Code of Life Via AI

Shraddha Kiran Aher[1]

[1]Pravara Institute of Medical Sciences (DU), Loni Bk., India

Corresponding author: shraddhaaher63@gmail.com
Doi: https://doi.org/10.5281/zenodo.17947034
Received: 10 December 2025
Accepted: 23 December 2025

Abstract:

Artificial intelligence (AI) is rapidly reshaping modern medicine, especially with the growing availability of digital health records, powerful computing systems, and large-scale genomic data. In clinical settings, AI tools now support diagnostic decision-making, help manage electronic health records, and even guide robotic systems used in surgery and patient care. In genomics, AI has become particularly valuable for interpreting the massive datasets produced by next-generation sequencing. These approaches help identify structural variants, analyze non-coding mutations, and integrate information from multiple omics layers to better understand disease mechanisms. Machine-learning algorithms are also improving the accuracy of read alignment, variant calling, and prediction of how specific genetic changes influence disease.

AI has shown impressive success in fields like oncology and medical imaging, often matching or surpassing expert-level performance. It is also accelerating biomarker discovery by allowing researchers to study genome, proteome, and metabolome data at the same time, rather than relying on traditional single-marker methods. In neurodevelopmental disorders, AI has helped analyze complex sequencing datasets and improved interpretation of genetic variants linked to conditions such as autism and epilepsy.

Despite its potential, the use of AI in genomics faces important challenges. These include the need for high-quality data, difficulties in interpreting complex model outputs, ethical concerns around privacy, risks of algorithmic bias, and the high computational resources required. While AI offers significant opportunities to advance precision medicine, its best use will come from combining powerful computational tools with skilled human expertise to ensure accuracy, fairness, and responsible clinical application.

Keywords:

Artificial intelligence; Machine learning; Genomics; Next-generation sequencing (NGS); Biomarkers; Precision medicine; Neurodevelopmental disorders; Deep learning

Introduction

Artificial Intelligence (AI) is a general term that implies the use of a computer to model intelligent behavior with minimal human intervention. AI, described as the science and engineering of making intelligent machines, was officially born in 1956. The term is applicable to a broad range of items in medicine such as robotics, medical

diagnosis, medical statistics, and human biology—up to and including today's "omics". AI in medicine has two main branches: virtual and physical. The virtual branch includes informatics approaches from deep learning information management to control of health management systems, including electronic health records, and active guidance of physicians in their treatment decisions. The physical branch is best represented by robots used to assist the elderly patient or the attending surgeon. Also embodied in this branch are targeted nanorobots, a unique new drug delivery system [1].

Artificial intelligence (AI) is gradually changing medical practice. With recent progress in digitized data acquisition, machine learning and computing infrastructure, AI applications are expanding into areas that were previously thought to be only the province of human experts [2].

In the last few decades, digitization of medical health records added a massive amount of data related to healthcare. Large digitization initiatives like EMERGE network, and 'All of Us' by NIH, USA; Electronic Health Record (EHR) initiatives by Canadian Institutes of Health Research, National Health Service, UK are some of the world's largest electronic health record databases. The application of AI algorithms will be greatly benefitted from these large digitization efforts that can help establish genotype-phenotype relationships for genetic diseases and have the potential to infer numerous phenotypic correlations and associations. Of course, collecting large scale digital data will only be helpful if the data comprise relevant clinical information to model AI algorithms.

The application of AI in medicine is a burgeoning area of development in light of the major impact it could potentially have on healthcare provision. The application of machine learning in medical imaging on skin lesions and treatable retinal diseases has been the most impactful, and demonstrates the potential for this technology in medical practice. Deep learning algorithm to diagnose heart attack using 549 ECG records shows a sensitivity of 93.3% and specificity of 89.7%, outperforming cardiologists.

