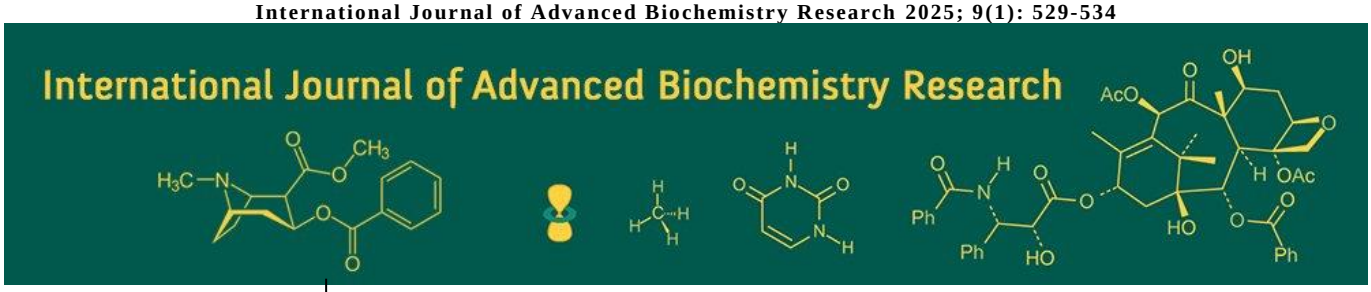

ISSN Print: 2617-4693
ISSN Online: 2617-4707
IJABR 2025; 9(1): 529-534
www.biochemjournal.com
Received: 22-10-2024
Accepted: 25-11-2024

Upti Aayer
Natubhai V. Patel College of Pure and Applied Sciences, The CVM University, Vallabh Vidyanagar, Gujarat, India

Dr. Pratibha Parihar
Natubhai V. Patel College of Pure and Applied Sciences, The CVM University, Vallabh Vidyanagar, Gujarat, India

Corresponding Author:
Dr. Pratibha Parihar
Natubhai V. Patel College of Pure and Applied Sciences, The CVM University, Vallabh Vidyanagar, Gujarat, India

Bioinformatics: The role of machine learning in revolutionizing biological data analysis

Upti Aayer and Pratibha Parihar

DOI: https://doi.org/10.33545/26174693.2025.v9.i1g.3551

Abstract
The advent of machine learning has revolutionized the field of bioinformatics, enabling unprecedented analysis and interpretation of complex biological datasets. This paper explores the integration of machine learning techniques in various bioinformatics applications. We discuss the potential of machine learning algorithms, such as deep learning, random forests, and support vector machines, to uncover insights from vast and diverse biological data. Key challenges, including data heterogeneity, interpretability of models, and ethical considerations, are highlighted. Additionally, future directions are outlined to emphasize the transformative impact of machine learning in advancing personalized medicine and understanding biological systems. This work underscores the necessity for interdisciplinary collaboration to harness the full potential of machine learning in bioinformatics.

Keywords: Machine learning, bioinformatics, deep learning, genomics, biological data analysis

Introduction
The increasing high-throughput nature of biological research has led to the production of extremely large and intricate datasets which demand new methodologies to produce useful inferences. Machine learning (ML), as a branch of artificial intelligence, has been recently recognized as an important technology in bioinformatics for high dimensional biological data analysis and interpretation. Machine Learning (ML) utilizes mathematical and computational approach to automate pattern recognition, modeling and predictive analytics based on statistical algorithms, thereby changing the way we understand biological systems and processes.

Genomic data analysis is one of the main application of ML in bioinformaics. Genomic datasets are big and complex generated by Next Generation Sequencers. Many ML algorithms (support vector machines [SVMs], random forests, and deep learning models) have been used to find disease-associated genetic variants, annotate genome functional regions, and predict gene expression levels.

Proteomics and protein structure prediction represent another critical domain where ML has made significant progress. Predicting protein folding has long been a complex challenge in bioinformatics. However, in the past few years progress has been made where deep learning has reached close to experimental accuracy in predicting structures such as DeepMind's AlphaFold, (Jumper *et al*., 2021) [14]. This identification has led to newer understanding of disease mechanisms and drug discovery.

