

Machine Learning Approaches in Genomics: An In-Depth Review of Predictive and Diagnostic Applications for Genetic Disorders

Sabisanthoshni Mary, and Sitalakshmi Thyagarajan*

Department of Bioinformatics, SIMATS Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Thandalam, Chennai-602 105, Tamil Nadu, India

*Corresponding author: sitalakshmithyagarajan.sse@saveetha.com

Abstract

The rapid growth of machine learning (ML) is transforming how we study and interpret the results of genomic data, opening new possibilities for predicting and diagnosing genetic disorders. Genetic disorders have long posed a challenge for medicine, but recent advances in machine learning are giving us powerful new tools to understand and combat these conditions. In this review, we explore the ML techniques from traditional algorithms to advanced deep learning models are being used to analyze genetic variations, improve disease detection, and pave the way for precision medicine. The various research analysis also gives a complete examination of diverse ML techniques that are revolutionizing the field of genetic research. This review also highlights the methods that facilitate the detection of disease-associated genetic variants, improve the accuracy of genomic sequencing, and enable the classification of complex genetic disorders. At the same time, it also addresses challenges like data quality, model interpretability, and ethical concerns that come with integrating AI into genomics. ML algorithm also has an current knowledge to improve the exactness of diagnosis by analyzing the genomic data and identifying the patterns that are associated with diseases.

Keywords: Machine learning, genomics, genetic disorders, AI in medicine, precision genomics, deep learning.

1. Introduction

The Artificial Intelligence (AI) have become an informative tool in the field of

biomedical research, along with Machine Learning (ML) at its core for extracting the complex and high-dimensional data. According to the research data, AI involves the science and engineering of developing intelligent systems capable of performing tasks that are typically require human intelligence (1). ML uses algorithms from the data to identify the patterns and make data-driven predictions or decisions. This ML basically divided into supervised learning (e.g., disease classification), unsupervised learning (e.g., patient clustering), and Deep Learning (DL) uses multi-layered neural networks to model complex relationships in data (2). Genetic disorders conditions caused by alterations in the genome range from single-nucleotide variants to large chromosomal abnormalities. These can be inherited in Mendelian patterns (e.g., cystic fibrosis or Huntington's disease), polygenic (e.g.,schizophrenia), or arise spontaneously through de novo mutations (e.g., CHARGE syndrome) (3). Despite affecting many people globally, these rare diseases remain difficult to diagnose, with patients often experiencing diagnostic delays of 5–7 years (4).

Rendering variations of uncertain significance (VUS) and integrating genetic with phenotypic data remains a major challenge in this modern world. The traditional rule systems used in clinical decision-making, which was developed in the mid-20th century, laid the groundwork for AI in medicine but limited by their inflexibility and reliance on hand-crafted rules (5). These systems were taken by data-driven ML models, which is better capture non-linear and context-dependent relationships in

genomic and clinical data (6). As such, AI/ML methods have become central to variant prioritization, genotype-phenotype mapping, and risk prediction, offering scalable, automated solutions for accelerating precision medicine.

2. Machine Learning Techniques in Genomics

ML is an essential tool in genomics, that enables the study of complex datasets produced through high output technologies. ML methods also enable researchers to derive meaningful patterns from gene expression profiles, single-nucleotide polymorphisms (SNPs), and whole-genome sequencing data for predicting disease susceptibility to identifying novel biomarkers (7).

Supervised Learning

The Supervised Learning involve in analysing of labelled input-output pairs, allowing to learn the function that maps all the inputs like genetic features to outputs which are disease status. Algorithms namely, support vector machines (SVMs) and random forests (RF) were widely used in genomics to perform classification, regression, and risk prediction tasks (8). These models are trained on curated datasets and then evaluated on unseen data to assess their ability to generalize.

