (ML) و یادگیری عمیق[4] (DL) از عناصر مهم هوش مصنوعی (شکل۱)، بر استفاده از داده‌های متفاوت تکیه دارند، داده‌هایی که مستقیم یا غیرمستقیم از هوش طبیعی استخراج شدند (۱).



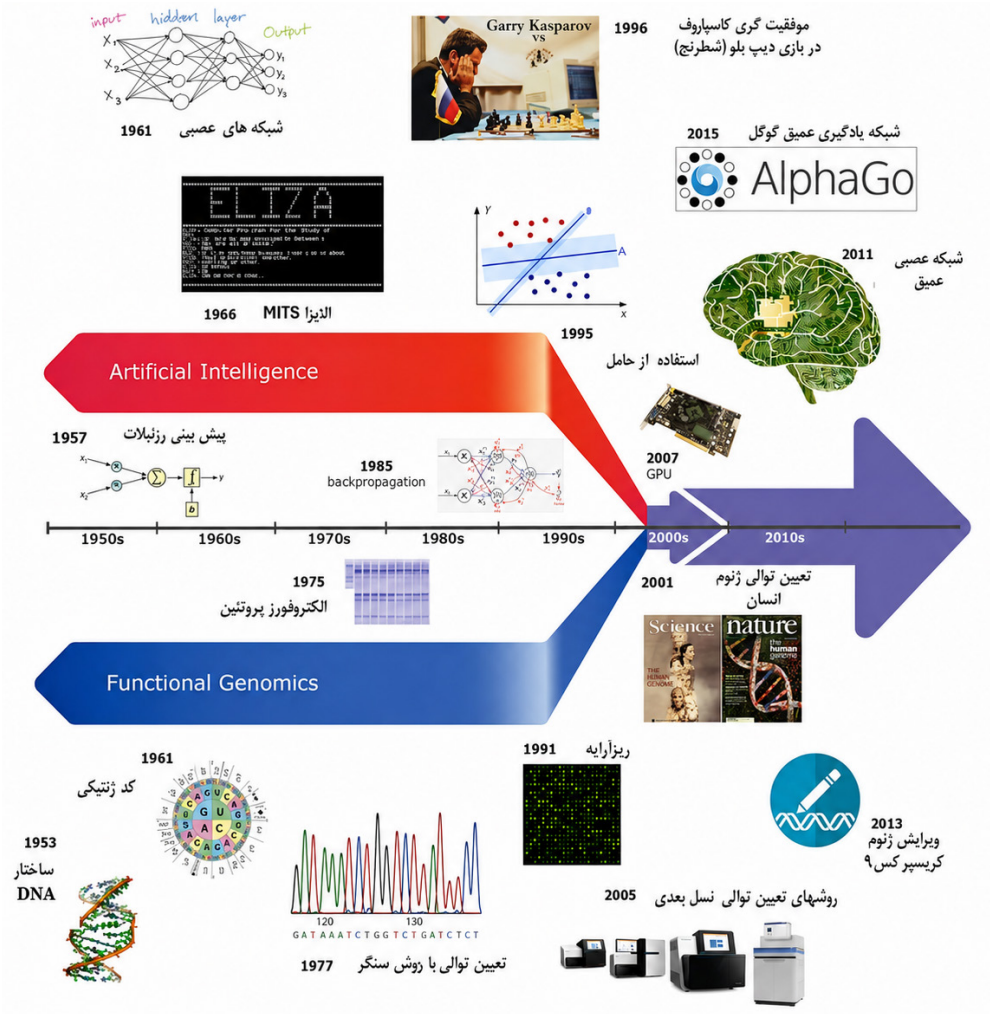
شکل ۱. یادگیری ماشین و یادگیری عمیق به عنوان بخشی از هوش مصنوعی (۲۶).

نخستین کاربردهای یادگیری ماشین در زیست‌شناسی به اوایل دهه ۱۹۸۰ برمی‌گردد. این فناوری برپایه الگوریتم‌هایی است که از روش‌های آماری برای کشف الگوهای داده و تحلیل آن‌ها با سطحی از اطمینان استفاده می‌کند. با گسترش اطلاعات در پایگاه‌های داده و گذر زمان، قابلیت استنتاج و دقت این الگوریتم‌ها ارتقا یافته و سامانه پیشرفته‌تر، هوشمندتر و خودآگاه‌تر شده است. روش‌های یادگیری ماشین در یکی از سه دسته زیر قرار می‌گیرند

- یادگیری نظارت‌شده[5] به الگوریتم‌هایی متکی است که توسط برنامه‌نویس انسانی تنظیم میشود و هدف آن‌ها تحلیل اطلاعات مشخص و شناخته‌شده برای رسیدن به نتیجه‌ای معین است.
- یادگیری بدون نظارت[6] به تحلیل اطلاعات ناشناخته توسط الگوریتم اشاره دارد؛ هدف این روش، شناسایی الگوهایی در پایگاه داده است که از پیش مشخص نیست.
- یادگیری تقویتی[7] را می‌توان ترکیبی از یادگیری نظارت‌شده و بدون نظارت دانست. هدف این روش، آزمودن الگوریتم پایه برای ارزیابی دقت پیش‌بینی‌های آن است.

1. John McCarthy
2. Artificial intelligence
3. Machine learning
4. Deep learning
5. Supervised leaning
6. Unsupervised learning
7. Reinforcement learning

مرحله بعدی، «یادگیری عمیق» به استفاده از الگوریتم‌هایی به نام شبکه‌های عصبی مصنوعی[1] اشاره دارد. «پرسپترون»[2] اولین معماری شبکه عصبی، که در سال ۱۹۵۸ توسط فرانک روزنبلات معرفی شد، محدودیت‌هایی در یادگیری داشت. علاوه بر این، کامپیوتری که الگوریتم پرسپترون را اجرا می‌کرد، گرچه بزرگ بود، توانایی پردازش نسبتاً کمی داشت و به دلیل هزینه‌های بالا مالی و محاسباتی، غیرقابل نگهداری بود. در دهه اخیر، از واحدهای پردازش‌دهنده گرافیک (GPUs)[3] در یادگیری عمیق استفاده شده است، بنابراین امکان گسترش عملکرد برای پردازنده‌ها در محاسبات وسیع فراهم گردید. این شبکه‌ها فرآیندهای پردازش اطلاعات و تصمیم‌گیری در انسان را شبیه‌سازی می‌کنند، مانند لایه‌هایی که ساختار و عملکرد مغز را تقلید می‌نمایند تا الگوهای شناخته‌شده یا ناشناخته را تشخیص دهند. بطومثال، هر لایه ممکن است ویژگی خاصی از یک شیء را تحلیل کند و پس از آن، نتیجه را به لایه بعدی ارسال کند تا با ویژگی‌های دیگر ترکیب یا پالایش شود. این روند تا زمانی ادامه می‌یابد که شیء به طور قاطع شناسایی یا دسته‌بندی شود. اخیراً، این فناوری در تمام حوزه‌های پژوهشی مرتبط با فعالیت‌های ژنگان در شرایط زیستی مختلف، مانند ژنومیکس، ترانسکریپتومیکس، پروتئومیکس و متابولومیکس، به کار گرفته میشود تا با انجام عملیات ریاضی، داده‌ها را در چندین لایه و متصل به یکدیگر (مانند شبکه عصبی) پردازش کنند (شکل۲).



شکل ۲. جدول زمانی از اطلاعات مهم فعالیت ژنگان در شرایط زیستی مختلف و هوش مصنوعی از زمان پایه‌گذاری تاکنون (۴).

سومین مرحله هوش مصنوعی، «رایانش شناختی»[4] است. این حوزه چند نوع یادگیری ماشینی را کنار هم قرار می‌دهد تا پرسش‌ها را بدون نیاز به راهنمایی مستقیم برنامه‌نویس پاسخ دهد. یکی از ویژگی‌های کلیدی آن، توانایی برقراری ارتباط مستقیم و یادگیری از تجربیات انسان‌ها است. عملکرد این سامانه‌ها با گذر زمان و بدون نیاز به برنامه‌ریزی خاص برای وظایف مشخص، بهبود می‌یابد. بنابراین، در برابر برنامه‌های مختلف بسیار انعطاف‌پذیر هستند (۲).

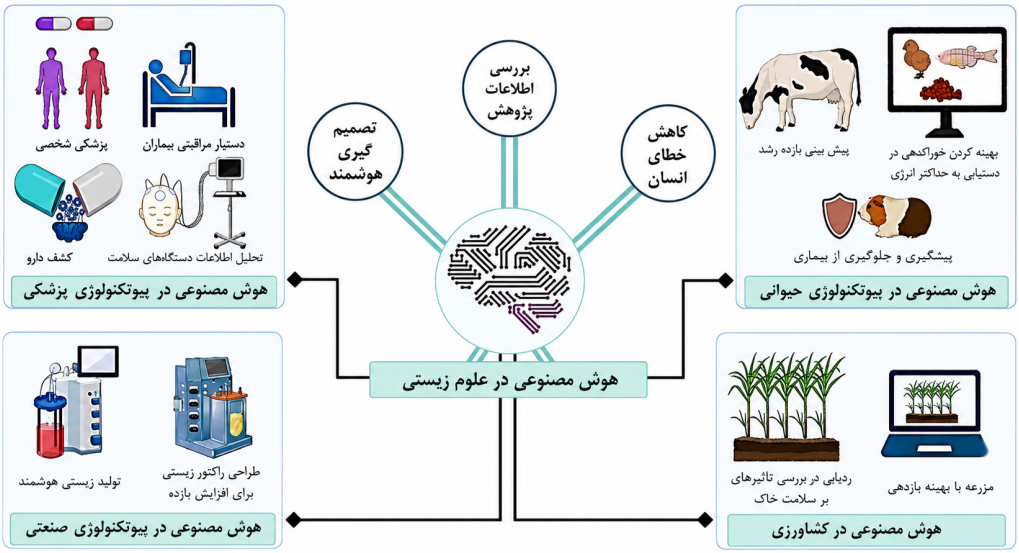
1. Artificial neural network
2. Perceptron
3. Graphical Processing Units
4. Cognitive computing

این نوع فناوری‌ها در مدیریت چالش‌های مرتبط با داده‌های زیستی با ابعاد بالا، روشی مؤثر ارائه می‌دهند که راه‌حل‌های سریع‌تر، دقیق‌تر و مقیاس‌پذیرتری انتخاب شود. زیرا در روش‌های سنتی، به دلیل نیاز به مداخله دستی در حجم بالای محاسبات، تحلیل داده‌های بزرگ با محدودیت‌های همراه بود. برای نمونه، الگوریتم‌های یادگیری ماشین در تحلیل داده‌های ژنی، پیش‌بینی ارتباط بین ژن‌ها و بیماری‌ها، بهینه‌سازی مسیرهای متابولیک در زیست‌فناوری مصنوعی، یا پیش‌بینی تاشدن پروتئین‌ها نیاز به آزمایش‌های آزمون و خطا را کاهش می‌دهند (۳).