In transcriptomics, clustering algorithms such as k-means and hierarchical clustering are widely used to analyze gene expression data. These methods assist in identifying groups of co-expressed genes, uncovering potential regulatory relationships, and classifying samples based on their expression patterns (Eisen *et al*., 1998) [8]. Supervised learning techniques, including decision trees and ensemble methods, are also extensively applied in cancer classification using gene expression profiles, contributing to advancements in personalized medicine (Golub *et al*., 1999; Libbrecht & Noble, 2015) [11, 18].

Machine learning (ML) is revolutionizing drug discovery and development by addressing the limitations of traditional methods that are often time-consuming and resource-intensive. Techniques such as generative adversarial networks (GANs) and reinforcement learning are being used to design novel molecules with specific desired properties, significantly accelerating the discovery of potential therapeutics (Zhavoronkov *et al*., 2019) [40].

Additionally, ML models are helpful in predicting drug-target interactions, optimizing drug formulations, and simulating drug responses in silico, enhancing the efficiency and precision of the drug development process (Ekins *et al*., 2019; Vamathevan *et al*., 2019) [9, 36].

The application of machine learning (ML) in bioinformatics faces several challenges. Biological datasets often have limitations such as small sample sizes, high heterogeneity, and experimental biases, which can negatively impact model performance and general application (Angermueller *et al*., 2016) [1]. Furthermore, the "black-box" nature of many ML models poses challenges related to interpretability and trustworthiness, particularly in clinical contexts where transparency is crucial (Topol, 2019) [34]. Addressing these issues through the development of interpretable ML models and strategies to mitigate data biases is essential for achieving reliable and meaningful results in bioinformatics applications (Tonekaboni *et al*., 2019) [33].

To summarise, machine learning is transforming bioinformatics on a comprehensive level as it provides new and advanced tools to address complex biological questions. Through genomics, proteomics, transcriptomics and drug discovery to harbours manifold implications for understanding life at the molecular level. ML has great potential to expand our understanding of biology and enhance healthcare if some existing challenges are addressed by an interdisciplinary research and methodological innovations.

Applications of Bioinformatics which uses ML

Disease Diagnosis and Prognosis

Machine learning (ML) is transforming the landscape of disease diagnosis and prognosis by providing robust tools to analyze and interpret complex, high-dimensional datasets such as medical imaging, electronic health records (EHRs), and genomic profiles. ML has shown its utility in early detection, outcome prediction, and the personalised treatments.

ML algorithms are being used in field of genomics to detect diseases with genetic or molecular approaches, such as identifying BRCA mutations for breast cancer predisposition (Libbrecht & Noble, 2015) [18]. Supervised ML models analyzing EHRs can integrate clinical, laboratory, and demographic data to diagnose diseases such as heart failure or sepsis at an early stage by identifying patterns missed by traditional methods (Shickel *et al*., 2018) [26]. Unsupervised ML algorithms like clustering have been employed to classify cancer subtypes based on gene expression profiles which enables personalized therapy selection (Libbrecht & Noble, 2015) [18]. Deep learning models, particularly convolutional neural networks (CNNs), are widely applied in radiology and pathology to detect diseases from medical images. For instance, CNNs have been used for detecting cancer in mammograms, classifying skin lesions, and identifying diabetic retinopathy with performance comparable to or exceeding human experts (Esteva *et al*., 2017; Rajpurkar *et al*., 2019) [10, 24].

ML-powered survival analysis tools, like DeepSurv, integrates neural networks with traditional Cox models to predict patient survival probabilities and recurrence risks. These tools assist in tailoring treatment plans to individual needs (Katzman *et al*., 2018) [16]. Predictive ML models can determine cancer progression stages or forecast metastasis based on histological and genomic data, aiding clinicians in making informed decisions. ML applications in Alzheimer's disease prognosis utilize MRI and PET imaging combined with cognitive test data to predict disease onset and progression (Jo *et al*., 2019) [13]. Predictive based models being used in analysing tumor mutational burden and gene expression data to forecast patient responses to immunotherapies, such as checkpoint inhibitors (Topol, 2019) [34].

Use of ML models in making of medical devices helped in patient real time monitoring, identifying diseases at an early stage, thus it proven to be helpful in critical care. For example, used of ML algorithms in devices used to detect rrhythmias or cardiac abnormalities from ECG data, reducing the risk of undetected cardiac events (Hannun *et al*., 2019) [12]. In intensive care units, ML systems predict sepsis risk by analyzing vitals and lab data, allowing for earlier interventions (Shickel *et al*., 2018) [26]. Wearables with ML models can predict seizure episodes in epileptic patients, offering better disease control and quality of life (Rajpurkar *et al*., 2019) [24].

