Shahid Bahonar University of Kerman
Agricultural Biotechnology Journal
Iranian Biotechnology Society
Print ISSN: 2228-6705 Online ISSN: 2228-6500

The role of artificial intelligence in genomics

Mohammadreza Mohammadabadi

*Corresponding Author. Professor, Animal Science Department, Faculty of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran. E-mail address: mrm@uk.ac.ir

Hamid Kheyrodin

Assistant Professor, Semnan University, Semnan, Iran. E-mail address: hamid.kheyrodin@semnan.ac.ir

Volodymyr Afanasenko

National University of Life and Environmental Sciences of Ukraine, Ukraine. E-mail address: afanasenko77@gmail.com

Olena Babenko

Assistant Professor, Department of Animal Science, Bila Tserkva National Agrarian University, Bila Tserkva, Ukraine. E-mail address: lelya.babenko1978@gmail.com

Nataliia Klopenko

Assistant Professor, Department of Animal Science, Bila Tserkva National Agrarian University, Bila Tserkva, Ukraine. E-mail address: dripa2604@ukr.net

Oleksandr Kalashnyk

Sumy National Agrarian University, Sumy, Ukraine. E-mail address: oleksandr.kalashnyk@snau.edu.ua

Yulia Ievstafiieva

Associate professor, Department of Technologies of Livestock Production and processing, Higher Educational Institution “Podillia State University”, Ukraine. E-mail address: pp.nika22@ukr.net

Vita Buchkovska

Associate Professor, Department of Technologies of Livestock Production and processing, Higher Educational Institution “Podillia State University”, Ukraine. E-mail address: vbutschk@ukr.net

Abstract

Objective

Data generation in biology and biotechnology has greatly increased in recent years due to the very rapid development of high-performance technologies. These data are obtained from studying

biological molecules, such as metabolites, proteins, RNA, and DNA, to understand the role of these molecules in determining the structure, function, and dynamics of living systems. Functional genomics is a field of research that aims to characterize the function and interaction of all the major components (DNA, RNA, proteins, and metabolites, along with their modifications) that contribute to the set of observable characteristics of a cell or individual (i.e., phenotype). Furthermore, in a breeding program, genetic improvement can be maximized through accurate identification of superior animals that are selected as parents of the next generation, thereby achieving breeding goals. Artificial neural networks have been proposed to alleviate this limitation of traditional regression methods and can be used to handle nonlinear and complex data, even when the data is imprecise and noisy. Omics data can be too large and complex to handle through visual analysis or statistical correlations. This has encouraged the use of machine intelligence or artificial intelligence. The objectives of this study was to review the main applications of artificial intelligence methods in functional genomics, cancer, agriculture, domestic animals and its intertwined fields, i.e. epigenomics, transcriptomics, epitranscriptomics, proteomics, and metabolomics, discuss important aspects of data management, such as data integration, cleaning, noise removal, balancing and ratio of missing data, functional genomics-system modeling, artificial intelligence and systems biology, addressing legal, ethical and economic issues related to the application of artificial intelligence methods in the field of genomics and presenting a view of possible scenarios in the future.

Materials and methods

In this review, all researches conducted in the field of artificial intelligence application in functional genomics, cancer, agriculture, domestic animals, and its intertwined fields, i.e. epigenomics, transcriptomics, epitranscriptomics, proteomics, and metabolomics, were tried, focusing on the applications of recent years after Increase production of big data to be studied and used.

Results

The studies showed that the application of artificial intelligence in all fields, including functional genomics, cancer, agriculture, domestic animals, and its intertwined fields, i.e., epigenomics, transcriptomics, epitranscriptomics, proteomics, and metabolomics, is increasing rapidly and has many benefits.

Conclusions

Considering the vital applications that are often addressed by biology and especially functional genomics, it is better to deal with artificial intelligence tools that are able to help mechanistic understanding of biological processes. In other words, enabling systems biology is important to

reap the benefits of AI results in genomics. Interpretability can certainly help AI to be more easily adopted in practical applications such as medicine. In our opinion, increasing the volume and diversity of reliable big data and integrating it with theoretical modeling will help increase human trust in AI-based predictions and decisions in the future. On the one hand, model-based approaches can provide knowledge-based constraints. On the other hand, the results of artificial intelligence can help to create the parameters of models of biological systems. Functional genomics, as well as all fields of medicine, biology, agriculture, animal science, and other sciences that involve both individual and collective rights, is a complex field of research. Those who want to use AI in this field will find it difficult to navigate, as the field is sensitive to legal, ethical, and spiritual aspects. Artificial intelligence opens many opportunities that we should not refuse for fear of not understanding all the steps. The path that artificial intelligence will follow is just beginning to unfold. It has many promises and many potential dangers ahead. This path will probably be long and irreversible. Artificial intelligence will change our lives and we need to change our minds as soon as possible to adapt, accept, and manage the resulting changes in the best possible way, to ensure that they will bring as many benefits as possible and will cause the least possible negative consequences for us.

Keywords: Agriculture, Artificial intelligence, Cancer, Domestic animals, Genomics

Paper Type: Review Paper.

Citation: Mohammadabadi M, Kheyrodin H, Afanasenko V, Babenko O, Klopenko N, Kalashnyk O, Ievstafiieva Y, Buchkovska V (2024) The role of artificial intelligence in genomics. *Agricultural Biotechnology Journal* 16 (2), 195-279.

Agricultural Biotechnology Journal 16 (2), 195-279. DOI: 10.22103/jab.2024.23558.1575

Received: March 04, 2024. Received in revised form: April 19, 2024.

Accepted: April 20, 2024. Published online: May 31, 2024.

Publisher: Faculty of Agriculture and Technology Institute of Plant Production, Shahid Bahonar University of Kerman-Iranian Biotechnology Society.

مجله بیوتکنولوژی کشاورزی
شاپا چاپی: ۶۷۰۵-۲۲۲۸ شاپا الکترونیکی: ۶۵۰۰-۲۲۲۸

نقش هوش مصنوعی در ژنومیکس

محمدرضا محمدآبادی

*نویسنده مسئول: استاد بخش علوم دامی، دانشکده کشاورزی، دانشگاه شهید باهنر کرمان، ایران. ایمیل: mrm@uk.ac.ir

حمید خیرالدین

استادیار، دانشکده کویرشناسی، دانشگاه سمنان، سمنان، ایران. رایانامه: hamid.kheyrodin@semnan.ac.ir

ولودیمیر آفاناسنکو

استادیار، دانشگاه ملی علوم محیطی و زیستی اوکراین، اوکراین. ایمیل: afanasenko77@gmail.com

اولنا بابنکو

استادیار، گروه علوم دامی، دانشگاه ملی کشاورزی بیلا تسرکوا، بیلا تسرکوا، اوکراین. ایمیل: lelya.babenko1978@gmail.com

ناتالیا کلوپنکو

استادیار، گروه علوم دامی، دانشگاه ملی کشاورزی بیلا تسرکوا، بیلا تسرکوا، اوکراین. ایمیل: dripa2604@ukr.net

الکساندر کلاشنیک

دانشگاه ملی کشاورزی سومی، سومی، اوکراین. ایمیل: oleksandr.kalashnyk@snau.edu.ua

یولیا ایوستافیوا

دانشیار، گروه فناوری‌های تولید و فرآوری دام، دانشگاه دولتی پودیلیا، اوکراین. ایمیل: pp.nika22@ukr.net

ویتا بوچکوفسکا

دانشیار، گروه فناوری های تولید و فرآوری دام، دانشگاه دولتی پودیلیا، اوکراین. ایمیل: vbutschk@ukr.net

تاریخ دریافت: ۱۴۰۲/۱۲/۱۴ تاریخ دریافت فایل اصلاح شده نهایی: ۱۴۰۳/۰۱/۳۱ تاریخ پذیرش: ۱۴۰۳/۰۲/۰۱

چکیده

هدف: تولید داده در زیست‌شناسی و زیست‌فناوری در سال‌های گذشته به دلیل توسعه بسیار سریع فناوری‌های با کارایی بالا بسیار زیاد شده است. این داده‌ها با مطالعه مولکول‌های زیستی، از قبیل متابولیت‌ها، پروتئین‌ها، RNA و DNA برای درک و فهم نقش

این مولکول‌ها در تعیین ساختار، عملکرد و دینامیک سیستم‌های زنده حاصل شده‌اند. ژنومیک عملکردی رشته‌ای از تحقیقات است که هدفش مشخص کردن عملکرد و تعامل همه اجزای اصلی (DNA، RNA، پروتئین‌ها و متابولیت‌ها، همراه با تغییرات آن‌ها) است. علاوه بر این، در یک برنامه اصلاح نژادی، پیشرفت ژنتیکی را می‌توان از طریق شناسایی دقیق حیوانات برتر که به عنوان والدین نسل بعدی انتخاب می‌شوند، به حداکثر رساند و در نتیجه به اهداف اصلاح نژادی دست یافت. شبکه‌های عصبی مصنوعی برای کاهش این محدودیت روش‌های رگرسیون سنتی پیشنهاد شده‌اند و می‌توانند برای مدیریت داده‌های غیرخطی و پیچیده، حتی زمانی که داده‌ها اریب و نویز هستند، استفاده شوند. داده‌های اومیکس بعضی اوقات به قدری بیش از حد بزرگ و پیچیده هستند که از طریق تجزیه و تحلیل بصری یا همبستگی‌های آماری قابل بررسی نیستند. این امر استفاده از هوش ماشینی یا هوش مصنوعی را تشویق کرده است. اهداف این مطالعه عبارتند از بررسی کاربردهای اصلی روش‌های هوش مصنوعی در ژنومیکس عملکردی، سرطان، کشاورزی، حیوانات اهلی و زمینه‌های درهم تنیده آن یعنی اپی‌ژنومیکس، ترانس‌کریپتومیکس، اپی‌ترانس‌کریپتومیکس، پروتئومیکس و متابولومیکس، بحث در مورد جنبه‌های مهم مدیریت داده‌ها، مانند یکپارچه‌سازی داده‌ها، جانهی یا انتساب (imputation)، تمیز کردن، حذف نویز، متعادل‌سازی و نسبت داده‌های از دست رفته، مدل‌سازی سیستم-ژنومیکس عملکردی، هوش مصنوعی و سیستم‌های بیولوژی، پرداختن به مسائل حقوقی، اخلاقی و اقتصادی مربوط به کاربرد روش‌های هوش مصنوعی در حوزه ژنومیکس و ارائه نمایی از سناریوهای احتمالی آینده است.

مواد و روش‌ها: در این بررسی سعی شد کلیه پژوهش‌های انجام شده در زمینه کاربرد هوش مصنوعی در ژنومیکس عملکردی، سرطان، کشاورزی، حیوانات اهلی و زمینه‌های درهم تنیده آن یعنی اپی‌ژنومیکس، ترانس‌کریپتومیکس، اپی‌ترانس‌کریپتومیکس، پروتئومیکس و متابولومیکس، با تمرکز بر روی کاربردهای سال‌های اخیر پس از افزایش تولید کلان داده مطالعه و مورد استفاده قرار گیرند.

نتایج: بررسی‌ها نشان داد که کاربرد هوش مصنوعی در همه رشته‌ها از جمله ژنومیکس عملکردی، سرطان، کشاورزی، حیوانات اهلی و زمینه‌های درهم تنیده آن یعنی اپی‌ژنومیکس، ترانس‌کریپتومیکس، اپی‌ترانس‌کریپتومیکس، پروتئومیکس و متابولومیکس با سرعت زیادی رو به افزایش است و فواید زیادی دارد.

نتیجه‌گیری: با توجه به کاربردهای حیاتی که اغلب توسط زیست شناسی و به ویژه ژنومیک عملکردی به آن پرداخته می شود، بهتر است با ابزارهای هوش مصنوعی که قادر به کمک به درک مکانیکی فرآیندهای بیولوژیکی هستند، سروکار داشته باشیم. به عبارت دیگر، فعال کردن بیولوژی سیستم‌ها برای به دست آوردن مزایایی از نتایج هوش مصنوعی در ژنومیکس مهم است. تفسیرپذیری مطمئناً می‌تواند به هوش مصنوعی کمک کند تا در کاربردهای عملی مانند پزشکی راحت‌تر پذیرفته شود. به نظر ما، افزایش حجم و تنوع داده‌های عظیم قابل اعتماد و ادغام آن با مدل‌سازی نظری به افزایش اعتماد انسان‌ها به پیش‌بینی‌ها و تصمیم‌گیری‌های مبتنی بر هوش مصنوعی در آینده کمک می‌کند. از یک سو، رویکردهای مبتنی بر مدل می‌توانند محدودیت‌های

مبتنی بر دانش را فراهم کنند. از سوی دیگر، نتایج هوش مصنوعی می‌تواند به ایجاد پارامترهای مدل‌های سیستم‌های بیولوژی کمک کنند. ژنومیکس عملکردی، و همچنین تمامی رشته‌های پزشکی، زیست شناسی، کشاورزی، علوم حیوانات اهلی و سایر علومی که هم حقوق فردی و هم حقوق جمعی در آن دخیل است، یک حوزه تحقیقاتی پیچیده است. کسانی که می‌خواهند از هوش مصنوعی در ژنومیکس استفاده کنند متوجه خواهند شد که مسیر دشواری در پیش دارند، چرا که این رشته نسبت به جنبه‌های قانونی، اخلاقی و معنوی بسیار حساس است. هوش مصنوعی فرصت‌های زیادی را می‌گشاید که از ترس درک نکردن همه مراحل نباید از آن‌ها امتناع کنیم. مسیر استفاده از هوش مصنوعی در علوم مختلف تازه شروع به باز شدن و توسعه کرده است. این مسیر مزایای زیادی را به همراه دارد و خطرات بالقوه زیادی را نیز پیش‌رو دارد. این مسیر احتمالا طولانی و غیرقابل برگشت خواهد بود. هوش مصنوعی زندگی ما را تغییر خواهد داد و ما باید در اسرع وقت نظر خود را تغییر دهیم تا تغییرات ناشی از آن را به بهترین شکل ممکن تطبیق دهیم، بپذیریم و مدیریت کنیم، تا اطمینان حاصل کنیم که آن‌ها تا حد امکان منافع بیشتری به همراه خواهند داشت و کمترین پیامدهای منفی ممکن را برای ما ایجاد خواهند کرد.

کلیدواژه‌ها: حیوانات اهلی، ژنومیکس، سرطان، کشاورزی، هوش مصنوعی

نوع مقاله: مروری.

استناد: محمدآبادی محمدرضا، خیرالدین حمید، آفاناسنکو ولودیمیر، بابنکو اولنا، کلوپنکو ناتالیا، کلاشنیک الکساندر، ایوستافیوا یولیا، بوچکوفسکا ویتا (۱۴۰۳) نقش هوش مصنوعی در ژنومیکس. *مجله بیوتکنولوژی کشاورزی*، ۱۶(۲)، ۲۷۹-۱۹۵.

Publisher: Faculty of Agriculture and Technology Institute of Plant Production, Shahid Bahonar University of Kerman-Iranian Biotechnology Society.

مقدمه

توسعه بسیار سریع فناوری‌های با کارایی بالا باعث تولید داده‌های حجیم در زیست‌شناسی و زیست‌فناوری شده است. این داده‌ها با مطالعه مولکول‌های زیستی، از قبیل متابولیت‌ها[1]، پروتئین‌ها[2]، RNA و DNA حاصل شده‌اند که برای درک و فهم نقش این مولکول‌ها در تعیین ساختار، عملکرد و دینامیک سیستم‌های زنده، مانند ارگانیسم، بافت یا سلول ضروری هستند. وقتی از اومیکس[3] در نام یک رشته یا عنوان استفاده می‌شود، بدین معنی است که قصد جمع‌آوری و تجزیه و تحلیل مجموعه‌های بزرگی از داده‌های بیولوژیکی وجود دارد. به علاوه، اومیکس برای نشان دادن کل مقدار DNA موجود در هر سلول یک ارگانیسم استفاده

[1] Metabolites
[2] Proteins
[3] Omics

می‌شود و چون ژنوم تحت تاثیر فاکتورهای زیادی قرار دارد هنگام بررسی آن با چالش‌های بزرگی روبرو می‌شویم (Chaudhary et al. 2018). این رشته‌ها عبارتند از متابولومیکس[1] (مطالعه متابولیت‌ها)، پروتئومیکس[2] (بررسی محصولات ترجمه‌ای رونوشت‌های کدکننده پروتئین)، اپی‌ترانسکریپتومیکس[3] (بررسی مجموعه و پویایی تغییرات RNA)، ترنسکریپتومیکس[4] (مطالعه رونوشت‌های RNA که منشا ژنومی دارند)، اپی‌ژنومیکس[5] (بررسی مدولاسیون‌هایی که DNA می‌تواند به طور برگشت پذیر متحمل شود) و ژنومیکس[6] (مطالعه محتوای اطلاعات DNA یک ارگانیسم یا سیستم). این موارد می‌توانند در یک ارگانیسم یا سیستم معین، در یک زمان و شرایط معین، در حالت‌های فیزیولوژیکی و پاتولوژیکی وجود داشته باشند. همه این رشته‌ها حوزه‌های مستقل مطالعاتی هستند، اما دانش و داده‌هایی که آن‌ها تولید می‌کنند با هدف بلندپروازانه ژنومیکس همگرا می‌شوند.

مفهوم "ژنوم" برای اولین بار در سال ۱۹۲۰ توسط‌هانس وینکلر، استاد وقت گیاه شناسی در دانشگاه‌هامبورگ، با اشاره به "تعدادهاپلوئید کروموزوم‌ها" واقع در هسته پیشنهاد شد (Noguera-Solano et al. 2013). ژنومیک عملکردی رشته‌ای است که هدفش مشخص کردن عملکرد و تعامل همه اجزای اصلی (DNA، RNA، پروتئین‌ها و متابولیت‌ها، همراه با تغییرات آن‌ها) است. این اجزای اصلی مجموعه‌ای از ویژگی‌های قابل مشاهده یک سلول یا فرد (یعنی فنوتیپ) را به تعامل عملکردی بین ویژگی‌های ژنتیکی زیربنایی (ژنوتیپ) و شرایط محیطی مرتبط می‌کنند. علاوه بر این، در یک برنامه اصلاح نژادی، پیشرفت ژنتیکی را می‌توان از طریق شناسایی دقیق حیوانات برتر که به عنوان والدین نسل بعدی انتخاب می‌شوند، به حداکثر رساند و در نتیجه به اهداف اصلاح نژادی دست یافت (Ahsani et al. 2010; Amiri Roudbar et al. 2017; Jafari Ahmadabadi et al. 2023; Mohamadipoor et al. 2021). یکی از مؤلفه‌های کلیدی این فرآیند، پیش‌بینی سریع و قابل اعتماد ارزش‌های اصلاحی برای نامزدهای انتخابی است. بااین‌حال، پیش‌بینی ارزش‌های اصلاحی اغلب یک کار محاسباتی چالش‌برانگیز و زمان‌بر است و بنابراین در بیشتر کشورها فقط به صورت دوره‌ای انجام می‌شود (Amiri Roudbar et al. 2018; Masoudzadeh et al. 2020; Mohammadabadi et al. 2021). جایگزین‌های سریع و کم‌هزینه‌ای که می‌توانند پیش‌بینی‌های تقریبی ارزش‌های اصلاحی را با دقت قابل‌قبول ارائه دهند، می‌توانند تصمیمات انتخاب و حذف به‌موقع‌تری را توسط شرکت‌های اصلاح‌کننده یا تولیدکنندگان محصولات لبنی فراهم کنند. شناسایی سریع نرهای برتر می‌تواند منجر به جمع آوری و توزیع زودتر منی و پیشرفت ژنتیکی سریع‌تر شود (Safaei et al. 2022; Mohammadabadi et al. 2023; Shokri et al. 2023). علاوه بر این، بررسی‌ها نشان داده‌اند که روش‌های رگرسیون مرسوم نمی‌توانند هم‌خطی (هم‌راستایی) چندگانه بین عوامل مستقل را ارزیابی کنند. از این رو، ممکن است منجر به نتایج اریب شود (Ghotbaldini et al. 2019). وقتی همبستگی بین متغیرها زیاد باشد، هم‌خطی (هم‌راستایی)

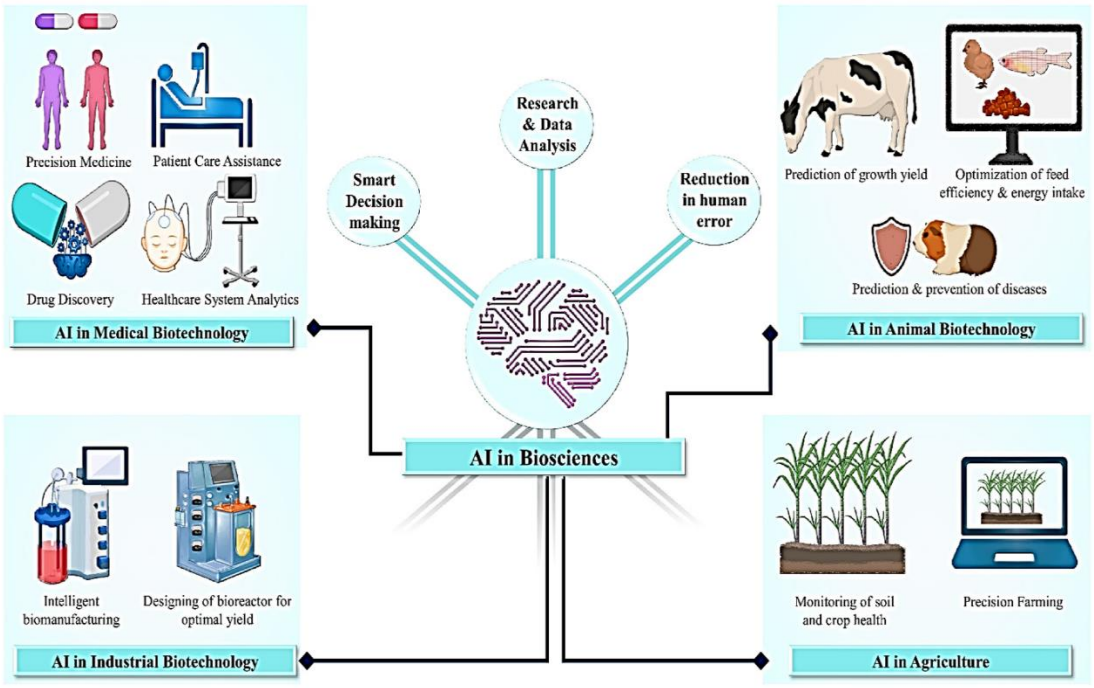
[1] Metabolomics
[2] Proteomics
[3] Epitranscriptomics
[4] Transcriptomics
[5] Epigenomics
[6] Genomics

چندگانه [1] اتفاق می‌افتد. بنابراین، به دست آوردن تخمین‌های قابل اعتماد از ضرایب رگرسیون فردی دشوار است (Khorshidi et al. 2019). در شرایطی که همبستگی برخی متغیرها بسیار زیاد است آن‌ها اساساً یک پدیده را اندازه گیری می‌کنند و اطلاعات مشابهی ارائه می‌دهند، از این رو این متغیرها می‌توانند به طور معکوس بر نتیجه رگرسیون تأثیر بگذارند. مشکلات ناشی از هم‌خطی (هم‌راستایی) بودن چندگانه در تحلیل رگرسیون مشخص شده است (Ghotbaldini et al. 2019). شبکه‌های عصبی مصنوعی برای کاهش این محدودیت روش‌های رگرسیون سنتی پیشنهاد شده‌اند و می‌توانند برای مدیریت داده‌های غیرخطی و پیچیده، حتی زمانی که داده‌ها نادقیق و نویز هستند، استفاده شوند (Pour Hamidi et al., 2017). این شبکه‌ها شامل مجموعه‌ای از اجزای پردازش هستند که به عنوان نورون‌ها یا گره‌ها نیز شناخته می‌شوند که عملکرد آن‌ها بر اساس نورون‌های بیولوژیکی است (Khorshidi et al. 2019). این واحدها در لایه‌هایی تشکیل می‌شوند که اطلاعات ورودی را پردازش کرده و به لایه‌های بعدی منتقل می‌کنند. توانایی شبکه در پردازش در نقاط قوت اتصال بین واحدی (یا وزن‌ها) جمع می‌شود و این توانایی از طریق فرآیند انطباق با مجموعه‌ای از الگوهای آموزشی به دست می‌آید (Ghotbaldini et al. 2019). علاوه بر این، روش شبکه عصبی مصنوعی کاملاً با رویکردهای آماری سنتی متفاوت است، که نیاز به یک الگوریتم مشخص برای تبدیل شدن توسط یک برنامه کامپیوتری دارد (Pour Hamidi et al., 2017). داده‌های اومیکس می‌توانند به قدری بیش از حد بزرگ و پیچیده باشند که از طریق تجزیه و تحلیل بصری یا همبستگی‌های آماری قابل بررسی نیستند. این امر استفاده از به اصطلاح هوش ماشینی یا هوش مصنوعی [2] (AI) را ترویج داده است (McCarthy et al. 2006). هوش مصنوعی نه تنها قادر به مدیریت حجم داده‌هایی است که برای ذهن انسان غیرقابل حل است، بلکه برای استخراج اطلاعاتی که فراتر از درک فعلی ما از سیستم تحت بررسی است نیز کاربرد دارد. نکته مهم این است که هوش مصنوعی به طور خودکار، از طریق تجربه به دست آورده از داده‌های آموزشی [3] عملکردش را بهبود می‌بخشد.

در هوش مصنوعی، نیاز به برنامه‌ریزی صریح برای انجام یک کار مشخص، که از ویژگی‌های متمایز الگوریتم‌های یادگیری ماشین [4] (ML) از جمله رگرسیون خطی، خوشه‌بندی و شبکه‌های بیزی [5] است وجود ندارد. اولین کاربرد روش‌های ML در زیست شناسی به اوایل دهه ۱۹۸۰ برمی‌گردد (Stormo et al. 1982). اخیراً، برنامه‌های ML در تمام زمینه‌های تحقیقاتی مرتبط با ژنومیک عملکردی، از جمله ژنومیکس (de Ridder et al. 2013; Angermueller et al. 2016; Camacho et al. 2018)، ترنسکریپتومیکس (Zhang et al. 2019)، پروتئومیکس (Park and Kellis 2015; Ragoza et al. 2017) و متابولومیکس (Grapov et al. 2018; Zampieri et al. 2019) به کار گرفته شده‌اند (شکل ۱).

[1] Multicollinearity
[2] Machine Intelligence or Artificial Intelligence
[3] Training data
[4] Machine Learning
5 Bayesian Networks

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500



شکل ۱. مدلی پیشنهادی از امکانات کاربردی هوش مصنوعی در رشته‌های بهداشت، کشاورزی، علوم دامی و بیوتکنولوژی صنعتی (Bhardwaj et al. 2022)

Figure 1. A proposed model of the practical possibilities of AI in the fields of health, agriculture, animal science and industrial biotechnology

در میان روش‌های ML، امیدوارکننده‌ترین روش‌ها برای پرداختن به پیچیدگی داده‌های اومیکس، روش‌هایی هستند که در مجموع به عنوان روش‌های یادگیری عمیق[۱] (DL) شناخته می‌شوند. این روش‌ها اطلاعات را با انجام عملیات ریاضی (که نورون[۲]، در قیاس با «عناصر محاسباتی» در مغز نام‌گذاری شده‌اند) که در لایه‌های متعدد (به همین دلیل عمیق نامیده می‌شوند) متصل به یکدیگر قرار گرفته‌اند (که به آن شبکه عصبی گفته می‌شود) پردازش می‌کنند. اگرچه اولین مدل‌های شبکه عصبی بیش از ۶۰ سال پیش پیاده‌سازی شدند، اما در ابتدا به دلیل هزینه‌های پولی و محاسباتی گزاف، منبعی جذاب اما ناپایدار بودند. پرسپترون[۳]، اولین معماری شبکه عصبی بود که در سال ۱۹۵۸ توسط فرانک روزنبلات[۴] به جامعه علمی معرفی شد (Rosenblatt 1958)، اما توانایی یادگیری محدودی داشت. علاوه بر این، علیرغم اینکه کامپیوتری که الگوریتم پرسپترون را اجرا می‌کرد اندازه یک اتاق بزرگ بود، اما قدرت پردازش کاملاً محدودی داشت. به طور کلی، به‌کارگیری مؤثر روش‌های DL تنها در دهه گذشته به لطف افزایش شدید عملکرد پردازنده که به تقاضای محاسباتی بالای آن‌ها رسیده است، به ویژه پس از استفاده مجدد از واحدهای پردازش گرافیکی[۵] (GPU) برای انجام بازی، امکان‌پذیر شده است. در همان سال‌ها، کاهش شدید هزینه‌های توالی‌یابی، در دسترس بودن سیل

[1] Deep Learning
2 Neurons
3 Perceptron
[4] Frank Rosenblatt
5 Graphical Processing Units

مجموعه‌های داده‌های بزرگ در مقیاس ژنومی را به ارمغان آورد و ژنومیک عملکردی را به زمینه‌ای مناسب برای کاربردهای DL تبدیل کرد (Eraslan et al. 2019; Esteva et al. 2019; Zou et al. 2019).

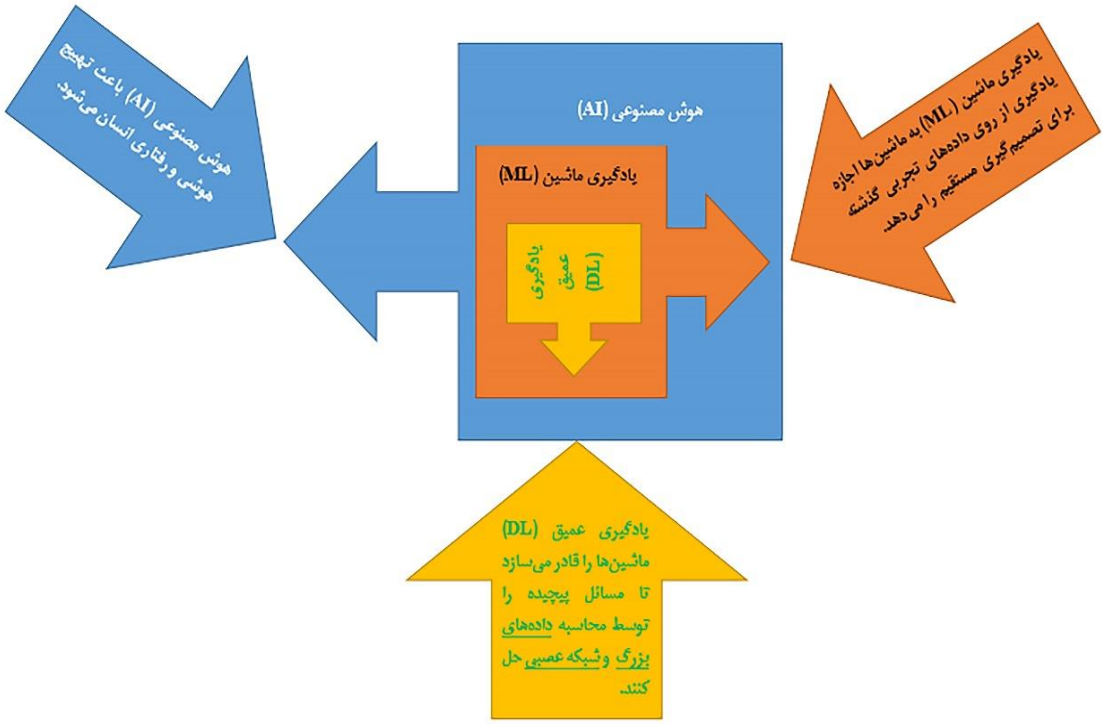
روش‌های DL در سال‌های اخیر دیدگاه‌های جالب و هیجان‌انگیزی را در حوزه‌های اصلی تحقیق (به عنوان مثال، تجزیه و تحلیل تصویر، تجزیه و تحلیل زبان و همچنین علوم اومیکس) باز کرده‌اند (Marx 2013; LeCun et al. 2015)، که دارای مزایای مهم زیادی نسبت به تکنیک‌های سنتی ML، مانند تجزیه و تحلیل مؤلفه‌های اصلی[1] (PCA) (Kramer 1991)، روش‌های بیزی[2] (BMs) (Weiss 2010)، ماشین‌های بردار پشتیبان[3] (SVMs) (Cortes and Vapnik 1995)، جنگل‌های تصادفی[4] (RF) و درخت‌های تصمیم‌گیری[5] (DTs) (Rokach and Maimon 2014) است. مزیت اصلی DL نسبت به روش‌های ML یادگیری سرتاسری است، یعنی امکان به دست آوردن نتایج طبقه بندی یا پیش بینی مستقیماً از داده‌های خام. در حالی که فرآیند را از منابع احتمالی سوگیری (به عنوان مثال، انتخاب داده‌های ورودی برای مرحله آموزش شبکه) نجات نمی‌دهد، یادگیری انتها به انتها از اجتناب از سوگیری (اریبی) بالقوه معرفی شده توسط مداخله دستی در مراحل مختلف پردازش داده سود می‌برد. همچنین، روش‌های DLادغام انواع داده‌های ورودی مختلف (متن، عددی، تصاویر، فایل‌های صوتی) را آسان می‌کنند. در نهایت، معماری‌های DL در مقایسه با تکنیک‌های سنتی ML قابلیت انتزاع[6] بسیار بالاتری دارند.

معماری‌های DL مدرن، مانند شبکه‌های عصبی عمیق[7] (DNNs) (Schmidhuber 2015)، شبکه‌های باور عمیق[8] (DBNs) (Hinton et al. 2006)، شبکه‌های عصبی تکراری[9] (RNNs) (Williams and Zipser 1989)، ماشین‌های عمیق بولتزمن[10] (DBMs) (Salakhutdinov and Hinton 2009)، شبکه‌های عصبی پیچیده یا کانولوشن[11] (CNN) (LeCun et al. 1998)، رمزگذارهای خودکار[12] (AEs) (Hinton and Salakhutdinov 2006; Vincent et al. 2010) و شبکه‌های متخاصم مولد[13] (GANs) (Goodfellow et al. 2014)، هم از نظر کارایی و هم عملکرد، از پرسپترون روزنبلات فاصله زیادی گرفته اند. با این حال، پیشرفت به قیمت کاهش شفافیت و از دست دادن توانایی ردیابی فرآیندهای استخراج و طبقه‌بندی ویژگی انجمنی[14] بود. عدم توضیح‌پذیری (قابلیت توضیح) از افزایش پیچیدگی معماری ناشی می‌شود که از تک لایه‌ی نورون‌های پرسپترون

[1] Principal Component Analysis
[2] Bayesian Methods
[3] Support Vector Machines
[4] Random Forests
[5] Decision Trees
[6] Abstraction
[7] Deep Neural Networks
[8] Deep Belief Networks
[9] Recurrent Neural Networks
[10] Deep Boltzmann Machines
[11] Convolutional Neural Networks
[12] AutoEncoders
[13] Generative Adversarial Networks
[14] Associative

به لایه‌های بسیاری از نورون‌های مخفی که بین لایه‌های ورودی و خروجی مدل‌های پیشرفته DL مداخله می‌کنند، می‌رود. توجه داشته باشید که از دست دادن قابلیت توضیح مستلزم خطرات جدیدی برای به دست آوردن نتایج مغرضانه مختلف است و بنابراین، در حال حاضر یکی از فعال‌ترین حوزه‌های تحقیقاتی در هوش مصنوعی است. توضیح‌پذیری در واقع یک موضوع اصلی برای بهره‌برداری از پتانسیل DL است، به‌ویژه در حوزه‌های تحقیقات زیست‌پزشکی و مراقبت‌های بهداشتی، که در آن ویژگی‌هایی که توسط سیستم یادگیری برای تصمیم‌گیری خروجی انتخاب می‌شوند باید از نظر انسانی قابل درک باشند. در واقع، توانایی معماری‌های DLبرای استخراج ویژگی‌های بسیار دقیق‌تر از استنتاج بصری و استنتاج ارتباط‌های مبتنی بر سطوح انتزاع بسیار بالا، استراتژی‌های تحقیقی جدید را تسهیل می‌کند. با این حال، به دلیل منطق رمزآلود حمایت از تصمیمات ماشین، که به عنوان یک جعبه سیاه مانع ارزیابی فرآیند و پاکسازی منابع احتمالی خطاها یا سوگیری‌ها می‌شود، مسائل اخلاقی و قانونی عمده‌ای را نیز مطرح می‌کند. در نهایت، معماری‌های DL پیچیده‌تر و پیچیده‌تر در معرض افزایش پیچیدگی آموزشی قرار می‌گیرند، که این هم به دلیل تعداد زیاد پارامترهای پیکربندی مدل (به عنوان مثال، وزن – یا سهم در پیش‌بینی – هر گره در یک شبکه عصبی مصنوعی) است که باید از داده‌های آموزشی تخمین زده شوند. علاوه بر این، معماری‌های DL به تنظیم دقیق و طولانی پارامترهای پیکربندی (مثلاً نرخ یادگیری برای آموزش شبکه عصبی) نیاز دارند که خارج از مدل هستند و ارزش آن‌ها را نمی‌توان از روی داده‌ها تخمین زد، اما می‌تواند به شدت بر سرعت آموزش و عملکرد مدل تأثیر بگذارد. درمجموع، این ملاحظات DL را به ابزاری قدرتمند تبدیل می‌کند که باید با احتیاط از آن استفاده کرد (Angermueller et al. 2016; Li et al. 2019; Mahmud et al. 2021). اصول AI، ML و DL نشان می‌دهد که چگونه حجم زیادی از داده‌ها پردازش و به یک سیستم هوش مصنوعی تبدیل می‌شوند. داده‌ها از طریق نرم افزارهای کامپیوتری در قالب یک شبکه عصبی پردازش می‌شوند و سپس ماشین‌ها بر اساس خروجی به دست آمده از شبکه عصبی به صورت هوشمند کار می‌کنند. سیستم کامل هوش مصنوعی یاد می‌گیرد که مانند مغز انسان رفتار کند و بر اساس آن به ورودی‌ها پاسخ دهد و کل سیستم را به یک هوش مصنوعی تبدیل کند (شکل ۲).