Support Vector Machine (SVM)

Support Vector Machine (SVMs) are powerful learning algorithms which aims to identify the parameters in separating the different classes in high-dimensional space. Their strength for overfitting and their ability to handle sparse, high-dimensional genomic data make them useful for applications such as gene expression-based classification and biomarker discovery (9). Zhi et al. (2017), in their research studies revealed that SVMs classified the metastatic and non-metastatic colorectal cancer (CRC) samples based upon RNA-Seq data. This model was trained on a 40 gene signature, which achieved 100% precision in distinguishing the metastatic CRC

and identifying genes such as *CREB1*, *CUL7*, and *SSR3* as key biomarkers. The tools like e-SVM have also been developed to detect gene fusions in RNA-Seq data, further showcasing the importance of SVMs in genomics (10).

Random Forest

Random Forest (RF) are multiple-based learning algorithm that builds the multiple decision trees and aggregates their predictions to improves in arrangement of accuracy and reduce overlapping. RFs are particularly well-suited for genomic data due to their ability to handle missing values, nonlinear interactions, and large feature sets (11). In genomic medicine, RFs are used widely for variant prioritization, prediction of variant pathogenicity and identifying disease-associated SNPs. For instance, the RF-DOS (Random Forest-Disease Ontology Score) algorithm mixes the features such as allele frequency, protein impact predictions, and evolutionary conservation to identify pathogenic variants from whole-exome sequencing data (12). Moreover, RF models have been instrumental in developing the polygenic risk scores (PRS) for complex disorders like schizophrenia and Alzheimer's disease. These models outperform traditional linear methods by capturing gene-gene interactions and epistatic effects (13).

RFs have also shown promise results in rare disease diagnostics. In neurodevelopmental disorders like Autism Spectrum Disorder (ASD), RF-based models combine the gene expression, protein interaction networks, mutation data and achieves the high predictive accuracy (14). The important measures in RFs also aids in interpreting model predictions, making them valuable for hypothesis generation and biological insight.

Unsupervised Learning in Genomic Medicine

The unsupervised learning is a type of machine learning where algorithms identify the patterns without labels or guidance. Unlike the supervised models, the algorithms

are separately uncovering the hidden structures, making them valuable for genomic medicine where the datasets are often high-dimensional and the ground truth is incomplete. The unsupervised techniques are used in various tasks like clustering, dimensionality reduction, and data visualization, which could help the researchers to identify the subtypes of diseases, group similar patients, and detect rare genetic signatures (15).

The most widely used techniques are k-means clustering, that partitions samples based on similarity in gene expression or genomic features. In the oncology studies, k-means clustering plays a key role in classifying cancer into molecular subtypes. The most prominent example is breast cancer, where the analysis of transcriptomic data reveals the intrinsic subtypes that are associated with distinct diagnosis and treatment strategies (16). Similar approaches have been applied in neurodevelopmental disorders, enabling stratification of genetically homogeneous groups from clinically overlapping phenotypes (17).

Hierarchical clustering is an important method for analyzing the complex datasets like metabolomics and transcriptomics. It is a tree-like dendrograms not only illustrate relationships between genes, but also integrate seamlessly with heatmap conceptions. In rare disease researches, the hierarchical clustering has uncovered previously unrecognized subtypes of metabolic disorders, aiding both diagnosis and treatment planning (18) Fig. 1).

The Density-based clustering algorithms, such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise), are used in detecting outlier patients or ultra-rare variants that might be missed by traditional methods. These tools have been successfully applied to integrate genomics and facial morphology data and also assisting in the diagnosis of rare dysmorphic syndromes (19).