مزیت اصلی یادگیری عمیق نسبت به روش‌های سنتی این است که امکان دریافت نتایج به صورت طبقه‌بندی یا پیش‌بینی به‌صورت مستقیم از داده‌های خام را فراهم می‌کند. هرچند در این فرآیند احتمال سوگیری[1] وجود دارد، اما با عدم نیاز به مداخله دستی در مراحل پیش‌پردازش داده، از بروز سوگیری ناشی از این مداخله تا حد زیادی جلوگیری می‌شود. همچنین، یادگیری عمیق امکان ترکیب انواع مختلف اطلاعات ورودی –از جمله متنی، عددی، تصاویر و فایل‌های صوتی– را به‌طور یکپارچه و کارآمد فراهم می‌آورد. از سوی دیگر، این رویکردها در استخراج ویژگی‌های پیچیده از چنین داده‌هایی، نسبت به روش‌های سنتی یادگیری ماشین توانمندتر هستند ولی موجب کاهش شفافیت مدل و محدودیت در قابلیت ردیابی مسیر استخراج ویژگی‌ها، می‌شوند. افزایش پیچیدگی معماری‌های یادگیری عمیق و وجود لایه‌های پنهان متعدد، هرچند برای استخراج ویژگی‌ها در سطوح بالا مؤثر است، اما به دلیل ماهیت این مدل (جعبه‌سیاه)[2]، تفسیرپذیری آن کاهش می‌یابد و ممکن است، مشکلات اخلاقی و حقوقی ایجاد کند. همچنین، به دلیل تعدد پارامترهای قابل تنظیم و نیاز به تنظیم دقیق و زمان‌بر بودن آن‌ها، این مدل‌ها در فرآیند آموزش با چالش‌هایی مواجه است. در نهایت، نیاز به منابع محاسباتی قابل توجه و زمان طولانی آموزش می‌تواند محدودیت‌هایی در کاربردهای عملی این روش‌ها ایجاد کند (۴). در ادامه، نقش هوش مصنوعی در روش‌های زیست‌فناوری برای حل مسائل پیچیده بررسی می‌شود.

هوش مصنوعی و نقش آن در زیست‌فناوری

ترکیب هوش مصنوعی با ابزارها و فناوری‌های زیستی، کارایی و اثربخشی در پاسخ به چالش‌های جهانی و مسائل زیست‌محیطی را افزایش داده است. این هم‌افزایی، مسیرهای نوینی برای مدیریت پایدار و نوآورانه محیط زیست در شاخه‌های مختلف زیست‌فناوری، در حوزه‌های پزشکی، دامپزشکی، گیاهی، صنعتی، غذایی، بهداشتی، کشاورزی، محیط‌زیست، دریا و نانوزیست‌فناوری[3] فراهم کرد (شکل۳).



شکل ۳. یک مدل مفهومی از کاربردهای هوش مصنوعی در رشته‌های سلامت، کشاورزی، علوم دامی و زیست‌فناوری صنعتی (۱).

در زیست‌فناوری پزشکی، هوش مصنوعی نقش کلیدی در توسعه پزشکی‌شخصی و بهبود سامانه‌های تشخیصی دارد؛ و امکان طراحی درمان‌هایی که دقیقاً با پروفایل ژنتیکی هر فرد تطابق دارند را فراهم می‌کند (۱). همچنین، در نظارت بر بیماری‌ها با توسعه روش‌های بهداشت عمومی، تحلیل‌های پیش‌بینی‌کننده در زیست‌فناوری سلامت به کار گرفته شده است. نانوزیست‌فناوری نیز با بهره‌گیری از هوش مصنوعی در بهبود سامانه‌های تحویل دارو و فناوری‌های تشخیصی، تحولی نوآورانه ایجاد کرد و آینده‌ای بسیار امیدوارکننده دارد.

علاوه بر این، ورود هوش مصنوعی در زیست‌فناوری صنعتی نقش مهمی در بهینه‌سازی فرآیندهای تولید و توسعه ترکیبات جدید ایفا می‌کند. سامانه‌های نظارتی و مدیریتی مبتنی بر این فناوری تاثیر

1. Biased
2. Black-box
3. Nanobiotechnology

زیادی در محیط‌زیست و کشاورزی داشته‌اند، و منجر به صرفه‌جویی در مصرف منابع، افزایش کارایی محصولات و بهبود شاخص‌های کلیدی شده است. توسعه سوخت‌های زیستی از جلبک‌ها یا دیگر زیست‌توده‌ها، با کمک زیست‌فناوری، به تأمین منابع پایدار و تجدیدپذیر انرژی کمک کرده و وابستگی به سوخت‌های فسیلی و انتشار کربن را کاهش می‌دهد. مدل‌های پیش‌بینی مبتنی بر یادگیری ماشین با استفاده از الگوریتم‌های هوش مصنوعی با دقت بالا، بهینه‌سازی شرایط رشد در فرآیندهای تولید سوخت‌های زیستی، زمان برداشت و شرایط فرآیند را بهبود می‌بخشند و با افزایش بهره‌وری، کاهش هزینه‌ها، این سوخت‌ها را به گزینه‌ای رقابتی‌تر تبدیل می‌کنند (۵).

در زمینه زیست‌پالایی[1]، برخی باکتری‌ها و گیاهان می‌توانند فلزات سنگین را از خاک‌ها و آب‌های آلوده جذب کرده و آن‌ها را غیرسمی سازند، به‌طوری که مناطق آلوده به مواد صنعتی مجدداً برای جانداران و انسان‌ها ایمن شوند. مدل‌های هوش مصنوعی اثربخشی میکروارگانیسم‌ها یا گیاهان خاص در جذب فلزات سنگین را پیش‌بینی می‌کنند و فرآیند زیست‌پالایی در مناطق آلوده را بهینه می‌سازند. علاوه براین، یادگیری عمیق برای شناسایی فلزات سنگین و پیش‌بینی تعاملات آن‌ها با عوامل زیست‌پالایی خاص به کار می‌رود تا کارایی سامانه را در جذب و تجزیه فلزات افزایش دهد.

سامانه‌های اطلاعات جغرافیایی[2] (GIS) یکپارچه‌شده با هوش مصنوعی برای شناسایی و پایش آلاینده‌های هوا، جهت انجام واکنش زیستی با استفاده از مواد زنده نظیر میکروارگانیسم‌ها و گیاهان به عنوان فیلترهای زیستی[3] رواج یافته است. هدف از کاربرد این فناوری، کنترل انتشار صنعتی[4] و در نتیجه کاهش سطح مواد مضر آزاد شده به اتمسفر است. هوش مصنوعی با بهینه‌سازی در طراحی، عملکرد و حفظ کارایی مطلوب فیلترها نقش کلیدی دارد. این فناوری با کنترل پارامترهایی مانند سرعت جریان هوا و رطوبت، اجزای تشکیل‌دهنده آلاینده‌های خاص را مشخص می‌کند. فیلترهای زیستی که با نرم‌افزارهای شبیه‌ساز مبتنی بر هوش مصنوعی ساخته شده‌اند، در مدل‌سازی جریان هوا و تجزیه آلاینده‌ها بهینه می‌شوند. همچنین، یادگیری ماشین برای انتخاب و ترکیب میکروارگانیسم‌ها به منظور حذف آلاینده‌های خاص به کار می‌رود.

الگوریتم‌های هوش مصنوعی با تنظیم پارامترهای شیمیایی، تبدیل روغن‌های ضایعاتی نباتی و چربی‌های حیوانی را به بیودیزل[5] را بهینه و به حفظ محیط‌زیست کمک می‌کنند. این امر منجر به افزایش بهره‌وری، کاهش مصرف انرژی، بهبود و ارتقاء کیفیت محصول نهایی می‌شود. همچنین، مدل‌های یادگیری ماشین برای پیش‌بینی کیفیت بیودیزل تولید شده از مواد اولیه مختلف مورد استفاده قرار می‌گیرند تا اطمینان حاصل شود که استانداردهای لازم رعایت می‌شوند (۶).

در انتخاب و توسعه‌ی تکنیک‌های ویرایش ژن و پیش‌بینی نتایج حاصل از تغییرات ژنتیک، ابزارهای ویرایش ژن کریسپر- کَس9[6] مورد استفاده محققان است. مهندسی ژنتیک و فناوری کریسپر، امکان معرفی ارقام گیاهیِ دارای مقاومت بیشتر در برابر آفات، بیماری‌ها و شرایط اقلیمیِ نامساعد را فراهم می‌سازند. این امر سبب کاهش وابستگی به آفت‌کش‌ها و کودهای شیمیایی، کاهش رواناب کشاورزی[7] و توسعه روش‌های سازگار با محیط زیست مانند استفاده از کودها، آفت‌کش‌های زیستی و افزایش سلامت اکوسیستم می شود. روش‌های یادگیری ماشین با تجزیه و تحلیل سریع‌تر داده ژنتیک، به شناسایی ژن‌های مؤثر بر مقاومت به انواع استرس کمک کرده و تولید محصولات مقاوم به آفات، بیماری‌ها و تغییرات اقلیمی[8] را سرعت می‌بخشند (۷). سامانه‌های هوش مصنوعی سلامت محصول و وضعیت خاک را تحلیل می‌کنند، خطرات مرتبط با روان‌آب در روش‌های مختلف کشاورزی را پیش‌بینی کرده و پیشنهاداتی برای استفاده دقیق‌تر از آفت‌کش‌ها و کودهای زیستی جهت کاهش روان‌آب‌ها و نشت مواد مغذی ارائه می‌دهند. همچنین، با بهره‌گیری از یادگیری ماشین، روش‌های بهینه چرخش محصولات را برای حفظ سلامت خاک و کاهش مصرف مواد شیمیایی تخمین می‌زنند.

توسعه‌ی پلاستیک‌های زیست‌تخریب‌پذیر[9]، با استفاده از مواد سازگار با محیط زیست نقش کلیدی در مبارزه با آلودگی پلاستیکی جهانی دارند. هوش مصنوعی در توسعه و ارزیابی بیوپلاستیک‌های جدید، بررسی قابلیت تجزیه، دوام و ویژگی‌های مختلف آن‌ها برای تعیین بهترین گزینه‌ها برای استفاده در صنعت، مفید است. این سامانه‌ها با مدل‌سازی و پیش‌بینی ویژگی‌های بیوپلاستیک‌ها، فرآیند جست‌وجوی پلیمرهای جدید و پایدار را سرعت می‌بخشند. همچنین ابزارهای هوش مصنوعی اثرات زیستگاه بیوپلاستیک‌ها را در طول چرخه زیستی (LCA)[10] ارزیابی می‌کند و راهنمایی‌هایی برای روش‌های تولید پایدار ارائه می‌دهند.

زیست‌فناوری با تبدیل ضایعات آلی به منابع ارزشمند مانند کمپوست و بیوسوخت، کمک می‌کند نیاز به محل‌های دفن زباله

1. Bioremediation
2. Geographic Information System
3. Biofilters
4. Industrial emissions
5. Biodiesel
6. Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR-associated protein 9 (CRISPR-Cas9)
7. Lowering agricultural runoff
8. Climate change
9. Biodegradable plastics
10. Lifecycle Analysis

و انتشار گازهای گلخانه‌ای کاهش یابد. هوش مصنوعی توانایی نظارت مداوم و تنظیم شرایط در فرآیندهای کمپوست‌سازی و هضم بی‌هوازی را دارد، تا بهترین عملکرد میکروارگانیسم‌ها برای تجزیه سریع‌تر زباله تضمین شود. سامانه‌های بیوراکتور هوشمند، با بهره‌گیری از سنسورها و الگوریتم‌های پیشرفته، شرایط محیطی را در زمان واقعی تطبیق می‌دهند، فعالیت میکروارگانیسم‌ها را افزایش داده و روند تجزیه زباله را تسریع می‌کنند. ماشین‌های با قابلیت‌های پیشرفته مبتنی بر هوش مصنوعی (ماشین‌های مرتب‌سازی زباله)[1] برای جداسازی مؤثر ضایعات آلی از سایر مواد به کار گرفته می‌شوند، این امر کارایی کلی فرآیندهای زیست‌تخریب را افزایش می‌دهد.