اهداف این مطالعه عبارتند از بررسی کاربردهای اصلی روش‌های هوش مصنوعی در ژنومیکس عملکردی، سرطان، کشاورزی، حیوانات اهلی و زمینه‌های درهم تنیده آن یعنی اپی‌ژنومیکس، ترانس‌کریپتومیکس، اپی‌ترانس‌کریپتومیکس، پروتئومیکس و متابولومیکس، با تمرکز بر روی کاربردهای سال‌های اخیر پس از افزایش تولید کلان داده در ژنومیکس و تلاقی طبیعی این رشته با حوزه شکوفایی هوش مصنوعی، بحث در مورد جنبه‌های مهم مدیریت داده‌ها، مانند یکپارچه‌سازی داده‌ها، جانهی، تمیز کردن، حذف نویز، متعادل‌سازی و نسبت داده‌های از دست رفته، مدل‌سازی سیستم–ژنومیکس عملکردی، هوش مصنوعی و سیستم‌های بیولوژی، پرداختن به مسائل حقوقی، اخلاقی و اقتصادی مربوط به کاربرد روش‌های هوش مصنوعی در حوزه ژنومیکس و ارائه نمایی از سناریوهای احتمالی آینده است.



شکل ۲. همبستگی هوش مصنوعی، یادگیری ماشین و یادگیری عمیق

Figure 2. Correlation of artificial intelligence, machine learning and deep learning

ژنومیکس عملکردی

ژنومیکس عملکردی علمی است که در مقیاس ژنومی، روابط بین اجزای یک سیستم بیولوژیکی–ژن‌ها، رونوشت‌ها، پروتئین‌ها، متابولیت‌ها و غیره– و نحوه کار این اجزا برای تولید یک فنوتیپ معین را مطالعه می‌کند. منشا اصطلاح "ژنومیکس عملکردی" به زمان ظهور اولین پروژه‌های توالی‌یابی ژنوم در جامعه علمی برمی‌گردد. این پروژه‌ها در نهایت با هدف تعیین توالی ژنوم کامل یک ارگانیسم مشخص و حاشیه‌نویسی ویژگی‌های مرتبط عملکردی در آن، مانند ژن‌های کدکننده پروتئین و غیر کدکننده و همچنین مناطق تنظیم‌کننده DNA، انجام می‌شوند. نقطه عطف چنین تلاشی پروژه ژنوم انسانی[1] (HGP) است، که یک پروژه مشترک جهانی بود که در سال ۱۹۹۰ راه اندازی شد و به طور رسمی در سال ۲۰۰۳ تکمیل شد (کنسرسیوم بین المللی توالی ژنوم انسانی[2] (Consortium IHGS 2004). با این حال، اولین ژنوم کاملاً توالی‌یابی شده از یوکاریوت‌ها، ژن مخمر جوانه‌دار ساکارومایسس سرویزیه[3] بود که در سال ۱۹۹۶ منتشر شد و اصولی را برای شروع کاوش در روابط پیچیده بین ژن‌ها و محصولات ژنی در مقیاس

[1] Human Genome Project
[2] International Human Genome Sequencing Consortium
3 Saccharomyces cerevisiae

ژنوم فراهم کرد (Goffeau et al. 1996). در واقع، یک تعریف آزمایشی از ژنومیکس عملکردی برای اولین بار در سال ۱۹۹۷ توسط هیتر و بوگوسکی[1] منتشر شد، که در ابتدای مقاله آن‌ها عبارت «یک نظرسنجی غیررسمی از همکاران نشان می‌دهد که اصطلاح "ژنومیکس عملکردی" به طور گسترده استفاده می‌شود، اما تعابیر مختلفی دارد (Hieter and Boguski 1997). حتی برخی احساس می‌کنند که این اصطلاح غیرضروری است و چیزی جز اشاره به تحقیقات بیولوژیکی به عنوان یک اصطلاح کلی نیست». با این حال، در همان مقاله، آن‌ها همچنین بیان می‌کنند که «مفهوم ژنومیک عملکردی در علم وارد شده است و باعث ایجاد ایده‌ها و رویکردهای جدید برای درک مکانیسم‌های بیولوژیکی در زمینه دانش ساختار کل ژنوم می‌شود». ژنومیکس عملکردی در نهایت توسط این نویسندگان به عنوان "مرحله جدید تجزیه و تحلیل ژنوم"، پس از پایان مرحله "ژنومیکس ساختاری" (یعنی ایجاد یک نقشه فیزیکی و توالی ژنوم) تعریف شد. این "مرحله جدید" شامل توسعه و بکارگیری رویکردهای تجربی و تکنیک‌های محاسباتی در سطح ژنوم برای استنتاج عملکردهای ژن بود. پیشرفت‌های چشمگیر از آغاز قرن حاضر در فناوری‌های توالی‌یابی موازی و پروتکل‌های مرتبط، چهره ژنومیک عملکردی را تغییر داده است. امروزه، می‌توانیم ادعا کنیم که این اصطلاح دیگر برای تفاسیر متفاوت باز نیست و به رشته‌ای اشاره دارد که طیف گسترده‌ای از داده‌های "اومیکس" را ادغام می‌کند و بر تعداد زیادی روش‌شناسی تجربی و رویکردهای محاسباتی برای درک رفتار سیستم‌های بیولوژیکی تکیه می‌کند. منظور از سیستم یک سلول، بافت یا کل ارگانیسم در شرایط سالم یا پاتولوژیک است. به طور خاص، داده‌های مورد استفاده در تحلیل‌های ژنومیک عملکردی در شرایط خاص و با فن‌آوری‌های رشته‌های اومیکس، از جمله ژنومیکس، اپی ژنومیکس، ترانس‌کریپتومیکس، اپی ترانس‌کریپتومیکس، پرواتئومیکس و متابولومیکس تولید می‌شوند.

کاربردهای هوش مصنوعی در ژنومیکس عملکردی

در دهه‌های گذشته، یادگیری ماشین[2] (ML) به طور گسترده در بسیاری از حوزه‌های علوم «اومیکس»، به‌ویژه آن‌هایی که با تولید مقادیر زیادی داده و/یا مکانیسم‌های پیچیده‌ای که توسط مشارکت هم‌افزایی عوامل مختلف مشخص می‌شوند، استفاده شده است. کاربردهای مهم هوش مصنوعی شامل: پیش بینی مناطق تنظیم کننده DNA، کشف مورفولوژی سلولی و سازمان فضایی، شناسایی ارتباط بین فنوتیپ‌ها و ژنوتیپ‌ها، طبقه بندی متیلاسیون DNA و تغییرات هیستون، کشف نشانگرهای زیستی، تشخیص تقویت کننده‌های رونویسی، تشخیص سرطان و تجزیه و تحلیل مکانیسم‌های تکاملی است (Min et al. 2016; Ravi et al. 2017; Cao et al. 2018; Ching et al. 2018; Miotto et al. 2018; Wainberg et al. 2018; Yue and Wang 2018). اولین تلاش‌ها برای اعمال تکنیک‌های آموزشی تحت نظارت در علوم اومیکس از دهه ۱۹۸۰ آغاز شده است. در سال ۱۹۸۲، Stormo et al. از الگوریتم پرسپترون برای تشخیص مکان‌های شروع ترجمه *E. coli* از تمام مکان‌های دیگر در

1 Hieter and Boguski
2 Machine Learning

کتابخانه‌ای با بیش از ۷۸۰۰۰ نوکلئوتید از توالی mRNA استفاده کرد. در سال ۱۹۹۳، Rost and Sander یک شبکه عصبی را برای پیش بینی ساختار ثانویه پروتئین اجرا کردند. تکنیک‌های DL به دلیل بهبود عملکرد رایانه شخصی و کاهش هزینه‌های توالی یابی ژنوم، تنها در دهه دوم دهه ۲۰۰۰ به طور گسترده در ژنومیک عملکردی مورد استفاده قرار گرفتند (Rost and Sander 1993; Eickholt et al. 2012; Asgari et al. 2015; Spencer et al. 2015).

در سال ۲۰۱۵، دو معماری عمیق مهم برای ژنومیک عملکردی پیاده‌سازی و اجرا شده‌اند و نتایج بسیار مفیدی را تولید کرده‌اند. یک نرم افزار مستقل کاملاً خودکار به نام DeepBind برای پیش بینی ویژگی‌های توالی پروتئین‌های اتصال دهنده DNAو RNA است (Alipanahi et al. 2015). نرم افزار دیگر تحلیلگر توالی مبتنی بر یادگیری عمیق[1] (DeepSEA) است که اثرات کروماتینی تغییرات توالی با وضوح تک نوکلئوتیدی را با یادگیری توالی‌های تنظیمی از روی داده‌های نمایه کروماتین در مقیاس بزرگ پیش بینی می‌کند (Zhou and Troyanskaya 2015). هر دو روش مبتنی بر معماری‌های عمیق، بر چالش‌های زیادی مانند پردازش میلیون‌ها دنباله، تعمیم داده‌ها از فناوری‌های مختلف، تحمل نویز و داده‌های از دست رفته و یادگیری سرتاسر و کاملاً خودکار، بدون نیاز به تنظیم دستی این رویکردها از سایر روش‌های پیشرفته برتری داشت و بسیاری از دانشمندان را تشویق کرد که مسیرهای هیجان انگیز مشابهی را دنبال کنند.

ژنومیکس: در عصر کنونی تحقیقات بیولوژیکی، با پیشرفت تکنولوژی در توالی‌یابی و کشف پیچیدگی DNA، این مفهوم به کل مجموعه توالی‌های DNA در یک سلول یا ارگانیسم (به عنوان مثال، تعداد نسخه‌های مجموعه اصلی کروموزوم‌ها یا پلوئیدی، و همچنین مواد DNA موجود در اندامک‌های خارج هسته‌ای مانند میتوکندری[2] و کلروپلاست[3]) بسط داده شده است. ژنومیک را می‌توان به عنوان "علم ژنوم" تعریف کرد. این اصطلاح در سال ۱۹۸۶ توسط Thomas Roderick برای توصیف رشته نوپای توالی یابی، نقشه برداری، حاشیه‌نویسی[4] و تجزیه و تحلیل ژنوم‌ها ابداع شد (Hieter and Boguski 1997). اولین توالی ژنوم کامل یک اندامک یوکاریوتی (میتوکندری انسانی به طول ۱۶٫۶ کیلوبایت) در سال ۱۹۸۱ تعیین توالی شد (Anderson et al. 1981)، اولین موجود زنده آزاد (*H*aemophilus *influenzae* به طول ۱/۸ کیلوبایت) در سال ۱۹۹۵ توالی‌یابی شد (Fleischmann et al. 1995) و اولین ژنوم یوکاریوتی (*S. cerevisiae* به طول ۱۲/۱ کیلوبایت) در سال ۱۹۹۶ تکمیل شد (Goffeau et al. 1996). با شروع تعیین توالی ژنوم انسان با استفاده از روش‌های اولیه کم توان (نسل اول فن آوری‌های توالی‌یابی) (Sanger et al. 1977; Sanger et al. 1997; Heather and Chain 2016)، در چند دهه این زمینه برای اولین بار با موازی سازی زیاد واکنش‌های توالی یابی متحول شد و به تولید میلیون‌ها قرائت کوتاه در چند ساعت اجرا رسید (نسل

[1] Deep learning-based sequence analyser
[2] mtDNA
[3] cpDNA
4 Annotating

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

دوم) (Margulies et al. 2005; Bentley et al. 2008). اخیراً تعیین توالی را بیشتر به سمت توالی‌یابی تک مولکولی و خواندن توالی بسیار طولانی (نسل سوم) تکامل داده‌اند (Braslavsky et al. 2003; Haque et al. 2013; Bleidorn 2016). امروزه، به لطف ظهور تکنیک‌های توالی یابی با توان عملیاتی بالا، صدها هزار ژنوم از سلسله‌های مختلف به طور کامل توالی یابی شده‌اند. از جمله، بیش از ۱۵۰۰۰ ژنوم یوکاریوتی (GENOME NCBI) و اصطلاح ژنومیکس اکنون گسترش یافته است تا بررسی، ساختار، عملکرد، تکامل و ویرایش DNA را نیز در برگیرد.

همانطور که توسط Libbrecht and Stafford Noble (2015) اشاره شده است، ML به طور گسترده‌ای در ژنومیکس برای حاشیه نویسی عناصر توالی، شناسایی مکان‌های اتصال، یافتن محرک‌ها و تقویت کننده‌ها و غیره استفاده شده است. تعداد زیادی از توالی‌های ژنوم برای آموزش مدل‌های ML برای تشخیص عناصر عملکردی خاص استفاده شده است. مقاله مهمی توسط Bucher (1990) منتشر شد که در آن یک الگوریتم ماتریس وزن بهینه برای صدها توالی پروموتر غیرمرتبط برای شناسایی عناصر پروموتور استفاده شده است. شایان ذکر است، این کاربرد ML، مانند سایر کاربردها در سال‌های بعد، با ایجاد پایگاه‌های اطلاعاتی مانند پایگاه داده‌های پروموتر یوکاریوتی[1] (EPD) یا آرشیو نوکلئوتیدی اروپا[2] (ENA) ممکن شده است. روش‌های پیش‌بینی SVM[3] و NB[4] (Degroeve et al. 2002) برای پیش‌بینی سایت اسپلایس[5] (محل اتصال) استفاده شده‌اند و نسبت به روش‌های سنتی انتخاب ویژگی‌های مرتبط، پیشرفت‌ها و مزایایی را نشان می‌دهند. در پژوهشی Segal et al. (2006) یک رویکرد تجربی و محاسباتی ترکیبی مهم برای بررسی سازمان نوکلئوزومی پیشنهاد کردند. در روش پیشنهادی، توالی‌های متصل به نوکلئوزوم از مخمر با وضوح بالا جدا شدند و برای ساخت یک مدل احتمالی برهمکنش نوکلئوزوم–DNA برای پیوند موقعیت‌های نوکلئوزوم به عملکردهای خاص کروموزوم و پیش‌بینی سازماندهی ژنومی نوکلئوزوم‌ها استفاده شدند. در تحقیقی Heintzman et al. (2007) پنج تغییر هیستون و چهار فاکتور رونویسی را بر روی ۳۰ مگابایت از ژنوم انسان با استفاده از رویکرد خوشه‌بندی ML ترسیم کردند. در پژوهشی Hoffman et al. (2012) از یک روش BN دینامیک بدون نظارت[6] برای تجزیه و تحلیل انواع مختلف داده‌های اومیکس (مانند علائم اصلاح هیستون و مکان‌های اتصال برای اصلاح‌کننده‌های ساختار کروماتین)، که همگی از یک رده سلولی لوسمی میلوئیدی مزمن انسانی[7] مشتق شده‌اند، برای تجزیه و تحلیل کل ژنوم، با وضوح ۱ جفت باز، علیرغم وجود نویز و داده‌های از دست رفته استفاده کردند. روش‌های DL تنها در سال‌های اخیر در زمینه ژنومی مورد استفاده قرار گرفته‌اند. یک پکیج منبع باز (open-source) مبتنی بر CNN به نام Basset[8] برای حاشیه نویسی و تفسیر ژنوم غیر کدکننده در سال ۲۰۱۶ پیشنهاد شد

[1] Eukaryotic Promoter Database
[2] European Nucleotide Archive
3 Support Vector Machines
4 Naïve Bayes
5 splice site
6 Unsupervised Dynamic BN method
7 Human chronic myeloid leukemia
[8] Open-source package based on CNNs named Basset

(Kelley et al. 2016). سال بعد، Killoran et al. (2017) یک GAN (شبکه متخاصم مولد[1]) را برای تولید توالی‌های DNAبا ویژگی‌های خاص پیشنهاد کردند. اخیراً، Avsec et al. (2021) یک رویکرد DL را برای کشف تأثیر فاصله موتیف[2] بین محل‌های اتصال فاکتور رونویسی همسایه[3] بر همکاری فاکتور رونویسی[4] معرفی کرده‌اند.

رویکردهای بدون نظارت، مانند شبکه‌های متخاصم مولد (GANها) و رمزگذارهای خودکار (AEها) توانایی زیادی برای استخراج ویژگی‌های بسیار معرف (نماینده) و یادگیری نمایش‌های پیچیده از داده‌های ورودی بدون هیچ نوع نظارت و آدرس‌دهی دارند. علاوه بر این، آن‌ها می‌توانند بدون از دست دادن اطلاعات، ابعاد را به طور موثری حذف و کاهش دهند. در سال‌های اخیر، مدل‌های نمایشی مورد استفاده برای پردازش زبان طبیعی[5] (NLP) برای پردازش داده‌های توالی بیولوژیکی استفاده شده است (Wu et al. 2020; Song et al. 2021). به یک معنا توالی‌های زیستی را می‌توان به عنوان جملات یک زبان در نظر گرفت. یک روش پرکاربرد در NLP روش LSTM است (Hochreiter and Schmidhuber 1997) که بر اساس معماری RNN (شبکه‌های عصبی تکراری[6]) است که برای استخراج اطلاعات معنایی و متنی از توالی‌های طولانی مناسب است. در پژوهشی Mikolov et al. (2013) روش Word2Vec را پیشنهاد دادند که یک روش جاسازی کلمه بدون نظارت برای انجام نمایش برداری کم بعدی از کلمات زبان طبیعی است. این روش می‌تواند زمینه یک کلمه را در یک سند به تصویر بکشد، روابط بین کلمات را خط بکشد و شباهت‌های معنایی و نحوی را به تصویر بکشد. در تحقیقی دیگر Vaswani et al. (2017) ترانسفورمر (تبدیل کننده) را پیشنهاد دادند، که یک معماری جدید مبتنی بر مکانیسم توجه است. ترانسفورمرها برای مدیریت داده‌های متوالی مانند LSTM و RNN (شبکه‌های عصبی تکراری[7]) طراحی شده‌اند و از این نظر برای ترجمه و تفسیر متن مناسب هستند. با این حال، ترانسفورمرها نه از تکرار و نه ورودی‌های پردازش به ترتیب خود استفاده می‌کنند. ترانسفورمرها از یک مقدار دهی اولیه تصادفی استفاده می‌کنند و بر اساس جاسازی کلمات پویا هستند (برخلاف سایر معماری‌های NLP که از جاسازی کلمه ایستا استفاده می‌کنند). سپس Devlin (2018) یک روش NLP جدید به نام BERT (نمایش رمزگذار دوطرفه از ترانسفورمرها[8]) را معرفی کردند که در آن نویسندگان آموزش دوطرفه ترانسفورمر را به کار بردند، که در گرفتن معنی معنایی[9] و کلمات متنی بسیار کارآمد بود. مقالات بسیاری به‌تازگی کاربردهای جالب روش‌های جاسازی کلمه بدون نظارت را در توالی‌های بیولوژیکی گزارش کرده‌اند. در پژوهشی Woloszynek et al. (2019) از Word2Vec برای جاسازی توالی‌های نوکلئوتیدی، به‌ویژه k-mers به‌دست‌آمده از

[1] Generative Adversarial Network
[2] Motif spacing
[3] Neighbor transcription-factor binding sites
[4] Transcription factor cooperativity
[5] Natural Language Processing
[6] Recurrent Neural Networks
[7] Recurrent Neural Networks
[8] Bidirectional Encoder Representations from Transformers
[9] Semantic meaning

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

بررسی‌های آمپلیکون 16S rRNA استفاده کرد و موفق شد ویژگی‌های مرتبط با بافت توالی، تاکسونومی و طبقه‌بندی را استخراج کند. در پژوهشی Ostrovsky-Berman et al. (2021) از روش Immune2vect، که اقتباسی از Word2Vec است برای داده‌های توالی‌یابی گیرنده‌های سلول B را ارائه کرد، که در آن داده‌های توالی‌یابی ایمنی را در بازنمایی‌های ناقل کم‌بعد[1] برای استخراج ویژگی‌های مرتبط مانند خواص n گرم و طبقه‌بندی ژن‌های متغیر زنجیره سنگین ایمونوگلوبولین[2] (IGHV) جاسازی کردند. اخیراً Le et al. (2021) یک تکنیک جدید متشکل از BERT و CNN برای پیش‌بینی تقویت‌کننده[3] DNA ارائه کرد. معلوم شد که این رویکرد نسبت به Word2Vec در گرفتن اطلاعات پنهان در توالی‌های DNA کارآمدتر است زیرا جاسازی کلمه ایجاد شده با BERT پویا است و نوکلئوتیدها را می‌توان در موقعیت‌های مختلف نشان داد و مقادیر برداری متفاوتی را در نظر گرفت. این یک مزیت نسبت به جاسازی کلمات ایستا است، که در آن بردارهای یکسان برای کلمات یکسان بدون توجه به بافت آن‌ها به دست می‌آید، زیرا نمایش‌های دقیق و با جزئیات بیشتری ارائه می‌دهد.

هدف اصلی در ژنومیکس، شناسایی واریانت‌های ژنتیکی است که ویژگی‌های انسانی، به‌ویژه بیماری‌ها را تشکیل می‌دهند. فن‌آوری‌های توالی‌یابی با توان عملیاتی بالا[4] (HTS) توانایی ما را برای شناسایی جهش‌های ژنی مسئول اختلالات انسانی که ناشی از تغییرات تأثیرات بزرگ در یک ژن واحد است (به عنوان مثال، بیماری‌هانتینگتون، دیستروفی عضلانی دوشن) بسیار تسریع کرد. علاوه بر این، هزاران مطالعه مرتبط با ژنوم گسترده[5] (GWAS) فهرست طولانی‌ای از واریانت‌های ژنتیکی مرتبط با بیماری‌های رایج (مانند آسم، دیابت، بیماری‌های قلبی) را تولید کرده‌اند که اغلب به دلیل مشارکت ضعیف ژن‌های متعدد و عوامل محیطی است. با این وجود، درک ما از عوامل ژنتیکی این بیماری‌های پیچیده هنوز محدود است. این محدودیت تا حدودی به دلیل بازخوانی‌های فنوتیپی غیرمنتظره است که از برهمکنش‌های عملکردی بین دو یا چند ژن نشات می‌گیرد، مانند جهش ژنتیکی که حضور آن می‌تواند اثرات یک آلل را در لوکوس دیگری (معروف به اپیستازی) بپوشاند. غربالگری‌های ژنتیکی سیستماتیک که در ارگانیسم‌های مدل انجام می‌شوند، درک بهتری از تعامل بین ژنوتیپ و فنوتیپ را تقویت کرده‌اند و چارچوبی را برای توسعه ژنتیک شخصی‌شده در انسان با نقشه‌برداری فنوتیپ‌ها بین ارگانیسم‌ها فراهم می‌کنند (Lehner 2013). جامع‌ترین تجزیه و تحلیل‌ها در مخمر جوانه‌زن ساکارومایسس سرویزیه[6] انجام شده است (Costanzo et al. 2010 and 2016) که منجر به اندازه گیری‌های کمی فنوتیپی برای ده‌ها میلیون جفت جهش در مخمر شده است. این تلاش‌های گسترده غربالگری همراه با روش‌های ML و DL برای حوزه‌های متعدد مانند پیش‌بینی خودکار تأثیر رشد فعل و انفعالات ژنتیکی منتخب در شبکه متابولیک مخمر، بر اساس رگرسیون و الگوریتم ژنتیک کابرد داشته است که موجب افزایش دقت پیش‌بینی (Szappanos et al. 2011)، ارتباط اثرات متقابل ژنتیکی با تاثیر

[1] Low-dimensional vector representations
[2] Immunoglobulin heavy-chain variable
[3] Enhancer
[4] High-Throughput Sequencing
[5] genome-wide association studies
[6] Budding yeast *Saccharomyces cerevisiae*

عملکردی، بر اساس رگرسیون RF (Yu et al. 2016) و ساخت یک NN قابل تفسیر یا "مرئی"[1]، به نام DCell، که رشد سلول یوکاریوتی پایه را شبیه سازی می‌کند (Ma et al. 2018) می‌گردد و پاسخ به آشفتگی ژنتیکی را از نظر تناسب سلولی پیش بینی می‌کند.

ظهور فن‌آوری‌های قدرتمند ویرایش ژنوم، مانند CRISPR-Cas9 (تکرارهای کوتاه پالیندرومیک خوشه‌ای با فاصله منظم–پروتئین مرتبط با CRISPR 9[2]) امکان دستکاری مقیاس‌پذیر DNA را برای مشخص کردن عملکرد ژن‌ها و عناصر تنظیم‌کننده ژن در تعدادی از ارگانیسم‌های مختلف و در انسان فراهم کرده است. یک نکته کلیدی برای کاربرد موفقیت آمیز CRISPR-Cas9، طراحی مناسب RNAهای کوتاه[3] (که به طور کلی به عنوان gRNA، مخففRNAها نامیده می‌شود[4]) است که داربستی[5] را برای مجموعه آنزیمی فراهم می‌کند و مجموعه آنزیمی را به سمت سایت‌های هدف برای ویرایش، بر اساس مکمل بودن توالی ۱۷–۲۰ نوکلئوتیدی در انتهای ´۵ رشته gRNA هدایت می‌کند. به طور خاص، در فرآیند طراحی gRNA، بهینه سازی توالی مهندسی شده به سمت اثرمتقابل خاص با هدف ویرایش (فعالیت روی هدف) و در عین حال به حداقل رساندن اثرات متقابل ناخواسته با سایر سایت‌های ژنومی (فعالیت خارج از هدف)، که ممکن است از تشابه توالی با هدف واقعی ناشی شود، بسیار مهم است. روش‌های مختلف ML و روش‌های DL برای بهینه‌سازی طراحی gRNA و پیش‌بینی فعالیت روی هدف و خارج از هدف توسعه یافته‌اند. از جمله: CRISTA (Abadi et al. 2017) یک مدل رگرسیون مبتنی بر RF که تمایل یک سایت ژنومی را برای شکافتن توسط یک gRNA معین ارزیابی می‌کند، DeepCRISPR (Chuai et al. 2018)، یک پلت فرم محاسباتی که از تکنیک تقویت داده برای گسترش مجموعه داده آموزشی توالی‌های gRNA معتبر آزمایشی استفاده می‌کند و دو CNN (یکی برای پیش بینی فعالیت روی هدف و دیگری برای پیش بینی فعالیت خارج از هدف)، با بازنمایی‌های gRNA تولید شده توسط رمزگذارهای خودکار از پیش آموزش دیده تغذیه می‌کند، CROTON (Li et al. 2021)، یک چارچوب سرتاسر مبتنی برCNNهای چند وظیفه‌ای عمیق و جستجوی معماری عصبی برای پیش بینی نتایج ویرایش CRISPR-Cas9 و ابزارهای مکمل CRISPR-ONT و CRISPR-OFFT (Zhang et al. 2021)، که CNNهای مبتنی بر توجه هستند که به ترتیب برای پیش بینی فعالیت‌های gRNA روی و خارج از هدف آموزش دیده اند.

ترکیب اغتشاش ژنی کارآمد[6] ارائه شده توسط فناوری CRISPR-Cas9 با فنوتیپ رونویسی چندگانه ارائه شده توسط توالی یابی RNA تک سلولی[7] (scRNA-seq) فرصتی بی سابقه برای کشف فعل و انفعالات ژنتیکی در سلول‌های پستانداران در مقیاس

[1] Interpretable or ‘visible’ NN
[2] Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR-associated protein 9
[3] Short RNAs
[4] gRNAs, acronym to guide RNAs
[5] Scaffold
[6] Efficient gene perturbation
[7] Single-cell RNA sequencing

بزرگ ارائه می‌دهد. این چارچوب تجربی اخیراً توسط Norman et al. (2019) بررسی شده است، که سیستم توصیه‌کننده ML را برای کاهش ابعاد نقشه با ابعاد بالا از حالات رونویسی (فنوتیپ‌ها) مرتبط با اغتشاش ژن اعمال کرد تا امکان تجزیه و تحلیل بصری و پیش‌بینی اثرات متقابل ژنتیکی را فراهم کند. چندین گروه رویکردهای ML و DL را هم برای شناسایی تعاملات ژنتیکی مرتبط با بیماری و هم برای پیش‌بینی خطر ژنتیکی بیماری‌های پیچیده در جمعیت‌ها از روی نقشه‌های ژنومی تنوع ژنتیکی، مانند وقوع پلی‌مورفیسم‌های تک‌نوکلئوتیدی[1] (SNPs) یا درج‌ها یا حذف‌های نوکلئوتیدی کوچک[2] (indels) در ژنوم انسان مورد بررسی قرار داده‌اند. در پژوهشی Kircher et al. (2014) تخلیه وابسته به حاشیه نویسی ترکیبی[3] (CADD) را پیشنهاد دادند که یک رویکرد SVM برای طبقه بندی واریانت‌های عملکردی، مضر و بیماری‌زا است، که با میلیون‌ها آلل مشتق شده از انسان با فرکانس بالا و واریانت‌های شبیه سازی شده[4] آموزش داده شد. این روش در تشخیص واریانت‌های بیماری زا که زمینه ساز بیماری‌ها هستند از واریانت‌های خوش خیم مجاور، از روش‌های موجود بهتر عمل کرد. در پژوهشی Quang et al. (2016) DANN را پیاده‌سازی کردند. DANN روشی برای حاشیه‌نویسی بیماری‌زایی واریانت‌های ژنتیکی است که با استفاده از مجموعه ویژگی‌ها و داده‌های آموزشی مشابه با CADD، اما مبتنی بر یک DNN، مناسب‌تر از SVM برای ثبت روابط غیرخطی بین صفات است. در همان سال، Ionita-Laza et al. (2016) یک روش جایگزین مبتنی بر رویکرد طیفی بدون نظارت (Eigen) پیشنهاد کرد که واریانت‌های ژنتیکی را برای ارتباط بیماری امتیاز می‌دهد. روش ExPecto توسط Zhou et al. (2018) پیشنهاد شد، که یک چارچوب end-to-end مبتنی برCNN، که بر روی داده‌های اومیکس متعدد به‌دست‌آمده از ۲۰۰ بافت و نوع سلول انسانی آموزش داده شد تا اثرات نوع سلولی تنوع توالی ژنتیکی بر بیان ژن و خطر بیماری را پیش‌بینی کند. در نهایت، با توجه به تجزیه و تحلیل داده‌های توالی‌یابی خام برای شناسایی وجود تنوع ژنتیکی، در سال ۲۰۱۸ تیم ژنومیک در Google Brain یک معماری یادگیری عمیق به نام DeepVariant را بر اساس CNN منتشر کرد که برای فراخوانی SNPها آموزش دیده بود و انواع ایندل را از انبوهی از خوانش‌های توالی‌یابی هم‌تراز می‌کند (Poplin et al. 2018). این روش برنده بالاترین جایزه عملکرد برایSNP ها در گونه تحت حمایت سازمان غذا و داروی ایالات متحده با نام چالش حقیقت شد.

ژنومیک سرطان: در دهه‌های گذشته، ظهور تکنیک‌های NGS رویکرد پزشکی به سرطان را متحول کرده است (Berger and Mardis 2018). ژنومیک در مطالعات بالینی، پیشگیری، درمان و اقدامات نظارتی اهمیت فزاینده‌ای پیدا کرده است. ژنومیکس سرطان تفاوت‌ها در توالی‌های DNA و بیان ژن بین سلول‌های تومور و طبیعی را با هدف درک پویایی‌های تشکیل و گسترش تومورها در سطوح ژنتیکی، متابولیکی، سیستمیک و محیطی مورد مطالعه قرار می‌دهد.

[1] Singlenucleotide polymorphisms

2 Small nucleotide insertions or deletions

[3] Combined Annotation-Dependent Depletion

[4] Simulated variants

پروژه اطلس ژنوم سرطان داده‌های NGS چند سطحی را برای ۳۳ نوع مختلف تومور رایج جمع‌آوری کرد. این داده‌ها یک منبع عظیم برای مطالعه مکانیسم‌های سرطان خاص و همچنین برگشت‌پذیر[1] در دسترس قرار دادند (TCGA 2020). در دسترس بودن و ادغام مقادیر زیادی از اطلاعات ژنومیک، پروتئومیک و اپی ژنومیک اجازه می‌دهد تا بازنمایی‌های جامع و فزاینده ای از پویایی‌های پیچیده، مانند تشکیل سرطان به دست آید (Sánchez-Vega et al. 2018). در واقع، ادغام داده‌های اومیکس متعدد می‌تواند به غلبه بر نویز احتمالی و/یا سوگیری لایه‌های داده منفرد کمک کند، بنابراین ارتباط ویژگی‌های نماینده استخراج‌شده را بهبود می‌بخشد. در این چارچوب، یکپارچه‌سازی داده‌ها یک زمینه تحقیقاتی فعال برای تکنیک‌های ML و DL بوده است که در داده‌های اومیکس، به‌ویژه ژنومیک سرطان اعمال می‌شود (Swan et al. 2013; Huang et al. 2018). به ویژه، معرفی رمزگذارهای خودکار، مانند حذف نویز از رمزگذارهای خودکار، امکان ارائه نمایش قوی از داده‌های ناهمگن را فراهم کرده است و استخراج ویژگی‌های بسیار نماینده و پیش‌بینی‌کننده را آسان‌تر انجام می‌دهد. در واقع، کاربردهای هوش مصنوعی در ژنومیک سرطان می‌تواند اطلاعات مفیدی را برای رشد سریع پزشکی دقیق و پیشگیری و نظارت بر بیماری فراهم کند (Chaudhary et al. 2018; Wang et al. 2020; Chai et al. 2021).

کاربردهای ML برای تشخیص و تفسیر جهش می‌تواند به شناسایی ژن‌های مستعد سرطان و پیش بینی خطر سرطان کمک کند (Li et al. 2017; Doncescu et al. 2009). عملکرد هوش مصنوعی در ژنومیک سرطان بسیار امیدوارکننده است. به عنوان مثال، نتایج هوش مصنوعی در تشخیص ملانوما و سرطان سینه بسیار قابل اعتماد است و اغلب از ارزیابی متخصصان پیشی می‌گیرد (Bejnordi et al. 2017; Haenssle et al. 2018). بسیاری از تکنیک‌های ML برای تشخیص و طبقه‌بندی سرطان و به‌ویژه برای شناسایی نشانگرهای زیستی استفاده شده‌اند. در سال ۲۰۰۳، با استفاده از درخت تصمیم[2] (Vlahou et al. 2003)، نتایج خوبی در طبقه بندی سرطان تخمدان به دست آمد. دو گروه، Abeel et al. (2010) و Chen et al. (2011)، SVM را برای شناسایی نشانگرهای زیستی سرطان به کار بردند. در سال‌های اخیر، معماری‌های عمیق برای فراخوانی و تشخیص جهش به کار گرفته شده‌اند. در پژوهش‌هایی Yuan et al. (2016) Deep-Gene، یک طبقه‌بندی کننده سرطان DNN، و Qi et al. (2021) از یک MVP برای اولویت بندی انواع بیماری زا استفاده کردند. در پژوهشی دیگر Malta et al. (2018) یک LR یک کلاسه برای استخراج ویژگی‌های رونویسی و اپی ژنتیکی مرتبط با حالت‌های انکوژنیک تمایز زدایی شده پیشنهاد کردند. مدل‌های بقا، مانند SurvivalNet (Yousefi et al. 2017)، یک رویکرد DL برای غربالگری مجموعه داده‌های بزرگ ژنومی سرطان، می‌تواند برای بهبود دقت پیش‌آگاهی و پیش بینی نتایج سرطان مفید باشد.