Other important powerful approach is Self-Organizing Maps (SOMs), a neural network-based technique which reduces high-

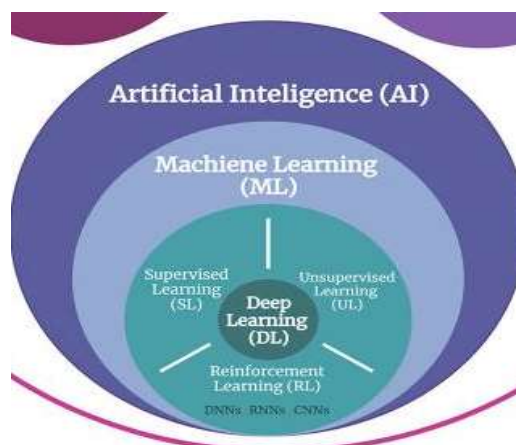


Fig 1: Shows the hierarchical relationship between Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL)

dimensional genomic data into low-dimensional representations by preserving topological structure. In autism research, SOMs have been employed to analyze DNA methylation profiles, revealing epigenetically distinct subgroups with different clinical features and developmental routes (20).

Spectral clustering have shown great promise in network-based genomics, particularly in Protein-Protein Interaction (PPI) networks. By identifying the tightly connected modules of genes or proteins, researchers have found a novel pathway involved in diseases like cardiomyopathies and congenital heart defects (21). These learning techniques empower researchers to navigate the complex genetic landscapes, especially in the context of rare diseases where the patient are heterogeneity and data sparsity make traditional approaches insufficient.

2.3 Principal Component Analysis

Principal Component Analysis (PCA) is a unsupervised learning technique, which are widely used in genomics to refine the complex from of high-dimensional data into a more manageable form. By transforming correlated variables into a set of linearly uncorrelated components, PCA seizes the most significant variance within the data, facilitating the identification of patterns,

population structures, and outliers that indicate pathogenic variations (22).

In Genome-Wide Association Studies (GWAS), PCA is an instrumental tool which helps in correcting the population stratification and confounding factors where genetic differences between populations can lead to false associations. The 1000 Genomes Project represented this application, using PCA to distinguish the global populations along with principal components, thereby enhancing the accuracy of genetic association findings (23). Apart from GWAS, the PCA also aids in the detection of rare genetic disorders. Individuals with unique combinations of rare variants often appear as outliers in PCA plots, enabling researchers to pinpoint novel disease-associated genes (24).

On oncology studies, PCA has been used to classify the tumors into molecular subtypes based upon the gene expression profiles, offering visions that are transcended the traditional histopathological methods (25). PCA also serves as a quality control measure in genomic studies. By visualizing the sample distributions, researchers detect the batch effects, sample contamination and technical artifacts that compromise the data integrity (26). In single-cell RNA sequencing (scRNA-seq), PCA also facilitates the identification of cell populations and developmental trajectories, focussing on cellular heterogeneity in diseases (27).

PCA methodologies, such as probabilistic PCA and kernel PCA have expanded its applicability. Probabilistic PCA addresses the missing data issues that are common in genomic datasets, while kernel PCA addresses the non-linear relationships thereby enhancing the analysis of complex biological systems (28).

The normalization techniques are very important for effective PCA application. For example, the quantile normalization is done for gene expression data and linkage disequilibrium trimming which are essential for SNP data to ensure meaningful principal components (29). Interpreting these components requires biological validation,

typically by examining the loading scores to identify relevant genes or pathways.

Integrating PCA with other machine learning approaches, such as t-distributed stochastic neighbor embedding (t-SNE) or uniform manifold approximation and projection (UMAP), enhances data visualization and interpretation. These hybrid methods are particularly powerful in multi-omics analyses, combining genomics, transcriptomics, and proteomics to uncover comprehensive molecular signatures of genetic disorders (30).

3. Deep Learning

Integrating PCA with other machine learning approaches, like t-distributed stochastic neighbor embedding (t-SNE) or uniform manifold approximation and projection (UMAP), improve the data visualization and interpretation. These methods are very powerful in multi-omics analyses, combining the genomics, transcriptomics and proteomics to discover comprehensive molecular signatures of genetic disorders (30).