Recently, DNA sequencing technology adopted machine learning to read out long stretches of DNA fragments from digital electronic signaling data. Long read technologies are important to resolve repetitive regions in the genome and detect complex structural variants. The current short reas technology can not resolve these issues and it is still unknown the disease risk contribution from repetitive region and structural variation of the genome. Nanopore sequencing technology in particular uses a neural network based deep learning method to 'call' DNA bases from the electronic signal produced by the nanopore flow cells. This method has accuracy over 98% and can produce mega base long DNA reads.

There has been an attempt to use AI in the clinical classification of genomic variation, based on the characterization of non-coding variants splicing code, DNA/RNA binding proteins and non-coding RNA (ncRNA) using large-scale molecular datasets. Classifying mutations according to their clinical relevance is very complex due to the largely unknown penetrance of individual variants, (i.e., the probability of diagnosis given a particular variant is identified, or mathematically, P(disease+|variant+)) Moreover, high penetrance variants are largely infrequent, with those of low penetrance much more common.

Although most of the variants are non-coding in our genomes, determining pathogenicity of rare or common non-coding variants still requires major advancement in genomics. It will require multidimensional biological data

and the use of artificial intelligence approaches to decipher the pathogenicity. Furthermore, many penetrant variants are also known to have more than one clinical manifestation, known as pleiotropy, and many diagnoses are characterized by variable presentation (phenotypic heterogeneity).

Despite this, however, recent deep learning methods have had some degree of success in the correct interpretation of phenotype and genomic data for disease risk in numerous types of cancer, diabetic retinopathy and pharmacogenomics. For example, in discriminating lymph node metastases, 7 independent deep learning implementations showed greater discrimination power (i.e., in relation to pathological versus non-pathological) compared to 11 pathologists. The best deep learning algorithm performed with an area under the curve (AUC) of 0.99, compared to 0.88 for 'best' clinician-derived score. The specificity found to be similar between AI and the diabetic retinopathy expert, AUC 0.96 and 0.98, respectively [3].

AI in Next Genomic Sequencing (NGS)

With the advancement in technology, the future of healthcare will be transformed due to the generation of big digital datasets acquired by means of next generation sequencing (NGS), use of algorithms for image processing, patient-related health records, data arising from large clinical trials and disease predictions. Oncology has been in the forefront to reap the benefits of AI for universal cancer management. This includes early detection, tailored or targeted therapy by obtaining genetic information of the patient and predictions of future outcomes. AI's capabilities of pattern recognition and complex algorithms can be employed to gain relevant clinical information that will decrease errors related to diagnostics and therapy. ML is a valuable tool in oncology with frequent applications in precision medicine. Diagnostic images and genetic analysis data are obtained from complex neural networks and can predict probability of disease and treatment outcomes. Deep learning is the most frequently used AI tool in radiomics, a field of machines that extracts diagnostic imaging to identify malignant tumours that fail to be identified by the human eye. The collective efforts of radiomics and deep learning will deliver increased accuracy in diagnostic image analysis. Combined, the applications of AI and ML in healthcare are implemented to improve disease management and provide effective medical care. Improved work in AI permits decision making in a human-like manner [4].

NGS generates large-scale, complex genomic data with capabilities to identify patterns and correlations using AI-enabled toolsets. Advanced bioinformatics infrastructures assist with decoding genomic data to unravel clinically relevant information required in implementing precision medicine. Big data arising from the NGS is described using 5 important characteristics that are: i) volume, ii) variety, iii) velocity, iv) verification and v) value. NGS produces volumes of data that are translated into gigabytes, terabytes or petabytes. For instance, over 100 gigabytes of data is produced from sequencing a single genome. The data produced is diverse and has variety which is typically presented in text form or images that require decoding, while maintaining authenticity and reliability.

