Modern American Journal of Medical and Health Sciences
ISSN (E): 3067-803X
Volume 01, Issue 09, December, 2025
Website: usajournals.org

AI ALGORITHMS FOR GENOMIC DATA ANALYSIS AND DISEASE RISK PREDICTION

Rahmatilla Amirqulov
Akbarshoh Abdusalomov
Qodirbek Isoqov
Abdullo Ganiboyev
Mirabbos Jafaraliyev
Tashkent State Medical University, Tashkent, Uzbekistan

Abstract

Over the past two decades, genomic and omics data generation has increased exponentially. High-throughput sequencing technologies, GWAS, and large-scale biobank projects have produced datasets containing millions of genetic variants across diverse human populations. These data promise deep insights into genetic contributions to disease risk, progression, and therapeutic response. However, conventional statistical models — like linear or logistic regression — frequently fail to capture complex genotype–phenotype relationships, epistasis (gene–gene interactions), nonlinearity, and the influence of regulatory or epigenetic factors (especially for complex diseases) (Cordell, 2009).

AI, encompassing ML and DL, offers powerful solutions. By learning from large-scale data, AI models can identify subtle patterns, nonlinear interactions, and complex dependencies that classical models may miss. Recent years have seen a surge in applying AI to human genomics for tasks such as variant effect prediction, polygenic risk scoring, disease classification, and multi-omics integration (Alharbi & Rashid, 2022). This review provides a comprehensive synthesis of these developments, outlines methodological frameworks, discusses empirical applications and performance, evaluates challenges, and surveys future directions in AI-enabled genomic analysis.

2. METHODS, DATA SOURCES, AND ANALYTICAL FRAMEWORKS

2.1 Types of Genomic Data and Typical Datasets

AI-based genomic analysis relies on diverse data sources:

1. Whole-genome sequencing (WGS) / Whole-exome sequencing (WES): Provides single nucleotide variants (SNVs), insertions/deletions (indels), structural variants, copy-number variants, enabling comprehensive variant discovery.
2. GWAS genotype data: Large cohorts with SNP arrays or imputed genotypes, often linked to phenotypic and clinical information.
3. Transcriptomics (e.g., RNA-seq), epigenomics (e.g., DNA methylation, chromatin accessibility), proteomics — used in multi-omics integration to capture regulatory and functional layers beyond static DNA variants.
4. Clinical / phenotypic data: From electronic health records (EHRs), biobank metadata, disease status, environmental exposures — enabling genotype–phenotype association studies, risk prediction models.

These data often require intensive preprocessing: quality control, imputation of missing genotypes, normalization (for expression data), encoding (e.g., genotype encoding as 0/1/2), and sometimes dimensionality reduction or feature selection (e.g., filtering variants, principal component analysis).

Large-scale public or semi-public datasets commonly used include UK Biobank (for European ancestry), disease-specific cohorts, and disease-control GWAS datasets.

2.2 Machine Learning and Deep Learning Methodologies

2.2.1 Traditional Machine Learning (ML) Approaches

Classical ML algorithms remain widely applied for genomic risk prediction, especially when interpretability or computational efficiency is desired, or sample sizes are modest. Typical methods include:

1. Random Forests (RF) and Gradient Boosting Machines (GBM / XGBoost): These can model non-linear relationships, interactions among variants, and work well with structured genotype features (e.g., summary scores, engineered features).

2. Support Vector Machines (SVMs): Employed for classification tasks, such as distinguishing cases vs controls, or pathogenic vs benign variants.
3. Feature selection methods (filter-based, wrapper-based, regularization like LASSO) are often paired with ML to reduce the dimensionality (from thousands–millions of variants) and avoid overfitting.

These methods provide balance between interpretability and performance, and are often used in polygenic risk scoring or variant prioritization.