Biomarker Discovery

The field of biomarker discovery has achieved newer heights by levaraging Machine learning (ML) based approaches. Use of ML approaches enabled the identification of biological molecules that indicate normal or diseased states. These biomarkers can be derived from genomics, proteomics, metabolomics, imaging, or other high-dimensional datasets. ML excels in analyzing high-dimensional datasets that are characteristic of biological data. Algorithms like LASSO (Least Absolute Shrinkage and Selection Operator) regression, random forests, and decision trees are commonly employed to identify relevant features (potential biomarkers) while excluding noise and redundant variables (Tibshirani, 1996) [32]. Feature selection methods are crucial in biomarker discovery as they reduce the dataset to its most informative elements. For example, LASSO regression applies regularization to select only those features strongly correlated with the target variable, such as a disease state (Simon *et al*., 2011) [27]. In diseases like Alzheimer's, ML has been pivotal in analyzing cerebrospinal fluid (CSF). For instance, ML algorithms have identified amyloid-beta and tau proteins as biomarkers for early diagnosis and disease progression (Sperling *et al*., 2011) [28, 29]. Additionally, ML-powered imaging techniques analyze structural changes in the brain, such as hippocampal atrophy, to predict the onset of Alzheimer's years before clinical symptoms manifest (Rathore *et al*., 2017) [25].

Biomarker discovery in oncology has greatly benefited from ML. Algorithms analyze mutational signatures, gene expression profiles, and liquid biopsy data to identify biomarkers like circulating tumor DNA (ctDNA) and specific miRNAs for early cancer detection and treatment monitoring (Chan *et al*., 2013) [6]. ML models have been applied to genomic and proteomic data to uncover biomarkers predicting conditions like myocardial infarction or heart failure. For example, studies have highlighted troponins and natriuretic peptides as predictive markers, identified and validated through ML approaches (Than *et al*., 2019) [31].

Early detection of diseases through biomarkers can enable timely interventions, potentially halting disease progression or improving patient outcomes. For example, in Alzheimer's disease, identifying biomarkers like amyloid-beta through

ML can lead to earlier therapeutic intervention during the pre-symptomatic stage (Sperling *et al*., 2011) [28, 29]. Biomarkers identified through ML facilitate tailored treatments by stratifying patients based on their specific molecular profiles. In cancer, for instance, ML-driven biomarkers help predict patient responses to targeted therapies, ensuring that treatments are optimized for individual needs (Libbrecht & Noble, 2015) [18]. ML-identified biomarkers offer a non-invasive and cost-effective alternative for disease screening compared to traditional diagnostic methods. For example, liquid biopsy biomarkers like ctDNA and proteins detected through ML are less invasive than tissue biopsies.

Drug Discovery and Development

The integration of machine learning (ML) in drug discovery and development has brought unprecedented advancements to a traditionally time-intensive and costly process. ML enhances the efficiency of identifying drug candidates by navigating through massive chemical libraries and biological datasets to pinpoint molecules with therapeutic potential. Generative adversarial networks (GANs) and reinforcement learning models have emerged as powerful tools for creating novel chemical compounds with desired pharmacological properties. GANs generate diverse chemical structures by mimicking the process of natural evolution, while reinforcement learning optimizes these structures to improve efficacy, safety, and bioavailability (Zhavoronkov *et al*., 2019) [40]. Traditional high-throughput screening of compounds is expensive and resource-intensive. ML models, such as deep learning frameworks, predict bioactivity and drug-likeness, reducing the need for exhaustive physical screening (Jumper *et al*., 2021) [14].

ML plays crucial nad important role in assessing a drug's efficacy and safety profile as it’s detection is crucial in early developmental stages. ML models predict off-target effects and toxicological risks by analyzing large-scale toxicology datasets. Algorithms like deep neural networks (DNNs) and support vector machines (SVMs) are employed to detect patterns indicative of hepatotoxicity, cardiotoxicity, or mutagenicity, reducing the likelihood of late-stage failures (Mayr *et al*., 2016) [21]. Predicting interactions between drugs and biological targets is another area where ML excels. Techniques such as graph neural networks and matrix factorization models analyze complex interactions and predict binding affinities, streamlining the hit-to-lead process (Zitnik *et al*., 2018) [44].