نوآوری در فناوری زیستی به‌طور مؤثر دی‌اکسیدکربن را از جو جذب و ذخیره می‌کند، بنابراین با کاهش گازهای گلخانه‌ای نقش مهمی در مقابله با تغییرات اقلیمی بر عهده دارد. هوش مصنوعی با تحلیل داده‌های زیستی و محیطی، بهترین روش‌ها را برای جذب زیستی کربن شناسایی می‌کند و به استفاده از مشارکت‌کننده‌های زیستی در مواجهه با تغییرات اقلیمی اشاره دارد. تحقیقات در زمینه اصلاحات ژنتیکی برای بهبود فرآیند فتوسنتز و افزایش جذب کربن، با بهره‌گیری از هوش مصنوعی، در حال انجام است. سامانه‌های هوشمند بهینه‌سازی‌شده‌ی جذب و ذخیره‌ی کربن (CCS)[2]، فرآیندهای جذب دی‌اکسیدکربن و ذخیره‌ی آن در زیرزمین یا در مواد را شبیه‌سازی و بهینه می‌کنند (۸).

کاربرد هوش مصنوعی در توسعه و بهینه‌سازی فرآیندهای زیستی

هم‌افزایی ماشین[3] و هوش مصنوعی، کاربردهای گسترده‌ای در خودکارسازی آزمایشگاه‌ها، مدیریت نمونه‌ها و انجام آزمایش‌های با توان عملیاتی بالا[4] دارد. بنابراین، این توسعه باعث ارتقای قابل‌توجه در کارایی، دقت و تکرارپذیری فرآیندهای زیست‌فناوری شده است. بسیاری از وظایف آزمایشگاهی که قبلاً نیازمند کارهای دستی و تکراری بودند، اکنون با دقت بالا توسط ماشین‌ها انجام می‌شوند. این وظایف شامل پیپت‌کردن دقیق، مدیریت هوشمند نمونه‌ها برای دستیابی به نتایج پایدار و قابل تکرار، و اجرای روش‌های خودکار مانند آماده‌سازی DNA است. عملکرد این ماشین‌ها در انجام این فعالیت‌ها، از نظر دقت و ثبات، نسبت به انسان‌ها برتری دارند. خودکارسازی فرآیندها، با ساده‌سازی مراحل و کاهش نیاز به نیروی کار دستی، ریسک خطای انسانی و مواجهه مستقیم با مواد خطرناک را کاهش می‌دهد.

یادگیری ماشین با استفاده از ابزارهای قدرتمند در تجزیه و تحلیل داده‌های پیچیده، نقش مهمی در پیشرفت زیست‌فناوری[5] ایفا می‌کنند. استفاده از الگوریتم‌های یادگیری ماشینی در تجزیه و تحلیل داده‌های حاصل از حسگرهای[6] تجهیزات و بیوراکتورها، امکان تشخیص زودهنگام اختلالات و برنامه‌ریزی برای نگهداری را فراهم می‌کند و با استفاده از مدل‌های ریاضی و آماری، بهینه‌سازی برای انجام فرآیندهای زیستی را با ترکیب داده‌های حسگرها، دوربین‌ها، پروب‌ها و توالی‌یاب‌ها تسهیل می‌نماید. این مدل‌ها توانایی پیش‌بینی تأثیر عوامل مهم، مانند دما، pH و غلظت مواد مغذی، بر عملکرد سامانه‌های زیستی را دارند. به‌کارگیری یادگیری ماشینی در بهبود فرآیندهای زیستی، کاهش نیاز به آزمایش‌های گران و زمان‌بر را ممکن می‌سازد و منجر به تولید زیستی کاراتر و اقتصادی‌تر می‌شود. این ابزارها، همچنین در نظارت و کنترل فرآیندهای زیستی در زمان واقعی[7] به کار می‌روند. تحلیل داده‌های حسگرها و سامانه‌های کنترل به بهبود شرایط داخلی بیوراکتورها کمک می‌کند و سبب افزایش راندمان و کیفیت محصول نهایی می‌شود (۳).

الگوریتم‌های یادگیری ماشین در بهینه‌سازی مسیرهای متابولیک و تولید ترکیبات مورد نظر در زیست‌شناسی مصنوعی، می‌توانند تأثیر تغییرات ژنتیکی بر مسیرهای متابولیسم در میکروارگانیسم‌ها را پیش‌بینی کرده و امکان طراحی سویه‌های تولیدکننده بهینه را فراهم سازند. این رویکرد برای افزایش تولید سوخت‌های زیستی[8]، داروها و سایر محصولات زیستی با ارزش بالا به کار گرفته شده است و مزایایی مانند توسعه سریع‌تر محصولات جدید، افزایش بهره‌وری، کاهش زمان چرخه تولید و ضایعات را به همراه دارد. علاوه بر این، تحلیل داده‌های حجیم آزمایش‌ها توسط هوش مصنوعی می‌تواند به کشف‌های نوین و راه‌حل‌های ارزشمند منجر شود و با رفع چالش‌ها و گسترش همکاری بین حوزه‌های مختلف، هوش مصنوعی، فرآیندهای تولید زیستی را کارآمدتر، مقاوم‌تر و اقتصادی‌تر کند و در نهایت تولید داروهای حیاتی، درمان‌ها و دیگر محصولات زیستی را افزایش دهد (۹).

قابلیت هوش مصنوعی در بررسی فعالیت ژنگان در شرایط زیستی متفاوت

پروژه‌های تحقیقاتی در ژنومیکس بر شناسایی لوکوس‌های مرتبط با یک ژنوتیپ خاص از طریق هوش مصنوعی تمرکز دارند. محققان در حال بررسی روش‌هایی هستند که قابلیت اطمینان، تفسیرپذیری و تبیین‌پذیری الگوریتم‌ها را بهبود بخشند. در تفسیرپذیری، اطمینان حاصل می‌شود که دقت مدل بر اساس نمای صحیح مسئله است و ناهنجاری‌های داده‌های آموزشی تأثیر منفی ندارند. در واقع، ممکن است سامانه هوش مصنوعی دقت آزمایش را نه بر اساس «درک واقعی»، بلکه با شناسایی روابط

1. Waste Sorting Robots
2. AI-optimized Carbon Capture and Storage
3. Robotics
4. High-Throughput
5. Biotechnology
6. Sensor
7. Real-time
8. Biofuels

پنهان و ناآشنا بین داده‌های آموزشی و آزمون ارزیابی کند. عدم وجود چنین روابطی ممکن است منجر به کاهش دقت و کاربرد نادرست فناوری‌های مبتنی بر داده شود.

تبیین‌پذیری اغلب با مفاهیمی مانند شفافیت و تفسیر همراه است، یعنی حق هر فرد برای آگاهی از دلایل تصمیمات الگوریتم که بر زندگی او تأثیر می‌گذارند. این اصل برای بسیاری از قوانین اهمیت دارد، بطورمثال، در ایالات متحده، در بخش‌هایی مانند امور مالی به رسمیت شناخته شده است. در حالیکه در اتحادیه اروپا، با حمایت از مقررات حفاظت اطلاعاتی عمومی (GDPR)[1]، به تمامی حوزه‌ها تعمیم یافته است. علاوه بر اعطای حقوق شخصی، تبیین‌پذیری می‌تواند از روش‌های مورد استفاده توسط هوش مصنوعی در اتخاذ تصمیمات خاص و استخراج دانش جدید از آن‌ها برای انجام تحلیل‌های پس از وقوع[2] مفید باشد. شفافیت در این حوزه همان توضیحات مدل است. این ویژگی در مورد طبقه‌بندی‌کننده‌های ساده بر اساس قواعد و مدل‌های خطی که تفسیرپذیر هستند، صدق می‌کند. در این حالت، توضیح دلایل یک تصمیم خاص و بررسی روند تصمیم‌گیری آسان است. مزیت این مدل‌ها، تفسیرپذیری و توضیح‌پذیری ذاتی آن‌ها است. اما این مدل‌های ساده قابلیت یادگیری روابط غیرخطی را ندارند و در برخی کاربردهای پیچیده ممکن است مناسب نباشند. اکثر سامانه‌های هوش مصنوعی بیش از حد پیچیده هستند که بتوان به مکانیسم آنها بطور مستقیم آگاهی یافت. در این حالت، باید به عنوان جعبه‌سیاه در نظر گرفته شوند و توضیحات آن‌ها از طریق تحلیل‌های پس از وقوع به‌دست آید. این کار می‌تواند به دو روش اصلی انجام شود: استفاده از روش‌های مستقل از مدل (یعنی روش‌هایی که صرفاً بر ورودی و خروجی تکیه دارند) یا روش‌های وابسته به مدل. مزیت روش اول، قابلیت استفاده همیشگی آن است، در حالیکه روش دوم می‌تواند از ویژگی‌های خاص سامانه هوش مصنوعی بهره‌مند شود (۴).

کاربرد هوش مصنوعی در بررسی توالی‌یابی DNA

فناوری‌های هوش مصنوعی با سرعت بخشیدن به توالی‌یابی و تفسیر اطلاعاتی ژنگان در مقیاس بزرگ، فرآیند تجزیه و تحلیل آن را را متحول کردند. فرآیند سنتی توالی‌یابی DNA، اگرچه بسیار مؤثر است، اما برای ژنگان بزرگ و پیچیده در موجودات بالاتر زمان‌بر و پرهزینه است. هوش مصنوعی توانایی غلبه بر این محدودیت‌ها را دارد، توالی‌یابی سریع‌تر و دقیق‌تری را فراهم می‌آورد. ابزارهای مبتنی بر هوش مصنوعی، مانند واریانت دیپ گوگل[3]، با خودکارکردن فرآیند شناسایی تغییرات ژنتیکی از اطلاعات توالی‌یابی، تجزیه و تحلیل ژنگان را تسهیل کردند. این ابزار توسعه‌یافته توسط هوش مصنوعی گوگل، از تکنیک‌های یادگیری عمیق برای طبقه‌بندی توالی‌های ژنگان و شناسایی پلی‌مورفیسم‌های تک‌نوکلئوتیدی (SNPs)[4] و درج‌ها/حذف‌ها (indels)[5] با دقت بیشتری نسبت به روش‌های سنتی استفاده می‌کند. توانایی یادگیری عمیق در پردازش داده‌های ژنگان در مقیاس بزرگ، ضمن افزایش سرعت و دقت، راه‌حل‌های جدیدی در مورد علل ژنتیکی بیماری‌ها ارائه می‌دهد (۱۰).

حوزه دیگری که هوش مصنوعی در آن تأثیر قابل توجهی دارد، ویرایش ژن مبتنی بر کریسپر است. مدل‌های هوش مصنوعی در حال توسعه‌اند تا دقت و کارایی سیستم‌های کریسپر را بهبود بخشند و به این ترتیب، ویرایش ژن را متحول کنند. با استفاده از هوش مصنوعی، دانشمندان می‌توانند RNAهای راهنمای مؤثر برای هدف‌گیری ژن‌های خاص پیش‌بینی کنند و تاثیرات خارج از هدف را به حداقل برسانند. به عنوان مثال، مدل‌های کریسپر مبتنی بر هوش مصنوعی، مانند کریسپر دیپ[6]، از یادگیری عمیق برای پیش‌بینی فعالیت هدف و خارج از هدف RNAهای راهنما[7] استفاده می‌کنند، بنابراین ایمنی و اثربخشی آزمایش‌های ویرایش ژن را به طور قابل توجهی افزایش می‌دهند (۳).