2.2.2 Deep Learning (DL) Approaches

Deep learning has proven particularly effective for large genomic datasets, with architectures able to learn hierarchical, non-linear representations:

1. Multilayer Perceptrons (MLPs): Applied to SNP-based genotype arrays or processed variant features for phenotype classification or risk scoring. Some studies report high AUC values when training on large cohorts.
2. Convolutional Neural Networks (CNNs): Useful for sequence-level data (raw DNA sequence, regulatory region sequences), capturing spatial patterns (e.g., motifs, regulatory elements). For example, CNNs have been used to analyze NGS data for variant effect prediction and capture regulatory features (Alharbi & Rashid, 2022).
3. Autoencoders / representation learning: Used for dimensionality reduction and denoising of high-dimensional omics data (e.g., gene expression), before downstream classification or regression tasks.
4. Multi-omics integration models: DL architectures that ingest data from multiple omics layers (e.g., genomics, transcriptomics, epigenomics) to build comprehensive predictive models — especially valuable in cancer genomics or complex diseases where regulation and expression matter, not just genotype (Alharbi & Rashid, 2022).
5. Transformer-based architectures and attention mechanisms: Recent work proposes transformer-based frameworks to model long-range dependencies in genomic data (e.g., over entire chromosomes) and handle very large numbers of variants. A 2025 study introduced Ge-SAND — a self-attention deep learning

model that captures complex genetic interactions at scale for disease risk prediction.

These DL methods excel at capturing complex, high-order interactions, non-linear dependencies, and multi-layered regulatory effects.

2.3 Evaluation Metrics and Validation Strategies

To assess model performance in genomic risk prediction or classification tasks, common metrics include:

1. Receiver-Operating Characteristic — Area Under Curve (ROC-AUC): Widely used for binary disease / control prediction.
2. Precision, recall, F1-score, area under precision-recall curve (PR-AUC): Especially relevant for imbalanced datasets (rare diseases).
3. Calibration metrics: Evaluating whether predicted risk probabilities match observed outcomes (important for clinical translation).
4. Cross-validation (k-fold, stratified), external validation on independent cohorts, and, where possible, replication across different datasets/populations.
5. Explainability measures: For DL models, methods like attention scores, feature importance (e.g., SHAP values), or saliency maps are used to understand which variants/regions drive predictions — a growing necessity for biological interpretability and clinical trust.

3.EMPIRICAL APPLICATIONS AND PERFORMANCE OF AI IN GENOMIC RISK PREDICTION

3.1 Overview of AI Use in Genomics: Recent Trends

A comprehensive review by Wardah S. Alharbi & Mamoon Rashid (2022) summarized broad application of deep learning across human genomics, including variant calling, regulatory region prediction, gene expression inference, epigenetic state prediction, and disease classification. They highlighted that DL has been adopted in multiple subfields — but some areas (e.g., structural variant interpretation, long-read sequencing) remain under-charted.

Key takeaway: AI is no longer niche but mainstream in genomic research, and its adoption keeps growing rapidly.

3.2 Case Study: Ge-SAND — Modeling Complex Genetic Interactions

A breakthrough example is the 2025 study introducing Ge-SAND, which employs a self-attention deep learning framework to detect large-scale, high-order genotype interactions and predict disease risk (e.g., Crohn's disease, schizophrenia, Alzheimer's disease

1. Performance gains: On real-world datasets, Ge-SAND exhibited up to ~20% improvement in AUC-ROC compared to conventional methods.
2. Interpretability: Through attention scores, the model identifies SNP–SNP interaction pairs potentially contributing to disease risk, offering biologically plausible hypotheses rather than "black-box" predictions.
3. Scalability: By embedding genomic loci and using attention mechanisms, Ge-SAND can handle large genotype datasets — addressing the "curse of dimensionality" common in genomics.

This work illustrates the potential of transformer-derived architectures for whole-genome risk modeling, particularly when interactions beyond additive effects matter.

3.3 Multi-omics Integration and DL in Disease Subtyping & Prognosis

Beyond genotype-based risk scoring, DL has been widely applied in multi-omics integration — combining genomics with gene expression, methylation, proteomics — to improve disease subtype classification, prognosis prediction, and biomarker discovery. For instance, Alharbi & Rashid (2022) report multiple studies in cancer genomics where DL models ingest both sequence and expression data to stratify tumor subtypes, predict survival, and suggest therapeutic targets.

These integrative models often outperform single-omic or classical models as they capture multilayer biological regulation (genetic, epigenetic, transcriptional) — especially important in complex diseases where non-coding variants and regulatory changes play critical roles.

3.4 Polygenic Risk Scores (PRS) Enhanced by AI

Traditional PRS methods — summing risk alleles weighted by effect sizes from GWAS — assume additive effects and ignore interactions, epistasis, and non-linear dependencies. AI methods (ML or DL) can extend PRS by modeling non-additive effects, interactions, and including rare variants, thus improving predictive power, especially in diseases with complex genetic architecture (e.g., autoimmune disorders, neurodegenerative diseases, metabolic conditions).
Although comprehensive systematic reviews remain limited, mounting empirical evidence supports AI-informed PRS as a promising direction for personalized risk stratification.