ML has revolutionized drug repurposing by enabling the identification of new therapeutic uses for existing drugs. During the COVID-19 pandemic, ML models rapidly screened libraries of approved drugs to identify compounds with potential antiviral activity against SARS-CoV-2. For example, ML tools identified remdesivir and baricitinib as promising candidates for clinical testing, significantly shortening the timeline from hypothesis generation to treatment evaluation (Beck *et al*., 2020) [3].

ML is increasingly used to design and manage clinical trials, improving patient selection, dosage optimization, and trial monitoring. ML facilitates adaptive trial designs where data collected during the trial is used to modify the study dynamically, improving success rates and reducing timelines. ML algorithms analyze patient genetic profiles and biomarkers to stratify populations, ensuring trials enroll those most likely to respond to treatment. This improves trial efficiency and reduces costs (Topol, 2019) [34].

Single-Cell Analysis

Single-cell RNA sequencing (scRNA-seq) is a transformative technology that provides unparalleled resolution for studying cellular heterogeneity. It captures the transcriptomic profile of individual cells, enabling researchers to investigate cellular diversity, identify rare populations, and uncover dynamic biological processes. However, the complexity and high dimensionality of scRNA-seq data necessitate advanced computational tools, and machine learning (ML) has emerged as a key enabler in extracting meaningful insights.

A major application of ML in scRNA-seq is clustering single cells into distinct populations based on their transcriptomic profiles. The high-dimensional nature of scRNA-seq data requires dimensionality reduction methods before clustering. Algorithms like t-SNE (t-distributed Stochastic Neighbor Embedding) and UMAP (Uniform Manifold Approximation and Projection) are widely used to project high-dimensional data into two or three dimensions for visualization and analysis. t-SNE is effective in preserving local data structure, making it ideal for identifying subpopulations of cells (Van Der Maaten & Hinton, 2008) [37]. UMAP goes a step further by preserving global structure and is computationally more efficient for large datasets (McInnes *et al*., 2018) [22]. ML algorithms like k-means, hierarchical clustering, and graph-based methods are employed to group cells into biologically meaningful clusters. Advanced methods like Louvain and Leiden algorithms leverage graph theory to improve the resolution of clusters, enabling the identification of rare cell types that traditional methods might overlook (Blondel *et al*., 2008) [4].

Rare cell populations often play crucial roles in tissue function and disease but are challenging to identify due to their low abundance. Although use of ML can enhances this detection. Autoencoders and variational autoencoders (VAEs) compress scRNA-seq data into lower-dimensional representations, helping uncover subtle patterns associated with rare cell types. For example, scVI (single-cell Variational Inference) integrates scRNA-seq data from different experimental batches to identify rare populations while addressing batch effects (Lopez *et al*., 2018) [19].

Understanding how cells transition through developmental or differentiation pathways is another critical aspect of scRNA-seq analysis. ML techniques infer "pseudotime," an abstract timeline of cellular differentiation based on gene expression patterns. Tools like Monocle and Slingshot utilize dimensionality reduction and graph-based methods to order cells along developmental trajectories (Trapnell *et al*., 2014; Street *et al*., 2018) [35, 30]. Deep learning-based approaches, such as recurrent neural networks (RNNs), model temporal gene expression changes to reconstruct complex lineage hierarchies. These methods provide a dynamic view of cellular transitions, offering insights into tissue development, regeneration, and disease progression.

The integration of ML with scRNA-seq has yielded significant biological insights in developmental biology. ML-powered scRNA-seq analysis has been instrumental in mapping cell fate decisions. For example, studies of early embryonic development have revealed lineage-specific transcriptional programs regulated at single-cell resolution (Pijuan-Sala *et al*., 2019) [23]. n oncology, ML methods have

identified tumor-specific cell populations, such as cancer stem cells, and delineated tumor microenvironment interactions. Similarly, scRNA-seq studies in neurodegenerative diseases have highlighted dysfunctional neuronal and glial subpopulations, advancing our understanding of disease pathogenesis (Mathys *et al*., 2019) [20].