کاربرد هوش مصنوعی در طراحی و بررسی فعالیت پروتئین‌ها

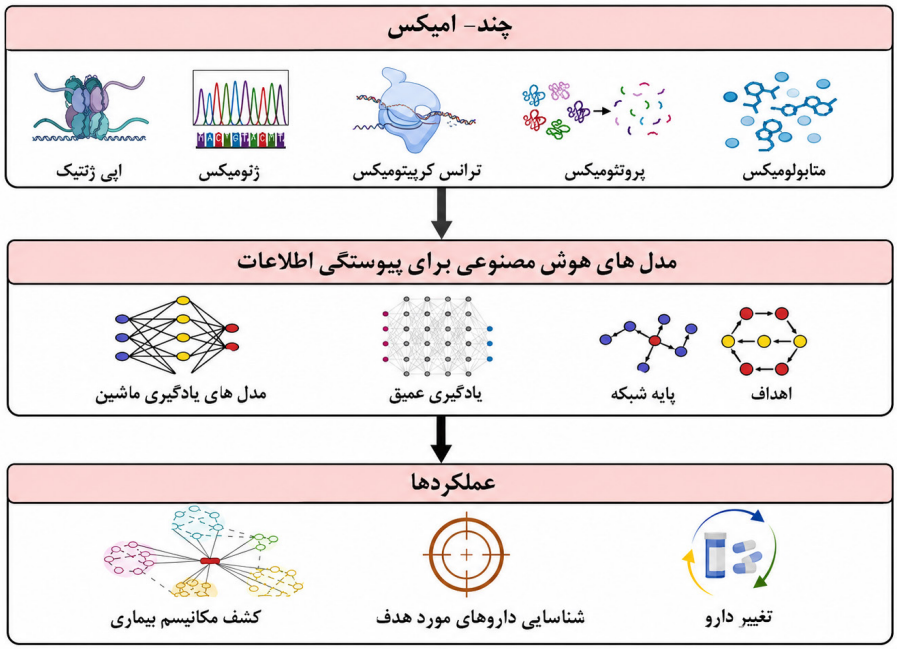
رویکرد طراحی«پروتئین از نو»[8] به جای اصلاح یا بازمهندسی پروتئین‌های طبیعی، بر ساخت پروتئین‌هایی کاملاً جدید و قابل برنامه‌ریزی مبتنی بر اصول مهندسی تمرکز دارد. اگرچه استفاده از روش‌های بهینه‌سازی پروتئین‌های طبیعی موفق بوده‌اند، طراحی از نو مزایای منحصربه‌فردی ازجمله؛ امکان ایجاد عملکردهایی از پروتئین که در طبیعت وجود ندارند و همچنین طراحی سامانه‌هایی با ویژگی‌های قابل تنظیم، کنترل‌پذیری ماشین‌های مولکولی دارد(۱۱)

در روش‌های سنتی، طراحی پروتئین‌ها با چالش‌های فنی متعددی همراه است. ابتدا ساختار سه‌بعدی پروتئین مشخص می‌شود، سپس توالی مناسب آن طراحی می‌گردد. برای رسیدن به عملکرد مطلوب، دقت در نحوه شکل‌گیری جایگاه فعال اهمیت زیادی دارد؛ حتی انحراف کم‌تر از یک آنگستروم در جایگاه اتم‌ها، در عملکرد پروتئین اختلال ایجاد می‌کند. علاوه براین، مطالعه دینامیک ساختاری و زمانی پروتئین‌ها نیز از عوامل پیچیده و ضروری هستند. بررسی ساختارهای پروتئین نیز وابسته به تکنیک‌های تجربی مانند کریستالوگرافی اشعه X[9]، طیف‌سنجی NMR[10] و کرایو میکروسکوپ الکترونی[11] است که زمان‌بر و پرهزینه هستند (۱۲).

1. General Data Protection regulation
2. Ex-post
3. Google Deep Variant
4. Single nucleotide polymorphisms
5. Small nucleotide insertions or deletions
6. Deep CRISPR
7. Guide RNAs
8. De novo
9. X-ray crystallography
10. Nuclear Magnetic Resonance
11. Cryo-Electron Microscopy

یادگیری ماشین با استفاده از نرم‌افزار آلفافولد[1]، از تکنیک‌های یادگیری عمیق برای پیش‌بینی ساختار پروتئین با دقت بالا بهره می‌گیرد. این الگوریتم بر اساس توالی‌های اسیدآمینه، ساختارهای پروتئین را مدل‌سازی می‌کند. با آموزش بر روی داده‌های ساختاری شناخته شده، ساختارهای جدید را با دقت قابل توجهی پیش‌بینی کند. این پیش‌بینی‌های ساختاری دقیق، به محققان کمک می‌کند تا مسیرهای نوینی در زیست‌شناسی مصنوعی، مهندسی آنزیم و طراحی پروتئین کشف کنند (۱۳).

امروزه، رویکردهای مبتنی بر یادگیری عمیق در پیش‌بینی ساختارهای پروتئین، امکان طراحی هم‌زمان ساختار، توالی، و عملکرد را فراهم کرده‌اند و پیشنهاداتی برای تولید ساختارهای متنوع از یک الگو یا طراحی پروتئین حول یک جایگاه خاص را مطرح می‌کنند (شکل ۴). اگرچه طراحی کامل و خودکار پروتئین‌های عملکردی با هوش مصنوعی همچنان در حال توسعه است، اما چشم‌اندازی روشن برای دستیابی به آن در آینده نزدیک وجود دارد (۱۲).



شکل ۴. استفاده از چند-اُمیکس در مدل‌های مختلف هوش مصنوعی برای کاربردهای گوناگون (۳).

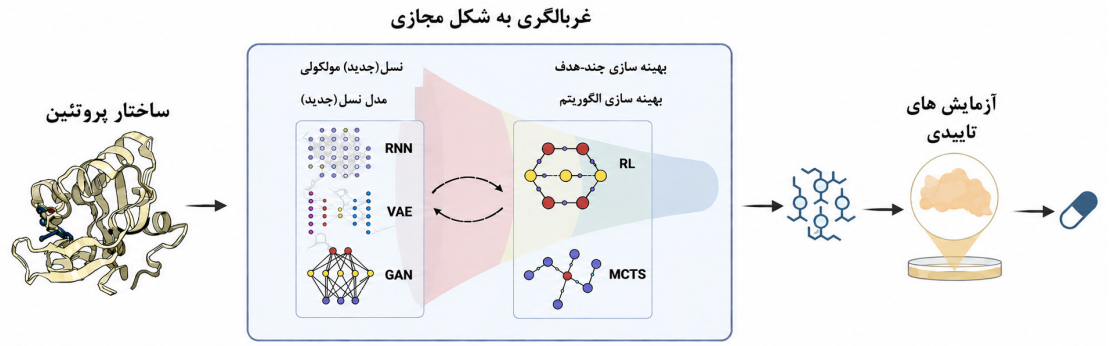
کاربرد هوش مصنوعی در کشف و توسعه داروها

در زیست‌شناسی ساختاری و توسعه دارو، یادگیری ماشین در پیش‌بینی ساختار پروتئین نقش کلیدی دارند. پروتئین‌ها که در بسیاری از فرآیندهای حیاتی نقش دارند، باید به شکل‌های سه‌بعدی مشخص تا شوند، تا عملکرد صحیح خود را نشان دهند. بررسی ساختار یک پروتئین، دیدگاه‌هایی در مورد عملکرد آن ارائه می‌دهد و برای طراحی داروهایی که با آن تعامل دارند، ضروری است و نیازمند آزمایش‌های گسترده برای اثبات اثربخشی و ایمنی است.

فرایند سنتی توسعه دارو با مشکلاتی مانند هزینه و زمان بالا، و محدودیت‌های استفاده از فناوری‌های پیشرفته مواجه است، هوش مصنوعی با مدیریت فرایند تولید محصول، چالش‌ها را برطرف و با شناسایی ترکیبات هدفمند، روند تأیید را تسریع و طراحی مولکولی دارو را بهبود می‌بخشد. نرم‌افزار آلفافولد برای پیش‌بینی ساختار پروتئین و «هوش مصنوعی در شیمی»[2] برای پیش‌بینی ویژگی‌های مولکولی و طراحی دارو، مورد استفاده هستند.

الگوریتم‌های یادگیری ماشین با تحلیل داده‌های موجود؛ مانند غربال‌گری مجازی برای کشف خواص و ارزیابی سمیت داروها با استفاده از ابزارهایی مانند ایکس جی باکس[3] و جنگل تصادفی[4] در طبقه‌بندی و رگرسیون، الگوهای قابل اعتماد را شناسایی و پیش‌بینی‌های دقیقی ارائه می‌دهند. شبکه‌های عصبی کانولوشنی[5]، شبکه‌های عصبی بازگشتی[6] و شبکه‌های مولد عمیق[7] در حوزه یادگیری عمیق با بهره‌گیری از داده‌های حجیم و قدرت محاسباتی بالا، در مدل‌سازی ساختار پروتئین، تحلیل داده‌های ژنومی، شناسایی اهداف دارویی (شکل۵) و طراحی مولکول‌های جدید پیشرفت‌های وسیع داشته است (۸،۱۳).

1. AlphaFold
2. DeepChem
3. XGBoost
4. Random Forest
5. CNN
6. RNN
7. GDNN



شکل ۵. فرآیند مبتنی بر هوش مصنوعی برای کشف داروهای کوچک مولکول. فرآیند کشف داروهای کوچک مولکول مبتنی بر هوش مصنوعی از اطلاعات ساختاری پروتئین شروع می‌شود و از مدل‌های یادگیری عمیق مانند شبکه‌های عصبی بازگشتی (RNN)، آواتورکدِر متناظر متفاوت (VAE) و شبکه‌های عصبی تولیدکننده متقابل (GAN) برای تولید مولکول‌ها استفاده می‌کند. این مدل‌ها با استراتژی‌های بهینه‌سازی مانند یادگیری تقویتی (RL) و جستجوی درخت مونت کارلو (MCTS) ترکیب می‌شوند تا امکان کاوش چند هدف شیمیایی فراهم شود. سپس ترکیبات نامزد حاصل از غربالگری مجازی، مورد تأیید آزمایشگاهی قرار می‌گیرند. شبکه عصبی تولیدکننده متقابل GAN:؛ جستجوی درخت مونت کارلو MCTS:؛ یادگیری تقویتی RL:؛ شبکه عصبی بازگشتی RNN:؛ آواتورکدِر متناظر متفاوت VAE (۱۳).

هوش مصنوعی و پزشکی فرد-مدار

در مطالعه ژنومیکس، هوش مصنوعی نقش مهمی در پیشرفت پزشکی فردی ایفا می‌کند و توانایی زیادی در ارتقای فرآیندهای تشخیص و درمان دارد. جایی که درمان‌ها براساس ساختار ژنتیکی هر بیمار تنظیم می‌شوند. هوش مصنوعی امکان تجزیه و تحلیل سریع ژنگان بیماران و شناسایی جهش‌های مرتبط با بیماری و پیش‌بینی پاسخ به داروهای خاص را فراهم می‌کند. با تحلیل داده‌های ژنتیکی، بالینی و سبک زندگی بیماران، این فناوری امکان توسعه راهبردهای درمانی شخصی‌سازی‌شده را فراهم می‌آورد. استفاده از الگوریتم‌های یادگیری ماشین در این حوزه کمک می‌کند تا پاسخ بیماران به درمان‌ها پیش‌بینی شده و عوارض جانبی احتمالی شناسایی شوند. در چند سال اخیر، روش‌های هوش مصنوعی، به‌ویژه شبکه‌های عصبی، در پزشکی فردی گسترش یافته است. این مدل‌ها با تحلیل داده‌های پیچیده، مثل داده‌های طیفی و شاخص‌های تشخیصی، توانایی تشخیص بیماری‌هایی مانند ملانومای بدخیم، اختلالات چشمی و سرطان‌های مختلف را نشان دادند. ترکیب داده‌های چندمنبعی، که شامل اطلاعات ژنتیکی، سوابق پزشکی و عوامل سبک زندگی است، امکان ارائه درمان‌های هدفمند و دقیق‌تر را فراهم می‌کند. این رویکرد منجر به بهبود نتایج درمانی، افزایش اثربخشی مداخلات و کاهش عوارض ناخواسته می‌شود و در نتیجه کیفیت خدمات پزشکی را ارتقا می‌دهد (۱۴).