4. CHALLENGES, LIMITATIONS, AND ETHICAL CONSIDERATIONS

Despite impressive advances, several limitations and challenges remain in applying AI to genomic data — particularly when translating to clinical or public health contexts.

4.1 High Dimensionality, Overfitting, and "Large p, Small n" Problem

Genomic datasets often include millions of variants (features) but relatively few samples ("p ≫ n"). Without careful regularization, dimensionality reduction, or sufficiently large sample sizes, ML/DL models risk severe overfitting. Feature selection, embedding strategies, and cross-validation are crucial, but even these may not fully mitigate overfitting for rare diseases or small cohorts.

4.2 Population Bias and Generalizability

Most large publicly available genomic datasets (e.g., UK Biobank) are skewed toward individuals of European ancestry. Models trained on such datasets may not generalize to other ancestries, limiting equity and applicability. Using AI may exacerbate these disparities if not carefully validated across diverse populations.

4.3 Interpretability and Biological Plausibility

Deep learning models — especially complex architectures — are often criticized as "black boxes." For adoption in clinical genomics, interpretability is critical:

researchers and clinicians need to know which variants, interactions, or regulatory features drive risk predictions, and whether these make biological sense. Without interpretability, AI predictions may lack trust and hinder translation.

Even models that offer some interpretability (e.g., via attention scores) must be validated carefully: attention does not always equal biological causality. Hypothesized variant interactions must ideally be tested via functional genomics or experimental validation.

4.4 Data Privacy, Sharing, and Ethical Concerns

Genomic data is sensitive. Aggregating, sharing, or centralizing raw genotype or sequence data raises privacy and consent issues. Moreover, individuals' risk predictions may impact insurance, employment, or psychological well-being.

Emerging solutions include federated learning (training models across decentralized datasets without sharing raw data), privacy-preserving architectures, and strict consent frameworks — but they remain under development, and regulatory/ethical frameworks often lag behind the technology.

4.5 Computational and Infrastructure Demands

Large-scale DL models (especially transformer-based) on whole-genome data require substantial computational resources (memory, GPUs), efficient data pipelines, and specialized expertise (bioinformatics + ML). This can limit applicability in resource-constrained settings or smaller research groups.

5. FUTURE DIRECTIONS & RECOMMENDATIONS

Based on current trends, literature, and existing gaps, we suggest several key research directions and best practices for future work:

5.1 Embrace Explainable AI (XAI) and Interpretability-First Models

Develop and employ models that prioritize interpretability — e.g., attention-based architectures, feature attribution methods (SHAP, Integrated Gradients), pathway-aware modeling — to ensure biological plausibility and clinical trust. Studies like Ge-SAND mark a promising direction.

5.2 Expand Diversity: Use Multi-ethnic / Global Cohorts

Construct and validate models on diverse population datasets to avoid ancestry bias, improve generalizability, and ensure equitable precision medicine. Encourage data sharing across consortia, with consent and ethical safeguards.

5.3 Adopt Federated Learning and Privacy-Preserving Frameworks

Use federated learning architectures or secure aggregation protocols to enable collaborative genomic modeling without raw data exchange, thus respecting privacy while leveraging large datasets.

5.4 Leverage Multi-omics Integration for Comprehensive Risk Modeling

Combine genomic, transcriptomic, epigenomic, proteomic, and clinical data to build richer predictive models that reflect multiple biological layers. Deep learning is particularly suited to integrate such heterogeneous data.

5.5 Explore Transformer / LLM-Based Genomic Models

Given the success of attention-based models in non-genomic domains, transformer architectures and large language models (LLMs) tailored for genome data represent an emerging paradigm. Recent studies propose using such models for variant effect prediction and risk scoring.

5.6 Rigorously Validate Models with Independent Cohorts & Functional Studies

Beyond statistical validation, candidate variant interactions or risk-predictive features identified by AI should be cross-validated in independent datasets and — when possible — subjected to functional validation (cellular assays, CRISPR screens) to confirm biological relevance.

6. CONCLUSION

AI algorithms — particularly deep learning and modern ML — have transformed genomic data analysis and disease risk prediction. They enable researchers to harness the complexity of high-dimensional genotype data, capture non-linear and high-order interactions, integrate multiple omics layers, and deliver predictive performance beyond classical statistical models.