Metagenomics and Machine Learning

Metagenomics focuses on studying the genetic material recovered directly from environmental samples, enabling the exploration of microbial communities in various habitats without requiring individual culturing. ML has become indispensable in analyzing metagenomic data, addressing challenges like taxonomic classification, functional annotation, and microbiome-host interaction studies. Algorithms such as Random Forest, Support Vector Machines (SVMs), and deep learning models classify DNA or protein sequences into taxonomic groups. Tools like Kraken2 use k-mer-based ML approaches to rapidly assign reads to specific taxa with high precision (Wood *et al*., 2019) [38]. ML methods identify marker genes, such as 16S rRNA, to infer microbial diversity and relative abundances in samples. Models like Naïve Bayes classifiers are implemented in tools like QIIME2 for taxonomic assignment (Bolyen *et al*., 2019) [5].

Deep learning models such as Convolutional Neural Networks (CNNs) predict GO terms for metagenomic reads, enabling annotation of microbial metabolic pathways (Zhou *et al*., 2020) [42]. ML models like XGBoost identify metabolic pathways from metagenomic data, linking microbial communities to ecological or health-related functions (Douglas *et al*., 2020) [7].

ML helps uncover the complex interplay between the microbiome and its host, particularly in health and disease contexts. ML models analyze microbiome profiles to predict diseases such as inflammatory bowel disease (IBD), diabetes, and cancers. For example, Random Forest classifiers have been applied to identify microbial biomarkers for colorectal cancer (Yu *et al*., 2021) [39]. redictive models integrate metagenomic data with patient health records to tailor probiotics or dietary interventions. Tools like MicrobiomeAI leverage neural networks for personalized microbiome management.

ML plays a pivotal role in predicting antibiotic resistance genes (ARGs) from metagenomic data, addressing the global challenge of antimicrobial resistance. For example, DeepARG ool uses deep learning to predict ARGs from metagenomic sequences, enabling rapid identification of resistant strains in clinical and environmental samples (Arango-Argoty *et al*., 2018) [2]. Graph-based ML models map the co-occurrence of ARGs with microbial taxa, helping trace the spread of resistance across ecosystems. ML models predict the impact of pollutants on microbial diversity in soil or water systems. Techniques like Gradient Boosting assess changes in microbial functions linked to environmental stressors (Karimi *et al*., 2021) [15].

Challenges in Machine Learning Applications in Bioinformatics

While machine learning (ML) has revolutionized bioinformatics, enabling breakthroughs in areas like genomics, transcriptomics, proteomics, and drug discovery, several challenges still remain unsolved such as data quality, algorithmic complexity, interpretability, scalability, and ethical considerations.

1. Data Quality and Availability

Biological experiments often involve small sample sizes due to costs and logistical constraints, making it difficult to train robust ML models. This is especially problematic in genomics and rare disease studies. High-throughput technologies generate large and complex data sets hence making it noisy as well with missing values, which can lead to biased predictions or reduced accuracy. Variability in experimental protocols, instruments, and sample types introduces batch effects, complicating cross-study data integration. For example RNA-seq data from different platforms may show significant variability that ML models must address.

2. Algorithmic Challenges

Omics datasets often have more features (e.g., genes, proteins) than samples, leading to the "curse of dimensionality," where ML algorithms struggle to generalize and overfit. Techniques like feature selection or dimensionality reduction (e.g., PCA, UMAP) are used but may lose biologically relevant information. Rare diseases or underrepresented biological regions may create imbalanced datasets, leading to biased model predictions favoring majority classes. Oversampling or specialized loss functions can be applied but remain challenging to implement effectively.

3. Interpretability and Trustworthiness

Many ML methods, particularly deep learning, are considered "black boxes," which makes it difficult for researchers to understand how predictions are made. This lack of transparency limits their adoption in clinical applications. Complex ML pipelines often lack standardization, leading to difficulties in reproducing results across studies. For example different preprocessing steps or hyperparameter settings can yield significantly different outcomes.

4. Integration, Scalability and Computational Demands

High-throughput technologies generate massive datasets that require significant computational resources for storage, processing, and analysis. For example Single-cell RNA-seq experiments often produce millions of data requiring scalable clustering algorithms. Many bioinformatics applications demand real-time analysis, such as metagenomics or pathogen surveillance. Scaling ML algorithms to meet these demands remains a challenge. Integrating diverse data types (e.g., genomics, proteomics, and transcriptomics) is essential to uncover biological insights but is computationally and algorithmically challenging. Different bioinformatics datasets often have varying formats, levels of resolution, and noise profiles, which complecates its integration.