With further advances in AI and computational methods, acquiring relevant information from NGS datasets is becoming more and more time effective with some platforms allowing real-time viewing. NGS uses file formats such as FASTQ (to align reference sequences), BAM (the binary version of sequence alignment/map) and VCF (Variant Call Format) to generate large datasets. The size of the file is dependent on the coverage and read length. The main challenge with such big datasets is analysing and interpreting for clinically relevant outcomes in the

presence of large sequencing data. NGS utilises ML-enabled tools to ensure accurate read alignment, reliable variant calling and variant annotation. These ML algorithms are developed to identify useful insights to distinguish and classify genotypes based on patterns. Moreover, ML enhanced bioinformatic tools are proficient in assigning clinical relevance and level of severity in correlated genetic variations. ML utilises two approaches to classify data known as supervised and unsupervised methods. In the supervised method, the system is trained to identify known genetic information such as regulatory regions, promoters, enhancers, active sites and splice sites. In the unsupervised method, unlabelled sequences are detected. Functional impact of missense variants is predicted by computational algorithms such as SIFT, PolyPhen2, PROVEAN, AlignGVGD and MutationTaster. Other *in silico* computational tools such as SpliceSiteFinder, MaxEntScan, NNSPLICE, GeneSplicer and Human Splicing Finder function as splice site prediction programs for intronic and silent variants. This way genetic variants and mutations are identified by leveraging ML algorithms. Despite advances in AI and ML, human input with adequate clinical and analytical knowledge remains essential [4].

Predicting and developing biomarkers using AI

Previous work in biomarker research involved the testing of suspected markers individually. Through this process, individual markers were selected from proposed biological processes or previous research results. This hypothesis-driven process is inefficient and costly in terms of money, time, and sample used for each experiment. With the advent of omics, experimenters can test multiple markers simultaneously. Beyond improving throughput and decreasing waste, omics reduce bias as it. In contrast to genetics (the study of heredity), genomics is the study of the entirety of all deoxyribonucleic acid (DNA) within an organism. In certain diseases with underlying genetic components, understanding the genome helps identify patients at risk for developing particular diseases. With improved DNA sequencing technologies, the genome can be characterized through genome-wide association studies (GWASs). GWAS aims to identify variations in genes, such as single nucleotides.

MS is a powerful analytical technique for identifying unknown samples. Unlike other proteomic techniques, MS is destructive, as the analyte is fragmented before being ionized and sorted by mass-to-charge (m/z) ratio. Using existing libraries, fragments can be pieced together to identify analytes. A benefit of MS is it is untargeted and proteins across the characterized proteome can be detected as well as posttranslational modifications. In the past, large and complicated data sets generated by omics required extensive statistical analysis. However, increased access to AI has improved analysis of omics data sets. Although there are several methods to analyze omics, the major methods include principal components analysis (PCA), random forest models, and support vector machines (SVMs). PCA allows for related groups of proteins, referred to as features, to be extracted using orthogonal (90°) or oblique (<90°) rotations.

When designing biomarker discovery experiments, there are several considerations. One decision is sample source. Although most proteomics studies test serum, it may be difficult to detect proteins with low expression.16,52 Other sample types, such as tissue, provide insights to the local proteome within the heart but are difficult to obtain.66 Urine samples, although easily obtained are not without their shortcomings. The application of proteomics in HF research has uncovered new insights into disease processes. The following studies demonstrate how proteomics of

various platforms can be used to discover biomarkers to aid in identifying those at risk for developing HF, predicting disease progression, and understanding mechanisms of disease progression and treatments.

Advancements in omics and AI have improved researcher's ability to discover HF biomarkers. Rather than relying on hypothesis-driven, individual marker testing, omics has opened the door to high-throughput analyses of genome, transcriptome, proteome, and metabolome to gain deeper understandings of pathophysiology in HF. With future validation studies, the characterized markers can be developed into panels to predict and monitor HF in the clinic [5].