[1] Recurrent
[2] Decision Tree

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

یک زمینه کاربردی اخیر و امیدوارکننده برای روش‌های هوش مصنوعی در ژنومیک سرطان، به بررسی محاسباتی فعل و انفعالات کشنده مصنوعی در رده‌های سلولی سرطانی برای هدایت طراحی داروهای ضد سرطان مربوط می‌شود. کشندگی مصنوعی به نوعی از تعامل ژنتیکی اشاره دارد که در آن اغتشاش همزمان دو ژن منجر به مرگ سلولی یا اختلال شدید در حیات سلولی می‌شود، در حالی که اختلال در هر یک از ژن‌ها به تنهایی چنین نیست. در دسترس بودن همزمان نقشه‌های کامل تعاملات ژنتیکی به‌دست‌آمده در ارگانیسم‌های مدل (Costanzo et al. 2016)، کاتالوگ‌های داده‌های ژنومیک سرطان (TCGA 2020)، ابزارهای قدرتمند برای ویرایش ژنوم (مانند سیستم ویرایش CRISPR-Cas9) و فناوری‌های توالی‌یابی با توان بالای تک سلولی راه را برای کشف فنوتیپی سیستماتیک در وضوح تک سلولی، که برای مقابله با ناهمگنی سلول‌های تومور بسیار مهم است باز کرده است. در سال ۲۰۱۷، Way et al. (2017) یک رویکرد ML مبتنی بر رگرسیون لجستیک گروهی را توسعه دادند، که بر روی پروفایل‌های جهش و رونویسی گلیوبلاستوما از اطلس ژنوم سرطان (TCGA 2020) آموزش داده شد تا ژن‌هایی را که ممکن است کشندگی مصنوعی را در سلول‌های سرطانی فاقد ژن سرکوبگر تومور نوروفیبرومین ۱ نشان دهند، پیش بینی کند. در مطالعه‌ای Das et al. DiscoverSL (2019) را پیاده‌سازی کردند. DiscoverSL یک طبقه‌بندی کننده RF چند پارامتری است که بر روی داده‌های سرطان چندامیکس از اطلس ژنوم سرطان (TCGA 2020) برای پیش‌بینی و تجسم مرگ و میر مصنوعی در سرطان‌ها آموزش دیده بود. در پژوهشی Wan et al. (2020)، یک روش مبتنی بر NN نیمه نظارتی به نام EXP2SL را توسعه دادند، که بر روی مجموعه بزرگی از امضاهای بیان رده سلولی سرطانی از برنامه LINCS1000 (Stathias et al. 2020) آموزش داده شد تا فعل و انفعالات کشنده مصنوعی خط سلول سرطانی را پیش بینی کند.

دیگر کاربردهای مهم هوش مصنوعی در ژنومیک سرطان مربوط به شناسایی انواع تنظیمی در حوزه‌های غیرکدکننده (Kalinin et al. 2018)، پیش‌بینی زیست فعالی (Chen et al. 2018)، اولویت‌بندی داروهای ضد سرطان (Gupta et al. 2016) و پیش‌بینی حساسیت (Hejase and Chan 2015; Zitnik et al. 2018) است. همه این کاربردها گام‌های مهمی را به سمت پزشکی شخصی‌سازی شده، افزایش مسیرهای پیشگیری، درمان و نظارت دقیق و کمتر تهاجمی بر اساس ویژگی‌های خاص بیماران و محیطی که در آن زندگی می‌کنند، نشان می‌دهند (Xu et al. 2019).

اپی‌ژنومیکس: اپی ژنومیکس رشته‌ای است که فرآیندهای اپی ژنتیکی را در مقیاس ژنوم مطالعه می‌کند. این فرآیندها شامل مکانیسم‌های تنظیمی فعالیت ژن و وراثت است که توسط معماری ژنوم و مستقل از تغییرات در توالی DNA دیکته می‌شود. اصطلاح اپی ژنتیک که در سال ۱۹۴۲ توسط زیست‌شناس بریتانیایی کنراد وادینگتون[1] ابداع شد، نشان‌دهنده یک لایه تنظیمی از بیان ژن است که عمدتاً توسط ترکیبات شیمیایی کوچک (مانند گروه‌های متیل، استیل یا فسفات) که می‌توانند به طور برگشت‌پذیر به DNA (مانند متیلاسیونDNA) یا پروتئین‌های کروماتین (به عنوان مثال، متیلاسیون، استیلاسیون، فسفوریلاسیون و سایر

[1] Conrad Waddington

تغییرات شیمیایی که در دم‌های پروتئین‌های هیستون رخ می‌دهد) متصل شوند. این نشانه‌های اپی‌ژنتیکی به‌طور پویا از پروتئین‌های «نویسنده[۱]»، «خواننده[۲]» و «پاک‌کن[۳]» تنظیم شده‌اند (یعنی لایه‌بندی[۴]، تفسیر[۵] یا حذف[۶] می‌شوند). آن‌ها باعث مدولاسیون DNA از نظر سازمان فضایی و ظرفیت تعامل با دستگاه تنظیم کننده ژن می‌شوند که در نهایت منجر به خاموش یا روشن شدن بیان ژن‌های آسیب دیده می‌شود. علاوه بر متیلاسیون DNA و تغییرات هیستون، کمپلکس‌های بازسازی کروماتین در هماهنگی با سایر پروتئین‌های متصل شونده به DNA (مانند پروتئین‌های باند شونده به تقویت‌کننده[۷] و واسطه‌های حلقه کروماتین دوربرد[۸]) مکانیسم‌های اپی ژنتیک بیشتری را ارائه می‌کنند که مجموعاً سازمان سه‌بعدی (3D) از ژنوم را تعریف می‌کنند. این به نوبه خود، نواحی کروماتینی با حالت‌های فعال (یعنی دارای قابلیت رونویسی) یا سرکوب شده (یعنی غیرقابل دسترس برای ماشین‌های رونویسی) را تعریف می‌کند. هدف اپی ژنومیکس ترسیم سیستماتیک مجموعه‌هایی از علائم اپی‌ژنتیکی و مناظری از نواحی ژنومی فعال و سرکوب شده (یعنی اپی‌ژنوم) در انواع و حالات مختلف سلولی، برای مشخص کردن اثر عملکردی بر بیان ژن است. در واقع، هر نوع سلول دارای یک اپی ژنوم منحصر به فرد است که اجازه تمایز خاصی را می‌دهد و حالت خاصی را برای سلول منعکس می‌کند (Stueve et al. 2016). ثابت شده است که شناسایی حالت‌های کروماتین، تراکم محلی علائم اپی‌ژنتیک، اتصال‌های کروماتین با برد بلند و الگوهای اصلاح هیستون برای مطالعه و تفسیر مناطق تنظیم‌کننده، فعالیت خاص سلول و الگوهای مرتبط با بیماری ارتباط دارد. برای این منظور، بسیاری از تکنیک‌های ML و DL برای تعریف پروفایل‌های نوع سلولی متیلاسیون DNA (یا متیلوم‌ها) و تغییرات هیستون، طبقه‌بندی نواحی کروماتین به حالت‌های فعال و سرکوب شده و اخیراً، طبقه‌بندی انواع تومور بر اساس داده‌های متیلوم با توان عملیاتی بالا و پیش‌بینی ساختار پیچیده سه بعدی ژنوم استفاده شده است (Holder et al. 2017; Belokopytova and Fishman 2020; Perez and Capper 2020). در سال ۲۰۱۵، Ernst J, Kellis روش ChromImpute را توسعه دادند، که یک رویکرد ML مبتنی بر درختان رگرسیون برای پیش‌بینی در مقیاس بزرگ علائم اپی ژنومیک (مانند متیلاسیون DNA و علائم هیستون) و حالت‌های کروماتین (مانند دسترسی به DNA) است. نویسندگان عملکرد روش استنتاج خود را در مجموعه بزرگی از نقشه‌های اپی‌ژنومیک در دسترس عموم، برای دستیابی به توافق قوی بین سیگنال‌های تجربی مشاهده شده و سیگنال‌های محاسباتی نسبت داده شده نشان دادند. در سال ۲۰۱۶، Zhang et al. سیستم IDEAS را توسعه دادند، که یک سیستم حاشیه نویسی اپی ژنوم یکپارچه بر اساس مدل‌های کمی پنهان مارکوف[۹] (HMMs) برای توصیف

[1] Writer
[2] Reader
[3] Eraser
[4] Layered
[5] Interpreted
[6] Removed
[7] Enhancer-binding proteins
[8] Mediators of long-range chromatin looping
[9] Quantitative Hidden Markov Models

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

پویایی اپی ژنتیک و تشخیص مناطق تنظیمی بود. روش پیشنهادی قادر به مدیریت چندین ژنوم و مقایسه رویدادهای اپی‌ژنومیک استنباط شده در وضوح پایه در انواع مختلف سلول، برای شناسایی الگوهای برگشت‌پذیر خاص سلولی است. یک معماری رمزگذار خودکار حذف نویز انباشته[1] به نام DeepMethyl توسط Wang et al. (2016) توسعه داده شد که از ویژگی‌های توالی DNA و ساختار ژنوم سه بعدی برای پیش بینی وضعیت متیلاسیون DNA سایت‌های CpG استفاده می‌کرد. اخیراً، Kelley et al. (2018) رویکرد Basenji را پیشنهاد کردند که یک رویکرد شبکه‌های عصبی پیچیده یا کانولوشن (CNN) برای پیش‌بینی پروفایل‌های اپی‌ژنتیکی مختص سلول و پروفایل‌های ترانسکریپتی با استفاده از فقط توالی DNA به عنوان ورودی می‌باشد. داده‌های اپی‌ژنومیک اغلب تحت تأثیر نویز و اریبی (سوگیری) قرار می‌گیرند، و روش‌های ML و DL به طور گسترده در سال‌های اخیر برای افزایش کیفیت داده‌ها مورد استفاده قرار گرفته‌اند. در سال ۲۰۱۷، Koh et al. از یک CNN برای حذف نویز و بهبود کیفیت داده داده‌های هیستون ChiP-seq (توالی‌یابی رسوب ایمنی کروماتین[2]) استفاده کردند. در سال ۲۰۱۹، Hiranuma et al. الگوریتم AIControl را پیشنهاد کردند، که یک الگوریتم رگرسیون برای تشخیص گسترده ژنوم مناطق غنی شده با اتصال است، و بسیاری از مجموعه داده‌های کنترلی در دسترس عموم را برای بهبود تفریق پس‌زمینه و تمایز سیگنال ادغام می‌کند. مزیت یکپارچه‌سازی داده‌هایی که توسط AIControl مورد سوء استفاده قرار می‌گیرد، توانایی تفریق انواع بایاس‌های مؤثر بر داده‌های ChiP-seq است، بنابراین روشی مؤثر برای حذف سیگنال‌های پس‌زمینه از آزمایش‌های فاقد نمونه‌های کنترل ارائه می‌کند. اخیراً، Lal et al. (2021) ابزار AtacWorks را معرفی کردند، که یک ابزار مبتنی بر DL است، و یک مدل NN باقیمانده متشکل از چندین بلوک باقیمانده انباشته را آموزش می‌دهد تا داده‌های توالی‌یابی تک سلولی با پوشش پایین یا با کیفیت پایین به‌دست‌آمده توسط ATAC-seq (آزمایش کروماتین قابل دسترسی به ترانسپوزاز با استفاده از توالی‌یابی) را حذف کند. ATAC-seq یک تکنیک با توان عملیاتی بالا است که مکان‌های کروماتین باز در سطح ژنوم را به عنوان یک پروکسی برای مناطق تنظیم‌کننده فعال می‌گیرد. چندین رویکرد ML برای مدل‌سازی ساختار کروماتین از داده‌های تجربی به‌دست‌آمده از جذب ساختار کروموزوم[3] (3C) و فناوری‌های مشتق شده از آن (مانند 4C، C5 و Hi-C) استفاده شده است (Dekker et al. 2013; Lin et al. 2019). در سال ۲۰۱۲، Ernst and Kellis روش ChromHMM را ارائه کردند، که یک روش خودکار مبتنی بر HMM چند متغیره برای استنتاج حالات کروماتین است که از مجموعه‌هایی از خواندن‌های هم‌تراز برای هر علامت اصلاح کروماتین[4] مورد بررسی شروع می‌شود. در سال ۲۰۱۴، Gusmao et al. یک HMM را برای تشخیص مکان‌های اتصال فاکتور رونویسی و نواحی کروماتین باز پیشنهاد کرد که اطلاعات ساختاری مانند حساسیت DNase I و تغییرات هیستون را یکپارچه می‌کند. چارچوب‌های Chrom3D (Paulsen et al.) و Chrom-Struct (Caudai et al. 2019 a and b) از بهینه سازی مونت کارلو با کمینه سازی تابع امتیاز از دست دادن برای

[1] Stacked denoising autoencoder
[2] Chromatin immune-precipitation sequencing
[3] Chromosome conformation capture
[4] Chromatin modification mark

تخمین ساختار کروماتین برای داده‌های Hi-C استفاده می‌کنند. بسیاری از چارچوب‌های محاسباتی برای مدل‌سازی سه‌بعدی کروماتین نیز ابزارهای تجسمی (Serra et al. 2017) را فراهم می‌کنند تا بتوان الگوهای ساختاری کروماتین را به صورت بصری تفسیر کرد و روابط بین حالت‌های کروماتین، موقعیت‌های ژنومی و تغییرات پاتولوژیک را آسان‌تر درک کرد. در سال ۲۰۲۰، Fudenberg et al. ابزار Akita را توسعه دادند، که یک CNN است که ساختارهای ژنوم سه بعدی محلی را بر اساس فرکانس‌های اتصال مختص مکان پیش بینی می‌کند. الگوریتم Akita، که بر روی مجموعه‌ای از نقشه‌های Hi-C با وضوح بالا آموزش داده شده است، یک ناحیه ژنومی یک میلیون جفت باز را به عنوان ورودی دریافت می‌کند و فرکانس‌های اتصال بین هر جفت پنجره‌های با طول ۲۰۴۸ جفت باز از توالی DNA در این ناحیه را پیش‌بینی می‌کند. در همان سال، Schwessinger et al. روش DeepC را توسعه دادند، که یک DNN است که از رویکرد یادگیری انتقال و داده‌های Hi-C مختص بافت استفاده می‌کند تا مدل‌هایی را آموزش دهد که فولدینگ (تاخوردگی)[1] ژنوم را در پنجره‌های DNA به اندازه مگاباز پیش بینی کند. سپس این مدل‌های آموزش‌دیده برای پیش‌بینی مرزهای دامنه کروماتین در وضوح بالا و تعیین‌کننده‌های توالی فولدینگ ژنوم مورد بهره‌برداری قرار می‌گیرند، که به DeepC اجازه می‌دهد تأثیر گونه‌های ژنتیکی با اندازه‌های مختلف (به عنوان مثال، از تغییرات ساختاری بزرگ تا SNP) را بر روی ساختار سه‌بعدی پیش‌بینی کند.

ترنسکریپتومیکس: ترنسکریپتوم مجموعه کاملی از ژن‌های رونویسی شده است که در یک سلول در یک نقطه زمانی مشخص وجود دارد. اولین استفاده و تعریف کلمه "ترنسکریپتوم" به سال ۱۹۹۷ در مقاله Velculescu et al. برمی‌گردد، در این پژوهش نویسندگان ژن‌های بیان شده در تنها یوکاریوتی که کل توالی ژنوم آن در آن زمان در دسترس بود (Goffeau et al. 1996)؛ یعنی مخمر را تجزیه و تحلیل و مشخص کردند. رونوشت‌ها با استفاده از یکی از اولین روش‌های رونویسی مبتنی بر توالی یابی که توسعه یافته بود، یعنی آنالیز سریالی بیان ژن[2] (SAGE) (Velculescu et al. 1995) کمی سازی شدند. آن‌ها (Velculescu et al. 1997) ترنسکریپتوم (رونوشت) را به عنوان "هویت هر ژن بیان شده و سطح بیان آن برای جمعیت مشخصی از سلول‌ها" تعریف کردند. این اصطلاح بعداً به معنای گسترده‌تری استفاده شد و اکنون می‌توان آن را برای جمعیت مشخصی از سلول‌ها، یک بافت، یک اندام یا یک موجود زنده به کار برد. ترنسکریپتوم شامل کل محتوای رونوشت، شامل ژن‌های رونویسی کدکننده پروتئین و غیر پروتئینی، از متداول‌ترین RNAهای زیرساختی (RNAهای ناقل و ریبوزومی) و RNAهای پیام‌رسان (درگیر در ترجمه پروتئین) تا جدیدترین RNAهای شناسایی شده کوچک و بلند غیر کدکننده (با طول ۲۰۰ جفت باز) (Nagano and Fraser 2011)، RNAهای حلقوی (Kristensen et al. 2019)، RNAهای برهمکنشی Piwi (Ozata et al. 2019) و بسیاری دیگر از انواع جدید RNA غیر کدکننده (ncRNA). در واقع، پروژه‌های مبتنی بر کنسرسیوم برای حاشیه

[1] Folding
[2] Serial analysis of gene expression

نویسی[1] و توصیف سیستماتیک عناصر عملکردی، مانند دایره المعارف عناصر DNA یعنی ENCODE (www.encodeproject.org) (Djebali et al. 2012; EPC 2012)، رونویسی فراگیر غیرمنتظره ای را در ژنوم‌ها شناسایی کردند. مشخص شده حدود ۸۰ درصد از DNA ژنومی پستانداران به طور فعال رونویسی می‌شود، اما اکثریت قریب به اتفاق آن به عنوان ncRNA طبقه بندی می‌شود. در مقایسه با ژنوم، رونوشت ذاتاً متغیر و پویا است و این امر تعریف و تجزیه و تحلیل آن را بسیار پیچیده‌تر می‌کند. ترانسکریپتومیکس مطالعه ترانسکریپتوم در شرایط فیزیولوژیکی یا پاتولوژیک مورد علاقه است که هدف آن به دست آوردن پیوند پویا بین ژنوم یک ارگانیسم و ویژگی‌های فنوتیپی آن است. در حالت ایده‌آل، سعی می‌کند تمام انواع RNA و توالی‌های موجود در یک سلول معین را در یک زمان معین شناسایی کند تا ساختار رونویسی ژن‌ها را از نظر مکان‌های شروع، انتهاهای ۵́ و ۳́، اگزون‌ها، اینترون‌ها و الگوهای اسپلایسینگ برای شناسایی سطوح بیان ژن و کشف مکانیسم‌های تنظیمی احتمالی در مقیاس کل ژنوم با استفاده از تکنیک‌های با توان عملیاتی بالا تعیین کند. به جای تمرکز بر عملکرد ژن‌ها یا رونوشت‌های فردی، ترانسکریپتومیکس جاه طلبی دارد که کل رونوشت و تغییرات آن را در سلول‌های مختلف، مراحل رشد، در شرایط مختلف بیولوژیکی و محیطی توصیف کند. از اواخر دهه ۱۹۹۰، تحقیقات ترنشکریپتومیکس (رونوشت شناسی) به طور مکرر توسط نوآوری‌های فن‌آورمحور جدید در این زمینه متحول شده است، و در هر مرحله مشخص می‌شود که چه چیزی ممکن است مورد بررسی قرار گیرد. توسعه ریزآرایه‌ها (Chang 1983; Schena et al. 1995) و بعدها، فناوری‌های NGS (Morozova and Marra 2008; Buermans and den Dunnen 2014) دو لحظه کلیدی در این فرآیند بوده‌اند. ریزآرایه‌ها امکان کمی سازی مجموعه‌ای از توالی‌های RNA از قبل شناخته شده و از پیش انتخاب شده را فراهم می‌کنند، زیرا سیگنال‌های خروجی آن‌ها به هیبریداسیون مولکول‌های هدف با پروب‌های طراحی شده موقتی که روی آرایه لنگر انداخته‌اند، متکی است. فن‌آوری‌های NGS به کار رفته در توالی‌یابی RNA (RNA-seq) (Wang et al. 2010; Stark et al. 2019) می‌توانند مولکول‌های رونویسی شده را مستقل از دانش قبلی ضبط کنند، زیرا توالی مولکول‌های RNA سنجش‌شده را به عنوان بخشی از مرحله تشخیص بازسازی می‌کنند (مانند رویکرد توالی‌یابی به وسیله سنتز، که در آن توالی هدف با سنتز رشته مکمل همراه با یک سیستم تشخیص نوکلئوتیدهای وارد شده در طول سنتز آشکار می‌شود). در نتیجه افزایش توان عملیاتی، دقت بالاتر و هزینه کمتر این فناوری‌های تخصصی NGS، در دو دهه گذشته شاهد رشد تصاعدی در مطالعات رونویسی بوده‌ایم که منابع ارزشمندی را برای تحقیقات گسترده در مورد مقررات رونویسی و پس از رونویسی فراهم کرده‌اند (Lowe et al. 2017). ترنسکریپتومیکس شامل موارد زیر است:

الف) طبقه بندی رونوشت کد کننده پروتئین و غیر کد کننده پروتئین: یکی از اهداف اصلی ژنومیک عملکردی، طبقه‌بندی عناصر رونویسی است، مانند حاشیه‌نویسی رونوشت‌ها به عنوان mRNA (یعنی کدکننده پروتئین) یا ncRNA، یا پیش‌بینی پتانسیل کدگذاری برای هر یک از محصولات رونویسی چندگانه (یعنی ایزوفرم‌ها) از یک مکان ژن به دلیل اسپلایسینگ

[1] Annotation

متناوب[1] (AS). بسیاری از روش‌های *in silico* (بیوانفورماتیک) برای حل این کار تلاش کرده‌اند، اما در عمل می‌تواند به طرز شگفت‌آوری دشوار باشد. در واقع، راه‌حل‌های پیشنهادی اغلب منجر به جریان‌های کاری دستی و زمان‌بر با تعدادی محدودیت می‌شود. پروژه‌های ENCODE و GENCODE (Djebali et al. 2012; EPC 2012; Frankish et al. 2019) نقش مهمی در این زمینه ایفا کردند. در اکثریت قریب به اتفاق موارد، توصیف رونوشت‌های جدید براساس مقایسه با مجموعه‌های فعلی حاشیه‌نویسی‌های ژنوم در دسترس از پایگاه‌های داده عمومی، مانند توالی‌های رونوشت و پروتئین جمع‌آوری‌شده از ارگانیسم‌های مختلف، دامین‌ها و ساختارهای پروتئینی شناخته‌شده، ادغام شده با داده‌های تجربی اومیکس چندگانه است. هر چه شواهد پشتیبان بیشتر باشد، اعتماد به نفس بالاتری برای فراخوانی رونوشت تحت بررسی به عنوان کدکننده پروتئین یا کدنکننده پروتئین وجود دارد (Fickett and Tung 1992; Frith et al. 2006; Leoni et al. 2011). طبقه‌بندی نوع رونوشت یک برنامه کاربردی را ارائه می‌دهد که در آن هوش مصنوعی می‌تواند بسیار مهم باشد. در واقع، این یک کار معمولی ML است که برای آن چندین روش و ابزار، مبتنی بر یادگیری نظارت شده و بدون نظارت، در دسترس قرار گرفته است. به عنوان مثال، روش‌های SVM (ماشین‌های بردار پشتیبان[2]) با موفقیت برای تخصیص پتانسیل کدگذاری به رونوشت‌ها با توجه به توالی‌های انتخابی و ویژگی‌های ساختاری استفاده شده است. به طور خاص، الگوریتم‌های طبقه‌بندی متنوع به طور متغیر ویژگی‌های مربوطه؛ از قبیل طول چارچوب قرائت باز[3] (ORF)، که زیر توالی mRNA خاصی است که مجموعه‌ای از اسیدهای آمینه را برای تولید پروتئین دیکته می‌کند را ادغام می‌کنند. مجموعه‌ای از قبیل، ترکیب اسید آمینه مربوطه؛ ساختار ثانویه پروتئین پیش بینی شده، نسبت پیش بینی شده باقی مانده‌های پروتئین در معرض حلال، وجود همولوگ متناظر در موجودات دیگر؛ و نرخ‌های جایگزینی مترادف در مقابل غیر مترادف (Liu et al. 2006; Kong et al. 2007; Li et al. 2014; Schneider and Hugo 2017). علاوه بر این، طبقه‌بندی‌کننده‌های مبتنی بر الگوریتم‌های ML نیز برای تشخیص ncRNAهای بلند (lncRNA) از رونوشت‌های کدکننده پروتئین پیشنهاد شدند. به عنوان مثال، Pian et al. (2016) از یک روش RF با برخی ویژگی‌های خاص جدید استفاده کردند. از آنجایی که به نظر می‌رسد رونوشت‌های کدکننده پروتئین دارای ORFهای طولانی‌تری در مقایسه باlncRNAها هستند، نویسندگان دو ویژگی خاص زیر را برای تمایز بهتر انتخاب کردند: الف) طول ORF بیشتر (MaxORF) به دست آمده در سه طرح سخنرانی ممکن (یعنی شروع ترجمه *in silico* هر سه نوکلئوتید به اسید آمینه مربوطه در موقعیت ۱، ۲ یا ۳ رونوشت داده شده و ب) مقدار MaxORF نرمال شده که با در نظر گرفتن طول کل رونوشت به دست می‌آید. به طور مشابه، الگوریتم‌های دیگری که مجموعه‌ای از ویژگی‌ها را از توالی‌ها استخراج می‌کنند و آن را به الگوریتم‌های سنتی ML می‌رسانند تا پتانسیل کدگذاری را ارزیابی کنند، توسعه یافته‌اند و در دسترس هستند (Wang et al. 2013; Kang et al. 2017). اگر چه ادغام اطلاعات اضافی که ذاتاً از توالی‌های رونوشت مشتق

[1] Alternative splicing
[2] Support Vector Machines
[3] Open reading frame

نشده‌اند، ممکن است طبقه‌بندی رونوشت را بهبود بخشد، این همچنین می‌تواند وابستگی به حاشیه‌نویسی قابل اعتماد را ایجاد کند و توسط دانش علمی کنونی محدود شود، که به سمت موضوعات یا گونه‌های جریان اصلی (مثلاً کمتر در دسترس برایlncRNAها در مقایسه باmRNAها، یا برای غیر مدل در مقایسه با ارگانیسم‌های مدل) اریب می‌شود. علاوه بر این، انتخاب ویژگی دستی که در ML سنتی انجام شده است، می‌تواند سوگیری‌هایی را در طبقه‌بندی ایجاد کند، زیرا آن‌ها با دست طراحی و انتخاب می‌شوند. برعکس، روش‌های DL با استفاده از شبکه‌های عصبی می‌توانند قوانین بیولوژیکی پیچیده و پنهان را در داده‌های خام رونویسی کشف کنند، در نتیجه تبدیل به ابزار قدرتمندتری برای بررسی رونوشت‌های بی‌شمار گونه‌هایی می‌شوند که توسط فناوری‌های توالی‌یابی با توان بالای فعلی در دسترس هستند (Baek et al. 2018; Hill et al. 2018; Amin et al. 2019; Camargo et al. 2020).

ب) تجزیه و تحلیل داده‌های بیان ژن: بیان ژن فرآیند پویایی است که اطلاعات رمزگذاری شده در ژنوم را به محصولات نهایی ژن عملکردی تبدیل می‌کند و باعث ایجاد طیف وسیعی از پروتئین‌ها یاncRNAها می‌شود. شناسایی مکانیسم‌های مولکولی که بیان ژن افتراقی را کنترل می‌کنند، هدف اصلی تحقیقات بیولوژیکی پایه و کاربردی است. داده‌های بیان ژن حاصل از ریزآرایه‌ها یا پلتفرم‌های RNA-seq به طور گسترده برای تشخیص بافت‌ها، شرایط بیولوژیکی در مقابل فیزیولوژیکی، فنوتیپ‌های بیماری و شناسایی نشانگرهای زیستی ارزشمند بیماری استفاده شده‌اند. یک مشکل معمولی با فناوری‌های با توان بالا، عدم تناسب ابعاد بین نمونه‌ها و متغیرها در مجموعه داده است. در واقع، تعداد متغیرهای سنجش شده با ابعاد زیاد، مانند سطوح بیان ده‌ها هزار ژن یا رونوشت، معمولاً بسیار بیشتر از نمونه‌های موجود تحت بررسی (مانند تکرارهای بیولوژیکی، افراد مبتلا به یک بیماری) است. علاوه بر این، این مجموعه داده‌های با ابعاد بالا اغلب پراکنده و نویزی هستند. در عمل، افزایش پراکندگی جمع آوری داده‌ها با قدرت آماری را مختل می‌کند، و به دست آوردن بینش بیولوژیکی از این داده‌ها با استفاده از رویکردهای تحلیلی سنتی بسیار دشوار است. این پدیده "نفرین ابعاد[1]" نامیده می‌شود. الگوریتم‌های تخصصی ML می‌توانند ابزار قدرتمندی برای رسیدگی به چنین مسائلی و سایر چالش‌های جدی باشند. رویکردهای یادگیری بدون نظارت مانند خوشه‌بندی و PCA به طور گسترده برای یافتن الگوهای ذاتی در داده‌ها بدون ارجاع به دانش قبلی، برای مثال، برای شناسایی سیگنیچرهای ژن در پروفایل‌های بیان ژن که ممکن است در غیر این صورت نادیده گرفته شوند مورد استفاده قرار گرفته‌اند. همبستگی‌های بیان ژن جهانی (یا متاآنالیزها) حتی با مقایسه مطالعات متعدد در سطح ژنوم امکان‌پذیر است. در پژوهشی Talavera et al. (2018) یک متاآنالیز از حدود ۱۵۰۰ مجموعه داده ریزآرایه مخمر حاوی چندین آزمایش مرتبط با استرس را انجام دادند. آن‌ها از یک الگوریتم خوشه‌بندی تجمعی برای شناسایی گروه‌های (بلوک‌های) رونوشت‌هایی که همبستگی بالایی از سطوح RNA را در شرایط مختلف نشان می‌دهند، استفاده کردند. تجزیه و تحلیل‌های غنی‌سازی عملکردی بعدی بلوک‌های رونویسی به‌دست‌آمده، که با استفاده از حاشیه‌نویسی‌های ژنوم مخمر فرآیندهای

[1] Curse of dimensionality

بیولوژیکی بر اساس زیرمجموعه‌های هستی‌شناسی ژن (همچنین به عنوان GO slims شناخته می‌شوند) انجام شد، نشان داد که آن گروه‌هایی از ژن‌های با تنظیم بالا یا پایین‌تر در واقع با فرآیندهای بیولوژیکی که با پاسخ به محرک‌های خارجی مختلف وابسته هستند (به عنوان مثال، استرس اکسیداتیو، استرس اسمزی، محرک آسیب DNA، محدودیت گلوکز) مرتبط هستند. این استراتژی نشان می‌دهد که چگونه اطلاعات عملکردی در سطح بلوک رونوشت، بهتر از در سطح تک ژنی، در تجزیه و تحلیل‌های بیان افتراقی می‌تواند به طور موثر به ایجاد فرضیه‌ها و مدل‌سازی مکانیسم‌های بیولوژیکی مولکولی سیستم مورد بررسی کمک کند. ریزآرایه‌ها یا داده‌های RNA-seq نیز می‌توانند توسط رویکردهای هوش مصنوعی به‌عنوان مجموعه‌های آموزشی برای یادگیری مؤثر نحوه تمایز گروه‌های بالینی مجزا و اختصاص صحیح بیماران به آن‌ها استفاده شوند (Myszczynska et al. 2020). در یک پژوهش (van IJzendoorn et al. 2019)، محققان حدود ۲۰۰ نمونه سارکوم بافت نرم را از پروژه اطلس ژنوم سرطان (NIH 2020) تجزیه و تحلیل کردند تا بینش جدیدی در مورد بسیاری از زیرگروه‌های متفاوت در پیش آگهی و درمان به دست آورند، که متأسفانه دارای همپوشانی مورفولوژیکی قابل توجهی با یکدیگر هستند و تشخیص افتراقی را واقعا دشوار می‌کند. برای این منظور، محققان الگوریتم‌های مختلف ML را اعمال کردند: الگوریتم PCA برای کاهش ابعاد؛ یک DNN برای بررسی همپوشانی الگوهای بیان ژن سارکوم‌های بافت نرم با الگوهای بیان ژن بافت‌های سالم انسانی؛ یک رویکرد RF برای شناسایی نشانگرهای تشخیصی جدید. در نهایت، ژن‌های پیش‌بینی کننده زیرگروه خاص تومور[1] مشخص شدند و با استفاده از آنالیز k-NN به‌عنوان پیش‌بینی‌کننده فاصله بدون متاستاز[2] مورد آزمایش قرار گرفتند. بسیار جالب است که در آخرین مرحله پروژه ENCODE (Breschi et al. 2020)، خوشه‌بندی سلسله مراتبی[3] برای تعریف مجموعه‌های ژنی هسته‌ای که با انواع سلول‌های اصلی در ۵۳ سلول اولیه از مکان‌های مختلف بدن انسان مطابقت دارند، استفاده شد. خوشه‌بندی این سلول‌های اولیه نشان داد که بیشتر سلول‌ها در بدن انسان چند برنامه رونویسی گسترده را به اشتراک می‌گذارند که پنج نوع سلول اصلی را تعریف می‌کنند و شامل: سلول‌های اپیتلیال، اندوتلیال، مزانشیمی، عصبی و سلول‌های خونی هستند. بر اساس پروفایل‌های بیان ژن، این مجموعه جدید از انواع سلول، پارادایم اولیه بافت شناسی را که بافت‌ها به طور سنتی طبقه بندی می‌شوند، دوباره تعریف کردند. با توجه به پیشرفت فناوری‌هایی که می‌توانند مولکول‌ها را در یک سلول منفرد، مانند scRNA-seq، نمایه کنند، وظیفه کاهش ابعاد برای امکان تجسم و تجزیه و تحلیل مجموعه‌های داده با ابعاد بالا به طور فزاینده‌ای سخت شده است. در نتیجه، روش‌های غیرخطی، مانند t-distributed Stochastic Neighbour Embedding (t-SNE) (van der Maaten and Hinton 2008) و Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018)، هنگام برخورد با نمونه‌های بزرگ و ناهمگن نسبت به روش‌های خطی

[1] Tumor subtype-specific prognostic genes
[2] Metastasis-free interval
[3] Hierarchical clustering

معمولی، مثل PCA شتاب بیشتری گرفتند (Yang et al. 2021). مقدار کمی از مواد RNA ذاتی[1] آزمایش‌های scRNA-seq در ماهیت بسیار نویزی و ناقص داده‌های خروجی منعکس می‌شود. به طور خاص، یکی از مشکلات عمده مربوط به آزمایش‌های scRNA-seq، درصد بالای مشاهدات با ارزش صفر (معروف به dropouts) است که باعث ایجاد چندین رویکرد مبتنی بر ML و DL برای انتساب داده‌ها شد. در سال ۲۰۱۸، Li and Li روش ScImpute را پیشنهاد دادند که یک رگرسیون LASSO تکراری برای برانگیختن مقادیر حذف در داده‌های scRNA-seq است. رویکرد DrImpute را Gong et al. (2018) ارایه کردند، که یک رویکرد مبتنی بر خوشه‌بندی را توسعه داد و از یک استراتژی اجماع برای منتسب کردن مقادیر گمشده برای یک ژن هدف معین در داده‌های scRNA-seq، بر اساس مقادیر بیان ژن سایر سلول‌های متعلق به همان خوشه استفاده می‌کند. معماری DeepImpute توسط Arisdakessian et al. (2019) اجرا شد، که یک معماری DNN است که یک رویکرد تقسیم و غلبه را برای استخراج الگوهای مرتبط مفید برای انتساب مقادیر بیان گمشده برای ژن‌های هدف تعبیه می‌کند. به طور خاص، با توجه به مجموعه‌ای از ژن‌های هدف با حذف در داده‌های scRNAseq ، معماری DeepImpute شبکه‌های زیرعصبی متعددی ایجاد می‌کند که هدف هریک یادگیری رابطه بین ژن‌های ورودی (ژن‌های پیش‌بینی‌کننده) و زیرمجموعه‌ای از ژن‌های هدف دارای dropouts (با مقادیر صفر بیان ژنی که باید نسبت داده شود) است، در نتیجه پیچیدگی را با یادگیری مسائل کوچکتر کاهش می‌دهد. در پژوهشی Ghahramani et al. (2018) از یک GAN (شبکه متخاصم مولد[2]) برای ادغام و حذف نویز مجموعه داده‌های scRNA-seqمختلف مشتق‌شده از آزمایشگاه‌ها و پروتکل‌های تجربی مختلف و انجام کاهش ابعاد استفاده کردند. در سال ۲۰۱۹، Grønbech et al. از یک رویکرد DL بدون نظارت مبتنی بر AEs متغیر (Heje Grønbech et al. 2020) برای تخمین سطوح بیان ژن به طور مستقیم از داده‌های خام scRNA-seq استفاده کردند.