Deep learning (DL) has become an important tool in genetic disorder research, offering advanced capabilities to analyze complex genomic data and discover the disease mechanisms that are traditional methods might miss. These sophisticated neural network architectures excel in identifying subtle patterns in high-dimensional biological data, enabling breakthroughs across the various aspects of genetic disease research right from variant detection and interpretation to disease prediction and personalized treatment strategies.

3.1 Convolutional Neural Networks (CNNs)

CNNs are advanced genomic sequence tool by the detecting local patterns and hierarchical features in DNA and RNA sequences. Google's Deep Variant employs CNNs to achieve high accuracy in variant calling, processing sequencing reads as images to identify single nucleotide polymorphisms (SNPs) and insertions/

deletions (indels) with precision. The DeepSEA utilizes this CNNs tool to predict, functional impact of non-coding variants by analyzing chromatin accessibility and transcription factor binding patterns of various cell types. In rare disease diagnostics, CNNs applied to whole-genome sequencing data identifies the complex structural variants and repetitive element insertions that are often challenging to detect with conventional methods (31).

3.2 Recurrent Neural Networks (RNNs)

RNNs along with advanced variants the Long Short-Term Memory (LSTM) network and Gated Recurrent Units (GRUs), has proven to be an effective tool for modeling sequential dependencies in genomic data. Various tools like SpliceAI, Long Short-Term Memory (LSTM) leverage this architecture to predict RNA splicing defects that contributes to genetic disorders with remarkable accuracy, identifying pathogenic splicing variants directly from DNA sequences. For dynamic processes such as gene expression regulation, RNNs can model temporal patterns in time-series omics data, aiding in understanding how mutations disrupt developmental pathways in disorders like autism or congenital heart defects (32).

The Transformer architecture was developed for natural language processing, which have been adapted to genomics with prominent success. Models like DNA-BERT learn contextual representations of DNA sequences by processing the vast amounts of nucleotide data during pretraining, enabling the prediction of regulatory elements, transcription factor binding sites, and variant effects with high accuracy. These models are particularly valuable for interpreting variants of uncertain significance (VUS) in clinical exomes, thereby improving diagnostic yields for rare genetic disorders.

3.3 Graph Neural Networks (GNNs)

GNNs has opened new avenues in analyzing biological networks pertinent to genetic disorder research. By representing

genes as nodes and their interactions as edges, GNNs predicts the novel disease genes by propagating information through protein-protein interaction networks. This approach has successfully identified candidate genes from unsolved rare diseases by finding network neighborhoods similar to known disease genes. GNNs also predict the drug-gene interaction by assisting the patient gene with genetic disorders to target therapies based on mutational profiles.

Autoencoder architecture are instrumental in dimensionality reduction and anomaly detection within genetic disorder research. These models learn compressed representations of high-dimensional genomic data, facilitating the identification of patients who have rare genetic conditions. Variational autoencoders (VAEs), in particular, have been successful in single-cell genomics, aiding in the identification of rare cell populations affected by pathogenic mutations in developmental disorders.

The integration of multimodal deep learning approaches represents the forefront of genetic disorder research. Hybrid architectures that combine CNNs for image data (e.g., histopathology), RNNs for time-series data (e.g., longitudinal clinical measurements), and GNNs for biological networks are providing comprehensive views of genotype-phenotype relationships. For example, deep learning models integrating retinal images with genomic data have improved the diagnosis of inherited retinal dystrophies, while similar approaches combining ECG traces with genetic information enhance the prediction of cardiac channelopathies.

The integration of multimodal Deep Learning approaches, represents the forefront of genetic disorder research. Hybrid architectures, combines the CNNs for image data (e.g., histopathology), RNNs for time-series data (e.g., longitudinal clinical measurements), and GNNs for biological networks for providing comprehensive views of genotype-phenotype relationships. Deep Learning model integrates with retinal images with genomic data and improved the

diagnosis of retinal dystrophies, while the similar approaches combine the ECG traces with genetic information which enhance the prediction of cardiac channelopathies.