مدیریت اطلاعات در هوش مصنوعی برای بررسی فعالیت ژنگان در شرایط مختلف زیستی

بیشتر الگوریتم‌های یادگیری ماشین و یادگیری عمیق ورودی خود را در قالب یک ماتریس می‌گیرند؛ در این ماتریس، هر ستون نشان‌دهنده‌ی یک نمونه و هر سطر، یک ویژگی مربوط به آن نمونه است. ماهیت این ماتریس‌ها نیز به کاربرد مورد نظر و زمینه‌ی خاص آن بستگی دارد. در بیوانفورماتیک، برای بررسی فعالیت ژنگان در شرایط زیستی مختلف[1]، این نوع نمایش بسیار کارآمد است، چون بیشتر داده‌ها به این شکل ارائه می‌شوند. به عنوان نمونه، اطلاعات توالی RNA معمولاً در قالب ماتریس‌هایی ساخته می‌شود که در آن‌ها، میزان ژن یا رونویسی (در ردیف‌ها) در مجموعه‌ای از نمونه‌ها با شرایط مختلف (در ستون‌ها) اندازه‌گیری شده است.

ارتباط بین ماتریس‌ها و حوزه‌های یادگیری ماشین و بیوانفورماتیک، براساس زبان‌های برنامه‌نویسی مانند آر[2] و پایتون است که از ماتریس‌های اطلاعات به عنوان ساختار اصلی استفاده می‌کنند. ولی در بسیاری از موارد، این ماتریس‌ها ممکن است ناقص و/ یا دارای خطا باشند. روش‌های مختلف، شرایط تجربی متفاوت و دستگاه‌های مختلف می‌توانند باعث ایجاد سوگیری‌ها و دستکاری‌های مصنوعی[3] شوند، علاوه بر این، ممکن است در برخی نقاط، اطلاعات برای برخی نمونه‌ها از بین رفته باشد. باتوجه به این، اجرای یک رویکرد جامع برای درک عملکرد هر عنصر ژنگان نیازمند در نظر گرفتن داده‌هایی با انواع مختلف است. در این شرایط، جایگذاری داده‌های گمشده، حذف اطلاعات نامرتبط[4] و یکپارچه‌سازی اطلاعات باید بعنوان جزییات ضروری و لاینفک طراحی الگوریتم‌های یادگیری ماشین و هوش مصنوعی در حوزه فعالیت ژنگان در شرایط زیستی مختلف محسوب شوند (۴).

جایگذاری اطلاعات

مواجهه با اطلاعات ناقص غیرعادی نیست. دلایل این نوع مقادیر گمشده متعدد هستند، از جمله: در دسترس نبودن اطلاعات،

1. Functional genomics
2. R
2. Artefacts
3. Data denoising

نقص در اندازه‌گیری، و ناسازگاری طرح‌واره[1] پایگاه‌های اطلاعات مختلف. با توجه به روش‌های مختلف از دست رفتن داده، در برخی از مطالعات سه نوع دسته‌بندی پیشنهاد گردید: کاملاً تصادفی گمشده[2] (MCAR)، تصادفی گمشده[3] (MAR)، و غیرتصادفی گمشده[4] (NMAR) اول، گمشده کاملاً تصادفی، وضعیتی را شرح می‌دهد که در آن احتمال گم شدن برای تمامی اندازه‌گیری‌ها برابر است؛ این حالت ممکن است، در داده‌های ریزآرایه[5] رخ دهد، موقعیتی که احتمال از دست دادن اطلاعات در هر نقطه آن وجود دارد. اگر احتمال گم شدن داده‌ها فقط درون یک گروه به شکل یکنواخت توزیع شده باشد، این داده‌ها به عنوان اطلاعات تصادفی گمشده محسوب می‌شوند. دسته سوم شامل مواردی است که در آن‌ها اصطلاح «کاملاً تصادفی گمشده» و «تصادفی گمشده» صحیح نیستند. قبل از به کارگیری الگوریتم‌های یادگیری ماشین یا هوش مصنوعی، باید مشخص شود کدام رویکرد برای استفاده از داده‌های گمشده مناسب‌تر است. از نظر برخی از پژوهشگران همراستا با این مطالعات دو رویکرد اصلی می‌تواند مدنظر قرار گیرد: ۱) حذف متغیرها یا نمونه‌های دارای مقادیر گمشده و ۲) جایگزینی داده‌های گمشده. حذف داده‌ها به دلیل احتمال سوگیری توصیه نمی‌شود. روش جایگزینی شامل پرکردن شکاف‌های داده با بهترین مقدار پیش‌بینی‌شده برای هر داده گمشده است، اغلب هدف این است که مقداری پیدا شود که بهترین برآورد برای مقدار گمشده باشد. بررسی این مفهوم به صورت تجربی در برخی مطالعات مشخص کرد که در خوشه‌بندی و طبقه‌بندی، روش‌های پیچیده‌تر جایگزینی نسبت به روش‌های ساده‌تر مانند میانگین برتری چندانی ندارند. این سری از مطالعات بر اهمیت حفظ معنا و مفهوم داده‌ها در فرآیند جایگزینی تأکید می‌کنند، به این صورت که جایگزینی نباید به صورت «جعبه سیاه» انجام شود، بلکه بر اساس هدف خاص باید صورت گیرد. نتایج آزمایش‌ها نشان می‌دهد که در مجموع، روش‌های چندمتغیره در پیش‌بینی مقادیر گمشده، دقت بالاتری دارند.

مسئله‌ای مرتبط، در یکپارچه‌سازی اطلاعات چند–اومیکس[6] وجود ستون‌های گمشده است. در این حالت، چندین جدول حاوی اطلاعات موجود است که هرکدام ویژگی‌های مربوط به گروهی از نمونه‌ها را شرح می‌دهند. برخی الگوریتم‌ها، مانند ترکیب موارد مشابه برای یکپارچه‌سازی این داده‌ها، به تطابق دقیق با گروه‌های نمونه نیاز دارند؛ بنابراین، حتی نبود یک ستون در یک جدول می‌تواند منجر به عدم اجرای این روش‌ها شود. افزودن یک ستون کامل بسیار دشوارتر از وارد کردن مقادیر جداگانه است و ممکن است روش‌های سنتی استفاده از آمار برای این کار مناسب نباشند. برای رفع این مشکل، روش‌های تخصصی پیشنهاد شده‌اند. پژوهشگران روشی را ارائه کردند که در آن‌ها با «قرض گرفتن»[7] ستون‌های گمشده از نمونه‌های دیگر، مجموعه‌ای از جایگزین‌های ممکن را تولید می‌کنند، سپس تحلیل را به طور مستقل انجام می‌دهند و جایگزین‌هایی را که بیشترین تطابق را با نتایج اجماع دارند، انتخاب می‌کنند (۱۵).

حذف اطلاعات نامرتبط

اطلاعات اُمیکس به کاهش فراوانی ویژگی‌های خاص در یک سلول یا گروه سلولی در اثر عوامل مختلفی مانند تغییرپذیری زیستی اشاره دارد. همچنین، تعداد صفات اندازه‌گیری‌شده اغلب بیش از حجم نمونه‌ها است. فناوری‌های پیشرفته و دسترسی به تکرارها می‌توانند در برآورد و اصلاح خطاهای اندازه‌گیری مؤثر باشند، اما در مواجهه با تغییرپذیری زیستی که نقش مهم‌تری دارد، تأثیر کمتری دارند (۹). پیش‌شرط اصلی برای اجرای برنامه‌های یادگیری ماشین و هوش مصنوعی، داده‌های صحیح[8] است (چیزی که نرم‌افزار برای آموزش نیاز دارد). حذف اطلاعات غیرمربوط شامل دو مرحله است: اصلاح داده‌ها (نرمال‌سازی) و حذف اندازه‌گیری‌های نادرست یا بی‌ربط (براساس ویژگی). دو روش کلاسیک برای حذف اطلاعات نادرست وجود دارد که برپایه دانش تجربی قرار دارند و بسته به اینکه از کدام مدل استفاده شود، متفاوت هستند. روش اول، بطورمثال برای نرمال‌سازی پروفایل بیان توالی RNA است، که در آن داده‌ها به صورت توزیع دوجمله‌ای منفی مدل‌سازی می‌شوند. برعکس، روش‌های بدون مدل که برای حذف اطلاعات بی‌ربط، بر تنظیم آستانه‌ها تمرکز دارند. این روش‌ها به جای اصلاح داده‌ها، سیگنال‌های ضعیف و ویژگی‌های نامعتبر را فیلتر می‌کنند. هر دو روش دارای مزایا و معایب خاص خود هستند. حذف اطلاعات نامرتبط مبتنی بر مدل می‌تواند دست‌سازها را تصحیح کند و در نتیجه امکان ترکیب مجموعه‌های ناهمگن اطلاعات را فراهم سازد. این امر، به عنوان مثال، در هنگام نرمال‌سازی اطلاعات بیان ژن به‌دست‌آمده با فناوری ریزآرایه رخ می‌دهد. چالش اصلی، وابستگی کاربرد و اثربخشی روش به میزان انطباق مدل با مجموعه داده‌های مورد نظر است. از سوی دیگر، روش‌های بدون مدل همواره قابل اجرا هستند. ولی آستانه‌ها به مقیاس اطلاعات وابسته‌اند و انتخاب نامناسب آن‌ها می‌تواند منجر به حذف ویژگی‌های مرتبط شود که به طور تصادفی از رسیدن به مقدار حد آستانه باز می‌مانند. برای رفع این مشکل، فیلترهایی با چند مقیاس[9] پیشنهاد می‌گردد.

رویکرد نوین در حذف داده‌های نامربوط، استفاده از یادگیری عمیق است. در رایج‌ترین روش پیشنهاد می‌گردد؛ ویژگی‌ها به‌عنوان

1. Schema
2. Missing Completely At Random
3. Missing At Random
4. Not Missing At Random
5. Microarray
6. Multi-omics
7. Borrowing
8. Clean Data
9. Multi-scale filters

بخش‌هایی از مدل یادگیری ماشین محاسبه شوند. کدگذارهای خودکار، به‌خصوص آن‌هایی که بر پایه شبکه‌های عمیق ساخته شده‌اند، اطلاعات غیرمرتبط را حذف می‌کنند. تفاوت این شبکه‌ها با کدگذارهای خودکار استاندارد در این است که برای بازسازی ورودی از نمونه‌های ناقص آن آموزش می‌بینند. این کدگذارهای خودکار –چه نوع استاندارد و چه حذف‌کننده‌های اطلاعات غیرمرتبط– قابلیت ترکیب با فناوری‌های دیگر را دارند. به عنوان مثال، کدگذارهای خودکار حذف‌کننده اطلاعات نامرتبط می‌توانند به صورت «انباشته‌شده»[1] استفاده شوند تا شبکه‌های پیچیده‌تری ایجاد شود، یا با مدل‌های بیزی ترکیب گردند تا اطلاعات مربوط به توالی‌یابی تک‌سلول از داده‌های نامرتبط پاک شوند (۴،۱۶).