Empirical successes (e.g., multi-omics cancer subtype classification, transformer-based risk models such as Ge-SAND) illustrate the potential for AI to underline precision medicine and personalized risk stratification.

However, significant challenges remain: overfitting, population bias, interpretability, computational demands, and ethical/privacy concerns. To realize the full potential of AI-driven genomics, future work must combine technical innovation (XAI, federated learning, multi-omics integration) with rigorous validation, inclusive datasets, and responsible ethical frameworks.

Overall, AI stands as a cornerstone technology for the future of genomics — but realizing its promise for human health requires careful, multidisciplinary, and ethically grounded efforts.

References

1. Alharbi, W. S., & Rashid, M. (2022). A review of deep learning applications in human genomics using next-generation sequencing data. Human Genomics, 16, 26.
2. Ge-SAND: an explainable deep learning-driven framework for disease risk prediction by uncovering complex genetic interactions in parallel. (2025). BMC Genomics, 26, 432.
3. Consens, M. E., Dufault, C., Wainberg, M., Forster, D., Karimzadeh, M., Goodarzi, H., Theis, F. J., Moses, A., & Wang, B. (2023). To Transformers and Beyond: Large Language Models for the Genome.
4. SunnatulloAmruloevich, G., & Rustambekovna, S. A. (2022). Dental status and diagnosis of children and adolescents suffering from various forms of connective tissue dysplasia. Galaxy International Interdisciplinary Research Journal, 10(11), 955-962.

5. Gafforov, S., Nazarov, U., & Khalimbetov, G. (2022). On the Pathogenesis of Periodontal Disease in Mineral Metabolism Disorders. Central Asian Journal of Medical and Natural Science, 3(2), 131-136.
6. Gafforov, S., Nazarov, U., & Khalimbetov, G. (2022). Diagnosis and treatment of chronic generalised periodontitis in connective tissue dysplasia pathologies.
7. Gafforov, S., Nazarov, U., Khalimbetov, G., & Yakubova, F. (2022). Centre for the Professional Development of Health Professionals under the Ministry of Health of the Republic of Uzbekistan. NeuroQuantology, 20(5), 1433-1443.
8. Базарбаев, М. И., & Сайфуллаева, Д. И. (2022). РахиĞ мов Б Т., Ж¸ раева З Р. Роль информационных техĞ нологий в медицине и биомедицинской инженеĞ рии в подготовке будущих специалистов в пеĞ риод цифровой трансформации в образовании. ТТА Ахборотномаси, 10(10), 8Ğ13.
9. Марасулов, А. Ф., Базарбаев, М. И., Сайфуллаева, Д. И., & Сафаров, У. К. (2018). Подход к обучению математике, информатике, информационным технологиям и их интеграции в медицинских вузах.
10. Bazarbaev, M. I., & Sayfullaeva, D. I. (2025). WHEN ALGORITHMS MEET ANATOMY: UZBEKISTAN'S MEDICAL EDUCATION IN THE AGE OF TECHNOLOGY. Central Asian Journal of Medicine, (4), 35-39.
11. Baxtiyorovna, E. D., Alisherovna, F. N., & Jurayeva, U. O. N. (2024). PROPERTIES OF ELECTRON AND NEUTRON THERAPY. Web of Medicine: Journal of Medicine, Practice and Nursing, 2(10), 137-141.
12. Fayziyeva, N. A. (2025). OLIY TA'LIMDA PEDAGOGIK TA'LIM-TARBIYANI TASHKIL ETISHNING AHAMIYATLARI VA ZAMONAVIY METODLARIDAN FOYDALANISH USULLARI.
13. Fayziyeva, N. (2025). THE EFFECT OF MAGNESIUM ON PREGNANT WOMEN. Web of Medicine: Journal of Medicine, Practice and Nursing, 3(5), 60-63.
14. Sobirova, D. R., Usmanov, R. D., Po'latov, X. X., Azizova, F. X., & Akbarova, M. N. (2023). QANDLI DABET KASALLIGIDA O 'PKA ENDOTELIYIDAGI GISTOLOGIK O 'ZGARISHLAR.