In spite these advancements, challenges still persist in applying Deep Learning to genetic disorders. The necessity for large training datasets poses various difficulties for ultra-rare conditions. However, techniques like transfer learning, pretraining on common variants and fine-tuning on rare disease data and generative adversarial networks (GANs) for synthetic data generation are helping to mitigate this limitation. The Model interpretability remain unexplored in clinical applications, but emerging techniques such as attention visualization and integrated gradient improves our ability to understand model decisions. The substantial computational resources required for training genome-scale models are becoming more accessible through cloud-based solutions and specialized hardware.

The development of foundation models for genomics akin to large language models in natural language processing holds a very good promise for adaptation to multiple genetic disorder tasks. The integration of deep learning with CRISPR screening data has anticipated to validate disease genes further. Moreover, the application of reinforcement learning to optimize personalized treatment strategies are limited. As these technologies evolve, Deep Learning is poised to transform every aspect of genetic disorder research and clinical care, from accelerating diagnosis, enabling the precision therapies tailored to an individual's unique genetic makeup.

4. AI Applications in Genetic Disorder Prediction

AI-driven Polygenic Risk Score (PRS) are progressive genetic risk prediction by analyzing thousands of Single-Nucleotide Polymorphisms (SNPs) to assess disease susceptibility. The models like random forests and logistic regression, genomic data with clinical variables like age and family history

were used to categorize patients according to their risk for complex disorders, including Alzheimer's disease and Type 2 diabetes (33).

The research studies revealed that PRS were used to identify the high-risk individuals for breast cancer (BRCA1/2-negative) with 80% accuracy, enabling earlier and targeted screening (34). Despite these advances, challenges such as population bias, due to most PRS being trained on European ancestry data and the need for greater interpretability of risk thresholds persist (35). The Federated learning and explainable AI (XAI) are an emerging trend to address these issues by diversifying training datasets and clarifying the factors influencing risk predictions (36).

Deep Learning models are used for screening and pre-symptomatic diagnosis of genetic disorders. This Collaborative Laboratory Integrated Reports (CLIR) system uses Whole-Genome Sequencing (WGS) to identify rare disorders like spinal muscular atrophy within hours, far outpacing traditional biochemical tests (37). The Retinal AI tools are capable of detecting the abnormalities and diagnosing the inborn errors of metabolism in Tay-Sachs disease before the symptoms appear (38). The Machine Learning models are trained for vocal patterns to identify the neurodevelopmental disorders including Angelman Syndrome with 90% sensitivity (39). However, these innovations face their key limitations, including false positives for ultra-rare diseases and ethical concerns surrounding pre-symptomatic testing in children (40).

AI is an advancing personalized medicine by linking genetic profiles to drug responses. Tools like PharmGKB uses to Support Vector Machines (SVMs) to predict adverse drug reactions, such as warfarin sensitivity based on VKORC1 variants (41). The reinforcement learning is being utilized to optimize the guide RNAs for gene editing in personalized treatments, such as correcting CFTR mutations in cystic fibrosis (42). Generative AI, namely ProteinGAN, were designed to study the novel enzymes for the

enzyme replacement therapies for lysosomal storage disorders. Research studies have demonstrated that AI-matched therapies significantly improved outcomes for patients with Duchenne Muscular Dystrophy by 40% compared to standard care (43).

Several AI tools have already been approved by the FDA for use in genetic medicine. Google Health's Deep Variant has received FDA clearance for variant calling in clinical next-generation sequencing (NGS) pipelines (44). AlphaMissense, developed by DeepMind, is currently under FDA review which classifies the variants of uncertain significance (VUS) with 90% accuracy (45). Fabric Genomics' Fabric GEM automates the application of ACMG guidelines for variant interpretation (46). However, barriers to broader adoption remains underrepresented populations, reimbursement challenges for AI-based diagnostics and the integration of AI tools with Electronic Health Record (EHR) systems (47). Emerging solutions like FHIR Genomics APIs are helping to address these integration issues (48).