یکپارچه‌سازی اطلاعات

هدف اصلی مراحل یکپارچه‌سازی داده، سازماندهی اطلاعات به گونه‌ای است که به راحتی قابل استفاده، بهره‌برداری و در صورت امکان، توزیع مجدد باشند. همان‌طور که در شماره‌های دسترسی پایگاه اطلاعات اسید نوکلئیک[2] به وضوح قابل مشاهده است. اطلاعات تنها زمانی مفید هستند که بتوان آن‌ها را به طور موثر جستجو کرد تا دانش جدید کسب شود. بنابراین، اطلاعات باید یا به صورت عمومی در دسترس باشند یا توسط مالکین‌شان به روشی کامل مورد بررسی قرار گیرند.

مرحله اول در تجزیه و تحلیل داده‌ها جمع‌آوری اطلاعات از منابع مختلف و ادغام آن‌ها با داده‌های تولیدشده در محل است. هرچند کاوش و یکپارچه‌سازی داده‌های عمومی می‌تواند به کشف نتایج جدید و مهم کمک کند، افزودن داده‌های جدید بر اساس پیشینه و تجربه پژوهشگر، ارزش مجموعه داده‌ها را افزایش می‌دهد. فرآیند یکپارچه‌سازی، شامل بازیابی، تمیزکاری و سازماندهی داده‌ها از منابع متفاوت است. در بسیاری موارد، بخش قابل توجهی از داده‌های مرتبط را می‌توان از پایگاه‌های داده موجود یا محلی دریافت نمود. این پایگاه‌ها حاوی اطلاعاتی مانند توالی‌های DNA ،RNA و پروتئین، بیان ژن، داده‌های فنوتیپ، ساختارهای پروتئین، تعاملات پروتئین–پروتئین و ترجمه پروتئین هستند. علاوه بر آن، داده‌های مربوط به مسیرهای پیام‌رسانی و متابولیک، تصویربرداری سلولی و منابع دیگر نیز می‌تواند به مجموعه داده‌ها افزوده شود. تمامی این داده‌ها باید به صورت منسجم و یکپارچه سازمان‌بندی شوند تا برای اهداف خاص مورد استفاده قرار گیرند.

روش‌های یکپارچه‌سازی اطلاعات را می‌توان به دو دسته مشتاق[3] و تنبل[4] طبقه‌بندی کرد. در روش مشتاق، اطلاعات معمولاً در یک مرکز ذخیره‌سازی متمرکز کپی می‌شوند. در روش تنبل، اطلاعات در منابع اصلی خود باقی می‌مانند و در صورت نیاز و بنا به درخواست[5] یکپارچه می‌شوند. هر دو روش برای دسترسی، به رابط‌های ماشین برای قابلیت خواندن تکیه دارند، چون دریافت[6] و سازماندهی حجم زیادی از داده‌ها به صورت دستی مشکل و اغلب ناممکن است. بسیاری از پایگاه‌های داده، که به عنوان سرویس‌های وب شناخته می‌شوند، این اطلاعات را ارائه می‌دهند و به کاربران امکان می‌دهند، درخواست‌های خود برای دریافت اطلاعات بفرستند. سپس، این اطلاعات می‌توانند با رویکردهای مشتاق یا تنبل در فرآیند ادغام قرار گیرند. سرویس‌های وب می‌توانند از روش‌های مختلفی از جمله AJAX[7]، REST[8] یا SOAP[9] استفاده کنند. اطلاعات مبادله‌شده توسط ماشین به یکی از زبان‌های استاندارد مانند XML[10] یا JSON[11] در قالب قابل خواندن منتقل می‌شوند.

مشکل دیگر این است که پروتئین‌ها در موجودات زیستی مختلف توسط پژوهشگران با نام‌های متفاوتی مورد اشاره قرار می‌گیرند که این امر منجر به عدم استانداردسازی و دشواری در یکپارچه‌سازی اطلاعات می‌شود. در مواردی که رابط‌های قابل خواندن توسط ماشین در پایگاه‌های اطلاعات، قابلیت دسترس عمومی ندارند، پژوهشگران از نویسندگان پایگاه، درخواست رونوشت اطلاعات[12] می‌کنند یا می‌توانند از خزنده‌های وب[13] اطلاعات را از وب‌سایت‌ها در قالب HTML[14] استخراج کنند. این داده‌ها سپس برای استخراج اطلاعات مرتبط پردازش می‌شوند. اما استفاده از خزنده‌ها و رونوشت‌ها در پایگاه داده ممکن است مسائل حقوقی مربوط به مالکیت فکری به وجود آورد، که معمولاً با ارائه داده‌ها از طریق سرویس‌های وب در قالب شرایط از پیش تعریف‌شده برطرف می‌شود (۹).

برای بهتر فهمیدن فرآیند یکپارچه‌سازی، می‌توان خطوط مختلف داده‌های اومیکس مانند ژنومیکس، پروتئومیکس، ترانسکریپتومیکس و دیگر موارد را همچون عناصر یک ارکستر سمفونیک تصور کرد. فنوتیپ سلول‌های مشاهده‌شده نتیجه نواختن همزمان این عناصر است. گوش دادن به تنها یک عنصر (تحلیل یک ردیف اومیکس) تصویری از ملودی ارائه می‌دهد، اما جزئیات ناپیدا هستند. منطق یکپارچه‌سازی داده‌ها در فعالیت ژنگان در حالات زیستی مختلف بر تحلیل جامع تمامی داده‌ها و تعاملات آن‌ها تأکید دارد (شکل۶)، زیرا می‌تواند دید کامل‌تری از وضعیت سلول نشان دهد. رابطه

1. Stacked
2. Nucleic Acids Research
3. Eager
4. Lazy
5. On-demand
6. Download
7. Asynchronous JavaScript And XML
8. RE presentational State Transfer
9. Simple Object Access Protocol
10. Extensible Markup Language
11. JavaScript Object Notation
12. Dump
13. Web crawlers
14. Hyper Text Markup Language

بین متیلاسیون DNA و بیان ژن نمونه‌ای از آن است (۴). سه نوع کلی یکپارچه‌سازی وجود دارد: اول، یکپارچه‌سازی افقی که شامل مجموعه‌های جداگانه‌ای از یک نوع اومیکس است؛ دوم، یکپارچه‌سازی عمودی که داده‌ها را در یک نمونه دسته‌بندی می‌کند؛ و سوم، یکپارچه‌سازی موازی که این دو نوع را ترکیب می‌کند. نوع اول برای ترکیب داده‌های منابع مختلف مفید است و هدف آن کاهش ناسازگاری بین تعداد نمونه‌ها و ویژگی‌های ژنگان است. نوع دوم بیشتر برای شناسایی ویژگی‌هایی است که فنوتیپ مشاهده‌شده را شکل می‌دهند. در نهایت، نوع سوم زمانی مناسب است که داده‌ها ناهمگن باشند.

از نظر ریاضی، ورودی‌های الگوریتم‌های یکپارچه‌سازی مجموعه‌ای از ماتریس‌ها هستند (هرکدام مربوط به نوع خاصی از داده‌های اومیکس)، که هر ردیف نشانگر یک ژن و هر ستون نشانگر نمونه است. هدف این است که یک ماتریس جدید ساخته شود که روابط چندسطحی در آن نشان داده شده باشد. در حالت اول، خروجی یک ماتریس مربعی است که می‌تواند به عنوان مجاورت یک ماتریس تفسیر شود؛ هر ردیف و ستون مربوط به ژن است و مقادیر میزان ارتباط هر رابطه را نشان می‌دهد. مزیت این روش، امکان تحلیل مسیر و الگوریتم‌های شناسایی برای تحلیل‌های بعدی است. رویکرد دوم، یک ماتریس با ردیف‌هایی که ویژگی‌های ژنگان را نشان می‌دهد، ایجاد می‌کند. این نوع الگوریتم‌ها امکان تحلیل تفاضلی یا خوشه‌بندی را برای تحلیل‌های زیرین فراهم می‌کنند، اما در فضای نمونه یا ویژگی محدودیت دارند. برای نمونه، روشی که بر اساس ترکیب خطی پروفایل‌های مرتبط، امتیازی به هر ژن می‌دهد، نیازمند این است که ماتریس‌های ورودی در یک فضای ویژگی مشترک قرار داشته باشند (۴،۱۷).

شکل ۶. جنبه‌های مختلف فعالیت ژنگان در شرایط متفاوت زیستی شامل رشته‌های «اومیکس» مشارکت‌کننده، ویژگی‌های زیستی هدف، و فناوری‌های اصلی با توان بالا برای تولید اطلاعات (۳).

جستجوی اطلاعات

هدف از جستجوی اطلاعات، کشف دانش جدید و ناشناخته است که می‌تواند در توسعه فرآیندها و محصولات نوآورانه کاربرد داشته باشد. کاوش داده‌ها ممکن است به روش سنتی، بر پایه شهود و جستجوی دستی، انجام شود، اما با توجه به رشد سریع حجم داده‌ها، این رویکرد در حال زوال است. به همین دلیل، پژوهشگران باید از مجموعه‌ای گسترده از تکنیک‌های آماری و هوش مصنوعی مرتبط با یادگیری ماشین بهره‌مند شوند. اطلاعات را می‌توان به روش‌های متنوعی از طریق یادگیری ماشین جستجو کرد، اما به طورکلی، دو روش مستقل و مکمل وجود دارند: نظارت‌شده[1] و بدون نظارت[2].

1. Supervised
2. Unsupervised

در یادگیری ماشین نظارت شده، هدف استنباط قوانین کلی از مجموعه‌ای برچسب‌خورده از نمونه‌ها است. در ساده‌ترین شکل، الگوریتم‌های یادگیری ماشین با جداولی از نمونه‌ها ارائه می‌شوند که هرکدام نشان‌گذاری شده‌اند. بنابراین الگوریتم‌ها معادلاتی را نشان می دهند که امکان استخراج برچسب‌ها از داده‌ها را فراهم می‌سازد. به عنوان مثال، نویسندگان از کدگذارهای خودکار برای ترکیب اطلاعات اومیکس در سرطان کبد و الگوهای موثر بر پروفایل‌های بیان متفاوت ژن استفاده می‌کنند. کاربردهای یادگیری نظارت‌شده در زیست‌فناوری بسیار متنوع است. داده‌های بیان ژن، تنها داده‌ای نیست که در برچسب‌گذاری سلول به کار می‌رود. اطلاعات مورد نیاز برای این برچسب‌گذاری می‌تواند شامل داده‌های بیان ژن، توالی DNA یا پروتئین، تصاویر گرفته شده، یا هر نوع گزارش آزمایش دیگری باشد. روش‌های یادگیری ماشین نظارت‌شده بسیار انعطاف‌پذیر هستند و امکان برچسب‌گذاری این سلول‌ها را در صورت وجود مجموعه‌ای از نمونه‌های قبلاً برچسب‌خورده، فراهم می‌کنند. با توسعه روش‌های یادگیری عمیق، پردازش خودکار داده‌های با ابعاد بالا مانند تصاویر و ویدئوها به صورت سریع‌تر و کارآمدتر انجام می‌شود. این روش «برای طبقه‌بندی داده‌ها استفاده می‌شوند»[1]، تا نمونه‌ها را در مجموعه‌های گسسته طبقه‌بندی کنند. روش‌هایی که در مدل‌سازی قابل استفاده هستند، برچسب‌های گسسته نیستند بلکه پیوسته‌اند. در شرایط کنونی، نیازی نیست که پژوهشگران، این روش‌ها را پیاده‌سازی کنند، چون بسیاری از بسته‌های نرم‌افزاری به صورت عمومی در دسترس هستند که مستقیماً از تکنیک‌های یادگیری ماشین نظارت‌شده برای تحلیل اطلاعات کاربران بهره‌برداری شوند. از جمله این بسته‌ها می‌توان به بسته‌های منبع باز همچون کراس[2] اشاره کرد. برخی از این بسته‌ها نیازمند برنامه‌نویسی با استفاده از زبان‌هایی مانند آر، مطلب[3] یا (پایتون) هستند، در حالی که برخی دیگر را می‌توان با حداقل آشنایی در دستکاری داده‌ها استفاده کرد، بسیاری از شرکت‌ها نرم‌افزارهای خاص و یکپارچه برای کاربردهای زیست‌فناوری و توسعه داده‌ها ارائه می‌دهند.