ج) تشخیص کد اسپلایسینگ متناوب: اسپلایسینگ متناوب mRNA یوکاریوتی منبع مهمی از تنوع پروتئین را ایجاد می‌کند (Maniatis and Tasic 2002). گزارش شده است که بیشتر (یعنی ۹۵٪) ژن‌های انسانی چند اگزونی می‌توانند تحت حوادث اسپلایسینگ متناوب قرار گیرند (Wang et al. 2008; Mollet et al. 2010). براساس نتایج مطالعات مختلف اسپلایسینگ متناوب نابجا می‌تواند موجب بیماری‌های مختلف شود (Garcia-Blanco et al. 2004; Shen et al. 2012; Jha et al. 2017; Bretschneider et al. 2018; Kahles et al. 2018; Jaganathan et al. 2019; Zhang et al. 2019; Salovska et al. 2020;). علاوه بر ارائه اطلاعات در مورد فراوانی RNA، داده‌های RNA-seq می‌توانند برای استنتاج الگوهای اسپلایسینگ متناوب و شناسایی رویدادهای افتراقی اسپلایسینگ متناوب مرتبط با شرایط نمونه مختلف، مانند درمان در مقابل کنترل، بیماری در مقابل سلامت، مراحل رشد متنوع و غیره استفاده شوند. کار اصلی برروی توسعه روش‌های DL برای رمزگشایی کد اسپلایسینگ توسط Leung et al. (2014) انجام شد. آن‌ها با استفاده از یک DNN، با میلیون‌ها متغیر که

1 RNA material inherent
2 Generative Adversarial Network

هم ویژگی‌های ژنومی و هم محتویات بافت را نشان می‌دهند، الگوهای اسپلایسینگ را از داده‌های RNA-Seq موش، پیش‌بینی کرده‌اند، که از تلاش‌های قبلی که بر اساس معماری‌های کم‌عمق‌تر بود، بهتر عمل کردند.

د) تشخیص رویداد پلی آدنیلاسیون متناوب: ابزارهای متعددی در منابع برای پیش‌بینی مکان‌های پلی‌آدنیلاسیون[1] (PAS)از توالی‌های ژنومی انسانی معرفی شده‌اند. روش DNAFSMiner (Liu et al. 2005) مکان‌های پلی‌آدنیلاسیون را از توالی‌ها با استفاده از ویژگی‌های k-mer در یک مدل ماشین‌های بردار پشتیبان[2] (SVMs) پیش بینی می‌کند. روش Dragon PolyA Spotter (Kalkatawi et al. 2012) همچنین مکان‌های پلی‌آدنیلاسیون را از روی توالی‌ها با استفاده از شبکه‌های عصبی مصنوعی (ANN) و جنگل‌های تصادفی[3] (RF) پیش بینی می‌کند. روش POLYAH (Salamov and Solovyev 1997) با استفاده از یک تابع تشخیص خطی، PASهای واقعی را از سایر سیگنال‌های هگزامر متمایز می‌کند. این الگوریتم فقط بر روی PAS متناوب (یعنی توالی موتیف AATAAA) در تجزیه و تحلیل تمرکز می‌کند، اگرچه PASهای متناوب (انواع توالی AATAAA) ممکن است بر شناسایی و تمایز سایت تأثیر بگذارند. رویکرد Polyadq (Tabaska and Zhang 1999) از یک تابع تفکیک درجه دوم برای پیش بینی مناطق PAS واقعی استفاده می‌کند. این ابزار دو سیگنال PAS را در تحلیل در نظر می‌گیرد. با این حال، زیست‌شناسی زیربنایی پلی‌آدنیلاسیون متناوب پیچیده‌تر است، و انتخاب ماشین‌های پلی‌آدنیلاسیون برای شناسایی یک PAS معین نه تنها به خود PAS بلکه به عناصر غنی از U/GU (AUEs و DAEs) بستگی دارد. روش Polyasvm (Cheng et al. 2006) سایت‌های polyA را از توالی‌ها با استفاده از یک مدل SVM پیش بینی می‌کند. روش PolyAR (Akhtar et al. 2010) همچنین مکان‌های polyA را از توالی‌ها با استفاده از یک تابع تشخیص خطی پیش بینی می‌کند. هر دوی این ابزارها از ویژگی‌های توالی دستچین شده استفاده می‌کنند. به منظور غلبه بر این محدودیت، مدل‌های DL مانند DeepPolyA (Gao et al. 2018)، DeeReCT-PolyA (Xia et al. 2019) و Conv-Net (Leung et al. 2017) اخیراً برای پیش‌بینی PAS ها و شناسایی PASهای ژن نسبتاً غالب (یعنیPASهایی که اغلب در یک ژن مشخص استفاده می‌شوند) معرفی شده‌اند. توجه داشته باشید، همه این مدل‌ها از شبکه‌های عصبی پیچیده یا کانولوشن[4] (CNN) برای استخراج ویژگی‌ها از توالی‌های ژنومی ورودی استفاده می‌کنند. اگرچه ساختار ثانویه در نزدیکی یک PAS نیز برای انتخاب PAS برای فرآیند پلی‌آدنیلاسیون (Bar-Shira et al. 1991; Brown et al. 1991; Wu and Alwine 2004) بسیار مهم است، هیچ یک از این ابزارها ساختارهای ثانویه RNAرا در روش‌های پیش بینی خود در نظر نمی‌گیرند.

[1] Polyadenylation sites
[2] Support Vector Machines
[3] Random Forests
[4] Convolutional Neural Networks

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

اپی‌ترنسکریپتومیکس: در میان مکانیسم‌های تنظیمی متنوع زیست‌شناسی مولکولی، مشخص شده است که همه کلاس‌های RNA سلولی در معرض تغییرات همزمان و پس از رونویسی هستند. وضعیت تغییر رونوشت پویا است و لایه جدیدی از پیچیدگی را در تنظیم بیان ژن آشکار می‌کند. شبیه به اپی ژنومیکس، به نظر می‌رسد این مکانیسم تنظیمی توسط پروتئین‌های متصل شونده به RNA «نویسنده»، «خواننده» و «پاک‌کن» تنظیم شده است، که می‌تواند به سرعت سطوح بیان رونوشت را بر اساس تغییرات محیطی و رشدی تغییر دهد. در مجموع، انبوهی از تغییرات RNA، از جمله تغییرات شیمیایی غیرجانشینی و رویدادهای ویرایشی، اپی‌ترنسکریپتوم (Saletore et al. 2012) را تشکیل می‌دهند. گزارش‌های اولیه در مورد تغییرات RNA ناشی از مطالعات بر رویRNAهای غیر کد کننده فراوان مانند RNAهای ناقل وRNAهای ریبوزومی در پروکاریوت‌ها و یوکاریوت‌های ساده به دهه‌ها قبل برمی‌گردد (Agris 1996; Marbaniang and Vogel 2016). با این حال، اخیراً، پیشرفت‌های فنی و رویکردهای محاسباتی اصلاح‌شده، هزاران سایت اصلاحی جدید را در گونه‌های مختلف RNA سلولی، از جملهmRNAها وlncRNAها، نشان داده است. در حال حاضر، بیش از ۱۵۰ تغییر متمایز پس از رونویسی در انواع مختلف RNA (Machnicka et al. 2013; Mathlin et al. 2020) شناخته شده است، و تعداد نشانگرهای اپی‌ترنسکریپتومیک کشف شده همواره در حال افزایش است. با این وجود، دانش ما در مورد عملکرد و مکان خاص تغییرات RNA تا کنون کم و ناقص است. بر این اساس، اپی‌ترنسکریپتومیکس زمینه تحقیقاتی است که به شناسایی طیف کامل تغییرات RNA و مشخص کردن آن‌ها در هر دو RNA کد کننده پروتئین و غیر کدکننده پروتئین، جایی که به نظر می‌رسد نقشی فراتر از تنظیم دقیق ساختار و عملکرد RNA دارند، همان‌طور که مطالعات متعدد در مورد سندرم‌های مختلف بیماری برجسته شده است اختصاص داده شده است. در سال ۲۰۱۲، دو گروه مستقل (Dominissini et al. 2012; Meyer et al. 2012) به یک نقشه‌برداری از نوع خاصی از تغییرات (به عنوان مثال، متیلاسیون در موقعیت ششم حلقه پورین در آدنین RNA یا m6A) دست یافتند. این نتایج امکان شناسایی تغییرات RNA را در کل رونوشت نشان داد و زمینه اپی‌ترنسکریپتومیکس را ایجاد کرد (Frye et al. 2016). در دسترس بودن مجموعه‌های بزرگی از سایت‌های تغییر m6A شناسایی‌شده به طور تجربی، توسعه بسیاری از الگوریتم‌های یادگیری تحت نظارت را برای پیش‌بینی سایت‌های اصلاح رونوشت تحریک کرد. در میان سایرین، بهترین عملکرد مربوط به SRAMP [293] (Zhou et al. 2016) بود، که پیش‌بینی‌کننده سایت‌های تغییر m6A بر اساس طبقه‌بندی‌کننده‌های RF چندگانه بود. در سال ۲۰۱۹، Chen et al. رویکرد WHISTLE را توسعه دادند، که یک رویکرد ML است که با یکپارچه‌سازی ویژگی‌های ژنومی متعدد (مانند پروفایل‌های بیان ژن، پروفایل‌های متیلاسیون RNA و شبکه‌های تعامل پروتئین–پروتئین) برای پیش‌بینی مکان‌های تغییر m6A به جای تکیه بر توالی‌های رونوشت، بهتر از سایر الگوریتم‌ها عمل کرد. سال بعد، Dao et al. (2020) iRNA-m6A را ایجاد کردند، که یک طبقه‌بندی مبتنی بر SVM برای شناسایی مکان‌های m6A در بافت‌های متعدد انسان، موش و موش صحرایی است. طبقه‌بندی‌کننده روی مجموعه‌ای از ویژگی‌های بهینه انتخاب شده از سه نوع ویژگی رمزگذاری توالی (مانند ماتریس ویژگی فیزیکی–شیمیایی، کدگذاری دوتایی مونوکلئوتیدی و ویژگی شیمیایی نوکلئوتیدی) که از توالی‌های RNA ورودی محاسبه شده‌اند، کار

می‌کند. اخیراً، Zhang et al. (2021) DNN-m6A را معرفی کردند، که یک روش مبتنی بر DNN است که از روش‌های از قبل موجود در همان کار (یعنی پیش‌بینی مکان‌های تغییر m6A در توالی‌های RNA بافت‌های مختلف پستانداران) بهتر عمل می‌کند. همانطور که مشخص شده است، رونوشت‌ها می‌توانند ویرایش شوند (به عنوان مثال، با جایگزینی پایه)، یا به صورت کووالانسی به مولکول‌های کوچک متصل شوند. مورد اول (یعنی معرفی تغییرات پایه) را می‌توان مستقیماً با استفاده از تکنیک‌های RNA-seq به دلیل عدم تطابق‌هایی که هنگام نگاشت خواندن توالی به ژنوم مرجع ظاهر می‌شود، شناسایی کرد. مورد دوم (یعنی پیوند کووالانسی با مولکول‌های کوچک) برای تشخیص پیچیده‌تر است، زیرا رویکردهای متداول NGS اطلاعات مربوط به اصلاحات شیمیایی را در طول آماده‌سازی نمونه، به‌ویژه در مرحله رونویسی معکوس، پاک می‌کنند. در این مرحله اجباری پروتکل‌های NGS، آنزیمی به نام نسخه‌بردار معکوس[1] (RT) با خواندن رونوشت به عنوان یک الگو و قرار دادن پایه به پایه نوکلئوتید DNA مکمل در رشته cDNA در حال رشد، RNA را به DNA مکمل (cDNA) تبدیل می‌کند. در نتیجه، تغییراتی که در طول سنتز cDNA بر جفت شدن پایه Watson-Crick تأثیر نمی‌گذارند، لغو خواهند شد. سنجش‌های تجربی اختصاص داده شده به تشخیص تغییرات RNA غیرجهشی، مانند رسوب ایمنی با آنتی‌بادی‌های ویژه توسعه یافته‌اند. نکته مهم این است که این روش‌ها را می‌توان برای تعداد محدودی از اصلاحات RNA اعمال کرد، زیرا آن‌ها بر در دسترس بودن آنتی‌بادی‌های مؤثر متکی هستند. روش‌های دیگر از پیامدهای طبیعی تعداد انگشت شماری از تغییرات RNA برای وادار کردن RT به توقف در طول سنتز cDNA یا ایجاد خطا (به عنوان مثال، ترکیب نوکلئوتیدهای غیر مکمل) در cDNA جدید استفاده می‌کنند. در هر دو مورد، اختلال در پردازش RT در به اصطلاح سیگنیچرهای RT قابل مشاهده خواهد بود، که برای یک اصلاح RNA مشخص هستند و با نگاشت مجموعه خوانش‌های توالی که موقعیت RNA اصلاح شده تحت بررسی را در بر می‌گیرد به مرجع قابل مشاهده می‌شود. ژنوم این سیگنیچرهای RT شامل انباشت توالی خواندن با انتهای یکسان، که با موقعیت RNA اصلاح شده که باعث توقف RT شده است، یا در الگوهای متغیر عدم تطابق، که از خواندن نادرست باقیمانده RNA اصلاح شده توسط RT منطبق است، مطابقت دارد. اخیراً، Werner et al. (2020) از یک رویکرد RF برای پیش‌بینی تغییرات RNA بر اساس سیگنیچرهای RT استفاده کردند. نتایج آن‌ها تنوع قوی در میزان موفقیت را نشان می‌دهد که نه تنها به نوع اصلاح RNA بلکه به آنزیم RT خاص مورد استفاده در مرحله سنتز cDNA نیز بستگی دارد. پیش‌بینی مکان‌های اصلاح رونویسی از توالی‌های رونوشت یک کار یادگیری نظارت شده اولیه است. با این حال، برای اکثر اصلاحات RNA، تعداد موارد مثبت شناخته شده (یعنی، مکان‌های اصلاح شده به طور تجربی روی رونوشت‌ها) برای آموزش مدل‌های پیش‌بینی قوی بسیار کمیاب است. اخیرا، Salekin et al. (2020) یک رویکرد مبتنی بر شبکه‌های متخاصم

[1] reverse transcriptase

مولد[1] (GANs) را برای غلبه بر این مشکل با تقلید موفقیت آمیز توزیع داده‌های اساسی و دستیابی به پیش بینی سایت اصلاح RNA توسط یادگیری ویژگی بدون نظارت از توالی‌های RNA ورودی پیشنهاد کردند.

پروتئومیکس: پروتئوم کل مجموعه‌ای از پروتئین‌ها است که پس از بیان توسط یک سیستم بیولوژیکی، که این یک سلول، یک بافت، یک اندام یا یک ارگانیسم است بیان و اصلاح می‌شوند. پروتئوم از یک سیستم به سیستم دیگر (به عنوان مثال، از سلولی به سلول دیگر) تغییر می‌کند. همچنین در طول زمان در همان سیستم تغییر می‌کند و منعکس‌کننده رونوشت زیربنایی و سیستم‌های تنظیمی پیچیده‌ای است که سطوح بیان پروتئین، حرکات درون سلول، تغییرات پس از ترجمه و مشارکت در مسیرهای متابولیک را کنترل می‌کنند. پروتئومیکس رشته‌ای است که پروتئوم‌ها را با استفاده از رویکردهای مقیاس بزرگ مطالعه می‌کند. اصطلاح "پروتئومیکس" برای اولین بار توسط Wilkins et al. (1996) برای نشان دادن "مکمل پروتئینی ژنوم" استفاده شد، اگرچه ماهیت بسیار پویای پروتئوم‌ها، پروتئومیکس را پیچیده‌تر از ژنومیک می‌کند (Lander et al. 2001). پروتئومیکس از تکنیک‌های مختلفی برای بررسی محتوای کلی پروتئین یک سیستم در یک زمان معین و همچنین تجزیه و تحلیل عملکرد پروتئین، تنظیم، تغییرات پس از ترجمه، نوسانات در سطوح بیان، حرکات و تعاملات استفاده می‌کند. به طور خاص، تکنیک‌های متعارف (به عنوان مثال، تکنیک‌های مبتنی بر کروماتوگرافی، وسترن بلات، کریستالوگرافی اشعه ایکس)، پیشرفته (برای نمونه، ریزآرایه‌های پروتئینی، رویکردهای مبتنی بر ژل)، کمی (از قبیل، برچسب زدن پروتئین ایزوتوپی)، و توان عملیاتی بالا (مانند، روش‌های مبتنی بر طیف‌سنجی جرمی) تکنیک‌های در دسترس برای بررسی پروتئوم‌ها هستند (Aslam et al. 2017). طیف سنجی جرمی (MS) تکنیک پیشرو با توان بالا برای مطالعه مخلوط پروتئین است. برای تعیین وزن مولکولی پروتئین‌ها از طریق اندازه گیری نسبت جرم مولکولی به شارژ یا بار (m/z) استفاده می‌شود. در طیف‌سنجی جرمی، مولکول‌ها به یون‌های فاز گاز تبدیل می‌شوند، سپس از یک تحلیلگر جرم برای جداسازی یون‌ها در میدان‌های الکتریکی یا مغناطیسی با مقادیر m/z آن‌ها استفاده می‌شود و در نهایت مقدار هر گونه یون اندازه‌گیری می‌شود (Yates et al. 2011). در طیف‌سنجی جرمی پشت سر هم[2] (MS/MS) (van Agthoven et al. 2019) دو یا چند آنالایزر جرم با هم جفت می‌شوند تا توانایی شناسایی و جداسازی یون‌های متمایز با وزن مولکولی مشابه را افزایش دهند. بسته به مولکول در دست، رویکردهای مختلفی برای جداسازی یون اتخاذ می‌شود: کروماتوگرافی مایع (LC) و به دنبال آن طیف سنجی جرمی[3] (LC-MS) و LC-MS پشت سر هم[4] (LC-MS/MS) تکنیک‌های شیمی تحلیلی هستند که در آن‌ها از LC برای جداسازی مخلوط‌ها، با گونه‌های مولکولی متعدد، و MS یا MS پشت سر هم استفاده شده برای شناسایی گونه‌های یونی استفاده می‌شود (Zhang et al. 2014). هم زمان matrix-assisted laser desorption ionization-time of flight mass spectrometer (MALDI-TOF) (Hillenkamp et al. 1991) (MALDI-TOF) و MALDI-TOF پشت

[1] Generative Adversarial Networks
[2] Tandem mass spectrometry
[3] Liquid chromatography followed by mass spectrometry
[4] Tandem LC-MS

سر هم (MALDI-TOF/TOF) (Gogichaeva et al. 2007) یک تکنیک یونیزاسیون (MALDI) تشکیل می‌دهند، که یون‌هایی را از مولکول‌های بزرگ، با حداقل تکه تکه شدن ایجاد می‌کند به یک طیف سنج جرمی (TOF)، که زمان رسیدن یون‌ها به آشکارساز، زمانی که از طریق همان پتانسیل یک میدان الکتریکی شتاب می‌گیرند را اندازه گیری می‌کند. پروتئومیکس سطوح مختلفی از بررسی، مانند ساختار اولیه پروتئین (به عنوان مثال، تشخیص همسانی و خانواده‌های پروتئین، تشخیص موتیف، هم‌ترازی چند توالی، طبقه بندی توالی)، ساختار ثانویه (SS) (به عنوان مثال، شناسایی زیرساختارهای محلی)، ساختار سه بعدی (به عنوان مثال، پیش بینی تاشو، مقایسه ساختار، طبقه بندی دامنه، شناسایی الگوهای سه بعدی، تجزیه و تحلیل ویژگی‌های شیمیایی و توپولوژیکی)؛ عملکرد پروتئین و برهم کنش‌های عملکردی (به عنوان مثال، طبقه‌بندی عملکرد، پیش‌بینی مکان‌های فعال و باقیمانده‌های حیاتی، پیش‌بینی مکان‌های اتصال، تجزیه و تحلیل بسترها و تعدیل‌کننده‌ها، تجزیه و تحلیل روابط ساختار–عملکرد، طراحی دارو) را در بر می‌گیرد. به طور خاص، مطالعه SS پروتئین‌ها معمولاً به دقت Q3 یا Q8، به ترتیب به عنوان درصد باقیمانده‌هایی که برای آن ۳ حالت کلی (هلیکس، رشته و کویل) یا فرهنگ لغت خوب با هشت تا از این حالت‌ها تعریف می‌شود اشاره دارد (Kabsch and Sander 1983). در سال ۲۰۰۳، Kim and Park روش SVMpsi را بر اساس SVMs برای به حداکثر رساندن اندازه گیری Q3 پیشنهاد کردند. در سال ۲۰۰۵، Garrow et al. برنامه TMB-Hunt را معرفی کردند، برنامه‌ای که از یک الگوریتم k-NN برای طبقه‌بندی توالی‌های پروتئین به‌عنوان trans-membrane beta-barrel (TMB) یا غیر TMBاستفاده می‌کند. روش‌های مبتنی بر DL برای مطالعه SS جدیدتر هستند، مانند DeepCNF (Wang et al. 2016)، که از CNN برای تشخیص رابطه پیچیده بین توالی و ساختار پروتئین‌ها و SSREDNها (Wang et al. 2017)، بر اساس معماری شبکه‌های رمزگذار–رمزگشای عمیق بازگشتی استفاده می‌کردند، تا ارتباط غیرخطی پیچیده بین ویژگی‌های پروتئین و SS و همچنین برهمکنش‌های بین باقیمانده‌های پیوسته را ثبت کنند. در همان سال‌ها، Sønderby and Winther (2015) یک LSTM، و Fang et al. (2018) یک شبکه عصبی Deep Inception-Inside-Inception را برای پیش‌بینی SS که از توالی‌های اسید آمینه شروع می‌شد، پیاده‌سازی کردند. روش‌های ML همچنین برای ساختار کلی ثالث (Bonnel and Marteau 2012)، زوایای پیچشی (Faraggi et al. 2012; Fang et al. 2018) و پیش‌بینی حلقه (Jacobson et al. 2004; Nguyen et al. 2017) استفاده شده‌اند. قابل ذکر است، در نسخه ۲۰۱۸ ارزیابی انتقادی پیش‌بینی ساختار پروتئین[1] (CASP)، یک نرم‌افزار مبتنی بر NN که توسط تیم تحقیقاتی هوش مصنوعی در Google DeepMind، به نام AlphaFold، توسعه یافته است، از همه روش‌های مشارکت‌کننده در پیش‌بینی دقیق تاخوردگی کلی و فاصله بین جفت‌های باقیمانده بهتر عمل کرد (Senior et al. 2020). در نسخه ۲۰۲۰، نسخه جدیدی از روش AlphaFold ارائه شد که از یک مدل کاملاً متفاوت استفاده می‌کرد (Jumper et al. 2021) و نتایج بی‌سابقه‌ای را ارائه می‌کرد که در کل جامعه محققان علوم زیستی طنین‌انداز شده است. به طور

[1] Critical Assessment of Protein Structure Prediction

خاص، آخرین معماری AlphaFold شامل یک ماژول یادگیری خود نظارت جدید (ماژول Evoformer) مبتنی بر دو ترانسفورماتور (a two tower architecture) برای جاسازی دو قطعه از اطلاعات اصلی زیر است که سیستم سعی می‌کند از پایگاه‌های داده عمومی جمع‌آوری کند. پایگاه داده‌ای که از یک توالی اسید آمینه‌ای، از قبیل: ۱- Multiple Sequence Alignment (MSA) و ۲- فهرستی از ساختارها یا الگوهای بالقوه مشابه شروع می‌شود. در ماژولEvoformer، دو نمایش مبتنی بر ترانسفورماتور از اطلاعات توالی و ساختار با یکدیگر در سراسر شبکه عصبی برای بسیاری از چرخه‌های یادگیری (۴۸ چرخه) ارتباط برقرار می‌کنند تا زمانی که به نمایش‌های جامد، که به آخرین ماژول (ماژول ساختار) منتقل می‌شود برسند. در نهایت، NN ماژول ساختار، ساختار پروتئین پیش‌بینی‌شده را با مپینگ (نگاشت) بازنمایی‌های انتزاعی دریافت‌شده از ماژول Evoformer به مختصات اتم‌های سه‌بعدی واقعی، خروجی می‌دهد (Jumper et al. 2021). نکته قابل توجه این است که کد منبع AlphaFold بلافاصله پس از انتشار به صورت رایگان در اختیار جامعه علمی قرار گرفته است و توانایی برجسته آن برای پیش‌بینی ساختارهای پروتئینی برای ایجاد یک پایگاه داده رایگان از ساختارها که تمام ۲۰۰۰۰ پروتئین انسانی به اضافه پروتئوم‌های کامل چند موجود دیگر که از نظر بیولوژیکی مهم هستند را پوشش می‌دهد استفاده شده است. پیش‌بینی عملکرد پروتئین اساساً شامل طبقه‌بندی عملکردهای پروتئینی مرتبط با ویژگی‌های ساختاری مختلف است. حاشیه نویسی عملکرد تجربی گران و زمان‌بر است و اطلاعات مربوط به دامین‌ها، موتیف‌ها، خانواده‌ها، اثرات متقابل و پیوستگی‌ها بزرگ و پیچیده است. به این دلایل، تجزیه و تحلیل این اطلاعات بدون رویکردهای محاسباتی غیرممکن است. در سال ۲۰۱۷، Liu یک رویکرد RNN (شبکه‌های عصبی تکراری[1]) کارآمد را برای طبقه‌بندی عملکردهای پروتئین مستقیماً از توالی‌های اولیه پیشنهاد کرد. رویکردهای DeepFunc (Zhang et al. 2019) و DeepPred (Rifaioglu et al. 2019) دو رویکرد DL اخیر برای پیش‌بینی عملکرد پروتئین هستند. نتایج امیدوارکننده آن‌ها نشان می‌دهد که DL به دلیل پیچیدگی کار و اندازه و تنوع مجموعه داده‌ها، پتانسیل قابل توجهی در پیش بینی عملکرد پروتئین دارد (Kelchtermans et al. 2013; Sonsare and Gunavathi 2019). فعل و انفعالات فیزیکی پروتئین، از جمله برهمکنش‌های پروتئین-پروتئین، برهمکنش‌های پروتئین-دارو، و اتصال پروتئین‌ها به DNA یاRNA ، تعیین کننده‌های اصلی عملکرد سلول هستند و ابزارهای موثر برای بررسی سیستماتیک آن‌ها برای به دست آوردن درک کاملی از زیست‌شناسی سلولی مطلوب و مکانیسم‌های بیماری است. علیرغم پیشرفت‌های تکنولوژیکی، بررسی تجربی برهمکنش‌های پروتئین-پروتئین هنوز گران، پرزحمت و در مقیاس محدود است، بنابراین از تلاش‌های بی‌طرفانه و سیستماتیک جلوگیری می‌کند. در سال‌های گذشته، انباشته شدن داده‌های توالی و ساختار، استفاده از روش‌های محاسباتی را برای رسیدگی به تحقیقات در مقیاس بزرگ برهم‌کنش‌های پروتئین-پروتئین ارتقا داده است. در سال ۲۰۰۴، Koike and Takagi یک رویکرد SVM برای پیش‌بینی برهم‌کنش‌های پروتئین-پروتئین بر اساس چندین ویژگی پروتئین، مانند حوزه‌های عملکردی مشروح شده، پیشنهاد کرد. در سال ۲۰۱۶، An et al. روش RVM-BiGP را پیشنهاد کردند،

[1] Recurrent Neural Networks

که یک روش ML برای پیش‌بینی برهم‌کنش‌های پروتئین-پروتئین از توالی پروتئین، بر اساس طبقه‌بندی‌کننده RVM همراه با احتمالات Bi-gram، برای نمایش ویژگی‌های توالی پروتئین، و PCA، برای کاهش ابعاد است. در سال ۲۰۱۸، Huang et al. یک روش DL مبتنی بر یک NN و یک معماری رمزگذار خودکار برای تکمیل شبکه‌های برهمکنش پروتئین-پروتئین پراکنده و قطع شده از طریق پیش‌بینی برهم‌کنش‌های از دست رفته، پیشنهاد کردند. در میان برهمکنش‌های پروتئین-DNA، فاکتورهای رونویسی که به نواحی تنظیم‌کننده در DNA متصل می‌شوند، با تنظیم برنامه‌های رونویسی خاص سلول، و تعدیل بیان ژن در پاسخ به محرک‌های داخلی و خارجی، نقش اصلی را در تنظیم فرآیندهای سلولی مختلف بازی می‌کنند. در سال ۲۰۱۷، Qin and Feng روش TFImpute را توسعه دادند، که یک DNN است که می‌تواند الگوهای اتصال خاص سلول را با یادگیری از داده‌های تجربی مربوط به فاکتورهای رونویسی مختلف و خطوط سلولی پیش بینی کند. در سال ۲۰۱۸، Shen et al. روش KEGRU را پیشنهاد کردند، که یک مدل DL برای پیش‌بینی مکان‌های اتصال TF بر اساس شبکه واحد بازگشتی دروازه‌دار دوطرفه[1] همراه با تعبیه k-mer توالی‌های DNA است. اخیرا، Rives et al. (2021) یک ترانسفورماتور را بر روی ۸۶ میلیارد اسید آمینه در ۲۵۰ میلیون توالی پروتئین آموزش دادند. این مدل از پیش آموزش‌دیده بدون نظارت، نمایشی چند مقیاسی از ساختارهای پروتئینی را ارائه می‌کند که حاوی اطلاعاتی درباره سازمان‌دهی زنجیره‌های ثانویه و سوم، همسانی، اتصال‌ها و اثرات جهشی است. نمایش‌های آموخته‌شده همچنین می‌توانند برای کاربردهای امیدوارکننده مانند تولید توالی‌های جدید و طراحی پروتئین‌های کاربردی مورد استفاده قرار گیرند.

متابولومیکس: متابولومیکس رشته‌ای است که هدف آن مطالعه مشخصات جامع متابولیت‌ها در یک سلول، یک بافت یا یک ارگانیسم کامل است. متابولیت‌ها مولکول‌های کوچکی هستند که در طی فرآیندهای متابولیک تولید می‌شوند و کل مجموعه تولید شده توسط یک سلول خاص (متابولوم) یک بازخوانی عملکردی از فعالیت بیوشیمیایی سلولی را ارائه می‌دهد (Patti et al. 2012). این رشته جدید در ابتدای این قرن ظهور کرد و به لطف پیشرفت در فناوری ابزار به سرعت رشد کرد. مطالعات متابولومیکس می‌تواند بر روی مجموعه خاصی از متابولیت‌ها و مسیرهای خاصی که در آن‌ها شرکت می‌کنند متمرکز شود (به عنوان متابولومیکس هدفدار یا هدفمند شناخته می‌شود)، یا با هدف تعیین پروفایل متابولیت جهانی (Zamboni et al. 2015) (به عنوان متابولومیکس غیرهدفمند یا تفنگ ساچمه‌ای یا کشفی[2] شناخته می‌شود). آزمایش‌های متابولومیکس هدفمند را می‌توان با استفاده از طیف‌سنجی جرمی (MS) و رزونانس مغناطیسی هسته‌ای (NMR) انجام داد، در حالی که LC-MS تکنیک انتخابی برای متابولومیکس غیر هدفمند است (Patti et al. 2012). در متابولومیکس، تکنیک‌های ML و DL عمدتاً برای پیش پردازش داده‌ها (مانند شناسایی پیک و ادغام پیک[3])، و شناسایی و کمی سازی ترکیبات استفاده شده است (Pomyen et al. 2020). در سال ۲۰۰۸، Yuan et

[1] Bidirectional Gated Recurrent Unit network
[2] Untargeted or shotgun or discovery metabolomics
[3] Peak identification and peak integration

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

al. از LDA برای اکتشاف داده‌های متابولومیکس استفاده کردند. در همان زمان، Cavill et al. (2009) یک الگوریتم ژنتیکی را برای تجزیه و تحلیل طیف NMR ادرار موش‌ها برای طبقه‌بندی سمیت کبد و کلیه اجرا کردند. در سال ۲۰۱۲، Hao et al. روش BATMAN را را پیشنهاد کردند که یک تحلیلگر متابولیت خودکار بیزی برای طیف‌های NMR است. فرآیند تنظیم دستی پیک (peak)، تراز (alignment) و باینینگ (binning) می‌تواند زمان بر باشد و می‌تواند مصنوعات یا خطاهایی را معرفی کند، بنابراین اتخاذ ML در این زمینه به رویکردهای کلاسیک ارجحیت دارد. روش BAYESIL (Ravanbakhsh et al. 2015) یک سیستم کاملاً خودکار و در دسترس عموم برای خودکار کردن شناسایی و تعیین کمیت ترکیبات از طیف NMR مخلوط‌های پیچیده، از جمله نمونه‌های بیولوژیکی است. به طور خاص، این الگوریتم یک سیستم دکانولوشن (deconvolution) طیفی است که تطابق طیفی را به عنوان یک مسئله استنتاج مونت کارلو در یک مدل گرافیکی احتمالی مشاهده می‌کند، که به سرعت طیف NMRورودی را با محتمل‌ترین مشخصات متابولیک تقریب می‌کند. پس از چندین مرحله پردازش طیفی، BAYESIL طیف داده شده را با یک کتابخانه ترکیبی مرجع حاوی سیگنیچرهای بیش از ۶۰ متابولیت جفت می‌کند. فرآیند دکانولوشن قادر است هم هویت و هم کمیت متابولیت‌های موجود در یک مخلوط پیچیده تحت بررسی، مانند سیال زیستی فردی (به طور خاص، سرم یا مایع مغزی نخاعی) را به دست آورد. در پژوهشی Alakwaa et al. (2018) از شبکه‌های پیش‌خور[1]، یک چارچوب DL، برای پیش‌بینی وضعیت گیرنده استروژن از داده‌های متابولومیکس سرطان پستان استفاده کردند. نویسندگان رویکرد DL خود را با شش روش دیگر مبتنی بر ML مقایسه کردند، که همگی بر روی گروهی از ۲۷۱ بافت سرطان پستان (یعنی ۲۰۴ گیرنده استروژن مثبت و ۶۷ گیرنده استروژن منفی) آموزش داده شدند که توسط کروماتوگرافی گازی و به دنبال آن time-of-flight mass spectrometry ارزیابی شدند و دریافتند که رویکرد DL طبقه بندی بهتری برای کار ارایه کرده است. تفسیر بیولوژیکی لایه‌های پنهان DL مسیرهای مهمی مانند متابولیسم کربن مرکزی در سرطان سینه و متابولیسم گلوتاتیون را نشان داد و اجازه داد آنزیم‌های بیوسنتزی درگیر در مسیرهای متابولومیک نقشه‌برداری شوند.