15. Собирова, Д. Р., Нуралиев, Н. А., Усманов, Р. Д., Азизова, Ф. Х., & Пулатов, Х. Х. (2023). СОЯ УНИНИНГ ОЗУҚАВИЙ ҚИЙМАТИ, МИКРОЭЛЕМЕНТЛАР ВА РАДИОНУКЛИДЛАР КЎРСАТГИЧЛАРИ (24-СОНЛИ). «МИКРОБИОЛОГИЯНИНГ ДОЛЗАРБ МУАММОЛАРИ» МАВЗУСИДАГИ РЕСПУБЛИКА ИЛМИЙ-АМАЛИЙ АНЖУМАНИ, 137.
16. Nishanov, D. A., Kh, P. K., Sobirova, D. R., & Matrasulov, R. S. (2023). MODERN DIAGNOSIS OF NEPHROBLASTOMA IN CHILDREN. Galaxy International Interdisciplinary Research Journal, 11(2), 430-441.
17. Закиров, А. У., Пулатов, Х. Х., & Исмалов, Д. Д. (2001). Изучение противовоспалительных свойств диклозана. Экспер. и клин. фарм, (5), 50-52.
18. Пулатов, Х. Х. (2022). Влияние экспериментального сахарного диабета на надпочечники: дис. Ўзбекистон, Самарқанд.
19. Adilbekova, D. B., Usmanov, R. D., Mirsharapov, U. M., & Mansurova, D. A. (2019). MORPHOLOGICAL STATE OF EARLY POSTNATAL FORMATION OF THE ORGANS OF THE GASTROINTESTINAL TRACT AND LIVER IN OFFSPRING BORN AND RAISED BY MOTHERS WITH CHRONIC TOXIC HEPATITIS. Central Asian Journal of Medicine, 4, 43-55.
20. Адилбекова, Д. Б., Хатамов, А. И., Мансурова, Д. А., & Пулатов, Х. Х. (2020). Морфологическое состояние сосудисто-тканевых структур желудка у потомства в условиях хронического токсического гепатита у матери. Морфология, 157(2-3), 10-11.
21. Шералиев, И. И., & Пулатова, Х. Х. (2017). Теорема Эссена для различно распределенных случайных величин. Научное знание современности, (3), 347-349.
22. Zakirov, A. U., KhKh, P., Ismatov, D. N., & Azizov, U. M. (2001). Anti-inflammatory effect of dichlotazole. Eksperimental'naia i Klinicheskaia Farmakologiia, 64(5), 50-52.

23. Nurmatova, F. B., Xuan, R., & Fazilova, L. A. (2024). The advantages of implementing digital technology in education. Innovations in Science and Technologies, 1(3), 192-195.
24. Фазилова, Л. А. Масофавий таълимни ташкил этишда онлайн маърузалардан фойдаланишнинг назарий-методологик ахдмияти. Замонавий тиббий олий таълим: муаммолар, хорижий тажриба, истикболлар" мавзусидаги VII укув-услубий анжуман туплами.-УДК, 004-37.
25. Каюмова, М. (2025). ЦИФРОВЫЕ ТЕХНОЛОГИИ В КЛИНИЧЕСКОЙ ПРАКТИКЕ: ПЕРСПЕКТИВЫ И ВЫЗОВЫ. Academic research in educational sciences, (Conference 1), 165-168.
26. Фазилова, Л. (2025). ТАЛАБАЛАРНИНГ АКТГА ОИД КОМПЕТЕНТЛИГИНИ РИВОЖЛАНТИРИШ. Academic research in educational sciences, (Conference 1), 96-101.
27. Lukmanovich, H. N., Olegovna, M. T., & Komilzhonovich, U. F. (2016). Densitometric study of degree of osteointegration of the dental implant “implant. Uz” in experimental conditions. European science review, (3-4), 244-245.
28. Khabilov, N. L., Mun, T. O., Usmonov, F. K., & Baybekov, I. M. (2015). The Study of Structural Changes in Bone Tissue of Alveolar Process of Jaws in Experimental Animals after Implantation of a New Construction of Dental Implant from Titanium Bt-1.00 Developed in Uzbekistan. European Medical, Health and Pharmaceutical Journal, 8(1).
29. Хабилов, Н., Мун, Т., & Усмонов, Ф. (2014). Конструкционные особенности дентального имплантата, разработанного в Узбекистане. Стоматология, 1(3-4 (57-58)), 53-58.
30. USMONOV, F., & INAGAKI, F. (2017). JAPAN AND WATER RESOURCES OF TAJIKISTAN: CONTRIBUTION, CHALLENGES, AND REALITIES. Central Asia & the Caucasus (14046091), 18(3).