روش‌های یادگیری ماشین بدون نظارت توانایی جستجو و کشف اطلاعات دارند، اما معمولاً کمتر از روش‌های نظارت‌شده کارآمد هستند. این روش‌ها شامل تصویربرداری فضای اومیکس در ابعاد کوچک‌تر و ترکیبی از داده‌کاوی و انتخاب ویژگی‌ها می‌شوند، که به‌ویژه در خوشه‌بندی مؤثر است. این رویکردها برچسب نمی‌خواهند و به‌جای آن، نمونه‌ها را بر اساس شباهت‌ها و ویژگی‌هایشان دسته‌بندی می‌کنند. در حالیکه خوشه‌بندی یکی از رایج‌ترین روش‌ها در یادگیری بدون نظارت است، رویکردهای دیگری هم وجود دارند که هدف مشابهی دارند. این فرآیندها می‌توانند الگوهای پنهان و دانش جدیدی را آشکار کنند، مانند دسته‌بندی خودکار پروتئین‌ها بر اساس ویژگی‌هایشان مثل آب‌دوستی یا آب‌گریزی، یا گروه‌بندی انواع پروتئین در مجموعه‌های مختلف که منجر به کشف ویژگی‌های نوین می‌شود. به دلیل اینکه کشف دانش در این روش‌ها غیرمستقیم است، احتمال کشف اطلاعات جدید نسبتاً کمتر است. بنابراین، الگوریتم‌های یادگیری ماشین نظارت‌شده بیش‌ترین پتانسیل را در کشف دانش جدید دارند (۹،۱۸).

مقابله با چالش‌های استفاده از نرم‌افزارهای بیوانفورماتیک

کاربردهای هوش مصنوعی در بیوانفورماتیک متفاوت است؛ زیرا از حجم بالای نرم‌افزارها و اطلاعات عمومی در طراحی الگوریتم‌ها و تولید دانش جدید استفاده میشود. ولی ماهیت داده‌ها و کاربردهای هوش مصنوعی، سؤالاتی درباره مزایا و معایب اشتراک‌گذاری داده‌ها و همچنین تداخل احتمالی بین ملاحظات اقتصادی و اخلاقی ایجاد می‌کند. مجموعه‌های نرم‌افزاری یکپارچه و فرمت‌های استاندارد فایل، مکانیسم تحلیل را ساده‌تر کردند و برنامه‌های کاربردی وب، دسترسی به ابزارهای تحلیلی را برای آزمایشگاه‌های با منابع محاسباتی محدود ممکن ساخته‌اند. این اشتراک‌گذاری، اثرات اقتصادی مثبتی نیز دارد، زیرا پروژه‌های تحقیقاتی می‌توانند با استفاده از ابزارهای عمومی، آزمایش‌های پرهزینه و زمان‌بر برون–تن[4] را محدود کنند. ولی موفقیت این پروژه‌ها به قابلیت اعتماد نرم‌افزار مورد استفاده بستگی دارد. اکثر نرم‌افزارهای بیوانفورماتیک تحت مجوز عمومی همگانی (GPL)[5] یا مجوز MIT منتشر می‌شوند که امکان استفاده مجدد از کد را فراهم کرده و توسعه ابزارهای جدید را تسریع می‌کنند. در عین حال، این مجوزها شامل سلب مسئولیت‌هایی هستند که مسئولیت و ضمانت را محدود می‌کنند و در نهایت، کاربر نهایی را مسئول هرگونه نقص یا عملکرد نادرست نرم‌افزار می‌دانند. هرچند این رویکرد در بسیاری موارد رایج است، اما در برنامه‌های که نیازمند تضمین هستند، نبود ضمانت‌ها باید با احتیاط صورت گیرد. بنابراین، استفاده از روش‌های برنامه‌نویسی برای ارتقاء کیفیت نرم‌افزار ضروری است و با توجه به اهمیت ابزارهای قابل اعتماد، انجام آزمایش‌های دقیق پیش از انتشار نرم‌افزار لازم است (۴).

بازنگری در فرآیند بازیافت اطلاعات: مزایا و چالش‌های حقوقی، اخلاقی، و اقتصادی

این رویکرد، ژنگان را به‌عنوان یک سامانه یکپارچه می‌بیند که امکان به اشتراک‌گذاری و بازیابی داده‌های فناوری‌های اومیکس برای پروژه‌های تحقیقاتی جدید را فراهم کرده است. اشتراک‌گذاری و بازیابی اطلاعات، علاوه بر مزایای اخلاقی، از نظر اقتصادی هم مقرون به‌صرفه است. از نظر اخلاقی، اشتراک‌گذاری داده‌ها ممکن

1. in vitro
2. General Public Licence
3. Classifiers
4. Keras
5. MatLab

است نیاز به آزمایش‌های حیوانی را کاهش دهد، مخصوصاً زمانی که این آزمایش‌ها به حیوانات رنج و درد وارد می‌کند. از دیدگاه اقتصادی، هزینه نگهداری، ذخیره‌سازی و به اشتراک‌گذاری داده‌ها با فناوری‌های جدید، بسیار کمتر از تولید داده‌های نوین است. همچنین، این فرآیند، دسترسی آزمایشگاه‌های کوچک با بودجه محدود را به منابع داده گسترش می‌دهد (۱۹).

یکی دیگر از مزایای مهم، ایجاد مجموعه‌های داده بزرگ است؛ این مجموعه‌ها برای آموزش و بهبود الگوریتم‌های هوش مصنوعی ضروری هستند. این موضوع، انگیزه‌ای قوی برای جذب سرمایه‌گذاری و ترغیب ناشران به ترویج اشتراک‌گذاری داده‌ها به وجود آورده است. اما در مسیر اشتراک‌گذاری، خطراتی نیز وجود دارند؛ نشت اطلاعات، ممکن است ناشی از سوءاستفاده یا خطا، حریم خصوصی افراد را تهدید کرده و منجر به تبعیض شود. برای کاهش این خطرات، در حوزه امنیت پروژه‌های بزرگ سرمایه‌گذاری می‌شود و از فناوری‌های ناشناس‌سازی و روش‌های دقیق برای کنترل دسترسی استفاده می‌شود (۲۰). هم‌اکنون، اطلاعات عمومی در قالب ویژگی‌های کمی‌سازی شده ژنگان ارائه می‌شود، در حالیکه داده‌های خام تنها در اختیار موسسات مجاز است. با این حال، تضمین کامل امنیت و حریم خصوصی بیماران، همچنان چالش‌برانگیز است. با افزایش حجم داده‌ها، روش‌های ناشناس‌سازی با انجام یکپارچه‌سازی داده، ممکن است آسیب‌پذیر شوند. همچنین، اصل «انصراف»، که حق توقف مشارکت و حذف داده‌ها است، در پروژه‌های بزرگ بین‌المللی، به دلیل پیچیدگی پیگیری، ممکن است عملی نباشد.

نگرانی دیگر، نفوذ و نقض حریم خصوصی با پیشرفت‌های الگوریتم‌های هوش مصنوعی است، بدون اینکه فرآیند رضایت آگاهانه درنظر گرفته شود. با این حال، نباید بین حریم خصوصی و فناوری‌های هوشمند، یکی را بر دیگری ترجیح داد. در حال حاضر، تلاش‌هایی انجام می‌شود تا راه‌حل‌هایی پیدا شود که هر دو نیاز را مانند حذف انتخابی اطلاعات برآورده کند (۱۹).

در ارزیابی مراقبت‌های بهداشتی، هوش مصنوعی معمولاً بر شناسایی الگوهای خاص تمرکز دارد که در فرآیندهای تشخیص و درمان توسط پزشکان استفاده می‌شود. البته این نوع طبقه‌بندی کامل نیست و می‌توانند دو نوع خطا ایجاد کنند: اشتباه در تخصیص عناصر به یک کلاس (مثبت دروغین) و ناتوانی در شناسایی عناصر مربوطه (منفی دروغین). وقتی هدف، تشخیص «بیماری» است، مثبت‌های دروغین ممکن است منجر به اضطراب و نگرانی بی‌مورد در افراد شوند و با افزایش نیاز به آزمایش‌های غربالگری پرهزینه و غیرضروری، بار مالی سامانه بهداشتی را افزایش دهند. برعکس، منفی‌های دروغین می‌توانند منجر به تأخیر در تشخیص صحیح شوند و در بیماری‌هایی مانند سرطان که زمان تشخیص در اثربخشی درمان اهمیت دارد، عواقب جدی دارند. به همین دلیل، اشتباه در طبقه‌بندی عمدتاً به مشکلات اقتصادی و حقوقی می‌انجامد. پیامدهای متفاوت این دو خطا باعث شده است برخی محققان پیشنهاد دهند که ارزیابی عملکرد طبقه‌بندی‌ها بر اساس هزینه انجام شود. اما اجرای چنین رویکردی پیچیدگی‌هایی دارد؛ درحالیکه هزینه مرتبط با مثبت‌های دروغین را می‌توان بر اساس هزینه‌های آزمایش‌های غربالگری برآورد کرد، کمی‌سازی هزینه‌های منفی‌های دروغین دشوارتر است، زیرا نیازمند ارزیابی ارزش مالی جان انسان است. برای کاهش این خطاها و بهبود دقت در طبقه‌بندی، دو روش پیشنهاد می‌شود: ۱) بهره‌گیری از دستورالعمل‌هایی که شامل بازبینی موارد مشکل‌دار توسط متخصصان است؛ و ۲) توسعه روش‌های هوش مصنوعی، که برای تشخیص و اصلاح خطاهای طبقه‌بندی آموزش دیده‌اند. علاوه بر دقت، یکی دیگر از چالش‌های مهم در کاربرد هوش مصنوعی در شرایط زیستی و پزشکی، مسئله انصاف در ارائه اطلاعات است. این موضوع غالباً بر اثر آموزش الگوریتم‌ها بر داده‌های مبتنی بر سوگیری یا نابرابری ناشی می‌شود. برای مثال، آموزش بر روی داده‌های جمعیتی با سوگیری می‌تواند منجر به شناسایی نشانگرهای زیستی مربوط به اقوام خاص شود، که ممکن است در شناسایی نادرترین بیماری‌ها مفید باشد، اما همزمان خطر نابرابری بین جمعیت‌های کم‌نماینده و پرنماینده نیز وجود دارد (۲۱،۲۲).