5. Diagnostic Tools and Techniques in Genetic Disorders

The Artificial Intelligence (AI) have revolutionized, the diagnosis of genetic disorders by using new tool to analyze and interpret the genetic data with unprecedented precision. One of the major advancements in genetic testing is the use of AI to analyze massive amounts of data produced by modern DNA sequencing machines. Google's DeepVariant, introduced in 2018, employs AI to detect tiny genetic errors with 99.9% accuracy, greatly by improving the reliability and speed of diagnostic tests for conditions like cystic fibrosis and rare childhood diseases (49). AI is also helping to interpret how genes work in concert, similar to how an orchestra functions, by analyzing gene activity patterns. This capability has been instrumental in identifying the distinct subtypes of complex diseases, such as autism, and predicting the most effective treatments for individual patients (50). AI tools are making strides in identifying early

warning signs of genetic diseases before symptoms manifest, thus facilitating the earlier intervention (51). When doctors encounter unusual genetic changes, AI systems like DeepMind's AlphaMissense predicts whether these variations are likely to be harmful, with 90% accuracy, by aiding geneticist in solving previously perplexing cases (52). One of the most exciting applications of AI is in personalized medicine wherein, it can analyze an individual's genetic profile to predict disease risks and help design tailored therapies. Moreover, AI is accelerating the development of new drugs by simulating their effects on specific genetic conditions, making treatments more targeted and effective (53).

6. Future Directions in AI for Genetic Disorder Diagnosis and Treatment

The integration of AI into genomics has transformed the diagnosis and treatment of genetic disorders to next level. One of the promising paths is multi-modal AI, which combines genomic data with other sources, such as wearable health monitors, to provide dynamic and continuous risk assessments (54). Such approaches allow real-time tracking of genetic risks and personalized interventions. In gene therapy design, transformer-based models, like GeneGPT, have being used to optimize the adeno-associated virus (AAV) vectors for more effective gene delivery (55).

There are growing focus of AI tools are accessible and equitable across the globe. For example, the use of Mozambique's pilot AI-based Next-Generation Sequencing (NGS-AI) project helps to deliver genomic diagnostics in resource-limited settings, and also to improve healthcare (56). As AI continues to advance, these developments are reshaping our ability to predict, detect, and treat genetic disorders while addressing regulatory and ethical challenges that accompany such technologies (57).

7. Challenges

Despite of AI's significant promise, several challenges remain unexplored

wherein, these fully integrated technologies converted into clinical practice. One of the primary issues is ensuring that AI models are equally effective across all ethnic groups. Currently, many genetic datasets are excessively derived from populations of various countries, which can lead to biases in diagnostic accuracy when applied to other ethnic groups (58). The safeguarding genetic data remains a major concern of breaches that could have a serious implication for patient privacy. The cost also remains another barrier, with AI tools often being prohibitively expensive for many healthcare systems, especially in low-resource settings (59, 60).

8. Conclusion

The future AI in genetic disorder diagnosis mainly has three key gaps which includes

Technical: Developing more interpretable, biologically informed models that integrate multi-omics and clinical data.

Equity: Ensuring AI benefits to all populations by diversifying training datasets and reducing computational barriers.

Translation: The arrangement of research path to the clinical use by using various clinician-friendly tools.

As this AI continues to evolve along the advances in sequencing and functional genomics, the role in genetic medicine will expand from early detection and risk prediction to therapies. By the addressing current limitations through collaborative efforts among researchers, clinicians, and policymakers, AI can also fulfill its potential as a cornerstone of precision healthcare, ultimately improving outcomes for patients with genetic disorders worldwide.