5. Ethical and Privacy Concerns

Biological data often involve sensitive patient information, raising ethical concerns about data sharing and use. Regulations like GDPR and HIPAA impose constraints on data handling. Data from specific populations or experiments may introduce biases which limits the generalizability of ML models. For example training

datasets biased toward Indian populations may not be applicable in global health studies.

6. Validation and Clinical Translation

ML models often show high accuracy in research settings but fail during real-world validation due to data variability and lack of external validation datasets. Translating ML-based bioinformatics tools into clinical practice requires meeting stringent regulatory standards for safety, efficacy, and transparency.

Conclusion

Machine learning (ML) has profoundly transformed bioinformatics by providing sophisticated tools to analyze and interpret complex biological data. Its applications span diverse domains, including genomics, proteomics, transcriptomics, drug discovery, single-cell analysis, and metagenomics. These advancements enable novel insights into disease mechanisms, personalized medicine, and therapeutic interventions. To harness the full potential of ML in bioinformatics, interdisciplinary collaboration and methodological innovations are essential. Addressing the challenges of data quality, algorithm interpretability, computational scalability, and ethical considerations will enable more robust, transparent, and impactful applications. By overcoming these hurdles, ML has the potential to revolutionize our understanding of biology, advance healthcare, and deliver personalized treatments on a global scale.

References

1. Angermueller C, *et al*. Deep learning for computational biology. Molecular Systems Biology. 2016;12(7):878. DOI:10.15252/msb.20156651.
2. Arango-Argoty G, *et al*. DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome. 2018;6(1):23. DOI:10.1186/s40168-018-0401-z.
3. Beck BR, *et al*. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Computational and Structural Biotechnology Journal. 2020;18:784-790. DOI:10.1016/j.csbj.2020.03.025.
4. Blondel VD, *et al*. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008;2008(10):P10008. DOI:10.1088/1742-5468/2008/10/P10008.
5. Bolyen E, *et al*. Reproducible, interactive, scalable, and extensible microbiome data science using QIIME 2. Nature Biotechnology. 2019;37(8):852-857. DOI:10.1038/s41587-019-0209-9.
6. Chan KC, *et al*. Plasma epidermal growth factor receptor mutations and their correlation with lung cancer. Nature Medicine. 2013;19(11):1348-1355. DOI:10.1038/nm.3380.
7. Douglas GM, *et al*. PICRUSt2: An improved and extensible approach for metagenome inference. Nature Biotechnology. 2020;38(6):685-688. DOI:10.1038/s41587-020-0548-6.
8. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences. 1998;95(25):14863-14868. DOI:10.1073/pnas.95.25.14863.
9. Ekins S, Puhl AC, *et al*. Exploiting machine learning for end-to-end drug discovery and development. Nature Reviews Drug Discovery. 2019;18(6):463-477. DOI:10.1038/s41573-019-0024-5.
10. Esteva A, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-118. DOI:10.1038/nature21056.
11. Golub TR, *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531-537. DOI:10.1126/science.286.5439.531.
12. Hannun AY, *et al*. Cardiologist-level arrhythmia detection with convolutional neural networks. Nature Medicine. 2019;25(1):65-69. DOI:10.1038/s41591-018-0268-3.
13. Jo T, *et al*. Deep learning-based prediction of Alzheimer's disease using neuroimaging data. Frontiers in Aging Neuroscience. 2019;11:220. DOI:10.3389/fnagi.2019.00220.
14. Jumper J, *et al*. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-589. DOI:10.1038/s41586-021-03819-2.
15. Karimi B, *et al*. Deterministic processes dominate soil microbial community assembly in agroecosystems. Microbial Ecology. 2021;81(4):974-986. DOI:10.1007/s00248-020-01566-6.
16. Katzman JL, *et al*. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018;18(1):24. DOI:10.1186/s12874-018-0482-1.
17. Lähnemann D, *et al*. Eleven grand challenges in single-cell data science. Genome Biology. 2020;21(1):31. DOI:10.1186/s13059-020-1926-6.
18. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics. 2015;16(6):321-332. DOI:10.1038/nrg3920.
19. Lopez R, *et al*. Deep generative modeling for single-cell transcriptomics. Nature Methods. 2018;15(12):1053-1058. DOI:10.1038/s41592-018-0229-2.
20. Mathys H, *et al*. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;570(7761):332-337. DOI:10.1038/s41586-019-1195-2.
21. Mayr A, *et al*. DeepTox: Toxicity prediction using deep learning. Frontiers in Environmental Science. 2016;3:80. DOI:10.3389/fenvs.2015.00080.
22. McInnes L, *et al*. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint. 2018:arXiv:1802.03426.
23. Pijuan-Sala B, *et al*. A single-cell molecular map of mouse gastrulation and early organogenesis. Nature. 2019;566(7745):490-495. DOI:10.1038/s41586-019-0933-9.
24. Rajpurkar P, *et al*. Appointments and Algorithmic Solutions in Healthcare. Nature Reviews Bioengineering. 2019;3:1-10. DOI:10.1038/s41592-019-0575-9.
25. Rathore S, *et al*. A review on neuroimaging-based Alzheimer's disease diagnosis using deep learning models. Frontiers in Neuroscience. 2017;11:814. DOI:10.3389/fnins.2017.00814.