بحث

با رشد داده‌های زیستی و پیچیدگی آن از منابع مختلف، استفاده از زیرساخت‌های محاسباتی اهمیت زیادی یافته‌اند. پیشرفت‌های اخیر در فناوری‌های توالی‌یابی و روش‌های با توان عملیاتی بالا، تغییرات مهمی در علوم زیستی و زیست‌فناوری ایجاد کرده است. الگوریتم‌های هوش مصنوعی که قادر به ذخیره‌سازی و پردازش داده‌های خام و بدون ساختار در مقیاس وسیع هستند، امکان استخراج سریع اطلاعات را فراهم می‌کنند و در توسعه سامانه‌های هوشمند با تصمیم‌گیری پیچیده نقش حیاتی دارند. این دستاوردها در تولید، نگهداری و تحلیل داده‌ها، مسیر را برای توسعه محصولات و خدمات متنوع در حوزه‌هایی مانند زیست‌شناسی هموار کرده است. انقلاب صنعتی سوم با پیشرفت‌های محاسباتی و اینترنت آغاز شد و زیرساخت‌های لازم برای ظهور هوش مصنوعی را فراهم ساخت، در حالی که حجم عظیم داده‌ها و تحلیل آن‌ها، توانایی‌های شناختی انسان را به سطوح جدیدی رساند (۹). اکنون، هوش مصنوعی به عنوان عنصر مرکزی در انقلاب صنعتی چهارم شناخته می‌شود. آزمایش‌هایی که قبلاً ماه‌ها یا سال‌ها طول می‌کشید، با فناوری‌های نوین جمع‌آوری و تحلیل داده‌ها، نه تنها ممکن، بلکه غالباً ارزان‌تر شده‌اند. تحلیل‌های آزمایشی منجر به تولید داده‌های خام در قالب‌های متنوع شده و ذخیره‌سازی و تحلیل این داده‌ها با ابزارهای هوش مصنوعی، فرصت‌های جدیدی برای محققان، دانشگاهیان و صنعت زیست‌فناوری به وجود آورده

است. آینده هوش مصنوعی در زیست‌شناسی دو رویکرد اصلی پیش رو دارد: رقابت یا همکاری بین رویکردهای اطلاعات‌محور و مدل‌محور. همکاری این دو رویکرد، شناخت دقیق‌تر از روابط و مکانیسم‌های سامانه‌ها را تسهیل می‌کند. روش‌های مبتنی بر مدل می‌توانند محدودیت‌های دانشی را تعریف و کنترل کنند، در حالیکه نتایج هوش مصنوعی در تعیین پارامترهای مدل‌های زیستی مؤثر است (۴).

کاربردهای گسترده هوش مصنوعی در زیست‌شناسی شامل تشخیص دقیق ساختار سه‌بعدی مولکول‌های زیستی مانند پروتئین‌ها است، که در تحقیقات زیستی نقش اصلی دارند. علاوه بر این، هوش مصنوعی، فراتر از محدودیت‌های آزمایشگاهی بر تمامی مراحل چرخه عمر یک محصول دارویی یا شیمیایی اثر می‌گذارد. ابزارها و برنامه‌های هوشمند، فرآیندهای تولید پیچیده را خودکار می‌کنند و نیاز روزافزون به داروها، مواد شیمیایی صنعتی و مواد خام زیستی را برطرف می‌سازند. در زیست‌فناوری، هوش مصنوعی دوگانگی است که بر توازن بین نوآوری و اقدامات حفاظتی تأکید دارد. نظارت زیاد ممکن است مانع پیشرفت، در حوزه‌هایی مانند کشف دارو و زیست‌شناسی مصنوعی شود، جایی که فناوری‌های پیشرفته می‌تواند منجر به پیشرفت‌های قابل توجهی گردد (۲۳). در کنار آن، نبود تدابیر امنیتی ممکن است منجر به سوءاستفاده‌هایی مانند توسعه سلاح‌های زیستی یا تولید عوامل آسیب‌زا به صورت تصادفی شود. دستیابی به این توازن نیازمند سیاست‌های انعطاف‌پذیر است که در برابر فناوری‌های نوظهور مقاوم باشد و تهدیدات جدید را درنظر گیرد (۲۴). رویکرد فعال و بین رشته‌ای در مواجهه با چالش‌های هوش مصنوعی در زیست‌فناوری ضروری است. همکاری میان پژوهشگران هوش مصنوعی، مهندسان زیستی و متخصصان امنیت، می‌تواند درک جامع‌تری از خطرات دوگانه ایجاد کند و امکان طراحی اقدامات حفاظتی عمل‌گرایانه را فراهم آورد. در مراحل اولیه، مشارکت ذینفعان مختلف موجب می‌شود که نگرانی‌های ایمنی و اخلاقی در کنار پیشرفت‌های علمی لحاظ شوند. برخلاف تلاش‌های اولیه، برنامه‌های جاری و کاربردی هوش مصنوعی باید خطرات احتمالی، مانند سوءاستفاده در سلاح‌های بیولوژیکی یا برنامه‌های مخرب، را در نظر داشته باشند. ارتباط مؤثر میان خطرات و مزایای هوش مصنوعی در زیست‌فناوری برای جلب اعتماد عمومی بسیار حیاتی است. ساده‌سازی مفاهیم پیچیده به زبان قابل فهم، بدون اغراق در نتیجه‌گیری‌ها، می‌تواند فاصله دانش بین متخصصان و مردم عادی را کاهش دهد و چارچوب‌های نظارتی مؤثر و متعادلی توسعه دهد که همزمان نوآوری را ترویج داده و خطرات را کاهش می‌دهد(۲۵).

نتیجه‌گیری

هوش مصنوعی، با بهره‌گیری از اطلاعات متنوع، در تمام جنبه‌های فعالیت‌های تحقیقاتی تأثیرگذار است. معماری‌های یادگیری عمیق به دلیل توانایی در پردازش داده‌های حجیم، متنوع و پیچیده به‌طور سلسله‌مراتبی، قادرند الگوهای پنهان در داده‌ها را استخراج و تفسیر کنند. اما این معماری‌ها اغلب توضیح‌پذیر نیستند و نمی‌توانند فرآیند شفاف نتیجه‌گیری را توضیح دهند. این ابهام در انتخاب ویژگی‌های پایه می‌تواند ارزش کاربردی نتایج را کاهش دهد. همچنین، دشواری در انتخاب معماری مناسب برای مسائل خاص، به‌ویژه در محیط‌های حساس، مربوط به نبود تفسیرپذیری در برخی مسیرهای یادگیری عمیق است. در این فضا، اشتراک‌گذاری رایگان نرم‌افزار، منابع آموزشی و پایگاه‌های داده معتبر، که پایه توسعه کاربردهای هوش مصنوعی در علوم زیستی و حوزه‌های پیچیده‌تر هستند، اهمیت می‌یابد. به‌ویژه در بخش‌هایی که هوش مصنوعی می‌تواند شناخت دقیق‌تری از مکانیسم‌های زیستی ارائه دهد، استفاده از این ابزارها ترجیح داده می‌شود. در واقع، تقویت دانش زیستی با نتایج هوش مصنوعی به یک نیاز علمی و عملی تبدیل شده است. این امر نیازمند امکاناتی برای اعتبارسنجی و ساخت مدل‌های نظری با توانایی پیش‌بینی بالا در رفتارهای استاتیک و دینامیک سامانه‌های زیستی با سطوح مختلف پیچیدگی است. افزایش کمی و کیفی اطلاعات معتبر و تلفیق آن‌ها با مدل‌سازی نظری، در آینده اعتماد بیشتری به پیش‌بینی‌ها و تصمیمات مبتنی بر هوش مصنوعی ایجاد خواهد کرد.

منابع

1. Bhardwaj A, Kishore S, Pandey DK. Artificial Intelligence in Biological Sciences. Life 2022; 12(9):1430.
2. Wanerman RE, Javitt GH, Shah AB. Artificial Intelligence in Biotechnology: A Framework for Commercialization:419–27.
3. Jarallah SJ, Almughem FA, Alhumaid NK, Fayez NAL, Alradwan I, Alsulami KA et al. Artificial intelligence revolution in drug discovery: A paradigm shift in pharmaceutical innovation. International Journal of Pharmaceutics 2025:680 ;125789.
4. Caudai C, Galizia A, Geraci F, Le Pera L, Morea V, Salerno E et al. AI applications in functional genomics. Computational and Structural Biotechnology Journal 2021;19: 5762-90.
5. Holzinger A, Keiblinger K, Holub P, Zatloukal K, Müller H. AI for life: Trends in artificial intelligence for biotechnology. New Biotechnology 2023; 74: 16-24.

6. Gul I, Adil M, Lu H, Lu S, Li H, Liu F et al. AI-Driven Omics for Smart Remediation of Heavy Metal Contaminated Soils. Physiol Plant 2025; 177(6):e70611.
7. Jaganathan D, Ramasamy K, Sellamuthu G, Jayabalan S, Venkataraman G. CRISPR for Crop Improvement: An Update Review. Front Plant Sci 2018;9 :985.
8. Ali M, Shabbir K, Ali S, Mohsin M, Kumar A, Aziz M et al. A New Era of Discovery: How Artificial Intelligence has Revolutionized the Biotechnology. Nepal J Biotechnol; 2024; 12(1): 1-11.
9. Oliveira AL. Biotechnology, Big Data and Artificial Intelligence. Biotechnology Journal 2019; 14(8).
10. Du Lee H, Park EG, Lee YJ, Jeong H-S, Roh H-Y, Jeong G-R et al. Advances in Artificial Intelligence (AI) Models and Generative Algorithms Represent a New Paradigm for Genomics Research. Int J Mol Sci 2025; 26(22)
11. Kortemme T. De novo protein design-From new structures to programmable functions. Cell 2024; 187(3): 526-44.
12. Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol 2019; 20(11): 681-97.
13. Liu Y-T, Zhang L-L, Jiang Z-Y, Tian X-S, Li P-L, Wu P-H et al. Applications of Artificial Intelligence in Biotech Drug Discovery and Product Development. MedComm 2025; 6(8).
14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019; 25(1): 44-56.
15. Flores JE, Claborne DM, Weller ZD, Webb-Robertson B-JM, Waters KM, Bramer LM. Missing data in multi-omics integration: Recent advances through artificial intelligence. Front Artif Intell 2023:6 ;1098308.
16. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014; 15(12): 550.
17. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A et al. A survey of best practices for RNA-seq data analysis. Genome Biol; 2016;17:13.
18. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet 2015; 16(2): 85-97.
19. Kaye J. The tension between data sharing and the protection of privacy in genomics research. Annu Rev Genomics Hum Genet 2012; 13: 415-31
20. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. PLoS Med 2018; 15(11):e1002689.
21. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med 2018; 378(11): 981-3.
22. Haro LP de. Biosecurity Risk Assessment for the Use of Artificial Intelligence in Synthetic Biology. Appl Biosaf 2024;29(2): 96-107.
23. Sanders LM, Scott RT, Yang JH, Qutub AA, Garcia Martin H, Berrios DC et al. Biological research and self-driving labs in deep space supported by artificial intelligence. Nat Mach Intell 2023; 5(3): 208-19.
24. O'Brien JT, Nelson C. Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology. Health Security 2020; 18(3): 219-27.
25. Wheeler NE. Responsible AI in biotechnology: balancing discovery, innovation and biosecurity risks. Front. Bioeng. Biotechnol. 13 ;2025.
26. Sahner D, Spellmeyer DC. Artificial Intelligence: Emerging Applications in Biotechnology and Pharma:399–417.