کاربرد هوش مصنوعی در کشاورزی

تشخیص چهره (Voulodimos et al. 2018)، پیش‌بینی سرطان در بافت (Paeng et al. 2017) و تجزیه و تحلیل شار متابولیک (Wu et al. 2016) تنها چند نمونه از پیشرفت‌های قابل توجهی است که با رویکردهای هوش مصنوعی انجام شده است، و پتانسیل دستیابی به انقلابی مشابه در زمینه کشاورزی وجود دارد. بر اساس گزارش منتشر شده توسط سازمان خواربار و کشاورزی ملل متحد (فائو)، جمعیت جهان تا سال ۲۰۵۰ به بیش از ۹ میلیارد نفر خواهد رسید (FAO 2022). افزایش جمعیت در نهایت بر توانایی بخش کشاورزی برای تامین غذا فشار می‌آورد. برای تغذیه جمعیت رو به رشد جهان و پیشرفت اقتصاد کشورها، کشاورزی

[1] Feed-forward networks

ضروری است (Eli-Chukwu 2019). این منبع درآمد قابل توجهی برای تعدادی از کشورها است. کشاورزی حدود ۳۸ درصد از کل سطح زمین را اشغال می‌کند (FAO 2022). اکثر فعالیت‌های کشاورزی در حال حاضر دستی هستند و کشاورزی ممکن است به طور قابل توجهی از اتوماسیون از نظر عملکرد به دست آمده و نهاده‌های سرمایه گذاری شده بهره‌مند شود. اجرای پیشرفت‌های تکنولوژیک در کشاورزی ممکن است به تغییر در اقتصاد روستایی و معیشت روستاییان کمک کند (Mogili and Deepak 2018; Shah et al. 2019). تکنیک‌های کشاورزی به طور کلی برای غلبه بر موانع مختلفی از جمله آلودگی به آفات، استفاده ناکارآمد از سموم و کودهای شیمیایی، علف‌های هرز، خشکسالی و عدم وجود سیستم آبیاری کافی، برداشت ناکارآمد، ذخیره‌سازی و در نهایت بازاریابی طراحی شده‌اند. بخش کشاورزی را می‌توان با مداخله هوش مصنوعی در زمینه‌های مدیریت خاک، ارزیابی نیاز آب، نقشه برداری دقیق نیاز کود، آفت کش، حشره کش، نیاز علف کش، علوم دامی، پیش بینی عملکرد و مدیریت کلی محصول تغییر داد (Talaviya et al. 2020; Aggarwal and Singh 2021; Klyushin and Tymoshenko 2021). با پیشرفت فناوری مبتنی بر هوش مصنوعی، هواپیماهای بدون سرنشین و ربات‌ها برای بهبود نظارت بر زمان واقعی محصولات، برداشت و پردازش بعدی استفاده می‌شوند (Liakos et al. 2018). تکنیک‌های هوش مصنوعی و ML در حال حاضر توسط شرکت‌های بیوتکنولوژی برای طراحی و آموزش ربات‌های مستقلی که قادر به انجام فعالیت‌های کشاورزی کلیدی مانند برداشت محصول با سرعت بسیار بیشتری نسبت به روش‌های سنتی هستند، استفاده می‌شوند (Talaviya et al. 2020). داده‌های جمع آوری شده توسط هواپیماهای بدون سرنشین با استفاده از تکنیک‌های یادگیری عمیق و بینایی کامپیوتری پردازش و ارزیابی می‌شوند (Linaza et al. 2021). رویکردهای یادگیری ماشین به دسترسی و پیش‌بینی طیف گسترده‌ای از متغیرهای محیطی که بر تولید کشاورزی تأثیر می‌گذارند، مانند نوسانات آب و هوایی و خشکسالی به کشورها کمک می‌کند (Dutta Majumder et al. 2007; Talaviya et al. 2020; Ben Ayed and Hanana 2021). راه حل‌های مبتنی بر هوش مصنوعی در صنعت کشاورزی به بهبود کارایی و کنترل جنبه‌های متعددی مانند عملکرد محصول، مشخصات خاک، آبیاری محصول، سنجش محتوا، وجین علف‌های هرز و نظارت بر محصول کمک می‌کنند (Kim and Gilley 2008; Talaviya et al. 2020). بازرسی خصوصیات مورفولوژیکی سنتی و قدیمی‌تر زمان‌بر، مستعد خطا و پرهزینه است. روش بینایی ماشین[1] ممکن است به راحتی در روش‌های کشاورزی به کار رود، که می‌تواند در عین اینکه صحیح‌تر و دقیق‌تر است، روند را تسریع و ساده‌تر کند (Linaza et al. 2021). شناسایی و انتخاب واریته‌های بهبودیافته ممکن است با استفاده از امتیازدهی خودکار غیرتهاجمی و سریع ویژگی‌های گیاهی مختلف از طریق روش‌های فنوتیپ‌سازی با توان عملیاتی بالا، روند را تسریع و آسان‌تر کند (Matias et al. 2020). با توجه به ابزارهای هوش مصنوعی و اینترنت اشیاء، اکنون می‌توان از فناوری‌های هوشمند و پهپادها برای چندین فعالیت کشاورزی استفاده کرد (Spanaki et al. 2022). پیشرفت‌های اخیر در طراحی الگوریتم‌های مبتنی بر DL و ML برای تخمین قیمت محصولات کشاورزی ممکن است

[1] Machine vision

کشاورزان را قادر سازد که بازدهی بالاتری از کار و سرمایه‌گذاری خود دریافت کنند (Mahto et al. 2021). برای آبیاری موثر، شبکه‌های عصبی مصنوعی، منطق فازی و الگوریتم‌های فراابتکاری اخیرا توسعه یافته‌اند (Jha et al. 2019; Pazouki et al. 2021). طبق یک مطالعه، شبکه‌های عصبی پیچیده یا کانولوشن[1] (CNN) که چندین متغیر محیطی را در نظر می‌گیرد، یکی از قابل اعتمادترین الگوریتم‌های ML برای تخمین عملکرد سویا و ذرت است (Ju et al. 2019). پیشرفت‌های اخیر در حسگرهای زیستی مبتنی بر هوش مصنوعی برای تشخیص زودهنگام بیماری در گیاهان زراعی، حتی در گیاهان بدون علامت، این پتانسیل را دارد که تا حد زیادی از دست دادن محصول ناشی از عوامل استرس‌زای زیستی را به حداقل برساند (Ali et al. 2021). فناوری‌های پهپاد مبتنی بر هوش مصنوعی مانند EfficientNetV2 که برای شناسایی و طبقه‌بندی بیماری‌های گیاهی با صحت و دقت ۹۹,۹۹ درصد و ۹۹,۶۳ درصد طراحی شده‌اند، یکی از فناوری‌های خودکار امیدوارکننده برای نظارت بر سلامت گیاه در صرفه‌جویی در زمان و هزینه است (Albattah et al. 2022). برای تشخیص بیماری لکه باکتریایی در گیاهان، یک مدل هوش مصنوعی ترکیبی مبتنی بر رمزگذار خودکار کانولوشن (CAE) و CNN نیز به ترتیب ۹۹/۳۵٪ و ۹۹/۳۸٪ در دوره‌های آموزش و آزمایش به دست آورده است (Bedi and Gole 2021). استفاده از هوش مصنوعی ممکن است شناسایی اهداف بالقوه در داده‌های ژنوم بزرگ را برای دستکاری ژنتیکی و طراحی محرک‌های مصنوعی موثر در تلاش برای بهبود صفات زراعی در گیاهان ساده‌تر کند (Pandey and Chaudhary 2016; Pandey and Chaudhary 2021). نیازهای رو به رشد برای کشاورزی هوشمند منجر به پیشرفت‌های اساسی در زمینه پیش‌گویی و پیش‌بینی کشاورزی مبتنی بر هوش مصنوعی شده است که بهره‌وری محصول را تا حد زیادی بهبود بخشیده است (Linaza et al. 2021). تلاش مشابهی در یک مطالعه انجام شده که در آن مجموعه داده‌های تصویر با استفاده از الگوریتم‌های هوش مصنوعی، یعنی ANN و پلت‌فرم‌های مبتنی بر الگوریتم ژنتیک (GA)، برای پیش‌بینی عملکرد محصول به شیوه‌ای بهینه مورد تجزیه و تحلیل قرار گرفتند (Sharma et al. 2022). در طول دوره آموزش، مدل حداکثر دقت اعتبار ۹۸/۱۹٪ را به دست آورد، در حالی که حداکثر دقت ۹۷/۷۵ ٪ در طول دوره آزمایش به دست آمد (Sharma et al. 2022). این مدل تحت محدودیت‌های منابع محدود و داده‌های کمتر به طور موثر عمل کرد و نتایج بهینه را تولید کرد (Sharma et al. 2022). در مطالعه مهم دیگری، یک روش جدید برای پیش‌بینی عملکرد کشاورزی در محصولات گلخانه‌ای با استفاده از الگوریتم‌های شبکه‌های عصبی تکراری[2] (RNNs) و شبکه کانولوشن موقت یا زمانی[3] (TCN) پیشنهاد شد (Gong et al. 2021). براساس داده‌های محیطی و تولید قبلی، این رویکرد می‌تواند برای تخمین عملکرد محصولات گلخانه‌ای با دقت بیشتری نسبت به همتایان استاندارد MLو یادگیری عمیق خود استفاده شود (Gong et al. 2021). علاوه بر این، این بررسی تجربی همچنین اهمیت حیاتی مجموعه داده‌های عملکرد قبلی را در پیش‌بینی صحیح بهره‌وری محصول آینده نشان داده است (Gong et al. 2021; Mohd et al.

[1] Convolutional Neural Networks
[2] Recurrent Neural Networks
[3] Temporal convolutional network

2022). چندین میلیون نفر در کشورهای در حال توسعه با جلوگیری و ترکیب محصولات پربازده، کودهای مصنوعی و آب از انقلاب سبز سود برده‌اند. با این حال، به دلیل استفاده نادرست گسترده از علف کش‌ها، آفت کش‌ها و کودها، انقلاب سبز را نمی‌توان به طور کامل "سبز" در نظر گرفت. روش‌های خاص برای محصولات پربازده معمولاً به مقدار زیادی از مواد شیمیایی کشاورزی و آب نیاز دارند (Seyhan et al. 2021). رویکردهای مبتنی بر هوش مصنوعی برای کاهش وابستگی به مواد شیمیایی زراعی مضر و دستیابی به وضعیت پایدار در کشاورزی در حال توسعه هستند (Wang et al. 2020). برای بهینه‌سازی منابع کشاورزی، یک سیستم کنترل به کمک سنجش از دور[1] (RSCS) توسعه یافته است (Zhou et al. 2022). این روش از فناوری AL و ML برای بهبود پایداری محیطی و در عین حال برنامه‌ریزی توسعه محصولات کشاورزی جدید استفاده می‌کند. هنگامی که با تکنیک‌های دیگر تجزیه و تحلیل شد، یافته‌ها نشان داد که RSCS بالاترین دقت، عملکرد، سرعت انتقال داده، بهره وری، مدیریت آبیاری و نسبت انتشار دی اکسید کربن را با به ترتیب ۹۵/۱، ۹۶/۳۵، ۹۲/۳، ۹۴/۲، ۹۴/۷ و ۲۱/۵ درصد نشان می‌دهد (Zhou et al. 2022). بنابراین، مدل‌های هوش مصنوعی پتانسیل مدیریت محصولات کشاورزی و بهره‌وری را به شیوه‌ای «سبز» دارند. در مطالعه دیگری، یک سمپاش هوشمند مبتنی بر هوش مصنوعی و بینایی ماشینی برای سمپاشی علف‌کش‌ها به‌طور خاص برای اهداف علف‌های هرز توسعه داده شد، بنابراین استفاده بیش از حد از علف‌کش‌ها و آلودگی محیط زیست کاهش می‌یابد. این فناوری پیچیده مفهوم پیشرفته تشخیص علف‌های هرز، یک روش سمپاشی سریع و دقیق منحصر به فرد و یک مدل نقشه برداری علف‌های هرز را به ترتیب با دقت ۷۱ و ۷۸ درصد و یادآوری ترکیب می‌کند (Partel et al. 2021). به دلیل محدود بودن تکنیک‌های جمع‌آوری و عدم ادغام منابع داده‌های متنوع، جمع‌آوری داده‌ها از مناطق کشاورزی مرتبط با هیدراتاسیون خاک، کیفیت محصول یا هجوم حشرات اغلب به تجزیه و تحلیل دستی بستگی دارد. در همین حال، با دیجیتالی‌تر شدن صنعت، ترکیب سنجش از دور برای غربالگری رایانه‌ای و تکنیک‌های تحلیلی با مجموعه داده‌ها برای مطالعات خاک، پیش‌بینی آب‌وهوا و غیره، و مدل‌های پیشرفته هوش مصنوعی نیاز به مواد شیمیایی کشاورزی را کاهش می‌دهد (Linaza et al. 2021). در این راستا، برنامه عملیاتی قابل توجه NaLamKI که به دنبال ایجاد نرم افزار دسترسی باز مبتنی بر هوش مصنوعی است که می‌تواند کمک زیادی به صنعت کشاورزی کند، از بودجه ساخت و استفاده دریافت کرده است. این طرح به دنبال توسعه مجموعه‌های داده با ترکیب اطلاعات از حسگرهای مختلف به منظور بهینه‌سازی شیوه‌های کشاورزی مختلف با کمک فناوری‌های هوش مصنوعی و ML است (Paraforos et al. 2016; Linaza et al. 2021). ابتکارات دولتی مشابه در تعداد زیاد مورد نیاز است تا کشاورزان را وادار به تطبیق هوش مصنوعی در مقیاس بزرگتر کند. در کشاورزی، ادغام ویژگی‌های دقیق مبتنی بر تصویر با داده‌های اومیکس ممکن است به یافتن ویژگی‌های حیاتی درگیر در تحمل استرس و مکانیسم‌های سازگاری (Marchetti et al. 2019) کمک کند، و همچنین به توسعه محصولات مقاوم به آب و هوا کمک نماید. کشاورزان با توجه به تطبیق فناوری مبتنی بر هوش مصنوعی، قادر خواهند بود با نهاده کمتر محصول بیشتری

[1] Remote sensing assisted control system

داشته باشند، کیفیت محصول خود را افزایش دهند و از زمان سریعتری برای بازاریابی محصولات برداشت شده خود اطمینان حاصل کنند (Linaza et al. 2021). اگرچه هوش مصنوعی نسل اول را می‌توان در بررسی و طبقه بندی داده‌های اومیکس استفاده کرد، اما برای رسیدگی به مشکلات خاص مربوط به مجموعه داده‌های تک اومیکس[1] بدون ادغام داده‌های حاصل از سایر روش‌ها طراحی شده است (Harfouche et al. 2019; Linaza et al. 2021). در بیوتکنولوژی کشاورزی، هوش مصنوعی نسل بعدی اساساً برای بهبود پویا و مدیریت مجموعه داده‌های چند اومیکس[2] بزرگ علاوه بر پیش‌بینی ارزش اصلاحی صفات پیچیده در شرایط مختلف محیطی پیش‌بینی شده است (Harfouche et al. 2019).

کاربرد هوش مصنوعی در حیوانات اهلی

سیستم‌های تولید دام در سال‌های اخیر از نظر بهره‌وری به ازای هر حیوان یا واحد زمین یا نیروی کار تشدید شده‌اند (Thornton 2010). این تشدید شیوه‌های مختلف دامپروری همچنین منجر به نگرانی‌های اجتماعی از نظر پذیرش مصرف کنندگان، در مورد امنیت غذایی و تغذیه، ایمنی مواد غذایی، پایداری، رفاه حیوانات، سلامت حیوانات و سلامت انسان شده است (Scholten et al. 2013). دام در سطح جهان بزرگترین مصرف کننده منابع زمین است، زیرا تقریباً ۸۰٪ از کل زمین‌های کشاورزی به تولید خوراک و مرتع اختصاص دارد، علیرغم نسبت زیاد زمین و همچنین انرژی (و آب) که برای تولید حیوانات استفاده می‌شود. پروتئین، با تبدیل محصولاتی که ۵۰ تا ۶۰ درصد به تغذیه دام، برای تولید خوراک در زمین‌های زراعی رو به کاهش به ازای هر نفر اختصاص داده شده‌اند (D'Agaro et al. 2021). در طول کل زنجیره ارزش، انتشار مستقیم و غیرمستقیم مربوط به دام ۱۶٪/۵ از کل یا ۸/۱ درصد معادل GtCO2 بر اساس FAO است. علاوه بر این، تغییرات اقلیمی، همراه با رویدادهای شدید آب و هوایی شدیدتر مانند خشکسالی و بارندگی‌های شدید، ممکن است منجر به اثرات نامطلوب بر سلامت و رفاه حیوانات و همچنین افزایش انتشار گازهای گلخانه ای، کاهش کیفیت و کمیت خوراک و غذا و در نتیجه سلامت انسان شود. برای مقابله با این چالش‌ها، سیستم‌های کشاورزی باید با استفاده از منابع کمتر، بیشتر تولید کنند و بر کاهش ضایعات مواد غذایی با زنجیره‌های تولید بسته‌تر تمرکز کنند. فن آوری‌های نوظهور مانند "پرورش دام دقیق[3]" می‌تواند برای سلامت و رفاه حیوانات همراه با اطلاعات در مورد جنبه‌های زیست محیطی و اقتصادی در تولید کشاورزی مفید باشد. نظارت مبتنی بر حسگر[4] در دامپروری می‌تواند داده‌های ورودی کمی بهینه[5] را برای تجزیه و تحلیل چرخه زندگی ارائه دهد. ترکیبی با فناوری‌های ارتباطات و اطلاعات، پایه‌ای برای کشاورزی مدرن می‌سازد (D'Agaro et al. 2021). تجزیه و تحلیل چرخه حیات (LCA) سیستم‌های تولید کشاورزی با توجه به محیط

[1] Single-omics datasets
[2] Multi-omics datasets
[3] Precision Livestock Farming
[4] Sensor-based monitoring
[5] Optimal quantitative input data

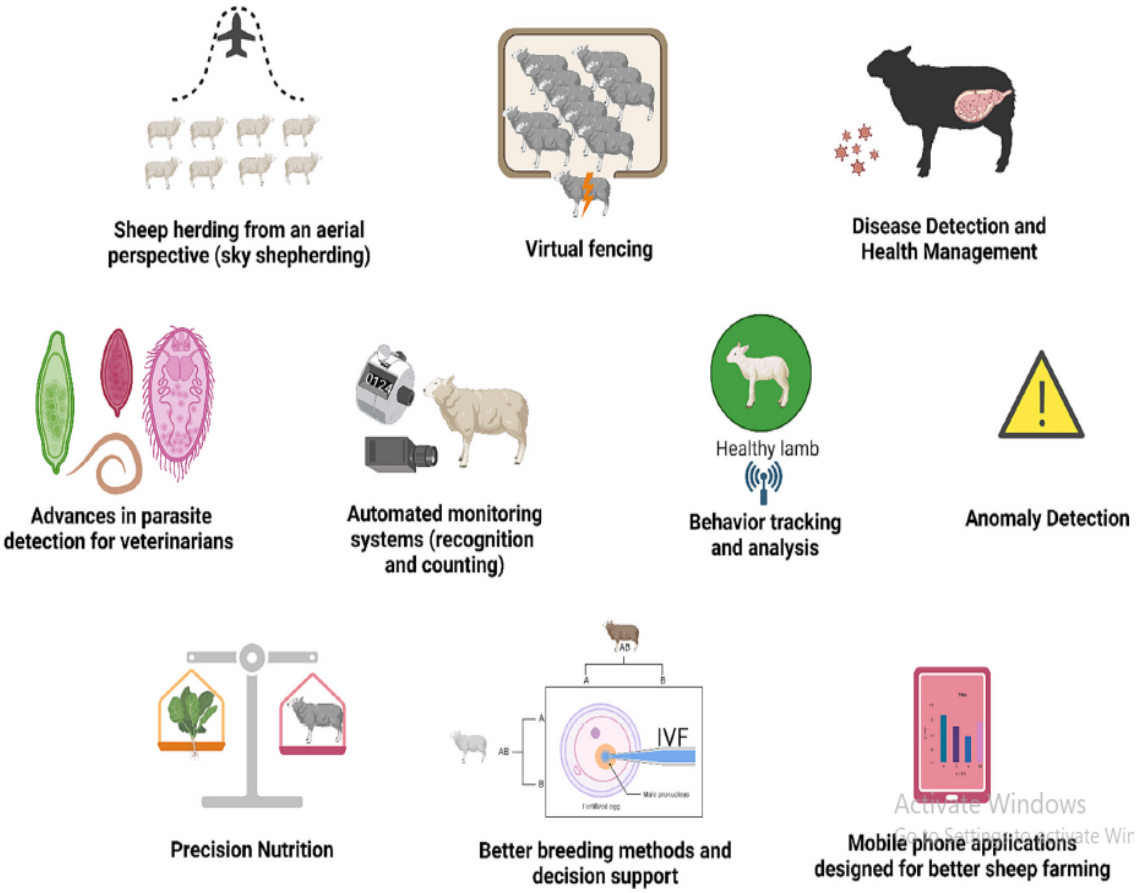
Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

زیست، منابع طبیعی و سلامت انسان که به عنوان مناطق حفاظت شده در ارزیابی تأثیر چرخه حیات (LCIA) طبقه بندی می‌شوند، اهمیت حیاتی دارند (De Vries and de Boer 2010). تحلیل‌های چرخه عمر مرتبط با زنجیره محصول همچنین شامل ارزیابی فناوری‌های دقیق کشاورزی دام می‌شود. این فناوری‌ها با بهبود عملکرد از جایگزینی کار دستی با تکنیک‌های کار هوشمند (D'Agaro et al. 2021) و با کاهش هزینه‌ها و اثرات زیست‌محیطی، پایداری تولید و رفاه حیوانات را افزایش می‌دهند، فرصتی را برای بخش بیوتکنولوژی حیوانات فراهم می‌کنند. دامپروری شامل ردیابی حیوانات است که امکان ردیابی کل فرآیند تولید را فراهم می‌کند و حتی برای مصرف کنندگان و سهامداران از طریق جمع‌آوری و ارزیابی داده‌ها در دسترس قرار می‌گیرد. ردیابی داده‌های کاملی که از طریق LCA ارزیابی می‌شود شامل تولید محصول پایدار برای خوراک، ردیابی موقعیت‌یابی بلادرنگ و نظارت بر سلامت حیوانات، همچنین حمل و نقل و پردازش و ذخیره‌سازی مواد غذایی است که منجر به ردیابی کامل از مزرعه تا سفره برای سلامت و ایمنی مصرف‌کننده می‌شود و می‌تواند آگاهی آن‌ها را از رفتارها بهبود بخشد. به طور خاص، داده‌های مربوط به شرایط محیطی و بهداشتی رفاه حیوانات را تضمین می‌کند. ردیابی محصولات حیوانی، داده‌های بزرگی را ایجاد می‌کند که می‌تواند چرخه‌های بسته‌تری را با کاهش ورودی‌ها و صرفه‌جویی در منابع و هزینه‌ها، و همچنین کاهش انتشار گازهای گلخانه‌ای (از جمله مدیریت کود) که به آب و هوای جهانی بازخورد می‌دهد، تسهیل کند (Cooper et al. 2013). به ویژه با در نظر گرفتن این‌که مالیات کربن قیمت گوشت گاو را بیش از ۱۰۰٪ افزایش می‌دهد (Doelman et al. 2020). این امر بدون در نظر گرفتن اثرات واقعی بحران انرژی است که هزینه‌های سوخت فسیلی و کود را نیز تسریع می‌کند. در پژوهشی، Pour Hamidi et al. (2017) با استفاده از شبکه‌های عصبی مصنوعی برای پیش‌بینی ارزش اصلاحی صفت تولید شیر در گاوهای هلشتاین استفاده کردند. آن‌ها نشان دادند که توانایی مدل شبکه عصبی مصنوعی، نسبت به روش‌های دیگر بالاتر و نزدیکتر به مقادیر تخمین زده شده است. بنابراین، می‌توان به جای روش‌های رایج برای پیش‌بینی ارزش‌های اصلاحی برای تولید شیر، از شبکه‌های عصبی مصنوعی استفاده کرد. در مطالعه‌ای دیگر Ghotbaldini et al. (2019) از شبکه‌های عصبی مصنوعی برای پیش‌بینی ارزش اصلاحی وزن بدن در سن ۶ ماهگی گوسفند سود بردند. نتایج آن‌ها نشان داد که شبکه عصبی مصنوعی به پیش‌بینی ارزش اصلاحی برای وزن بدن در ۶ ماهگی در گوسفند کرمانی است و سرعت و دقت قابل قبولی دارد. بنابراین، این شبکه به جای روش‌های رایج می‌تواند برای تخمین ارزش‌های اصلاحی صفات تولیدی و تولیدمثلی در حیوانات اهلی استفاده شود. در مطالعه‌ای مدل‌های شبکه عصبی مصنوعی و رگرسیون برای پیش بینی وزن بدن بز استفاده و مقایسه شدند (Khorshidi et al. 2019). مقایسه دو مدل نشان داد که هر دو مدل می‌توانند وزن بدن را به خوبی و نزدیک به وزن واقعی بدن پیش بینی کنند، اما قابلیت مدل شبکه عصبی مصنوعی بالاتر از رگرسیون چندگانه و نزدیک به وزن واقعی بدناست. با این حال، اگر اندازه گیری‌های مرتبط بیشتری ثبت شود، شبکه عصبی مصنوعی می‌تواند نتایج مطلوبی را ارائه دهد. بنابراین، می‌توان به جای روش‌های مرسوم برای پیش بینی وزن واقعی بدن با استفاده از اندازه گیری‌های بدن،

از شبکه‌های عصبی مصنوعی استفاده کرد. با توسعه روش‌های دامپروری دقیق[1] (PLF)، کشاورزان تشویق می‌شوند تا راه‌حل‌های دیجیتالی را برای مزارع خود اتخاذ کنند. هدف PLF مدیریت حیوانات فردی با نظارت مستمر بر تولید، سلامت، تولید مثل، رفاه و اثرات زیست محیطی در زمان واقعی با مجموعه‌ای از ابزارهای الکترونیکی است (Berckmans 2017). در سیستم‌های پرورش فشرده، اتخاذ مدل‌های مختلف هوش مصنوعی مانند یادگیری ماشین (ML)، یادگیری عمیق (DL) و شبکه‌های عصبی مصنوعی[2] (ANN) پرورش معمولی را پایدارتر، کارآمدتر و سودآورتر کرده است (Bao and Xie 2022). تمرین کشاورزی دیجیتال با استفاده از ابزارهای هوش مصنوعی راه را برای بهبود رفاه حیوانات و افزایش بهره‌وری کشاورزی هموار کرده است. هوش مصنوعی همچنین اتوماسیون را در نظارت بر رفتار حیوان، سلامت، تغذیه، مدیریت، عملکرد و تخصیص منابع فعال کرده است (Bao and Xie 2022). با حسگرهای متعددی که متغیرهای حیوانی مختلف مانند دما، ضربان قلب و هضم را ردیابی می‌کنند، مدل‌های هوش مصنوعی کشاورزان را قادر می‌سازند تا دام‌های خود را مدیریت کنند (Basciftci and Gunduz 2019). تقاطع کشاورزی دیجیتال و هوش مصنوعی (AI) افق‌های جدیدی را در حوزه رفاه حیوانات باز کرده است. هوش مصنوعی در صنایع مختلف نفوذ کرده است و توسعه و سازگاری آن مدیون تلاش مشترک و چند رشته‌ای است که شامل متخصصانی مانند دانشمندان علوم دامی، دانشمندان کامپیوتر، مهندسان کشاورزی و دانشمندان محیط زیست می‌شود. ادغام آخرین فناوری‌های هوش مصنوعی در شیوه‌های رفاه حیوانات، یک پیشرفت قابل توجه است که هوش مصنوعی را به عنوان راه‌حلی پایدار برای رفع نیازهای در حال تحول صنعت قرار می‌دهد. در این زمینه، تحقیقات سریع در زمینه هوش مصنوعی برای رفاه حیوانات اهلی از این قاعده مستثنی نیست. تحقیقات جهانی ادغام بالقوه هوش مصنوعی را برای افزایش کارایی در جنبه‌های مختلف پرورش حیوانات اهلی، شامل رفاه، مدیریت بیماری، نظارت بر رفتار، بهینه‌سازی فرآیندهای تغذیه و نظارت محیطی آشکار کرده است (شکل ۳). بنابراین، پوشش جامع چشم‌انداز پویا فناوری‌های نوظهور، با تمرکز بر تکنیک‌های دیجیتالی‌سازی و هوش مصنوعی پیشرفته، و تأثیر جمعی آن‌ها بر افزایش رفاه گوسفند ضروری است. علاوه بر این، شناسایی چالش‌های بزرگ بسیار مهم است و باعث می‌شود که جهت‌گیری‌های جدیدی برای تحقیقات آینده آشکار شود. لذا، تقویت ادغام کارآمد و پایدار هوش مصنوعی در شیوه‌های مدرن پرورش حیوانات اهلی ضروری است. یک جایگزین امیدوارکننده برای استفاده از سگ‌ها، ادغام سیستم‌های چوپانی مستقل با هوش مصنوعی است. پهپادها برای این منظور بررسی شده‌اند. به آن‌ها "چوپان آسمانی" هم گفته می‌شود. در پژوهشی .Yaxley et al (2021) پتانسیل پهپادها را به عنوان جایگزینی برای سگ‌ها در کاهش استرس در میان گوسفندان مزرعه برجسته کرده‌اند. پهپادها را می‌توان طوری برنامه‌ریزی کرد که نشانه‌های شنیداری مختلفی را برای برانگیختن پاسخ‌های خاص از گله‌های گوسفند منتشر کنند. علاوه بر این، در طول عملیات شبانه، پهپادهای مجهز به دوربین‌های مادون قرمز حرارتی را می‌توان برای شناسایی تهدیدات احتمالی و بالا بردن آلارم در صورت وجود مزاحمان

[1] Precision livestock farming
[2] Artificial Neural Networks

استفاده کرد (Bondi et al., 2019). سیستم نظارتی بیومیمتیک[1] دیگر از الگوریتم‌های چوپانی استفاده می‌کند و می‌تواند ضمن هدایت و محافظت از گوسفندان، با شرایط محیطی متفاوت سازگار شود و از ورود آن‌ها به مناطق حساس یا محدود جلوگیری کند (Strombom and King 2018).



شکل ۳. سیستم‌ها و رویکردهای پیشرفته پرورش گوسفند با استفاده از تکنیک‌های دیجیتال و هوش مصنوعی برای پرورش بهتر گوسفند (Arshad et al. 2024)

Figure 3. Advanced sheep breeding systems and approaches using digital techniques and AI for better sheep breeding (Arshad et al. 2024)

مدل‌سازی سیستم-ژنومیکس عملکردی، هوش مصنوعی و سیستم‌های بیولوژی

دو رویکرد اصلی برای استفاده از نتایج تجربی و غنی سازی درک ما از فرآیندهای بیولوژیکی در حال حاضر اتخاذ شده است: داده محوری و مدل محوری[2]. امروزه، رویکرد مبتنی بر داده عمدتاً در حوزه DL استفاده می‌شود که برای تصمیم‌گیری خودکار به

[1] Biomimetic
[2] Data-driven and model-based

سیستم‌های جعبه سیاه متکی است. به طور معمول، مدل‌های ML و DL ویژگی‌های یک سیستم را در یک کلاس یا امتیاز بدون افشای دلایل راهنمایی یا توضیح ساختار و پویایی سیستم زیربنایی ترسیم می‌کنند. این یکی از موانع کلیدی در برابر استفاده گسترده از هوش مصنوعی برای درک زیست شناسی است. در واقع، برای مثال، در تصمیم‌گیری‌های بالینی، افراد تمایل کمی به نتایجی دارند که مکانیسم پیش‌بینی آن‌ها مشخص نیست. رویکرد مبتنی بر مدل، برعکس حوزه سنتی سیستم‌های زیست شناسی است که هدف آن رمزگشایی پیچیدگی سیستم‌های بیولوژیکی و درک ساختار، اجزای آن‌ها، روابط و پویایی مبتنی بر اختلالات بیولوژیکی، ژنتیکی یا شیمیایی و نظارت اثرات بر روی سیستم است (Ideker et al. 2001). در این چارچوب، برای درک ساختار سیستم و حالت عملکرد، اجزای سیستم و ویژگی‌های آن‌ها باید شناسایی شوند، و تلاش برای استنباط چگونگی تعامل و تکامل پویا برای ایجاد رفتار بیولوژیکی قابل مشاهده (Kitano 2002) انجام شود. به طور خاص، دینامیک معمولاً توسط مجموعه‌ای از معادلات دیفرانسیل معمولی مدل‌سازی می‌شود که چگونگی تکامل گونه‌های شیمیایی و مولکولی در سیستم را در طول زمان توصیف می‌کند. ذکر این نکته حائز اهمیت است که ارزش پیش‌بینی واقعی نتایج به شدت به تخمین دقیق و مؤثر پارامترهای مدل بستگی دارد و مدل‌های دیفرانسیل به بسیاری از پارامترهای ناشناخته (مانند ثابت‌های سرعت و غلظت‌های اولیه) بستگی دارند که به طور کلاسیک نسبتاً از چند اندازه گیری تجربی استنباط می‌شوند. به خاطر این محدودیت، مدل‌سازی موفقیت‌آمیز فقط برخی از سیستم‌های بیولوژیکی نسبتاً ساده (مانند سیستم‌های استفاده از لاکتوز و گالاکتوز در باکتری‌ها، مانند اشریشیاکلی[1] (Khodayari and Maranas 2016) و استرپتوکوکوس[2] (Zeng et al. 2010)) امکان‌پذیر بوده است، در حالی که مدل سازی سیستم‌های پیچیده‌تر هنوز بازدارنده است. با توجه به موفقیت روزافزون تکنیک‌های هوش مصنوعی برای تولید و تجزیه و تحلیل داده‌های بیولوژیکی در مقیاس بزرگ، طراحی آزمایشی و اعتبارسنجی مدل، محققان به دنبال این هستند که چگونه رویکردهای داده‌محور را می‌توان در استراتژی‌های مبتنی بر مدل برای حل مشکل تخمین پارامتر و استنتاج دینامیک پنهان را، برای کمک به روشن شدن ساختار سیستم بیولوژیکی، مکانیسم‌ها و پویایی ادغام کرد. این احتمال در پژوهش .Wang et al (2018) مورد بحث قرار گرفته است، که در آن همگرایی رویکردهای داده محور و نظری به عنوان گام مهمی برای تکمیل چرخه داده–مدل–داده، که نمونه ای از علوم تجربی مانند فیزیک و زیست‌شناسی است در نظر گرفته شد. یک رویکرد امیدوارکننده برای تقویت این همگرایی، استفاده از مدل‌های نظری موجود برای محدود کردن نتایج هوش مصنوعی است. امروزه، در دسترس بودن داده‌های با توان عملیاتی بالا و ابزارهایی برای مدیریت آن، اعتبارسنجی و/یا اصلاح کارآمد مدل را امکان‌پذیر می‌سازد. در این راستا، برخی از راه حل‌های اخیر ارائه شده در نتایج منتشر شده مستحق توجه هستند، زیرا آن‌ها اجازه می‌دهند رویکردهای مبتنی بر مدل و داده محور به طور مؤثری به سمت درک عمیق فرآیندهای بیولوژیکی همگرا شوند. در پژوهشی Costello and Martin Garcia (2018) پویایی مسیر لیمونن و ایزوپنتنول را با حل یک مشکل شناسایی سیستم پیش بینی می‌کنند که در آن مناسب‌ترین مدل توسط یک استراتژی ML آموزش داده شده توسط داده‌های

[1] *Escherichia coli*
2 *Streptococcus*

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

پروتئومیکس سری زمانی انتخاب می‌شود. در پژوهشی دیگر Yazdani et al. (2018) از اصل شبکه‌های عصبی مبتنی بر فیزیک (Raissi et al. 2019) استفاده کردند، که بیشتر از داده محور بودن منحرف می‌شود، زیرا یک مدل ریاضی (با پارامترهایی که باید شناسایی شوند) به عنوان یک محدودیت قوی در آموزش شبکه استفاده می‌شود. بر این اساس، نویسندگان یک روش DL مبتنی بر زیست‌شناسی را توسعه دادند که قادر به تخمین پارامترهای مدل و همچنین استنتاج پویایی سیستم پنهان است. این رویکرد با موفقیت بر روی گلیکولیز مخمر، آپوپتوز سلولی و مدل‌های غدد درون ریز اولترادیان آزمایش شد. همچنین Fortelny and Bock (2020) از دانش زیست‌شناسی سیستم‌ها برای محدود کردن نتایج خود استفاده می‌کنند، بنابراین از همان اصل شبکه‌های عصبی مبتنی بر فیزیک پیروی می‌کنند. نویسندگان یک شبکه بیولوژیکی را به یک شبکه عصبی ترسیم می‌کنند که در آن هر گره (node) نشان دهنده یک مولکول و هر لبه (edge) نشان دهنده یک اثر متقابل است که وجود و قدرت آن، زمانی که مشخص شود، از یک مدل مکانیکی مشتق شده است. سپس از شواهد تجربی می‌توان تعاملات جدیدی را کشف کرد و برای اصلاح ساختار شبکه استفاده کرد. در یک بررسی، Muzio et al. (2021) ساختار شبکه‌های کانولوشن گراف و شبکه‌های عصبی گراف را توضیح می‌دهند و مجموعه‌ای از کاربردها را فهرست می‌کنند که در آن‌ها این شبکه‌ها با موفقیت برای تجزیه و تحلیل شبکه‌های بیولوژیکی، از جمله پیش‌بینی عملکرد پروتئین، برهم‌کنش‌های پروتئین-پروتئین و در کشف و توسعه دارو *in silico* استفاده می‌شوند. سایر بررسی‌ها، به عنوان مثال توسط Eraslan et al. (2019)، Zampieri et al. (2019)، Antonakoudis et al. (2020) و Gilpin et al. (2020) موضوع ماهیت رقابتی در مقابل ماهیت مشارکتی رویکردهای داده محور و مبتنی بر مدل را مطرح می‌کنند و بر اهمیت استفاده از مدل‌های ریاضی محدود شده مناسب برای کمک به ابزارهای هوش مصنوعی برای تولید دانش جدید از مکانیسم‌های بیولوژیکی تأکید می‌کنند.