References

1. Boycott KM, Rath A, Chong JX, et al (2017) International cooperation to enable the diagnosis of all rare genetic diseases. *Am J Hum Genet* 100:695–705
2. Esteva A, Robicquet A, Ramsundar B, et al (2019) A guide to deep learning in healthcare. *Nat Med* 25:24–29
3. FDA (2021) Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. In: FDA. <https://www.fda.gov>. 2021
4. Engel C, Gubbey T, Neyer S, et al (2017) Structural basis of RNA polymerase I transcription initiation. *Cell* 169:120-131.e22
5. Miotto R, Wang F, Wang S, et al (2018) Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 19:1236–1246
6. Obermeyer Z, Emanuel EJ (2016) Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med* 375:1216–1219
7. Shortliffe EH, Sepúlveda MJ (2018) Clinical decision support in the era of artificial intelligence. *JAMA* 320:2199–2200
8. Wright CF, FitzPatrick DR, Firth HV (2018) Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet* 19:253–268
9. Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. *Nat Rev Genet* 16:321–332
10. Zou J, Huss M, Abid A, et al (2019) A primer on deep learning in genomics. *Nat Genet* 51:12–18
11. 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:. <https://doi.org/10.1098/rsif.2017.0387>
12. Brown MP, Grundy WN, Lin D, et al (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* 97:262–267
13. Michl M, Taverna F, Woischke C, et al (2024) Identification of a gene expression signature associated with brain metastasis in colorectal cancer. *Clin Transl Oncol* 26:1886–1895
14. Liu X, Wu C, Li C, Boerwinkle E (2016) DbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum Mutat* 37:235–241
15. Cutler DR, Edwards TC Jr, Beard KH, et al (2007) Random forests for classification in ecology. *Ecology* 88:2783–2792

16. Laman Trip DS, van Oostrum M, Memon D, et al (2025) A tissue-specific atlas of protein-protein associations enables prioritization of candidate disease genes. *Nat Biotechnol*. <https://doi.org/10.1038/s41587-025-02659-z>
17. Agrawal P, Mohanty D (2021) A machine learning-based method for prediction of macrocyclization patterns of polyketides and non-ribosomal peptides. *Bioinformatics* 37:603–611
18. Krishnan A, Zhang R, Yao V, et al (2016) Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci* 19:1454–1462
19. Niu M, Ju Y, Lin C, Zou Q (2022) Characterizing viral circRNAs and their application in identifying circRNAs in viruses. *Brief Bioinform* 23:. <https://doi.org/10.1093/bib/bbab404>
20. Perou CM, Sørlie T, Eisen MB, et al (2000) Molecular portraits of human breast tumours. *Nature* 406:747–752
21. Satterstrom FK, Kosmicki JA, Wang J, et al (2020) Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* 180:568-584.e23
22. Patti GJ, Yanes O, Siuzdak G (2012) Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* 13:263–269
23. Gurovich Y, Hanani Y, Bar O, et al (2019) Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med* 25:60–64
24. Torres EB (2021) Precision Autism: Genomic stratification of disorders making up the broad spectrum may demystify its “epidemic rates.” *bioRxiv*
25. Choobdar S, Ahsen ME, Crawford J, et al (2019) Assessment of network module identification across complex diseases. *Nat Methods* 16:843–852
26. Jolliffe IT, Cadima J (2016) Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* 374:20150202
27. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65
28. Lek M, Karczewski KJ, Minikel EV, et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–291
29. Verhaak RGW, Hoadley KA, Purdom E, et al (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17:98–110
30. Leek JT, Scharpf RB, Bravo HC, et al (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11:733–739
31. Luecken MD, Theis FJ (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 15:e8746
32. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. *J R Stat Soc Series B Stat Methodol* 61:611–622
33. Schölkopf B, Smola A, Müller K-R (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput* 10:1299–1319
34. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193
35. Price AL, Patterson NJ, Plenge RM, et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909
36. Stuart T, Satija R (2019) Integrative single-cell analysis. *Nat Rev Genet* 20:257–272
37. Poplin R, Chang P-C, Alexander D, et al (2018) A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* 36:983–987
38. Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 12:931–934

39. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al (2019) Predicting splicing from primary sequence with deep learning. *Cell* 176:535-548.e24
40. Ji Y, Zhou Z, Liu H, Davuluri RV (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* 37:2112–2120
41. Jha K, Saha S, Singh H (2022) Prediction of protein-protein interaction using graph neural networks. *Sci Rep* 12:8360
42. Svensson V, Gayoso A, Yosef N, Pachter L (2020) Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* 36:3418–3421
43. Wang X, Li F, Zhang Y, et al (2024) Deep learning approaches for non-coding genetic variant effect prediction: current progress and future prospects. *Brief Bioinform* 25:. <https://doi.org/10.1093/bib/bbae446>
44. Lewis JT, Stephens J, Musick B, et al (2022) The leDEA harmonist data toolkit: A data quality and data sharing solution for a global HIV research consortium. *J Biomed Inform* 131:104110
45. Benegas G, Ye C, Albors C, et al (2025) Genomic language models: opportunities and challenges. *Trends Genet* 41:286–302
46. Huang D, Wang Z, Zhou Y, et al (2021) vSampler: fast and annotation-based matched variant sampling tool. *Bioinformatics* 37:1915–1917
47. Reddy S (2025) AI-Driven Early Detection of Rare Genetic Disorders in Neonates. *Journal Publication of International Research for Engineering and Management (JOIREM)* 5:
48. Cen L-P, Ji J, Lin J-W, et al (2021) Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat Commun* 12:4828
49. Duarte V, Gómez A, Corchado JM (2024) Analysis Pipelines and a Platform Solution for Next-Generation Sequencing Data. *Big Data Analysis and Artificial Intelligence for Medical Sciences* 219–251
50. Montanucci L, Brünger T, Boßelmann CM, et al (2024) Evaluating novel in silico tools for accurate pathogenicity classification in epilepsy-associated genetic missense variants. *Epilepsia* 65:3655–3663
51. Koh MY (2023) Vocal pattern recognition in the diagnosis of neurodevelopmental disorders. *Science Translational Medicine* 15:
52. Mavaddat N, Michailidou K, Dennis J, et al (2019) Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics* 104:21–34
53. Gordon AS, Rosenthal EA, Carrell DS, et al (2019) Rates of actionable genetic findings in individuals with colorectal cancer or polyps ascertained from a community medical setting. *Am J Hum Genet* 105:526–533
54. Adler-Milstein J, Aggarwal N, Ahmed M, et al (2022) Meeting the moment: Addressing barriers and facilitating clinical adoption of artificial intelligence in medical diagnosis. *NAM Perspect* 2022:. <https://doi.org/10.31478/202209c>
55. Joshi M, Pal A, Sankarasubbu M (2022) Federated learning for healthcare domain-pipeline, applications and challenges. *ACM Transactions on Computing for Healthcare* 3:1–36
56. Cacabelos R, Cacabelos N, Carril JC (2019) The role of pharmacogenomics in adverse drug reactions. *Expert Rev Clin Pharmacol* 12:407–442
57. Dolin RH, Heale BSE, Alterovitz G, et al (2023) Introducing HL7 FHIR Genomics Operations: a developer-friendly approach to genomics-EHR integration. *J Am Med Inform Assoc* 30:485–493
58. Woerner AC, Gallagher RC, Vockley J, Adhikari AN (2021) The use of whole genome and exome sequencing for newborn screening: Challenges and opportunities for population health. *Front Pediatr* 9:663752
59. Sun Q, Hao Q, Prasanth KV (2018) Nuclear long noncoding RNAs: Key regulators of gene expression. *Trends Genet* 34:142–157
60. Parisi GF, Terlizzi V, Manti S, et al (2025) Cutting-edge advances in cystic fibrosis: From gene therapy to personalized medicine and holistic management. *Genes (Basel)* 16:402