26. Shickel B, *et al*. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. Journal of Biomedical Informatics. 2018;83:168-185. DOI:10.1016/j.jbi.2018.04.007.
27. Simon N, *et al*. Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of Statistical Software. 2011;39(5):1-13. DOI:10.18637/jss.v039.i05.
28. Sperling RA, *et al*. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia. 2011;7(3):280-292. DOI:10.1016/j.jalz.2011.03.003.
29. Sperling RA, *et al*. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups. Alzheimer's & Dementia. 2011;7(3):280–292. https://doi.org/10.1016/j.jalz.2011.03.003.
30. Street K, *et al*. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19(1):477. DOI:10.1186/s12864-018-4772-0.
31. Than M, *et al*. Troponin and BNP as ML-driven biomarkers in cardiovascular disease prediction. JAMA Cardiology. 2019;4(3):251-259. DOI:10.1001/jamacardio.2018.4100.
32. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-288. DOI:10.1111/j.2517-6161.1996.tb02080.x.
33. Tonekaboni S, *et al*. What clinicians want: Contextualizing explainable machine learning for clinical end use. NPJ Digital Medicine. 2019;2:108. DOI:10.1038/s41746-019-0181-5.
34. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019;25(1):44-56. DOI:10.1038/s41591-018-0300-7.
35. Trapnell C, *et al*. Pseudotemporal ordering of individual cells reveals dynamic gene expression programs during cell differentiation. Nature Biotechnology. 2014;32(4):381-386. DOI:10.1038/nbt.2859.
36. Vamathevan J, *et al*. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. 2019;18(6):463-477. DOI:10.1038/s41573-019-0024-5.
37. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9:2579-2605.
38. Wood DE, *et al*. Improved metagenomic analysis with Kraken 2. Genome Biology. 2019;20(1):257. DOI:10.1186/s13059-019-1891-0.
39. Yu J, *et al*. Metagenomic analysis of colorectal cancer reveals key microbial interactions. Microbiome. 2021;9(1):20. DOI:10.1186/s40168-020-00989-9.
40. Zhavoronkov A, *et al*. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature Biotechnology. 2019;37(9):1038-1040. DOI:10.1038/s41587-019-0191-4.
41. Zhou J, *et al*. Deep learning sequence-based ab initio prediction of variant effects on protein function. Nature Machine Intelligence. 2018;1(5):271–279. https://doi.org/10.1038/s41592-018-0136-3.
42. Zhou N, *et al*. DeepGOPlus: Improved protein function prediction from sequence. Bioinformatics. 2020;35(14):5190-5198. DOI:10.1093/bioinformatics/btz864.
43. Zhou X, *et al*. Multi-omics profiling reveals widespread epigenomic differences in colorectal cancer tissue compared to normal adjacent tissue. Molecular Cancer Research. 2019;17(4):860–870. https://doi.org/10.1158/1541-7786.
44. Zitnik M, *et al*. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34(13):i457-i466. DOI:10.1093/bioinformatics/bty294.