مسائل مدیریت داده برای کاربردهای هوش مصنوعی در ژنومیک عملکردی

اکثر الگوریتم‌های ML یا DL ورودی خود را به شکل یک ماتریس می‌گیرند، جایی که به هر ستون یک نمونه و به هر ردیف یک متغیر (یعنی ویژگی) که نمونه‌ها را توصیف می‌کند، تعلق می‌گیرد. ماهیت این ماتریس‌ها هم به زمینه و محتویات و هم به کاربرد خاص بستگی دارد. در ژنومیک عملکردی و به طور کلی در بیوانفورماتیک، چنین نمایشی سودمند است زیرا بیشتر داده‌ها به طور طبیعی به این شکل درمی‌آیند. برای مثال، داده‌های RNA-seq معمولاً به‌صورت ماتریس‌هایی مرتب می‌شوند که حاوی کمیت ژن یا رونوشت فراوانی (ردیف‌ها یا سطرها) در مجموعه‌ای از نمونه‌ها هستند که شرایط مختلف (ستون‌ها) را نشان می‌دهند. در نتیجه این تناسب متقابل ماتریس‌ها برای ML و بیوانفورماتیک، زبان‌های برنامه نویسی مانند R و python که از ماتریس‌های داده (فریم‌های داده) به عنوان ساختار داده اصلی استفاده می‌کنند، به ابزارهای محبوبی تبدیل شده‌اند. با این حال، در بسیاری از کاربردهای واقعی بیوانفورماتیک، ماتریس‌های داده ممکن است ناقص و/یا حاوی خطا باشند. پروتکل‌های مختلف، شرایط آزمایشی

و ماشین آلات ممکن است باعث سوگیری‌ها (اریبی) و مصنوعات (آرتیفکت) شوند. علاوه بر این، برخی از نقاط داده ممکن است برای برخی از نمونه‌ها از دست رفته باشد. با این حال، نیاز به اجرای یک رویکرد کل‌نگر برای درک هر تابع عنصر ژنومی مستلزم در نظر گرفتن داده‌هایی با ماهیت متفاوت است. در این سناریو، انتساب (جانهی) داده‌ها، حذف نویز و یکپارچه سازی باید بخشی از طراحی ML و AI برای ژنومیک عملکردی باشد.

جانهی (انتساب) داده‌ها[1]: همانطور که در بالا ذکر شد، تجربه مزاحمت ناشی از برخورد با داده‌های ناقص غیر معمول نیست. دلایل این مقادیر از دست رفته شامل: در دسترس نبودن، خطاهای اندازه‌گیری، و ادغام پایگاه‌های داده با طرحواره‌های مختلف هستند. با توجه به احتمال از دست رفتن داده‌ها، Little and Rubin (1987) سه کلاس را تعریف کرده‌اند که شامل: به طور کاملا" تصادفی از دست رفته[2] (MCAR)، از دست رفته به طور تصادفی[3] (MAR) و از دست نرفته به طور تصادفی[4] (NMAR) هستند. کلاس اول، یعنی MCAR، موردی را توصیف می‌کند که در آن همه اقدامات احتمال یکسانی برای از دست رفتن دارند. این به احتمال زیاد، برای مثال، در داده‌های ریزآرایه‌ای که در آن شکست خواندن ممکن است در همه جا اتفاق بیفتد، وجود دارد. اگر احتمال داده‌های از دست رفته فقط در یک گروه به طور مساوی توزیع شود، داده‌ها MAR هستند. آخرین کلاس مواردی را پوشش می‌دهد که در آن نه MCAR و نه MAR اعمال نمی‌شود. در هر صورت، قبل از تغذیه الگوریتم‌های ML یا AI، باید به این سوال که کدام استراتژی برای مقابله با داده‌های از دست رفته مناسب‌تر است، پاسخ داده شود. همانطور که توسط Cismondi et al.(2013) توضیح داده شده است، دو گزینه اصلی ۱) حذف متغیرها یا نمونه‌هایی که مقادیر گمشده دارند، ۲) جانهی (imputation) داده‌های گمشده وجود دارد. اولین مورد به طور گسترده نباید استفاده شود زیرا باعث اریبی می‌شود (Heitjan 1997). رویکرد دوم، یعنی جانهی داده، شامل پر کردن شکاف‌های ماتریس داده با پیش‌بینی مناسب‌ترین مقدار برای هر اندازه‌گیری از دست رفته است. چندین استراتژی در بیش از سی سال تحقیق در مورد این موضوع (Chen and Shao 2000; Kim et al. 2020) پیشنهاد شده است. یکی از عوامل مشترک بیشتر رویکردهای پیشنهادی این است که مفهوم "مقدار مناسب[5]" با تقریب دقیق مقدار گمشده منطبق است. این مفهوم به طور تجربی توسط De Souto et al. (2015) به چالش کشیده شده است، که نشان می‌دهند کارهای خوشه‌بندی و طبقه‌بندی از استراتژی‌های جانهی پیچیده بهره نمی‌برند. برعکس، روش‌های ساده، مانند جایگزینی مقادیر گمشده با مقادیر متوسط، عملکرد مشابهی دارند. نویسندگان توجه را به توانایی روش‌های جانهی برای حفظ اهمیت جلب می‌کنند. این دیدگاه حاکی از آن است که از انتساب داده‌ها نمی‌توان به عنوان یک جعبه سیاه[6] استفاده کرد، اما روش باید مطابق با وظیفه خاص انتخاب شود. Van Buuren (2018) در کتاب اصلی خود، مروری بر تکنیک‌های جانهی آن‌ها را به سه

[1] Data imputation
[2] Missing completely at random
[3] Missing at random
[4] Not missing at random
[5] Appropriate value
[6] Black-box

دسته با تقسیم‌بندی می‌کنند: ۱- جانهی آماری، که از آماره‌های تک متغیره (به عنوان مثال، میانگین) یا چند متغیره (مثلاً k-NN (Malarvizhi and Thanamani 2012) استفاده می‌کند؛ ۲- جانهی چندگانه، که بیشتر از یک ($n > 1$) مجموعه داده کامل ایجاد می‌کند و سپس آن‌ها را با به حداقل رساندن یک تابع هدف معین ادغام می‌کند؛ و ۳- جانهی مبتنی بر مدل، که از رویکردهای ML(به عنوان مثال، خوشه بندی (Gautam and Ravi 2014)) استفاده می‌کند. به دلیل سادگی و کاربرد گسترده، روش‌های کلاس اول رایج‌تر هستند. با این حال، در غیاب دستورالعمل‌ها، تحلیل‌گران داده‌ها باید تنها بر تجربه خود تکیه کنند تا روش مناسب‌تری را انتخاب کنند. برای پرداختن به این موضوع، در یک بررسی (Petrazzini et al. 2021)، پژوهش‌گران آزمایش گسترده‌ای از ۶ روش آماری تک متغیره و چند متغیره را ارائه می‌کنند که برای انواع مختلف داده‌های گمشده (مانند MAR، NMAR، MCAR) اعمال می‌شود. آزمایش‌های آن‌ها نشان داد که به طور کلی، روش‌های چند متغیره در پیش‌بینی مقادیر گمشده دقیق‌تر هستند. یک مشکل مرتبط با موارد فوق، عدم وجود ستون‌ها در یکپارچه‌سازی داده‌های اومیکس چندگانه است. در اینجا، ما با جداول داده‌های متعددی سروکار داریم که هر کدام مجموعه‌ای از ویژگی‌های یک گروه از نمونه‌ها را توصیف می‌کنند. برای یکپارچه‌سازی این داده‌ها، برخی از الگوریتم‌ها، مانند همجوشی شبکه شباهت[1] (Wang et al. 2014)، به تطابق دقیق بین گروه‌های نمونه نیاز دارند. در نتیجه، وجود تنها یک ستون از دست رفته در جدول برای به خطر انداختن کاربرد روش کافی است. وارد کردن کل یک ستون بسیار چالش برانگیزتر از وارد کردن مقادیر نقطه‌ای است و روش‌های آماری ممکن است برای این کار کافی نباشند. برای رویارویی با این مشکل، چند روش تخصصی پیشنهاد شده است. یک رویکرد توسط Voillet et al. (2016) پیشنهاد می‌شود که در آن انتساب چندگانه با PCA مخلوط می‌شود. به طور خلاصه، آن‌ها مجموعه‌ای از انتساب‌های قابل قبول را با «قرض گرفتن (borrowing)» ستون‌های گمشده از نمونه‌های دیگر ایجاد می‌کنند، سپس آنالیز PCA جداگانه را انجام می‌دهند و گزینه‌ای را انتخاب می‌کنند که به بهترین وجه با PCA اجماع مطابقت دارد. سایر روش‌های جانهی عمدتاً مبتنی بر PCA هستند (Husson and Josse 2013; Josse and Husson 2016).

حذف نویز داده‌ها: داده‌های اومیکس اغلب شامل کمی کردن فراوانی ویژگی‌های خاص در یک سلول یا در تعداد زیادی از سلول‌ها است. مقادیر فراوانی در معرض چندین منبع سوگیری (اریبی)، از جمله تنوع بیولوژیکی و مصنوعات (آرتیفکت) تکنیکی هستند. علاوه بر این، ویژگی‌های اندازه‌گیری شده معمولاً بسیار بیشتر از نمونه‌های موجود است. در برخی موارد، پیشرفت‌های فناوری و در دسترس بودن تکرارهای فنی می‌تواند به تخمین و تصحیح خطاهای اندازه‌گیری کمک کند، اما نمی‌تواند در برابر تنوع بیولوژیکی (Hansen et al. 2011) که به طور کلی تأثیر بیشتری دارد (McIntyre et al. 2011) کمک کند. همانطور که ضرب المثل قدیمی در علوم کامپیوتر "زباله داخل، زباله بیرون[2]" یادآوری می‌کند، به طور کلی داده‌های پاک (clean) یک پیش نیاز ضروری

[1] Similarity network fusion
[2] Garbage in, garbage out

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

برای برنامه‌های ML و AI است که برای یادگیری به این داده‌ها متکی هستند. حذف نویز داده‌ها وظیفه تصحیح آرتیفکت‌ها (یعنی نرمال‌سازی داده‌ها) و حذف اندازه‌های ناصحیح یا نامربوط (به عنوان مثال، انتخاب ویژگی[1]) است. دو رویکرد کلاسیک برای حذف نویز (Nounou et al. 2013) وجود دارد که در استفاده یا عدم استفاده از یک مدل (که اغلب تجربی است) متفاوت هستند. نمونه ای از کلاس اول نرمال سازی داده‌های RNA-seq (Love et al. 2014) است، که در آن پروفایل بیان به عنوان یک توزیع دو جمله‌ای منفی مدل سازی می‌شود. در مورد کلاس دوم، اکثر روش‌های نویززدایی که از یک مدل استفاده نمی‌کنند، مبتنی بر تعیین آستانه‌ها هستند (Van Hulse et al. 2012). این روش‌های اخیر به جای اعمال تصحیح مقدار، سیگنال‌های ضعیف و ویژگی‌های بی‌معنی و غیرمهم را فیلتر می‌کنند. هر دو رویکرد مزایا و معایب خود را دارند. حذف نویز مبتنی بر مدل می‌تواند آرتیفکت‌ها را تصحیح کند، بنابراین امکان ادغام مجموعه داده‌های ناهمگن را فراهم می‌کند. برای مثال، این امر هنگام نرمال‌سازی داده‌های بیان ژن به‌دست‌آمده با فناوری ریزآرایه (Lazar et al. 2013) اتفاق می‌افتد. مشکل در این مورد این است که هم کاربرد و هم اثربخشی روش به میزان پایبندی مدل به مجموعه داده‌های داده شده بستگی دارد. از سوی دیگر، رویکردهای بدون مدل همیشه قابل اجرا هستند. با این حال، آستانه‌ها وابسته به مقیاس هستند و انتخاب ناکافی می‌تواند باعث حذف ویژگی‌های مربوطه، که به طور تصادفی به مقدار برش (cut-off) نرسد شود. فیلترهای چند مقیاسی، مانند آنچه توسط Nounou et al. (2012) ارائه شده است می‌تواند این مشکل را کاهش دهد. یک رویکرد اخیر برای حذف نویز داده‌ها از DL استفاده می‌کند. این ایده که توسط Eraslan et al. (2019) ارائه شده است، قرار دادن محاسبات ویژگی‌ها در مدل ML است. حذف نویز رمزگذارهای خودکار[2] (Vincent et al. 2008) (و به ویژه آن‌هایی که مبتنی بر شبکه‌های عمیق (Saad and Chen 2020) هستند) یکی از رایج‌ترین راه حل‌ها در این مسیر هستند. به طور خلاصه، این شبکه‌ها با رمزگذارهای خودکار استاندارد تفاوت دارند. زیرا برای بازسازی ورودی از نسخه خراب آن آموزش دیده‌اند. رمزگذارهای خودکار (اعم از استاندارد یا حذف نویز) مزیت اضافی ماژولار بودن (modular) را دارند که به آن‌ها اجازه می‌دهد با تکنیک‌های دیگر ترکیب شوند. برای مثال، رمزگذاری خودکار حذف نویز می‌تواند برای ساخت شبکه‌های پیچیده‌تر انباشته شود (Vincent et al. 2010) یا با یک مدل بیزی ترکیب شود تا نویز داده‌های توالی‌یابی سلول منفرد را حذف کند (Wang et al. 2019). [365].

یکپارچه‌سازی داده‌ها[3]: با استفاده از یک استعاره، می‌توانیم خطوط مختلف داده‌های اومیکس را در یک سلول (ژنومیکس، پروتئومیکس، ترنسکریپتومیکس و غیره) به عنوان عناصری در یک ارکستر سمفونیک در نظر بگیریم، جایی که فنوتیپ سلولی نمایش داده شده نتیجه نواختن همزمان همه عناصر است. گوش دادن به یک عنصر منفرد (به عنوان مثال، تجزیه و تحلیل یک آهنگ (تراک) داده اومیکس) می‌تواند ایده‌ای از ملودی ارائه دهد، اما بسیاری از تفاوت‌ها و نکات ظریف نادیده گرفته می‌شوند. منطق

[1] Feature selection
[2] Autoencoders
[3] Data integration

پشت یکپارچه‌سازی (ادغام) داده‌ها در ژنومیک عملکردی این است که تنها تجزیه و تحلیل جامع همه رویدادهای اومیکس و اثرات متقابل آن‌ها می‌تواند دید کاملی از وضعیت سلول ارائه دهد. یک مثال شناخته شده که این دیدگاه را تایید می‌کند با رابطه بین متیلاسیون DNA و بیان ژن ارائه شده است (Razin and Cedar 1991; Bell et al. 2011). به طور کلی، سه نوع یکپارچه‌سازی (ادغام) داده‌ها (Richardson et al. 2016) می‌تواند وجود داشته باشد: ۱- یکپارچه‌سازی افقی، که شامل مجموعه‌های داده جدا از یک نوع اومیکس[1] است. ۲- یکپارچه‌سازی عمودی، که انواع مختلف داده‌های اومیکس را در یک گروه از نمونه‌ها ترکیب می‌کند. ۳- یکپارچه‌سازی موازی که دو مورد بالا را با هم مخلوط می‌کند. اولین نوع یکپارچه‌سازی به ویژه برای ترکیب مجموعه داده‌های بدست آمده از منابع مختلف، با هدف کاهش عدم تعادل بین تعداد نمونه‌ها و تعداد ویژگی‌های ژنومی مفید است. رویکرد نوع دوم بیشتر برای مشخص کردن ویژگی‌هایی استفاده می‌شود که یک فنوتیپ مشاهده شده را القا می‌کند. در نهایت، رویکرد سوم در حضور داده‌های ناهمگن بهترین گزینه است. از نقطه نظر ریاضی، ورودی الگوریتم‌های یکپارچه‌سازی شامل مجموعه‌ای از ماتریس‌ها (یکی برای هر نوع داده اومیکس) است که هر ردیف مربوط به یک ژن و ستون‌ها نمونه‌ها هستند. هدف در اینجا تولید یک ماتریس جدید است که در آن روابط چند سطحی برجسته می‌شود. در پژوهشی (Bersanelli et al. 2016)، یک طبقه‌بندی از الگوریتم‌های یکپارچه‌سازی ارائه می‌دهند. یک ویژگی متمایز اصلی برای این طبقه‌بندی این است که آیا آن‌ها خروجی خود را در قالب یک شبکه برمی‌گردانند یا نه. در حالت اول، خروجی یک ماتریس مربع است که می‌تواند به راحتی به عنوان یک ماتریس مجاورت[2] تفسیر شود که در آن هر سطر/ستون با یک ژن مطابقت دارد و مقادیر نشان دهنده قدرت هر رابطه (ارتباط) است. مزیت این الگوریتم‌ها این است که امکان استفاده از مسیرها[3] (Gui et al. 2011) و الگوریتم‌های تشخیص جامعه[4] (Pellegrini et al. 2016) را برای تحلیل‌های بعدی فراهم می‌کنند. رویکرد دوم ماتریسی را برمی‌گرداند که ردیف‌های آن نمایانگر ویژگی‌های ژنومی هستند. مزیت این دسته از الگوریتم‌ها در امکان استفاده از تحلیل افتراقی یا خوشه‌بندی[5] برای تحلیل‌های پایین دستی است. از طرف دیگر، این روش‌ها می‌توانند محدودیت‌هایی را بر روی فضای نمونه یا فضای ویژگی[6] اعمال کنند. برای مثال، طرح وزن‌های ژنی[7] پیشنهاد شده توسط (Louhimo and Hautaniemi 2011) (که در آن به هر ژن امتیازی اختصاص داده می‌شود که به صورت ترکیب خطی از پروفایل‌های مربوطه محاسبه می‌شود) نیاز دارد که ماتریس‌های ورودی در همان فضای ویژگی (یعنی ژن‌ها) باشند. این محدودیت را می‌توان در موارد خاصی با استفاده از روش‌های پیش‌بینی ژن هدف مانند TargetScan (Agarwal et al. 2015) دور زد. در پژوهشی Huang et al. (2017) الگوریتم‌های یکپارچه‌سازی را بر اساس یک تاکسون

[1] Disjoint data sets of the same omics type
[2] Adjacency matrix
[3] Pathways
[4] Community detection algorithms
[5] Differential analysis or clustering
[6] Feature space
[7] Gene-wise weights schema

متعامد (اورتاگونال) تقسیم کردند که بر اساس این‌که برچسب‌های فنوتیپ نمونه‌ها (به عنوان مثال، بیماری یا نرمال) استفاده می‌شوند یا خیر، متمایز می‌شوند. دسته دوم (اخیر) شامل روش‌های فاکتورسازی ماتریس بدون نظارت[1]، مانند روش‌های بیزی و روش‌های مبتنی بر شبکه است. فاکتورسازی ماتریس بدون نظارت شامل نمایش فضاهای اومیکس در یک فضای با ابعاد پایین‌تر است (Shen et al. 2009; Zhang et al. 2011). این روش‌ها یکپارچه‌سازی داده‌ها و انتخاب ویژگی را ترکیب می‌کنند، و به ویژه برای برنامه‌های کاربردی خوشه‌بندی مفید هستند (Wu et al. 2019). روش‌های بیزی (Zhao et al. 2012; Zhang et al. 2011) از مفروضات پیشینی[2] در توزیع داده‌ها و روابط بین مجموعه‌های داده استفاده می‌کنند. به لطف قانون بیز، این روش‌ها می‌توانند به راحتی احتمال تعلق الگوهای خاص به یک فنوتیپ خاص را تخمین بزنند. روش‌های نظارت شده همان اهداف روش‌های بیزی، یعنی شناسایی تعاملات و/یا پروفایل‌های پیچیده را دارند. با این حال، آن‌ها از برچسب‌ها (لیبل‌ها) در مجموعه آموزشی[3] برای یادگیری[4] مدل استفاده می‌کنند. این امر به رویکردهای ML و AI منتهی می‌شود. به عنوان مثال، Chaudhary et al. (2018) از رمزگذارهای خودکار[5] برای ادغام سه نوع داده اومیکس سرطان کبد و الگوهای یادگیری که منجر به پروفایل‌های مختلف بقا می‌شوند، استفاده کردند. ما انتظار داریم که یکپارچه‌سازی تحت نظارت به لطف ML و AI (Hamamoto et al. 2020) پیشرفت‌های بزرگی را تجربه کند.

قابلیت توضیح‌پذیری و تفسیرپذیری هوش مصنوعی در ژنومیک عملکردی

دانشمندان اعتماد را فرآیند روانی تعریف می‌کنند که در آن یک فاعل (آزمودنی[6]) تصمیم می‌گیرد یک رابطه وابستگی[7] روی یک متولی[8] پس از بررسی ویژگی‌های آن ایجاد کند (Israelsen and Ahmed 2019). این تعریف هیچ محدودیتی برای ماهیت متولی که می‌تواند هر چیزی، از جمله نرم افزار باشد، ایجاد نمی‌کند. علاوه بر تمایل طبیعی افراد، اعتماد به دو جنبه اساسی بستگی دارد: شهرت متولی و هزینه ناشی از شکست آن. کاربردهای بیوانفورماتیک اغلب با هزینه‌های بالای خرابی مشخص می‌شوند. برای مثال، پروژه‌های تحقیقاتی ژنومی از هوش مصنوعی برای یافتن جایگاه‌های مورد علاقه برای یک ژنوتیپ استفاده می‌کنند. شکست مدل هوش مصنوعی در این حالت می‌تواند باعث شکست کل پروژه شود. جای تعجب نیست که جامعه هوش مصنوعی شروع به سرمایه‌گذاری روی تلاش‌های قابل توجه برای توسعه تکنیک‌هایی برای افزایش قابلیت اطمینان الگوریتم به منظور بهبود شهرت و

[1] Unsupervised matrix factorization methods
[2] Priori assumptions
[3] Training set
[4] Learning
[5] Autoencoders
[6] Subject
[7] Relationship of dependence
[8] Trustee

به نوبه خود اعتماد کرده است. هم تفسیرپذیری[1] و هم توضیح‌پذیری[2] الگوریتم‌های هوش مصنوعی در جهت اعمال اعتماد است(Samek and Müller 2019). اگرچه یک تعریف رسمی مشترک از این مفاهیم ایجاد نشده است، تفسیرپذیری به این موضوع مربوط می‌شود که چگونه مدل هوش مصنوعی به نتایج خود می‌رسد، در حالی که توضیح‌پذیری بر این تمرکز دارد که چرا مدل در یک مورد خاص به نتایج خاصی می‌رسد. با توجه به این تعاریف غیررسمی (Montavon et al. 2018)، تفسیرپذیری هدف نهایی مدل هوش مصنوعی است که تأیید کند که دقت مدل از نمایش صحیح مشکل و نه از آرتیفکت‌های موجود در داده‌های آموزشی ناشی می‌شود. در واقع، یک سیستم هوش مصنوعی ممکن است به دقت تست‌کردن بالایی دست یابد، نه بر اساس "درک واقعی"، بلکه به این دلیل که قادر به یافتن روابط پنهان ناشناخته بین داده‌های آموزشی و آزمایشی است. با این حال، در غیاب چنین روابطی، دقت این طبقه‌بندی‌کننده‌ها می‌تواند به میزان قابل توجهی کاهش یابد، در نتیجه آن‌ها را برای کارهای واقعی بیوانفورماتیکی نامناسب می‌سازد (Lapuschkin et al. 2019). توضیح پذیری عمدتاً به اصل حق توضیح دادن مرتبط است، یعنی حق یک فرد برای توضیح دلایلی که چرا یک الگوریتم تصمیمی گرفته است که بر زندگی او تأثیر می‌گذارد. این اصل در بسیاری از قوانین بیشتر و بیشتر مورد توجه قرار گرفته است. به عنوان مثال، در ایالات متحده، حداقل برای بخش‌های تجاری خاص (به ویژه امور مالی) شناخته شده است، در حالی که در اتحادیه اروپا، به لطف قانون حفاظت از داده‌های عمومی[3] (GDPR)، به هر زمینه‌ای گسترش یافته است. علاوه بر اعطای حقوق شخصی، توضیح‌پذیری می‌تواند برای انجام تحلیل‌های پسینی از استراتژی‌های مورد استفاده توسط یک سیستم هوش مصنوعی برای انتخاب‌های خاص و استخراج دانش جدید از آن‌ها مفید باشد. یک مثال معروف در این راستا، مسابقه Go بود که در آن AlphaGo در برابر یک قهرمان چندگانه جهان با حرکتی که قبلاً هرگز دیده نشده بود، پیروز شد (Silver et al. 2016). تجزیه و تحلیل‌های قبلی (بر اساس داده‌های واقعی) استراتژی برنده، اتخاذ شده توسط شبکه را نشان داد. به طور کلی، توضیح‌پذیری و تفسیرپذیری را می‌توان با بهره‌برداری از دو استراتژی اصلی ۱- شفافیت یا تجزیه و ۲- تحلیل تعقیبی[4] به دست آورد (Došilovic et al. 2018). ایده (استراتژی) شفافیت این است که توضیح مدل، خود مدل است. این مورد در مورد طبقه‌بندی‌کننده‌های ساده مانند درختان، طبقه‌بندی‌کننده‌های مبتنی بر قانون و طبقه‌بندی‌کننده‌های خطی است که به راحتی می‌توان آن‌ها را مستقیماً تفسیر کرد. در درختان تصمیم می‌توان با بررسی مسیری که از آن تصمیم گرفته شده است، توضیح دلایلی را که منجر به تصمیم خاصی شده است به دست آورد. مزیت مدل‌های این کلاس در تفسیرپذیری و توضیح‌پذیری طبیعی آن‌هاست. از سوی دیگر، این مدل‌های ساده قادر به یادگیری روابط غیر خطی نیستند و در نتیجه ممکن است برای کاربردهای پیچیده مناسب نباشند. اکثریت قریب به اتفاق سیستم‌های هوش مصنوعی بیش از حد پیچیده هستند که به طور مستقیم قابل درک نیستند. در این

[1] Interpretability
[2] Explainability
[3] General Data Protection regulation
[4] Post hoc analysis

مورد، سیستم‌های هوش مصنوعی باید به عنوان جعبه سیاه در نظر گرفته شوند و توضیح باید از طریق تجزیه و تحلیل‌های تعقیبی (پسا هوک) استخراج شود. تجزیه و تحلیل‌های تعقیبی (پسا هوک) را می‌توان بر اساس دو استراتژی اصلی؛ استفاده از روش‌های مدل-آگنوستیک[1] (یعنی روش‌هایی که فقط بر روی ورودی و خروجی اعمال می‌شوند) یا روش‌های وابسته به مدل انجام داد. مزیت استراتژی "استفاده از روش‌های مدل-آگنوستیک" این است که همیشه قابل اجرا است، در حالی که "روش‌های وابسته به مدل" می‌تواند از ویژگی‌های خاص سیستم هوش مصنوعی مورد علاقه استفاده کند. در برخی موارد، توضیح با استخراج یک مدل شفاف[2] که از مدل اصلی تقلید می‌کند، به دست می‌آید. به عنوان مثال، گروهی از پژوهشگران (Martens et al. 2007) یک طبقه‌بندی مبتنی بر قانون را از یک SVM استخراج کردند، در حالی که گروهی دیگر (Zhou et al. 2003) چیزی از نظر مفهومی مشابه با یک شبکه عصبی ساختند. با این حال، این به طور کلی امکان پذیر نیست و توضیح اغلب در قالب تجسم ابعاد پایین‌تر به عنوان مجموعه‌ای از مثال‌ها ارائه می‌شود.

مشکلات نرم‌افزاری و اشتراک‌گذاری داده‌ها

خوانندگانی که از رشته علوم کامپیوتر می‌آیند اغلب به اشتراک‌گذاری نرم‌افزار و داده عادت ندارند. علیرغم این واقعیت که جوامعی مانند مجموعه داده‌های ارزشمندی را در اختیار جامعه هوش مصنوعی قرار می‌دهند که برای طراحی الگوریتم مفید هستند، برای بسیاری از برنامه‌ها بار جمع‌آوری داده‌ها بر عهده توسعه دهندگان است. کاربردهای هوش مصنوعی در بیوانفورماتیک از این نظر متفاوت است، زیرا هم طراحی الگوریتم و هم تولید دانش جدید می‌توانند از مقدار زیادی نرم‌افزار و داده‌های در دسترس عموم بهره ببرند. با این حال، ماهیت این داده‌ها و کاربردهای هوش مصنوعی در مراقبت‌های بهداشتی (که در نهایت یکی از اهداف نهایی اصلی همه زیرشاخه‌های بیوانفورماتیک، از جمله ژنومیکس عملکردی است)، مسائلی را در مورد مزایا و معایب اشتراک‌گذاری داده‌ها و همچنین در مورد تضادهای احتمالی بین جنبه‌های اقتصادی و اخلاقی ایجاد می‌کند.

به اشتراک‌گذاری داده‌ها و حفظ حریم خصوصی: مستقیماً از تعریف ژنومیک عملکردی، الگوی جدیدی از تحقیق که از بازرسی چند ژن هدف فاصله می‌گیرد و ژنوم را به عنوان یک کل در نظر می‌گیرد، پدید می‌آید. به عنوان یک نتیجه مستقیم، ما اکنون فرصت «بازیافت» (recycling) مجموعه داده‌های به دست آمده با استفاده از فناوری اومیکس برای پروژه‌های تحقیقاتی جدید را داریم. بازیافت (و به اشتراک‌گذاری یا sharing) داده‌ها دارای مزایای اخلاقی و اقتصادی است. از نقطه نظر اخلاقی، گردش داده‌ها در بین آزمایشگاه‌ها نیاز به آزمایش حیوانات را کاهش می‌دهد و نیازی نیست سلامت آن‌ها را به خطر انداخت. وقتی آزمایش‌ها منبع ناراحتی یا درد هستند، استفاده مجدد اهمیت خاصی پیدا می‌کند. تا آنجا که به مزیت‌های اقتصادی مربوط می‌شود، هزینه ذخیره سازی، حفاظت و به اشتراک‌گذاری داده‌ها بسیار کمتر از تولید داده‌های جدید، با توجه به کاهش هزینه فناوری‌های با توان بالا است.

[1] Model-agnostic
[2] Transparent model

ممکن است ادعا شود که اشتراک‌گذاری نیز دموکراتیک است، زیرا به آزمایشگاه‌های کوچکی که بودجه محدود آن‌ها اجازه تولید داده‌های خود را نمی‌دهد، به داده‌ها دسترسی می‌دهد. مزیت دیگر اشتراک گذاری این است که به ایجاد مجموعه‌های بزرگ داده کمک می‌کند، که هم برای آموزش الگوریتم‌های هوش مصنوعی و هم افزایش دقت آن‌ها مورد نیاز است (Halevy et al. 2009). در این چارچوب، تعجب آور نیست که سهامداران (از سرمایه‌گذاران و ناشران شروع می‌شوند) انگیزه زیادی برای به اشتراک‌گذاری داده‌ها نشان می‌دهند. روی دیگر سکه این است که درزهای ناشی از استفاده مخرب یا بی‌احتیاط از داده‌های ژنومی (Harmanci and Gerstein 2018) ممکن است پیامدهای شدید حریم خصوصی داشته باشد و حتی منجر به پدیده‌های تبعیض[1] شود (Joly et al. 2017). برای حفظ حریم خصوصی، پروژه‌های اشتراک‌گذاری مشترک بزرگ با بکارگیری الگوریتم‌های پیچیده ناشناس‌سازی و تعریف پروتکل‌های دسترسی دقیق به داده، روی امنیت سرمایه‌گذاری کرده‌اند. داده‌های عمومی فقط به صورت کمی‌سازی ویژگی‌های ژنومی در دسترس هستند، در حالی که خواندن خام به موسسات واجد شرایط محدود می‌شود. با این حال، همه این اقدامات احتیاطی لزوماً نمی‌توانند امنیت و انتظارات بیماران را از نظر حریم خصوصی برآورده کنند (Kaye 2012). برای مثال، ناشناس بودن با راهبردهای پیوند[2] (Harmanci and Gerstein 2018) به چالش کشیده می‌شود، که با افزایش ابعاد مجموعه داده، شناسایی مجدد را به تدریج آسان‌تر می‌کند. در مطالعات مختلف نشان داده شده که تحت شرایط خاص، شناسایی مجدد می‌تواند اکثریت هویت‌ها را در مجموعه داده‌های بزرگ با موفقیت آشکار کند (Sweeney et al. 2013; de Montjoye et al. 2017). از جنبه حقوق کاربر، اصل اساسی انصراف[3]، که حق قطع مشارکت در یک مطالعه تحقیقاتی را با حذف متعاقباً همه داده‌های شخصی (هم به صورت خام و هم به صورت انبوه) اعطا می‌کند، در مقیاس بزرگ بین‌المللی به اشتراک‌گذاری پروژه‌ها به دلیل غیر عملی بودن ردیابی داده‌ها قابل ردکردن است. مشکل بزرگ‌تر، نقض احتمالی حریم خصوصی است که می‌تواند از پیشرفت‌های الگوریتم‌های هوش مصنوعی ناشی شود. پیشرفت بیشتر ممکن است روش‌های هوش مصنوعی را قادر به استخراج اطلاعات ژنوتیپ‌سازی جدید و تصفیه‌شده‌تر کند، که در زمان امضای رضایت آگاهانه در نظر گرفته نشده بود (Greenbaum et al. 2011). در نگاه اول، به نظر می‌رسد که ما فرا خوانده شده ایم تا تصمیم بگیریم که آیا حریم خصوصی را قربانی مراقبت‌های بهداشتی مبتنی بر هوش مصنوعی کنیم یا برعکس. در واقع، تلاش گسترده‌ای برای یافتن راه‌حل‌هایی وجود دارد که هر دو نیاز را متعادل کند (Azencott 2018). شناسایی هویت[4] را می‌توان با سرکوب انتخابی داده‌ها[5] اجرا کرد. به عنوان مثال، روش k-anonymity(Sweeney 2002) روشی برای حذف یا تعمیم انتخابی داده‌ها است تا زمانی که هیچ ترکیبی از ویژگی با کمتر از

[1] Discrimination phenomena
[2] Linking strategies
[3] Withdrawal
[4] De-identification
[5] Selective data suppression

kرکورد به اشتراک گذاشته نشود. حریم خصوصی دیفرانسیل[1] (Wood et al. 2018) نویز کنترل شده را در داده‌ها معرفی می‌کند تا احتمال خروجی یک پرس و جو داده شده از پایگاه داده n رکورد را به حداکثر برساند، تا شبیه به پایگاه داده با n-1 رکورد باشد. با توجه به تضمین‌های ریاضی ارائه شده توسط این روش، الگوریتم‌های تخصصی هوش مصنوعی توسعه یافته است. به عنوان مثال، Abadi et al. (2016) چارچوبی برای DL با حریم خصوصی دیفرانسیل معرفی می‌کند. راه حل‌های دیگر حفظ حریم خصوصی شامل یادگیری از داده‌های رمزگذاری شده یا استفاده از مدل‌های شبکه عصبی مولد برای شبیه‌سازی داده‌های واقعی است (Beaulieu-Jones et al. 2019).