Artificial intelligence in Neurodevelopmental disorders (NDDs)

The availability of fMRI that enabled the high-resolution capture of brain activity was a major milestone in NDD diagnosis and therapeutics in the 90s. Since then, the human genome has been mapped and exome and whole-genome sequencing technologies have led to the detection of hundreds of disease causal genes and loci for ASD and other NDDs. Indeed, conducting exome or genome sequencing for newborn babies at high risk of genetic abnormalities is now becoming more frequent and cost effective. Subsequently, the advent of transcriptome sequencing dependent technologies led to the establishment of the Allen developmental human brain atlas in 2011, ENCODE database profiling the non-coding elements in the human genome in 2012, and the Human Cell Atlas in 2017. Multiple sequencing consortiums focussed on the NDDs were also started during the period of 2012 and 2014 with the aim of identifying disease-implicated variants, and making exome and WGS data available to the scientific community for further study. Bearing in mind that most identified genetic variation is of unknown pathogenicity, and little is known about functional consequences, the discovery of CRISPR/Cas as a gene editing tool in 2012 has allowed scientists to better characterize identified genetic variants. In recent years, artificial intelligence approaches have been used in autism spectrum disorder, epileptic encephalopathy, intellectual disability, attention deficit hyperactivity disorder (ADHD),and rare genetic disorders [3].

Limitations of AI-based tools in genomics

While AI has great potential in the field of genomics, there are also some limitations that need to be considered. Therefore, the accuracy and reliability of the genomic data used to train AI algorithms is critical. Inaccurate or incomplete data can lead to misleading results. AI algorithms can generate complex and sometimes difficult-to-interpret results. While AI can enhance genomic analysis and decision-making, it should not replace human expertise. Clinical decisions based on genomic data should involve a combination of AI-driven insights and the expertise of healthcare professionals. It is crucial to maintain a balance and ensure that AI is used as a tool to augment human capabilities rather than replace them. Therefore, it is important to have experienced and knowledgeable professionals who can interpret the results and make informed decisions based on them. AI algorithms can sometimes become overfit to the training data, leading to poor performance on new data. This can be particularly problematic in genomics research, where there is a high degree of genetic variation. The use of AI in genomics raises ethical concerns, including privacy and data security. Genomic data is highly sensitive and can reveal sensitive information about individuals and their families. It is crucial to ensure proper data protection and privacy regulations are in place to prevent unauthorized access or misuse of genomic information. AI algorithms can be computationally expensive and require significant resources, including high-performance computing and large

data storage capabilities. AI models are trained on existing datasets, which may contain biases and disparities. If these biases are not properly addressed, AI algorithms may perpetuate existing inequalities in healthcare. It is important to ensure diverse and representative datasets and address potential biases in algorithm development to prevent exacerbating healthcare disparities [6].

References

1.Hamet, P., & Tremblay, J. (2017). *Artificial intelligence in medicine*. Metabolism, 69(Suppl.), S0–S40. https://doi.org/10.1016/j.metabol.2017.01.011

2.Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). *Artificial intelligence in healthcare*. Nature Biomedical Engineering, 2, 719–731. https://doi.org/10.1038/s41551-018-0305-z

3.Uddin, M., Wang, Y., & Woodbury-Smith, M. (2019). *Artificial intelligence for precision medicine in neurodevelopmental disorders*. NPJ Digital Medicine, 2, 112. https://doi.org/10.1038/s41746-019-0191-0

4.Dlamini, Z., Francies, F. Z., Hull, R., & Marima, R. (2020). *Artificial intelligence (AI) and big data in cancer and precision oncology*. Computational and Structural Biotechnology Journal, 18, 2300–2311. https://doi.org/10.1016/j.csbj.2020.08.019

5.Michelhaugh, S. A., & Januzzi, J. L., Jr. (2022). *Using artificial intelligence to better predict and develop biomarkers*. Heart Failure Clinics, 18, 275–285. https://doi.org/10.1016/j.hfc.2021.11.004

6.Reddy, S. S. P., Francis, D. L., Thirumoorthi, H., Manohar, B., Krishnan, S. A., & Chopra, S. S. (2023). *Artificial intelligence in the genomics era: A blessing or a curse?* Journal of Regenerative Biology and Medicine, 5(3). https://doi.org/10.37191/Mapscj-2582-385X-5(3)-134