نرم افزار منبع باز-مسئولیت و قابلیت اطمینان[2]: به اشتراک گذاری نرم‌افزار، همانند به اشتراک‌گذاری داده توسعه ژنومیک عملکردی را تقویت کرده است. مجموعه‌های نرم‌افزاری یکپارچه و فرمت‌های فایل استاندارد شده، ایجاد خطوط لوله تجزیه و تحلیل[3] را کمتر استرس‌زا کرده است، در حالی که برنامه‌های کاربردی وب به آزمایشگاه‌هایی با منابع محاسباتی محدود نیز دسترسی به آزمایش‌های بزرگ را داده‌اند. در واقع، اشتراک نرم‌افزار اثرات اقتصادی مثبتی نیز دارد. امروزه، چندین پروژه تحقیقاتی به فراوانی ابزارهای در دسترس عموم برای غربالگری در مقیاس بزرگ، با هدف محدود کردن آزمایش‌های گران‌قیمت و زمان‌بر در شرایط آزمایشگاهی تنها به ویژگی‌های ژنومی امیدوارکننده، تکیه می‌کنند. با این حال، در این موارد، موفقیت پروژه به شدت به قابلیت اطمینان نرم‌افزار موجود بستگی دارد. اکثر نرم‌افزارهای بیوانفورماتیک تحت مجوز GPL[4] یا مجوز MIT منتشر می‌شوند، که هر دو فرصت عالی برای استفاده مجدد از کد را ارائه می‌دهند و در نتیجه از توسعه سریع ابزارهای جدید پشتیبانی می‌کنند. در عین حال، هر دو دارای یک سلب مسئولیت هستند که ضمانت و مسئولیت را به شدت محدود می‌کنند، بنابراین تمام مسئولیت خرابی نرم‌افزار را به کاربر نهایی واگذار می‌کنند. اگرچه کلیات آن قابل قبول است، اما عدم وجود تضمین نیاز به احتیاط دارد. برای مثال، در برنامه‌های کاربردی حیاتی که در آن پروتکل‌های برنامه نویسی مانند برنامه نویسی تدافعی (Yang and Lodgher 2019) باید برای افزایش کیفیت نرم‌افزار استفاده شود. مشاهدات فوق این سوال طبیعی را مطرح می‌کند که آیا می‌توانیم به ابزارهای بیوانفورماتیک اعتماد کنیم یا خیر. همانطور که توسط Lawlor and Walsh (2015) بحث شده، چندین موضوع حیاتی باید برای بهبود قابلیت اطمینان نرم افزار بیوانفورماتیک مورد بررسی قرار گیرند. با این حال، یک انتخاب دقیق به ما اجازه می‌دهد تا به این سوال پاسخ مثبت بدهیم. به عنوان مثال، در پژوهشی نویسندگان (Giannoulatou et al. 2014) یک ارزیابی از پیش (ex-post) از سه ابزار همسویی رایج (bwa، bowtie و bowtie2) انجام می‌دهند. در همین پژوهش، محققان با مشکل ارزیابی در غیاب استاندارد طلایی مواجه هستند که پیشنهادات مفیدی برای توسعه دهندگان ارائه می‌دهد. به طور کلی، جامعه بیوانفورماتیک در حال آگاه شدن از نیاز به ابزارهای قابل اعتماد است، همانطور که هم درخواست ناشران برای آزمایش‌های دقیق قبل از انتشار نرم‌افزار و هم تعداد پیشنهادات

[1] Differential privacy
[2] Open-source software: Liability and reliability
[3] Analysis pipelines
[4] General Public Licence

بهترین شیوه‌ها که در حال بحث است، گواه آن است (Seemann 2013; Leprevost et al. 2014). انتظار می‌رود که اتخاذ این شیوه‌ها در آینده نزدیک با افزایش قابلیت اطمینان ابزارها رایج شود. در پژوهشی جدید (Mahmud et al. 2021) بهره برداری از مجموعه‌ای از ابزارهای منبع باز DL و داده‌های دسترسی باز را مورد بحث قرار دادند و این ابزارها را از دیدگاه کیفی، کمی و معیار[1] مقایسه کردند.

مسائل حقوقی، اخلاقی و اقتصادی

کاربردهای هوش مصنوعی در ژنومیک عملکردی، و به ویژه در مراقبت‌های بهداشتی، اغلب به شناسایی الگوهای خاصی اختصاص دارد که متعاقباً توسط تصمیم گیرندگان برای تشخیص و درمان استفاده می‌شود. با این حال، طبقه‌بندی‌کننده‌ها خطاناپذیر نیستند و می‌توانند مرتکب دو نوع خطا شوند: اختصاص دادن عناصر غیرعضو به یک کلاس (مثبت نادرست[2]) یا عدم تشخیص اینکه برخی از عناصر متعلق به یک کلاس هستند (منفی‌های نادرست). در صورتی که کلاس "بیمار" باشد، نتایج مثبت کاذب ممکن است افراد را در معرض پریشانی غیرضروری قرار دهد و همچنین هزینه‌ها را با ترویج غربالگری‌های غیرضروری و اغلب پرهزینه افزایش دهد (Wolff et al. 2020). از سوی دیگر، منفی کاذب منجر به تأخیر قبل از تدوین تشخیص صحیح می‌شود و تأثیر چشمگیری در آسیب شناسی‌هایی مانند سرطان دارد، جایی که زمان تشخیص به شدت بر نتیجه طولانی مدت درمان‌ها تأثیر می‌گذارد. جای تعجب نیست که طبقه بندی اشتباه منبع اصلی مسائل اقتصادی و حقوقی است. با توجه به سناریوهای مختلف ناشی از دو نوع خطا، برخی از پژوهشگران (Landgrebe et al. 2004) یک ارزیابی مبتنی بر هزینه از عملکرد طبقه‌بندی را پیشنهاد کرده‌اند. با این حال، استفاده از روش‌هایی از این نوع پیچیده است. در واقع، اگرچه هزینه مثبت کاذب را می‌توان به طور منطقی بر اساس هزینه‌های غربالگری تخمین زد، اما تعیین کمیت منفی‌های کاذب قطعا مشکل سازتر است (Rao and Makkithaya 2017) زیرا به معنای تخصیص ارزش هزینه به زندگی انسان است. در حالی که روش‌های بدون خطا را دنبال می‌کنیم، می‌توان دو استراتژی را برای مقابله با طبقه‌بندی اشتباه اتخاذ کرد: ۱) اتخاذ دستورالعمل‌هایی که شامل بررسی مجدد توسط متخصصان انسانی برای پیچیده‌ترین موارد می‌شود و ۲) روش‌های هوش مصنوعی ویژه آموزش دیده برای تشخیص خطاهای طبقه‌بندی را توسعه دهیم. به عنوان نمونه‌ای از مورد اخیر، Aboutalib et al. (2018) یک شبکه عمیق طراحی شده برای تجزیه و تحلیل مجدد نمونه‌هایی که با روش دیگری به عنوان مثبت طبقه بندی شده بودند، پیشنهاد کرد. هدف در این مورد، یافتن موارد مثبت کاذب بود. یکی دیگر از مسائل مهم در مورد استفاده از هوش مصنوعی در ژنومیک عملکردی یا پزشکی دقیق، انصاف داده‌ها[3] است. روزنامه‌ها داستان‌های زیادی در مورد رفتار تبعیض آمیز سیستم‌های هوش مصنوعی دارند. این رفتار معمولاً از این واقعیت ناشی می‌شود که الگوریتم‌ها بر

[1] Benchmarking
[2] False positives
[3] Data fairness

Agricultural Biotechnology Journal; Print ISSN: 2228-6705, Electronic ISSN: 2228-6500

روی داده‌های مغرضانه یا نامتعادل آموزش داده می‌شوند. برای مثال، یادگیری از داده‌های اریب جمعیت می‌تواند به شناسایی واریانت‌ها و نشانگرهای زیستی مختص اجداد[1] منجر شود. اگر این می‌تواند برای شناسایی انواع نادر مفید باشد (Sidore et al. 2015)، جایی که جمعیتی از افراد ساردینیا برای یافتن جایگاه‌های جدید مرتبط با سطح لیپیدهای خون و نشانگرهای التهابی مورد استفاده قرار گرفتند)، همچنین خطر بالقوه افزایش نابرابری را در میان جمعیت‌هایی که نماینده کمی دارند و جمعیت‌هایی که بیش از حد نماینده دارند در بر دارد. به منظور کاهش این خطر در میان جمعیت‌های کم نماینده و آن‌هایی که بیش از حد نماینده دارند (Martin et al. 2019)، طرح‌های محلی جمع‌آوری داده در مقیاس بزرگ (به عنوان مثال به نسخه ژاپنی اطلس ژنوم سرطان (Nagashima et al. 2020) مراجعه کنید) باید در سطح بین‌المللی پشتیبانی و با بانک‌های داده موجود ادغام (یکپارچه) شوند.

نتیجه‌گیری: شیوع هوش مصنوعی تقریباً تمام زمینه‌های تحقیقاتی، به ویژه آن‌هایی که با کلان داده‌ها سروکار دارند، مانند ژنومیک عملکردی را تحت تأثیر قرار داده است. در میان شاخه‌های مختلف ژنومیک عملکردی، فناوری‌های توالی‌یابی نسل بعدی و نسل سوم، حجم وسیعی از داده‌ها را در سال‌های گذشته تولید کرده‌اند. کشف روابط بین واریانت‌ها و بیماری‌ها، جهش‌های اپی ژنتیک و بیان ژن، موقعیت‌های محل اتصال و فرآیندهای نظارتی بیشتر و بیشتر به دلیل توسعه و در دسترس بودن ابزارهای هوش مصنوعی جذاب‌تر شده است. به‌ویژه، معماری‌های عمیق می‌توانند به سطوح بالایی از انتزاع و توانایی سازمان‌دهی سلسله مراتبی مقادیر زیادی از داده‌ها با ماهیت متفاوت اما بسیار به هم پیوسته دست یابند، و آن‌ها را قابل تفسیرتر سازند. از جنبه منفی، معماری‌های عمیق هوش مصنوعی باعث می‌شوند که ما نتوانیم توضیح دهیم که در نهایت چگونه نتایج صحیح در کارهای حیاتی به دست آمده است. ارتباط شناسایی پروفایل بیان ژن با سرطان مسلم است، اگرچه مبهم بودن فرآیند انتخاب ویژگی اساسی ممکن است ارزش عملی این یافته را تضعیف کند. عدم توضیح برخی از مسیرهای DL همچنین انتخاب بهترین معماری برای استفاده برای یک کار مشخص را دشوار می‌کند، به همین دلیل است که به اشتراک‌گذاری و در دسترس بودن رایگان نرم افزار، منابع و پایگاه داده برای افزایش کاربردهای هوش مصنوعی در ژنومیک عملکردی بسیار مهم است. با توجه به کاربردهای حیاتی که اغلب توسط زیست شناسی و به ویژه ژنومیک عملکردی به آن پرداخته می‌شود، بهتر است با ابزارهای هوش مصنوعی که قادر به کمک به درک مکانیکی فرآیندهای بیولوژیکی هستند، سروکار داشته باشیم. به عبارت دیگر، فعال کردن بیولوژی سیستم‌ها برای به دست آوردن مزایایی از نتایج هوش مصنوعی در ژنومیک عملکردی مهم است. این امر مستلزم توانایی کمک به اعتبارسنجی یا حتی ساختن مدل‌های نظری با هدف داشتن ارزش پیش‌بینی در رفتار استاتیک و/یا دینامیکی سیستم‌های بیولوژیکی با پیچیدگی‌های مختلف است. تفسیرپذیری، به معنایی که در بالا توضیح داده شد، مطمئناً می‌تواند به هوش مصنوعی کمک کند تا در کاربردهای عملی مانند پزشکی راحت‌تر پذیرفته شود. به نظر ما، افزایش حجم و تنوع داده‌های عظیم قابل اعتماد و ادغام آن با مدل‌سازی نظری به افزایش اعتماد انسان‌ها به پیش‌بینی‌ها و تصمیم‌گیری‌های مبتنی بر هوش مصنوعی در آینده کمک می‌کند. تا آنجا که به آینده هوش

[1] Ancestry-specific variants and biomarkers

مصنوعی در سیستم‌های بیولوژی مربوط می‌شود، دو سناریو مختلف پیشنهاد شده است که شامل رقابت یا همکاری بین داده‌ها و رویکردهای مدل محور است. ما معتقدیم که همکاری و ادغام بین این دو رویکرد برای دستیابی به درک مناسبات و مکانیسم‌های سیستم بسیار مفید خواهد بود. در واقع، از یک سو، رویکردهای مبتنی بر مدل می‌توانند محدودیت‌های مبتنی بر دانش را فراهم کنند. از سوی دیگر، نتایج هوش مصنوعی می‌تواند به ایجاد پارامترهای مدل‌های سیستم‌های بیولوژی کمک کنند. یکی از چشمگیرترین دستاوردهای روش‌های هوش مصنوعی در ژنومیکس عملکردی تا کنون، بدون شک انقلابی است که با روش DeepMind AlphaFoldدر زمینه پیش‌بینی ساختار پروتئین ارائه شده است، تا جایی که این یکی از بزرگترین چالش‌های زیست‌شناسی، به عنوان یک مشکل حل شده است که باید مورد توجه قرار گیرد (Callaway 2020). موفقیت چشمگیر نرم‌افزار AlphaFold تا حد زیادی به ترانسفورماتورها متکی است که در بازنمایی توالی و تفسیر زمینه و محتویات از معماری‌های عمیق پیشرفته بهتر عمل می‌کنند. در واقع، معرفی آن‌ها کاربردهای بسیاری را در زمینه زیست شناسی برانگیخته است، زیرا مدل‌های بدون نظارت مبتنی بر مکانیسم توجه، از پیش آموزش داده شده بر روی مجموعه داده‌های بزرگ، به طور موثر در تشخیص همسانی، نمایش برهمکنش رزیدو–رزیدو، پیش بینی ساختار ثانویه و زیست شناسی مولد، علاوه بر پیش‌بینی‌های ساختار سوم پروتئین عمل می‌کنند. ژنومیکس عملکردی، و همچنین تمامی رشته‌های پزشکی، زیست شناسی و سایر علومی که هم حقوق فردی و هم حقوق جمعی در آن دخیل است، یک حوزه تحقیقاتی پیچیده است. کسانی که می‌خواهند از هوش مصنوعی در این زمینه استفاده کنند متوجه خواهند شد که پیمایش دشوار است، نه تنها به دلیل تنوع زیاد و حجم زیاد داده‌های موجود، نه فقط برای فهمیدن این‌که چه سوالی بپرسند و از چه ابزاری استفاده کنند، بلکه همچنین به دلیل این که این رشته نسبت به جنبه‌های قانونی، اخلاقی و معنوی بسیار حساس است. این بررسی برای کمک به افرادی که مایل به نزدیک شدن به کاربرد روش‌های هوش مصنوعی در ژنومیک عملکردی هستند، برای شناسایی جنبه‌های مختلف موضوع و یافتن ابزارهای مفید برای جهت‌یابی در نظر گرفته شده است. هوش مصنوعی فرصت‌های زیادی را می‌گشاید که از ترس درک نکردن همه مراحل نباید از آن‌ها امتناع کنیم. راهی که توسعه هوش مصنوعی را دنبال می‌کند، تازه شروع به باز شدن کرده است. وعده‌های زیادی را به همراه دارد و خطرات بالقوه زیادی را پیش‌رو دارد. این مسیر احتمالا طولانی و غیرقابل برگشت خواهد بود. هوش مصنوعی زندگی ما را تغییر خواهد داد و ما باید در اسرع وقت نظر خود را تغییر دهیم تا تغییرات ناشی از آن را به بهترین شکل ممکن تطبیق دهیم، بپذیریم و مدیریت کنیم، تا اطمینان حاصل کنیم که آن‌ها تا حد امکان منافع بیشتری به همراه خواهند داشت و کمترین پیامدهای منفی ممکن را برای ما ایجاد خواهند کرد.

سپا سگزاری: نگارندگان بر خود لازم می‌دانند از داوران و سردبیر محترم مجله به خاطر ارائه نظرهای ساختاری و علمی سپاسگزاری نمایند.

References

Abadi M, Chu A, Goodfellow I, et al. (2016) Deep learning with differential privacy. ArXiv 1607, e00133.

Abadi S, Yan WX, Amar D, Mayrose I (2017) A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. PLoS Comput Biol 13, e1005807.

Abeel T, Helleputte T, de Peer YV, et al. (2010) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26(3), 392-398.

Aboutalib SS, Mohamed AA, Berg WA, et al. (2018) Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clin Cancer Res 24, 5902-599.

Agarwal V, Bell GW, Nam JW (2015) Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. 4, e05005.

Aggarwal N, Singh D (2021) Technology Assisted Farming: Implications of IoT and AI. IOP Conf Ser Mater Sci Eng 1022, e012080.

Agris PF (1996) The importance of being modified: roles of modified nucleosides and Mg^{2+} in RNA structure and function. Prog Nucleic Acid Res Mol Biol 53, 79-129.

Ahsani MR, Mohammadabadi MR, Shamsaddini MB (2010) Clostridium perfringens isolate typing by multiplex PCR. J Venom Anim Toxins Includ Trop Dis 16 (4), 573-578.

Akhtar N, Rasheed Z, Ramamurthy S, et al. (2010) MicroRNA-27b regulates the expression of matrix metalloproteinase 13 in human osteoarthritis chondrocytes. Arthritis Rheum 62(5), 1361-1371.

Alakwaa FM, Chaudhary K, Garmire LX (2018) Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. J Proteome Res 17, 337-347

Albattah W, Javed A, Nawaz M, et al. (2022) Artificial Intelligence-Based Drone System for Multiclass Plant Disease Detection Using an Improved Efficient Convolutional Neural Network. Front Plant Sci 13, e808380.

Ali Q, Ahmar S, Sohail MA, et al. (2021) Research Advances and Applications of Biosensing Technology for the Diagnosis of Pathogens in Sustainable Agriculture. Environ Sci Pollut Res Int 28, 9002-9019.

Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnol 33, 831–838.

Amin N, McGrath A, Chen YPP (2019) Evaluation of deep learning in non-coding RNA classification. *Nat Mach Intell* 1, 246-256.

Amiri Roudbar M, Abdollahi-Arpanahi R, Ayatollahi Mehrgardi A, et al. (2018). Estimation of the variance due to parent-of-origin effects for productive and reproductive traits in Lori-Bakhtiari sheep. Small Rumin Res 160, 95-102.

Amiri Roudbar M, Mohammadabadi MR, Mehrgardi AA, Abdollahi-Arpanahi A (2017) Estimates of variance components due to parent-of-origin effects for body weight in Iran-Black sheep. Small Rumin Res 149, 1-5

An J-Y, Meng FR, You ZH, et al. (2016) Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model. Protein Sci 25, 1825-1833.

Anderson S, Bankier AT, Barrell BG, et al. (1981) Sequence and organization of the human mitochondrial genome. Nature 290, 457-465.

Angermueller C, Pärnamaa T, Parts L, Stegle O (2016) Deep learning for computational biology. Mol Syst Biol 12(7), e 878.

Antonakoudis A, Barbosa R, Kotidis P, Kontoravdi C (2020) The era of big data: Genome-scale modelling meets machine learning. Comput Struct Biotechnol J 18, 3287-3300.

Arisdakessian CG, Poirion OB, Yunits B, et al. (2019) Deepimpute: an accurate, fast and scalable deep neural network method to impute single-cell rna-seq data. Genome Biol 20(1), e211.

Arshad MF, Burrai GP, Varcasia A, et al. (2024) The groundbreaking impact of digitalization and artificial intelligence in sheep farming. Res Vet Sci 170, e105197.

Asgari E, Mofrad M, Kobeissy F (2015) Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS ONE 10(11), e0141287.

Aslam B, Basit M, Nisar MA, et al. (2017) Proteomics: Technologies and their applications. J Chromatogr Sci 55(2), 182-196.

Avsec Z, Weilert M, Shrikumar A, et al. (2021) Base-resolution models of transcription factor binding reveal soft motif syntax. Nat Genet 53, 354-366.

Azencott CA (2018) Machine learning and genomics: precision medicine versus patient privacy. Philosophical Transactions of the Royal Society A: Math Phys Engin Sci 376, e20170350.

Baek J, Lee B, Kwon S, Yoon S (2018) LncRNAnet: Long non-coding RNA identification using deep learning. Bioinformatics 34(22), 3889-3897.

Bao J, Xie Q (2022) Artificial intelligence in animal farming: a systematic literature review. J Clean Prod 331, e12995.

Bar-Shira A, Panet A, Honigman A (1991) An RNA secondary structure juxtaposes two remote genetic signals for human t-cell leukemia virus type I RNA 3’-end processing. J Virol 65(10), 5165-5173.

Basciftci F, Gunduz KA (2019) Identification of acidosis disease in cattle using IoT. In: 4th International Conference on Computer Science and Engineering (UBMK). IEEE, pp. 58–62.

Beaulieu-Jones BK, Wu ZS, Williams C, et al. (2019) Privacy-preserving generative deep neural networks support clinical data sharing. Circ Cardiovasc Qual Outcomes 12(7), e005122.

Bedi P, Gole P (2021) Plant Disease Detection Using Hybrid Model Based on Convolutional Autoencoder and Convolutional Neural Network. Artif Intell Agric 5, 90-101.

Bejnordi BE, Veta M, van Diest PJ, et al. (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. J Am Med Assoc 318, 2199-2210.

Bell JT, Pai AA, Pickrell JK, et al. (2011) DNA methylation patterns associate with genetic and gene expression variation in hapmap cell lines. Genome Biol 12(1), R10.

Belokopytova P, Fishman V (2020) Predicting genome architecture: Challenges and solutions. Frontiers Genet 11, e 617202.

Ben Ayed R, Hanana M (2021) Artificial Intelligence to Improve the Food and Agriculture Sector. J Food Qual 2021, e5584754.

Bentley DR, Balasubramanian S, Swerdlow HP, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53-59.

Berckmans D (2017) General introduction to precision livestock farming. Anim Front 7, 6-11.

Berger MF, Mardis ER (2018) The emerging clinical relevance of genomics in cancer medicine. Nat Rev Clinic Oncol 15, 353-365.

Bersanelli M, Mosca E, Remondini D, et al. (2016) Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics 17, S15.

Bhardwaj A, Kishore S, Pandey DK (2022) Artificial Intelligence in Biological Sciences. Life 12, e1430.

Bleidorn C (2016) Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Syst Biodivers 14, 1-8.

Bondi E, Oh H, Xu H, et al. (2019) Using Game Theory in Real Time in the Real World: A Conservation Case Study. Proc 18th Int Conf Autonomous Agents MultiAgent Syst, pp. 2336-2338.

Bonnel N, Marteau PF (2012) Lna: fast protein structural comparison using a laplacian characterization of tertiary structure. IEEE/ACM Trans. Comput Biol Bioinform 9, 1451-1458.

Braslavsky I, Hébert B, Kartalov E, Quake S (2003) Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci USA 100, 3960-3964.

Breschi A, Muñoz-Aguirre M, Wucher V, et al. (2020) A limited set of transcriptional programs define major histological types and provide the molecular basis for a cellular taxonomy of the human body. Genome Res 30(7), 1047-1059.

Bretschneider H, Gandhi S, Deshwar AG, et al. (2018) COSSMO: Predicting competitive alternative splice site selection using deep learning. Bioinformatics 34(13), i429-i437.

Brown PH, Tiley L, Cullen B (1991) Effect of RNA secondary structure on polyadenylation site selection. Genes Dev 5(7), 1277-1284.

Bucher P (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol 212, 563-578.

Buermans HP, den Dunnen JT (2014) Next generation sequencing technology: Advances and applications. Biochim Biophys Acta 1842(10), 1932-1941.

Callaway E (2020) It will change everything: DeepMind′s AI makes gigantic leap in solving protein structures. Nature 588, 203-204.

Camacho DM, Collins KM, Powers RK, et al. (2018) Next-generation machine learning for biological networks. Cell 173, 1581-1592.

Camargo AP, Sourkov V, Pereira GAG, Carazzolle MF (2020) RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. NAR Genom Bioinform 2(1), lqz024.

Cao C, Liu F, Tan H, et al. (2018) Deep learning and its applications in biomedicine. Genom Proteom Bioinform 16, 17-32.

Caudai C, Salerno E, Zoppè M, et al. (2019a) Chromstruct 4: A python code to estimate the chromatin structure from hi-c data. IEEE/ACM Trans Comput Biol Bioinform 16, 1867-1878.

Caudai C, Salerno E, Zoppè M, Tonazzini A (2019b) Estimation of the spatial chromatin structure based on a multiresolution bead-chain model. IEEE/ACM Trans Comput Biol Bioinform 16, 550-559.

Cavill R, Keun HC, Holmes E, et al. (2009) Genetic algorithms for simultaneous variable and sample selection in metabonomics. Bioinformatics 25(1), 112-118.

Chai H, Zhou X, Zhang Z, et al. (2021) Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Computers in biology and medicine. 134, e104481.

Chang TW (1983) Binding of cells to matrixes of distinct antibodies coated on solid surface. J Immunol Methods 65, 217-223.

Chaudhary K, Poirion OB, Lu L, Garmire LX (2018) Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res 24, 1248–59.

Chaudhary K, Poirion OB, Lu L, Garmire LX (2018) Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res 24, 1248-1259.

Chen H, Engkvist O, Wang Y, et al. (2018) The rise of deep learning in drug discovery. Drug Discov Today 23(6), 1241-1250.

Chen J, Shao J (2000) Nearest neighbor imputation for survey data. J Off Stat 16(2), 113-131.

Chen K, Wei Z, Zhang Q, et al. (2019) Whistle: a high-accuracy map of the human n6-methyladenosine (m6a) epitranscriptome predicted using a machine learning approach. Nucleic Acids Res 47, e41.

Chen L, Xuan J, Riggins RB, et al. (2011) Identifying cancer biomarkers by network-constrained support vector machines. BMC Syst Biol 5, e161.

Cheng Y, Miura R, Tian B (2006) Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics 22(19), 2320-2325.

Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. (2018) Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 15(141), e20170387.

Chuai G, Ma H, Yan J, et al. (2018) DeepCRISPR: optimized crispr guide rna design by deep learning. Genome Biol 19, e80.

Cismondi F, Fialho AS, Vieira SM, et al. (2013) Missing data in medical databases: Impute, delete or classify? Artif Intell Med 58(1), 63-72.

Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature 431, 931-945.

Cooper J, Noon M, Jones C, et al. (2013) Big data in life cycle assessment. J Ind Ecol 17(6), 796-799.

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20, 273-297.

Costanzo M, Baryshnikova A, Bellay J, et al. (2010) The genetic landscape of a cell. Science 327(5964), 425-431.

Costanzo M, VanderSluis B, Koch EN, et al. (2016) A global genetic interaction network maps a wiring diagram of cellular function. Science 353(6306), aaf1420,

Costello Z, Martin Garcia H (2018) A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. Npj Syst Biol Appl 4, e19.

D'Agaro E, Rosa F, Akentieva NP (2021) New technology tools and life cycle analysis (LCA) applied to a sustainable livestock production. Eur J 5(3), 130-141.

Dao F-Y, Lv H, Yang Y, et al. (2020) Computational identification of n6-methyladenosine sites in multiple tissues of mammals. Comput Struct Biotechnol J 18, 1084-1091.

Das S, Deng X, Camphausen K, et al. (2019) An R package for multi-omic data driven prediction of synthetic lethality in cancers. Bioinformatics (Oxford, England) 35, 701-702.

de Montjoye Y-A, Farzanehfar A, Hendrickx J, Rocher L (2017) Solving artificial intelligence′s privacy problem. Field Action Sci Rep 2017, 80-83.

de Ridder D, de Ridder J, Reinders MJT (2013) Pattern recognition in bioinformatics. Brief Bioinform 14, 633-647.

De Souto MC, Jaskowiak PA, Costa IG (2015) Impact of missing data imputation methods on gene expression clustering and classification. BMC Bioinform 16, e64.

De Vries M, de Boer IJ (2010) Comparing environmental impacts for livestock products: a review of life cycle assessments. Livest Sci 128(1–3), 1-11.

Degroeve S, De Baets B, Van de Peer Y, Rouzé P (2002) Feature subset selection for splice site prediction. Boinformatics 18, 75-83.

Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14, 390-403.

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Pre-training of deep bidirectional transformers for language understanding. arXiv 1810, e04805.

Djebali S, Davis CA, Merkel A, et al. (2012) Landscape of transcription in human cells. Nature 489(7414), 101-108.

Doelman JC, Stehfest E, van Vuuren DP, et al. (2020) Afforestation for climate change mitigation: potentials, risks and trade-offs. Glob Change Biol 26(3), 1576-1591.

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, et al. (2012) Topology of the human and mouse m6a RNA methylomes revealed by m6a-seq. Nature 485, 201-216.

Doncescu A, Tauzain B, Kabbaj N (2009) Machine learning applied to BRCA1 hereditary breast cancer data. In: 2009 International Conference on Advanced Information Networking and Applications Workshops. p. 942–947.

Došilovic FK, Brcic M, Hlupic N (2018) Explainable artificial intelligence: A survey. In 41st Int Conven Inform Commun Technol Elect Microelec (MIPRO). IEEE 2018, 0210–0215.

Dutta Majumder D, Ulrichs C, Majumder D, et al. (2007) Current Status and Future Trends of Nanoscale Technology and Its Impact on Modern Computing, Biology, Medicine and Agricultural Biotechnology. In Proc Int Conf Comput Theory Appl, Kolkata, pp. 563–572.

Eickholt J, Cheng J (2012) Predicting protein residue-residue contacts using deep networks and boosting. Bioinformatics 28(23), 3066-3072.

Eli-Chukwu CN (2019) Applications of Artificial Intelligence in Agriculture: A Review. Eng Technol Appl Sci Res 9, 4377-4383.

EPC (The ENCODE Project Consortium) (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.

Eraslan G, Avsec Z, Gagneur J, Theis FJ (2019) Deep learning: new computational modelling techniques for genomics. Nat Rev Genet 20, 389-403.

Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 9, 215–6.

Ernst J, Kellis M (2015) Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nature Biotechnology 33, 364-376.

Esteva A, Robicquet A, Ramsundar B, et al. (2019) A guide to deep learning in healthcare. Nat Med 25, 24-29.

Fang C, Shang Y, Xu D (2018) MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction. Proteins 86, 592-598.

Fang C, Shang Y, Xu D (2018) Prediction of protein backbone torsion angles using deep residual inception neural networks. IEEE/ACM Trans Comput Biol Bioinform 10, e1109.

FAO (2022) How to Feed the World in 2050: Global Agriculture Towards 2050. Available online: https://www.fao.org/fileadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture.pdf (accessed on 21 July 2022).

Faraggi E, Zhang T, Yang Y, et al. (2012) Spine x: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J Comput Chemy 33, 259-267.

Fickett JW, Tung CS (1992) Assessment of protein coding measures. Nucleic Acids Res 20, 6441-6450.

Fleischmann R, Adams MD, White O, et al. (1995) Whole-genome random sequencing and assembly of haemophilus influenzae rd. Science 269(5223), 496-512.

Fortelny N, Bock C (2020) Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. Genome Biol 21, e190.

Frankish A, Daiekhans M, Ferreira M, et al. (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47(D1), D766-773.

Frith MC, Bailey TL, Kasukawa T, et al. (2006) Discrimination of Non-Protein-Coding Transcripts from Protein-Coding mRNA. RNA Biol 3(1), 40-48.

Frye M, Jaffrey SR, Pan T, et al. (2016) RNA modifications: What have we learned and where are we headed? Nat Rev Genet 17(6), 365-372.

Fudenberg G, Kelley DR, Pollard KS (2020) Predicting 3d genome folding from DNA sequence with akita. Nat Methods 17, 1111-1117.

Gao J, Yang Y, Zhou Y (2018) Grid-based prediction of torsion angle probabilities of protein backbone and its application to discrimination of protein intrinsic disorder regions and selection of model structures. BMC Bioinformatics 19, e29.

Garcia-Blanco M, Baraniak AP, Lasda EL (2004) Alternative splicing in disease and therapy. Nat Biotechnol 22, 535-546.

Garrow AG, Agnew A, Westhead DR (2005) Tmb-hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. Nucleic Acids Res 33, W188-W192.

Gautam C, Ravi V (2014) Evolving clustering based data imputation. In Int Conf Circuits Power Comput Technol (ICCPCT-2014). IEEE 2014, 1763-1769.

Ghahramani A, Watt FM, Luscombe NM (2018) Generative adversarial networks simulate gene expression and predict perturbations in single cells. BioRxiv 2018, e262501.

Ghotbaldini H, Mohammadabadi MR, Nezamabadi-pour H, et al. (2019) Predicting breeding value of body weight at 6-month age using Artificial Neural Networks in Kermani sheep breed. Acta Scientiarum Anim Sci 41, e45282.

Giannoulatou E, Park SH, Humphreys DT, Ho JW (2014) Verification and validation of bioinformatics software without a gold standard: a case study of bwa and bowtie. BMC Bioinformatics 15, S15.

Gilpin W, Huang Y, Forger DB (2020) Learning dynamics from large biological data sets: Machine learning meets systems biology. Curr Opin Syst Biol 22, 1-7.

Goffeau A, Barrell BG, Bussey H, et al. (1996) Life with 6000 Genes. Science 274, 546-567.

Gogichaeva NV, Williams T, Alterman M (2007) Maldi tof/tof tandem mass spectrometry as a new tool for amino acid analysis. J Am Soc Mass Spectrom 18(2), 279-284.

Gong L, Yu M, Jiang S, et al. (2021) Deep Learning Based Prediction on Greenhouse Crop Yield Combined TCN and RNN. Sensors 21, e4537.

Gong W, Kwak I-Y, Pota P, et al. (2018) Drimpute: imputing dropout events in single cell rna sequencing data. BMC Bioinformatics 19(1), e220.

Goodfellow I, Pouget-Abadie J, Mirza M, et al. (2014) Generative adversarial nets. ArXiv 1406, e2661.

Grapov D, Fahrmann J, Wanichthanarak K, Khoomrung S (2018) Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. Omics 22, 630-136.

Greenbaum D, Sboner A, Mu XJ, Gerstein M (2011) Genomics and privacy: implications of the new reality of closed data for the field. PLoS Comput Biol 7, e1002278.

Gui H, Li M, Sham PC, Cherny SS (2011) Comparisons of seven algorithms for pathway analysis using the wtccc crohn's disease dataset. BMC Res Notes 4, e386.

Gupta S, Chaudhary K, Kumar R, et al. (2016) Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine. Sci Rep 6, e23857.

Gusmao EG, Dieterich C, Zenke M, Costa IG (2014) Detection of active transcription factor binding sites with the combination of dnase hypersensitivity and histone modifications. Bioinformatics 30, 3143-3151.

Haenssle HA, Fink C, Schneiderbauer R, et al (2018) Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 29, 1836-1842.

Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. IEEE Intelligent Syst 24, 8-12.

Hamamoto R, Komatsu M, Takasawa K, et al. (2020) Epigenetics analysis and integrated analysis of multiomics data, including epigenetic data, using artificial intelligence in the era of precision medicine. Biomolecules 10, e62.

Hansen KD, Wu Z, Irizarry RA, Leek JT (2011) Sequencing technology does not eliminate biological variability. Nat Biotechnol 29, 572-573.

Hao J, Astle W, De Iorio M, Ebbels T (2012) Batman-an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a bayesian model. Bioinformatics 28, 2088-2090.

Haque F, Li J, Wu H, et al. (2013) Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. Nano Today 81, 56-74.

Harfouche AL, Jacobson DA, Kainer D, et al. (2019) Accelerating Climate Resilient Plant Breeding by Applying Next-Generation Artificial Intelligence. Trends Biotechnol 37, 1217-1235.

Harmanci A, Gerstein M (2018) Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nat Commun 9, 1-10.

Heather J, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8.

Heintzman ND, Stuart RK, Hon G, el al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39, 311-318.

Heitjan DF (1997) Annotation: what can be done about missing data? approaches to imputation. Am J Public Health 87, 548-450.

Hejase H, Chan CY (2015) Improving drug sensitivity prediction using different types of data. Pharmacometrics Syst Pharmacol 4(2), e2.

Heje Grønbech C, Vording MF, Timshel PN, et al. (2020) scvae: Variational auto-encoders for single-cell gene expression data. Bioinformatics 36(16), 4415-4422

Hieter P, Boguski M (1997) Functional genomics: it’s all how you read it. Science 278(5338), 601-602.

Hill ST, Kuintzle R, Teegarden A, et al. (2018) A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. Nucleic Acids Res 46, 8105-8113.

Hillenkamp F, Karas M, Beavis R, Chait B (1991) Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. Analytical chemistry 63(24), 1193A-1203A.

Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18, 1527-1554.

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313, 504-507.

Hiranuma N, Lundberg SM, Lee SI (2019) AIControl: replacing matched control experiments with machine learning improves ChIP-seq peak identification. Nucleic Acid Res 47, e58.

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9, 1735-1780.

Hoffman MM, Buske OJ, Wang J, et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods 9, 473-476.

Holder LB, Haque MM, Skinner MK (2017) Machine learning for epigenetics and future medical applications. Epigenetics 12, 505-514.

Huang L, Liao L, Wu CH (2018) Completing sparse and disconnected protein-protein network by deep learning. BMC Bioinformatics 19, e103.

Huang S, Cai N, Pacheco PP, et al. (2018) Applications of support vector machine (svm) learning in cancer genomics. Cancer Genomics Proteomics 15, 41–51.

Huang S, Chaudhary K, Garmire LX (2017) More is better: recent progress in multiomics data integration methods. Front Genet 8, e84.

Husson F, Josse J (2013) Handling missing values in multiple factor analysis. Food Qual Prefer 30, 77-85.

Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2, 343-272.

Ionita-Laza I, McCallum K, Xu B, Buxbaum JD (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 48, 214-220.

Israelsen BW, Ahmed NR (2019) dave...i can assure you...that it’s going to be all right...a definition, case for, and survey of algorithmic assurances in humanautonomy trust relationships. ACM Comput Surv (CSUR) 51, 1-37.

Jacobson MP, Pincus DL, Rapp CS, et al. (2004) A hierarchical approach to all-atom protein loop prediction. Proteins 55(2), 351-367.

Jafari Ahmadabadi SAA, Askari-Hemmat H, Mohammadabadi M, et al. (2023) The effect of Cannabis seed on DLK1 gene expression in heart tissue of Kermani lambs. Agric Biotechnol J 15(1), 217-234.

Jaganathan K, Panagiotopoulou SK, McRae JF, et al. (2019) Predicting Splicing from Primary Sequence with Deep Learning. Cell 176, 535–548.e24.

Jha A, Gazzara MR, Barash Y (2017) Integrative deep models for alternative splicing. Bioinformatics 33(14), i274-i282.

Jha K, Doshi A, Patel P, Shah MA (2019) Comprehensive Review on Automation in Agriculture Using Artificial Intelligence. Artif Intell Agric 2, 1-12.

Joly Y, Feze IN, Song L, Knoppers BM (2017) Comparative approaches to genetic discrimination: chasing shadows? Trends Genet 33, 299-302.

Josse J, Husson F (2016) Missmda: a package for handling missing values in multivariate data analysis. J Stat Softw 70, 1-31.

Ju S, Lim H, Heo J (2019) Machine Learning Approaches for Crop Yield Prediction with MODIS and Weather Data. In Proceedings of the 40th Asian Conference on Remote Sensing, ACRS 2019: Progress of Remote Sensing Technology for Smart Future, Daejeon, Korea.

Jumper J, Evans R, Pritzel A, et al. (2021) Highly accurate protein structure prediction with alphafold. Nature 596, 583-589.

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577-2637.

Kahles A, Lehman KV, Toussaint SC, et al. (2018) Comprehensive analysis of alternative splicing across tumors from 8,705 patients. Cancer Cell 34(2), 211–224.e6.

Kalinin AA, Higgins GA, Reamaroon N, et al. (2018) Deep learning in pharmacogenomics: from gene regulation to patient stratification. Pharmacogenomics 19(7), 629-650.

Kalkatawi M, Rangkuti F, Schramm M, et al. (2012) Dragon polya spotter: predictor of poly(a) motifs within human genomic DNA sequences. Bioinformatics 28(1), 127-129.

Kang YJ, Yang DC, Kong L, et al. (2017) CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res 45(W1), W12-W16.

Kaye J (2012) The tension between data sharing and the protection of privacy in genomics research. Ann Rev Genom Human Genet 13, 415-431.

Kelchtermans P, Bittremieux W, Grave KD, et al. (2013) Machine learning applications in proteomics research: How the past can boost the future. Proteomics 14, 353-366.

Kelley DR, Reshef YA, Bileschi M, et al. (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res 28(5), 739-750.

Kelley DR, Snoek J, Rinn JL (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research 26, 990-999.

Khodayari A, Maranas CD (2016) A genome-scale *escherichia coli* kinetic metabolic model k-*E* .*coli*457 satisfying flux data for multiple mutant strains. Nat Commun 7, e13806.

Khorshidi M, Mohammadabadi MR, Esmailizadeh AK, et al. (2019) Comparison of artificial neural network and regression models for prediction of body weight in Raini Cashmere goat. Iran J Appl Anim Sci 9 (3), 453-461.

Killoran N, Lee LJ, Delong A, et al. (2017) Generating and designing dna with deep generative models. arxiv 1712, e06148.

Kim H, Park H (2003) Protein secondary structure prediction based on an improved support vector machines approach. Protein Eng 16, 553-560.

Kim J, Tae D, Seok J (2020) A survey of missing data imputation using generative adversarial networks. In Int Conf AI Inform Commun (ICAIIC). IEEE 2020, 454–456.

Kim M, Gilley JE (2008) Artificial Neural Network Estimation of Soil Erosion and Nutrient Concentrations in Runoff from Land Application Areas. Comput Electron Agric 64, 268-275.

Kircher M, Witten DM, Jain P, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46, 310-315.

Kitano H (2002). Systems biology: a brief overview. Science 295, 1662-1664.

Klyushin D, Tymoshenko A (2021) Optimization of Drip Irrigation Systems Using Artificial Intelligence Methods for Sustainable Agriculture and Environment. Stud Comput Intell 912, 3-17.

Koh PW, Pierson E, Kundaje A (2017) Denoising genome-wide histone chip-seq with convolutional neural networks. Bioinformatics 33, 225-233.

Koike A, Takagi T (2004) Prediction of protein-protein interaction sites using support vector machines. Protein Eng Des Sel 17(2), 165-173.

Kong L, Zhang Y, Ye ZQ, et al. (2007) CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35(W), W345-349.

Kramer MA (1991) Nonlinear principal component analysis using autoassociative neural networks. AIChE J 37, 233-243.

Kristensen LS, Andersen MS, Stagsted LVW, et al. (2019) The biogenesis, biology and characterization of circular RNAs. Nat Rev Genet 20(11), 675-691.

Lal A, Chiang ZD, Duarte FM, et al. (2021) Deep learning-based enhancement of epigenomics data with atacworks. Nat Commun 12, e1507.

Lander E, Linton LM, Birren B, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409(6822), 860-921.

Landgrebe T, Paclík P, Tax DM, et al. (2004) Cost-basedmclassifier evaluation for imbalanced problems. In: Fred A, Caelli TM, Duin RPW, Campilho AC, de Ridder D (eds) Structural, Syntactic, and Statistical Pattern Recognition. SSPR /SPR 2004. Lecture Notes in Computer Science, vol 3138. Springer, Berlin,

Lapuschkin S, Waldchen S, Binder A, et al. (2019) Unmasking clever hans predictors and assessing what machines really learn. Nat Commun 10, 1-8.

Lawlor B, Walsh P (2015) Engineering bioinformatics: building reliability, performance and productivity into bioinformatics software. Bioengineered 6, 193-203.

Lazar C, Meganck S, Taminau J, et al. (2013) Batch effect removal methods for microarray gene expression data integration: a survey. Briefings in Bioinformatics 14, 469-490.

Le NQK, Ho QT, Nguyen TTD, Ou YYA (2021) Transformer architecture based on bert and 2d convolutional neural network to identify dna enhancers from sequence information. Brief Bioinform 22(5), bbab005.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521, 436-444.

LeCun Y, Buttou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings IEEE 86, 2278-2324.

Lehner B (2013) Genotype to phenotype: lessons from model organisms for human genetics. Nature reviews. Genetics 14, 168–78.

Leoni G, Le Pera L, Ferrè F, et al. (2011) Coding potential of the products of alternative splicing in human. Genome Biol 12(1), R9.

Leprevost FDV, Barbosa VC, Francisco EL, et al. (2014) On best practices in the development of bioinformatics software. Front Genet 5, e199.

Leung M, Delong A, Frey B (2017) Inference of the human polyadenylation code. Bioinformatics 34, 2889-2798.

Leung MK, Xiong HY, Lee LJ, Frey BJ (2014) Deep learning of the tissue-regulated splicing code. Bioinformatics 30(12), i121–i129.

Li A, Zhang J, Zhou Z. PLEK (2014) A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics 15(1), e311.

Li H, Giger ML, Huynh BQ, Antropova N (2017) Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. J Med Imaging 4(4), e 041304.

Li VR, Zhang Z, Troyanskaya OG (2021) CROTON: an automated and variant-aware deep learning framework for predicting CRISPR/Cas9 editing outcomes. Bioinformatics 37, i342-348.

Li WV, Li JJ (2018) An accurate and robust imputation method scimpute for single cell rna-seq data. Nat Commun 2018;9(1), e997.

Li Y, Huang Ch, Ding L, et al. (2019) Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. Methods 166, 4-21.

Liakos KG, Busato P, Moshou D, et al. (2018) Machine Learning in Agriculture: A Review. Sensors 18, e2674.

Libbrecht MW, Stafford Noble W (2015) Machine learning applications in genetics and genomics. Nat Rev Genet 16, 321-332.

Lin D, Bonora G, Yardimci GG, Noble WS (2019) Computational methods for analyzing and modeling genome structure and organization. Wiley Interdiscip Rev Syst Biol Med 11(1), e1435.

Linaza MT, Posada J, Bund J, et al. (2021) Data-Driven Artificial Intelligence Applications for Sustainable Precision Agriculture. Agronomy 11, e1227.

Little R, Rubin D (1987) Statistical Analysis with Missing Data. John Wiley & Sons Publication.

Liu H, Han H, Li J, Wong L (2005) DNAFSMiner: a web-based software toolbox to recognize two types of functional sites in DNA sequences. Bioinformatics 21(5), 671-673.

Liu J, Gough J, Rost B (2006) Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet 2(4), e29.

Liu X (2017) Deep recurrent neural network for protein function prediction from sequence. Arxiv 1701, e08318.

Louhimo R, Hautaniemi S (2011) CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics 27, 887-888.

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol 15, e550.

Lowe R, Shirley N, Bleackley M, et al. (2017) Transcriptomics technologies. PLOS Computational Biology 13, e1005457.

Ma J, Yu MK, Fong S, et al. (2018) Using deep learning to model the hierarchical structure and function of a cell. Nat Methods 15, 290-298.

Machnicka MA, Milanowska K, Oglou OO, et al. (2013) Modomics: a database of RNA modification pathways–2013 update. Nucleic Acids Res 41, D262-D267.

Mahmud M, Kaiser MS, McGinnity TM, Hussain A (2021) Deep learning in mining biological data. Cognit Comput 2021, 1-33.

Mahto AK, Alam MA, Biswas R, et al. (2021) Short-Term Forecasting of Agriculture Commodities in Context of Indian Market for Sustainable Agriculture by Using the Artificial Neural Network. J Food Qual 2021, e9939906.

Malarvizhi MR, Thanamani AS (2012) K-nearest neighbor in missing data imputation. Int J Engin Res Develop 5, 5-7.

Malta TM, Sokolov A, Gentles AJ, et al. (2018) Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell 173(2), 338–354.e15.

Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418, 236-243.

Marbaniang CN, Vogel J (2016). Emerging roles of RNA modifications in bacteria. Curr Opin Microbiol 30, 50-57.

Marchetti CF, Ugena L, Humplík JF, et al. (2019) A Novel Image-Based Screening Method to StudyWater-Deficit Response and Recovery of Barley Populations Using Canopy Dynamics Phenotyping and Simple Metabolite Profiling. Front Plant Sci 10, e1252.

Margulies M, Egholm M, Altman WE, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380.

Martens D, Baesens B, Van Gestel T, Vanthienen J (2007) Comprehensible credit scoring models using rule extraction from support vector machines. Eur J Operation Res 183, 1466-1476.

Martin AR, Kanai M, Kamatani Y, et al. (2019) Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 51(4), 584-591

Marx V (2013) The big challenges of big data. Nature 498, 255-260.

Masoudzadeh SH, Mohammadabadi MR, Khezri A, et al. (2020). Dlk1 gene expression in different Tissues of lamb. Iran J Appl Anim Sci 10, 669-677 .

Mathlin J, Le Pera L, Colombo T (2020) A census and categorization method of epitranscriptomic marks. Int J Mol Sci 21(13), e4684.

Matias FI, Caraza-Harter MV, Endelman JB (2020) FIELDimageR: An R Package to Analyze Orthomosaic Images from Agricultural Field Trials. Plant Phenome J 3, e20005.

McCarthy J, Minsky M, Rochester N, Shannon CEA (2006) Proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Magazine 27, 12-14.

McInnes L, Healy J, Melville J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. ArXiv e1802.03426.

McIntyre LM, Lopiano KK, Morse AM, et al. (2011) RNA-seq: technical variability and sampling. BMC Genomics 12, e293.

Meyer KD, Saletore Y, Zumbo P, et al. (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3’ UTRs and near stop codons. Cell 149, 1635-1646.

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv 1301, e3781

Min S, Lee B, Yoon S (2016) Deep learning in bioinformatics. Briefings in Bioinformatics 18, 851-869.

Miotto R, Wang F, Wang S, et al. (2018) Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics 19, 1236-1246.

Mogili UR, Deepak BBVL (2018) Review on Application of Drone Systems in Precision Agriculture. Procedia Comput Sci 133, 502-509.

Mohamadipoor L, Mohammadabadi M, Amiri Z, et al. (2021) Signature selection analysis reveals candidate genes associated with production traits in Iranian sheep breeds. BMC Vet Res 17(1), 1-9.

Mohammadabadi M, Golkar A, Askari Hesni M (2023) The effect of fennel (Foeniculum vulgare) on insulin-like growth factor 1 gene expression in the rumen tissue of Kermani sheep. Agric Biotechnol J 15(4), 239-256.

Mohammadabadi M, Masoudzadeh SH, Khezri A, et al. (2021) Fennel (Foeniculum vulgare) seed powder increases Delta-Like Non-Canonical Notch Ligand 1 gene expression in testis, liver, and humeral muscle tissues of growing lambs. Heliyon 7(12), e08542 .

Mohd NFN, Ahamed HMNH, Abdullah R, et al. (2022) A Review of an Artificial Intelligence Framework for Identifying the Most Effective Palm Oil Prediction. Algorithms 15, e218.

Mollet I, Ben-Dov C,Felicio-Silva D, et al. (2010) Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome. Nucleic Acids Res 38(14), 4740-4754.

Montavon G, Samek W, Müller KR (2018) Methods for interpreting and understanding deep neural networks. Digit Signal Process 73, 1-15.

Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics 92(5), 255-264.

Muzio G, O’Bray L, Borgwardt K (2021) Biological network analysis with deep learning. Brief Bioinform 22, 1515-1530.

Myszczynska MA, Ojamies PN, Lacoste AMB, et al. (2020) Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol 16(8), 440-456.

Nagano T, Fraser P (2011) No-Nonsense Functions for Long Noncoding RNAs. Cell 145, 178-181.

Nagashima T, Yamaguhi K, Urakami K, et al. (2020) Japanese version of the cancer genome atlas, JCGA, established using fresh frozen tumors obtained from 5143 cancer patients. Cancer Sci 111(2), 687-699.

Nguyen SP, Li Z, Xu D, Shang Y (2017) New deep learning methods for protein loop modeling. IEEE/ACM Trans Comput Biol Bioinform 16, 596-606.

NIH (2020) The Cancer Genome Atlas - Cancer Genome – TCGA.

Noguera-Solano R, Ruiz-Gutiérrez R, Rodriguez-Caso JM (2013) Genome: twisting stories with DNA. Endeavour 37(4), 213–219.

Norman TM, Horlbeck MA, Replogle JM, et al. (2019) Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science 365, 786-793.

Nounou M, Nounou H, Meskin N, Datta A (2012) Wavelet-based multiscale filtering of genomic data. In IEEE/ACM Int Conf Adv Social Netw Anal Mining IEEE 2012, 804–809.

Nounou MN, Nounou HN, Mansouri M (2013) Model-based and model-free filtering of genomic data. Network Modeling Analysis in Health Informatics and Bioinformatics 2, 109-121.

Ostrovsky-Berman M, Frankel B, Polak P, Yaari G (2021) Immune2vec: Embedding b/t cell receptor sequences in RN, using natural language processing. Front Immunol 12, e680687.

Ozata DM, Gainetdinov I, Zoch A, et al. (2019) PIWI interacting RNAs: small RNAs with big functions. Nat Rev Genet 20, 89-108.

Paeng K, Hwang S, Park S, Kim MA (2017) Unified Framework for Tumor Proliferation Score Prediction in Breast Histopathology. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support; Springer: Cham, Switzerland, ISBN 9783319675572, Volume 10553 LNCS, pp. 231–239.

Pandey DK, Chaudhary B (2016) Domestication-Driven Gossypium Profilin 1 (GhPRF1) Gene Transduces Early Flowering Phenotype in Tobacco by Spatial Alteration of Apical/Floral-Meristem Related Gene Expression. BMC Plant Biol 16, e201310.

Pandey DK, Chaudhary B (2021) Transcriptional Loss of Domestication-Driven Cytoskeletal GhPRF1 Gene Causes Defective Floral and Fiber Development in Cotton (Gossypium). Plant Mol Biol 107, 519-532.

Paraforos DS, Vassiliadis V, Kortenbruck D, et al. (2016) A Farm Management Information System Using Future Internet Technologies. IFAC-PapersOnLine 49, 324–329.

Park Y, Kellis M (2015) Deep learning for regulatory genomics. Nature Biotechnology 33, 825-826.

Partel V, Costa L, Ampatzidis Y (2021) Smart Tree Crop Sprayer Utilizing Sensor Fusion and Artificial Intelligence. Comput Electron Agric 191, e106556.

Patti G, Yanes Ó, Siuzdak G (2012) Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 13, 263-269.

Paulsen J, Sekelja M, Oldenburg AR, et al. (2017) Chrom3d: three-dimensional genome modeling from hi-c and nuclear lamin-genome contacts. Genome Biol 18(1), e21.

Pazouki E (2021) A Practical Surface Irrigation Design Based on Fuzzy Logic and Meta-Heuristic Algorithms. Agric. Water Manag 256, e107069.

Pellegrini M, Baglioni M, Geraci F (2016) Protein complex prediction for large protein protein interaction networks with the core&peel method. BMC Bioinformatics 17, 37-58.

Perez E, Capper D (2020) Invited review: DNA methylation-based classification of paediatric brain tumours. Neuropathol Appl Neurobiol 46, 28-47.

Petrazzini BO, Naya H, Lopez-Bello F, et al. (2021) Evaluation of different approaches for missing data imputation on features associated to genomic data. BioData Mining 14, 1–13.

Pian C, Zhang G, Chen Z, et al. (2016) LncRNApred: Classification of Long Non-Coding RNAs and Protein-Coding Transcripts by the Ensemble Algorithm with a New Hybrid Feature. PLOS ONE 11(5), e0154567.

Pomyen Y, Wanichthanarak K, Poungsombat P, et al. (2020) Deep metabolome: Applications of deep learning in metabolomics. Comput Struct Biotechnol J 18, 2818-2825.

Poplin R, Chang PC, Alexander D, et al. (2018) A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36, 983-987.

Pour Hamidi S, Mohammadabadi MR, Asadi Foozi M, Nezamabadi-pour H (2017) Prediction of breeding values for the milk production trait in Iranian Holstein cows applying artificial neural networks. J Livestock Sci Technol 5 (2), 53-61.

Qi H, Zhang H, Zhao Y, et al. (2021) Mvp: predicting pathogenicity of missense variants by deep learning. Nat Commun 12(1), e510.

Qin Q, Feng J (2017) Imputation for transcription factor binding predictions based on deep learning. PLoS Comput Biol 13, e1005403.

Quang D, Xie X, Dan Q (2016) a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. Nucleic Acids Res 44, e107.

Ragoza M, Hochuli J, Idrobo E, et al. (2017) Protein-ligand scoring with convolutional neural networks. J Chem Inf Model 57, 942-957.

Raissi M, Perdikaris P, Karniadakis E (2019) Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput Phys 378, 686-707.

Rao RR, Makkithaya K (2017) Learning from a class imbalanced public health dataset: A cost-based comparison of classifier performance. Int J Elect Comput Engin 7, e2215.

Ravanbakhsh S, Liu P, Bjordahl TC, et al. (2015) Accurate, fully-automated NMR spectral profiling for metabolomics. PLOS ONE 10(5), e0124219.

Ravi D, Wong Ch, Deligianni F, et al. (2017) Deep learning for health informatics. IEEE J Biomed Health Inform 21(1), 4-21.

Razin A, Cedar H (1991) DNA methylation and gene expression. Microbiol Mol Biol Rev 55, 451-458.

Richardson S, Tseng GC, Sun W (2016) Statistical methods in integrative genomics. Annual Rev Stat Appl 3, 181-209.

Rifaioglu AS, Dogˇan T, Martin MJ, et al. (2019) Deepred: Automated protein function prediction with multi-task feed-forward deep neural networks. Sci Rep 9, e7344.

Rives A, Meier J, Sercu T, et al. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 118(15), e2016239118.

Rokach L, Maimon O (2014) Data mining with decision trees - theory and applications. In: Series in Machine Perception and Artificial Intelligence (2nd Ed.), Pages 328.

Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev 65, 386-408.

Rost B, Sander C (1993) Secondary structure prediction of all-helical proteins in two states. Protein Eng. 6, 831-836.

Saad OM, Chen Y (2020) Deep denoising autoencoder for seismic random noise attenuation. Geophysics 85, V367-V376.

Safaei SMH, Dadpasand M, Mohammadabadi M, et al. (2022) An Origanum majorana Leaf Diet Influences Myogenin Gene Expression, Performance, and Carcass Characteristics in Lambs. Animals 13(1), e14.

Salakhutdinov R, Hinton G (2009) Deep Boltzmann machines. Proc 20th Int Conf Artif Intell Stat, PMLR 5, 448-455.

Salamov A, Solovyev V (1997) Recognition of 3'-processing sites of human mRNA precursors. Comput Appl Biosci 13(1), 23-28.

Salekin S, Mostavi M, Chiu Y, et al. (2020) Predicting sites of epitranscriptome modifications using unsupervised representation learning based on generative adversarial networks. Front Phys 8, e196.

Saletore Y, Meyer K, Korlach J, et al. (2012) The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome Biol 13(10), e175.

Salovska B, Zhu H, Gandhi T, et al. (2020) Isoform-resolved correlation analysis between mRNA abundance regulation and protein level degradation. Mol Syst Biol 16(3), e9170.

Samek W, Müller KR (2019) Towards explainable artificial intelligence. In Explainable AI: interpreting, explaining and visualizing deep learning. Springer, 5-22.

Sánchez-Vega F, Mina M, Armenia J, et al. (2018) Oncogenic signaling pathways in the cancer genome atlas. Cell 173, 321-37.

Sanger F, Air GM, Barrell BG, et al. (1977) Nucleotide sequence of bacteriophage phi x174 DNA. Nature 265, 687-695.

Sanger F, Nicklen S, Coulson A (1997) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74(12), 5463-5467.

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science 270, 467-470.

Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Netw 61, 85-117.

Schneider EA, Hugo WA (2017) Support Vector Machine based method to distinguish long non-coding RNAs from protein coding transcripts. BMC Genomics 18, e804,

Scholten MT, De Boer I, Gremmen B, Lokhorst C (2013) Livestock farming with care: towards sustainable production of animal-source food. NJAS: Wagening J Life Sci 66(1), 3-5.

Schwessinger R, Gosden M, Downes D, et al. (2020) Deepc: predicting 3d genome folding using megabasescale transfer learning. Nat Methods 17, 1118-1124.

Seemann T (2013) Ten recommendations for creating usable bioinformatics command line software. Gigascience 2(1), e15.

Segal E, Fondufe-Mittendorf Y, Chen L, et al. (2006) A genomic code for nucleosome positioning. Nature 442, 772-778.

Senior AW, Evans R, Jumper J, et al. (2020) Improved protein structure prediction using potentials from deep learning. Nature 577, 706-710.

Serra F, Bau D, Goodstadt M et al. (2017) Automatic analysis and 3d-modelling of hi-c data using tadbit reveals structural features of the fly chromatin colors. PLOS Comput Biol 13(7), e1005665.

Seyhan K, Nguyen TN, Akleylek S, et al. (2021) Bi-GISIS KE: Modified Key Exchange Protocol with Reusable Keys for IoT Security. J Inf Secur Appl 58, e102788.

Shah G, Shah A, Shah M (2019) Panacea of Challenges in Real-World Application of Big Data Analytics in Healthcare Sector. J Data Inf Manag 1, 107-116.

Sharma R, Kumar N, Sharma BB (2022) Applications of Artificial Intelligence in Smart Agriculture: A Review. In Recent Innovations in Computing. Lecture Notes in Electrical

Engineering; Springer Science and Business Media Deutschland GmbH: Berlin/Heidelberg, Germany, Volume 832, pp. 135–142.

Shen R, Olshen AB, Ladanyi M (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25, 2906-2912.

Shen S, Park JW, Huang J, et al. (2012) MATS: A Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. Nucleic Acids Res 40(8), e61

Shen Z, Bao W, Huang DS (2018) Recurrent neural network for predicting transcription factor binding sites. Sci Rep 8(1), e15270.

Shokri S, Khezri A, Mohammadabadi M, Kheyrodin H (2023) The expression of MYH7 gene in femur, humeral muscle and back muscle tissues of fattening lambs of the Kermani breed. Agric Biotechnol J 15(2), 217-236.

Sidore C, Busonero F, Maschio A, et al. (2015) Genome sequencing elucidates sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat Genet 47(11), 1272-1281.

Silver D, Huang A, Maddison CJ, et al. (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529, 484-489.

Sønderby SK, Winther O (2015) Protein secondary structure prediction with long short term memory networks. Arxiv e1412.7828.

Song B, Li Z, Lin X, et al. (2021) Pretraining model for biological sequence data. Brief Funct Genomics 20(3), 181-195.

Sonsare PM, Gunavathi C (2019) Investigation of machine learning techniques on proteomics: A comprehensive survey. Prog Biophys Mol Biol 149, 54-69.

Spanaki K, Karafili E, Sivarajah U, et al. (2022) Artificial Intelligence and Food Security: Swarm Intelligence of AgriTech Drones for Smart AgriFood Operations. Prod Plan Control 33(16), 1498-1516.

Spencer M, Eickholt J, Cheng J (2015) A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics 12, 103-112.

Stark R, Grzelak M, Hadfield J (2019) RNA sequencing: the teenage years. Nat Rev Genet 20(11), 631-656.

Stathias V, Turner J, Koleti A, et al. (2020) Lincs data portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Res 48(D1), D431-D439.

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the ‘perceptron’ algorithm to distinguish translational initiation sites in *E. coli*. Nucleic Acids Research 10, 2997-3011.

Strombom D, King AJ (2018) Robot collection and transport of objects: a biomimetic process. Front Robot AI 5, e48.

Stueve TR, Marconett CN, Zhou B, et al. (2016) The importance of detailed epigenomic profiling of different cell types within organs. Epigenomics 8, 817-829.

Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J (2013) Application of machine learning to proteomics data: Classification and biomarker identification in postgenomics biology. OMICS 17, 595–610.

Sweeney L (2002) k-anonymity: A model for protecting privacy. Int J Uncertain Fuzz Knowledge-Based Syst 10, 557-570.

Sweeney L, Abu A, Winn J (2013) Identifying participants in the personal genome project by name (a re-identification experiment). ArXiv 1304, e7605.

Szappanos B, Kovács K, Szamecz B, et al. (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet 43, 656-662.

Tabaska JE, Zhang M (1999) Detection of polyadenylation signals in human DNA sequences. Gene 231(1–2), 77-86.

Talavera D, Kershaw CJ, Costello JL, et al. (2018) Archetypal transcriptional blocks underpin yeast gene regulation in response to changes in growth conditions. Sci Rep 8(1), e7949.

Talaviya T, Shah D, Patel N, et al. (2020) Implementation of Artificial Intelligence in Agriculture for Optimisation of Irrigation and Application of Pesticides and Herbicides. Artif Intell Agric 4, 58-73.

TCGA 2020. The Cancer Genome Atlas Program. Cancer Genome. National Cancer Institute, USA.

Thornton PK (210) Livestock production: recent trends, future prospects. Philos Trans R Soc B: Biol Sci 365(1554), 2853-2867.

van Agthoven MA, Lam YPY, O'Connor P, et al. (2019) Two dimensional mass spectrometry: new perspectives for tandem mass spectrometry. European Biophys J 48, 213-229.

van Buuren S (2018) Flexible imputation of missing data. CRC Press.

van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Machine Learning Res 9, 2579-2605.

van Hulse J, Khoshgoftaar TM, Napolitano A, Wald R (2012) Threshold-based feature selection techniques for high-dimensional bioinformatics data. Netw Model Anal Health Inform Bioinform 1, 47-61.

van IJzendoorn DG, Szuhai K, Brujin IHB, et al. (2019) Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. PLoS Comput Biol 15(2), e1006826.

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. ArXivabs 1706, e03762.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial Analysis of Gene Expression. Science 270, 484-487.

Velculescu VE, Zhang L, Zhou W, et al. (1997) Characterization of the Yeast Transcriptome. Cell 88, 243-251.

Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: Proc 25th Int Conf Machine Learn 1, 1096-1103.

Vincent P, Larochelle H, Lajoie I, et al. (2010) Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 11, 3371-3408.

Vlahou A, Schorge JO, Gregory BW, Coleman RL (2003) Diagnosis of ovarian cancer using decision tree classification of mass spectral data. J Biomed Biotechnol 2003, 308-314.

Voillet V, Besse P, Liaubet L, et al. (2016) Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. BMC Bioinformatics 17, 1-16.

Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E (2018) Deep Learning for Computer Vision: A Brief Review. Comput Intell Neurosci 2018, e7068349.

Wainberg M, Merico D, Delong A, Frey BJ (2018) Deep learning in biomedicine. Nat Biotechnol 36, 829-838.

Wan F, Li S, Tian T, et al. (2020) Exp2sl: A machine learning framework for cell-line-specific synthetic lethality prediction. Frontiers Pharmacol 11, e112.

Wang B, Mezlini AM, Demir F, et al. (2014) Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11(3), 333-337.

Wang E, Sandberg R, Khrebtukova I, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-476.

Wang J, Agarwal D, Huang M, et al. (2019) Data denoising with transfer learning in single-cell transcriptomics. Nat Methods 16, 875-878.

Wang J, Xie X, Shi J, et al. (2020) Denoising autoencoder, a deep learning algorithm, aids the identification of a novel molecular signature of lung adenocarcinoma. Genom Proteom Bioinform 18, 468-480.

Wang L, Park HJ, Dasari S, et al. (2013) CPAT: Coding-potential assessment tool using an alignmentfree logistic regression model. Nucleic Acids Res 41(6), e74.

Wang P, Liu X, Berzin TM, et al. (2020) Effect of a Deep-Learning Computer-Aided Detection System on Adenoma Detection during Colonoscopy (CADe-DB Trial): A Double-Blind Randomised Study. Lancet Gastroenterol Hepatol 5, 343-351.

Wang S, Peng J, Ma J, Xu J (2016) Protein secondary structure prediction using deep convolutional neural fields. Sci Rep 6, e18962.

Wang Y, Liu T, Xu D, et al. (2016) Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. Sci Rep 22(6), e19598.

Wang Y, Mao H, Yi Z (2017) Protein secondary structure prediction by using deep learning method. Knowledge-Based Syst 118, 115-123.

Wang Y, Zhang X-S, Chen L (2018) Integrating data- and model-driven strategies in systems biology. BMC Systems Biol 12, e38.

Wang Z, Gerstein M, Snyder M (2010) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57-63.

Way GP, Allaway RJ, Bouley SJ, et al. (2017) A machine learning classifier trained on cancer transcriptomes detects nf1 inactivation signal in glioblastoma. BMC Genomics 18(1), e127.

Weiss R (2010) Bayesian methods for data analysis. Am J Ophthalmol 149(2), 187-188.

Werner S, Schmidt L, Marchand V, et al. (2020) Machine learning of reverse transcription signatures of variegated polymerases allows mapping and discrimination of methylated purines in limited transcriptomes. Nucleic Acids Res 48, 3734-3746.

Wilkins M, Sanchez JC, Gooley AA, et al. (1996) Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol Genet Eng Rev 13, 19-50.

Williams RJ, Zipser D (1989) A learning algorithm for continually running fully recurrent neural networks. Neural Comput 1, 270-280.

Wolff J, Pauling J, Keck A, Baumbach J (2020) The economic impact of artificial intelligence in health care: Systematic review. J Med Internet Res 22, e16866.

Woloszynek S, Zhao Z, Chen J, Rosen G (2019) 16s rRNA sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses. PLoS Comput Biol 15(2): e1006721.

Wood A, Altman M, Bembenek A, et al. (2018) Differential Privacy: A Primer for a Non-Technical Audience. Vand J Ent Technol Law 21(1), 209-275.

Wu C, Alwine J (2004) Secondary structure as a functional feature in the downstream region of mammalian polyadenylation signals. Molecular and Cellular Biology 24, 2789-2796.

Wu C, Zhou F, Ren J, et al. (2019) A selective review of multi-level omics data integration using variable selection. High-Throughput 8(1), e4.

Wu ST, Roberts K, Datta S, et al. (2020) Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc 27(3), 457-470.

Wu Y, Schuster M, Chen Z, et al. (2016) Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv 1609, e08144.

Xia Z, Li Y, Zhang B, et al. (2019) Deerect-polya: a robust and generic deep learning method for pas identification. Bioinformatics 35(14), 2371-2379.

Xu J, Yang P, Xue S, et al. (2019) Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. Hum Genet 138(2), 109-124.

Yang J, Lodgher A (2019) Fundamental defensive programming practices with secure coding modules. In: 2019 International Conference on Security and Management. Las Vegas, Nevada.

Yang Y, Sun H, Zhang Y, et al. (2021) Dimensionality reduction by umap reinforces sample heterogeneity analysis in bulk transcriptomic data. Cell Rep 36(4), e109442.

Yates JA (2011) A century of mass spectrometry: from atoms to proteomes. Nat Methods 8, 633–637 (2011).

Yaxley KJ, Joiner KF, Abbass H (2021) Drone approach parameters leading to lower stress sheep flocking and movement: sky shepherding. Sci Rep 11, e7803.

Yazdani A, Lu L, Raissi M, Karniadakis GE (2020) Systems biology informed deep learning for inferring parameters and hidden dynamics. PLOS Comput Biol 16, e1007575.

Yousefi S, Amrollahi F, Amgad M, et al. (2017) Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci Rep 7 (1), e11707.

Yu MK, Kramer M, Dutkowski J, et al. (2016) Translation of genotype to phenotype by a hierarchy of cell subsystems. Cell Systems 2, 77-88.

Yuan Y, Liang Y, Yi L, et al. (2008) Uncorrelated linear discriminant analysis (ULDA): A powerful tool for exploration of metabolomics data. Chemom Intell Lab Syst 93(1), 70-79.

Yuan Y. Li C, Kim J, et al. (2016) Deepgene: an advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinformatics 17, e476.

Yue T, Wang H (2018) Deep learning for genomics: A concise overview. arXiv 1802, e00810.

Zamboni N, Saghatelian A, Patti G (2015) Defining the metabolome: size, flux, and regulation. Mol Cell 58(4), 699-706.

Zampieri G, Vijayakumar S, Yaneske E, Angione C (2019) Machine and deep learning meet genome-scale metabolic modeling. PLoS Comput Biol 15, e1007084.

Zeng L, Das S, Burne RA (2010) Utilization of lactose and galactose by streptococcus mutans: Transport, toxicity, and carbon catabolite repression. J Bacteriol 192, 2434-2444.

Zhang F, Song H, Zeng M, et al. (2019) Deepfunc: A deep learning framework for accurate prediction of protein functions from protein sequences and interactions. Proteomics 19(12), e1900019.

Zhang G, Zeng T, Dai Z, Dai X (2021) Prediction of CRISPR/Cas9 single guide RNA cleavage efficiency and specificity by attention-based convolutional neural networks. Comput Struct Biotechnol J 19, 1445-1457.

Zhang L, Qin X, Liu M, et al. (2021) DNN-m6A: A cross-species method for identifying RNA N6-Methyladenosine sites based on deep neural network with multi-information fusion. Genes 12(3), e354.

Zhang S, Li Q, Liu J, Zhou XJ (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics 27, i401-i409.

Zhang Y, An L, Yue F, Hardison RC (2016) Jointly characterizing epigenetic dynamics across multiple human cell types. Nucleic Acids Research 44, 6721-6731.

Zhang Z, Pan Z, Ying Y, et al. (2019) Deep-learning augmented RNA-seq analysis of transcript splicing. Nat Methods 16(4), 307-310.

Zhang Z, Wu S, Stenoien D, Pasa-Tolic L (2014) High-throughput proteomics. Annu Rev Anal Chem 7, 427-454.

Zhang Z, Zhao Y, Liao X, et al. (2019) Deep learning in omics: a survey and guideline. Brief Funct Genomics 18, 41-57.

Zhao B, Rubinstein BI, Gemmell J, Han J (2012) A bayesian approach to discovering truth from conflicting sources for data integration. ArXiv 1203, e0058

Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods 12, 931–934.

Zhou Y, Xia Q, Zhang Z, et al. (2022) Artificial Intelligence and Machine Learning for the Green Development of Agriculture in the Emerging Manufacturing Industry in the IoT Platform. Acta Agric Scand Sect B Soil Plant Sci 72, 284-299.

Zhou Y, Zeng P, Li Y-H, et al. (2016) Sramp: prediction of mammalian n6-methyladenosine (m6a) sites based on sequence-derived features. Nucleic Acids Res 44, e91.

Zhou ZH, Jiang Y, Chen SF (2003) Extracting symbolic rules from trained neural network ensembles. AI Communications 16, 3-15.

Zitnik M, Agrawal M, Leskovec J (2018) Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34(13), i457-466.

Zou J, Huss M, Abid A, et al. (2019) A primer on deep learning in genomics. Nature Genet 51, 12-18.