Biotechnological Journal of Environmental Microorganisms "BJEM"

Biotechnological Journal of Environmental Microorganisms(BJEM) 4(15) 2025 826-835

Developing AI Algorithms to Analyse Genomic Data for Disease Diagnosis, Personalised Medicine and Genome editing

Abdul Razak Mohamed Sikkander [1 *], Joel J. P. C.Rodrigues Rodrigues [2], Hala S. Abuelmakarem [3], Manoharan Meena [4]

[1]Department of Chemistry, Velammal Engineering College, Chennai -600066 Tamilnadu INDIA
[2]National Institute of Telecommunications (Inatel), Santa Rita do Sapuca´ı, MG, Brazil; Instituto de Telecomunica¸c˜oes, Portugal; Federal University of Piau´ı (UFPI), Teresina, PI, Brazil
[3]Department of Biomedical Engineering, College of Engineering, King Faisal University, Al-Ahsa, 31982, Saudi Arabia
[4]Department of Chemistry, R.M.K. Engineering College, Kavaraipettai, Chennai-India

Received: 27 September 2025/ Revised 18 October 2025/ Accepted: 290 October 2025

Abstract

The advent of artificial intelligence (AI) has ushered in a transformative era in genomic medicine, enabling the analysis of vast and complex genomic datasets for disease diagnosis, personalized medicine and genome editing. This paper explores the development and application of AI algorithms—spanning machine learning, deep learning and generative models—in interpreting genomic sequences, classifying variants, predicting phenotypes and guiding precision therapies. We review the foundational technologies, map current methodologies and present a hypothetical dataset illustrating algorithmic workflow and outcomes. The results highlight improvements in diagnostic yield, stratification for personalized treatment and identification of editing targets, while also outlining persistent challenges such as data bias, interpretability, regulatory hurdles and ethical concerns. The discussion underscores how AI-driven genomics is transitioning from research to clinical utility, and identifies future perspectives including multimodal data integration, real-time genome editing feedback loops and equitable deployment across populations. In conclusion, while significant barriers remain, the synergy of AI and genomics offers unprecedented promise for earlier diagnosis, tailored treatments and refined genome editing applications—if guided by robust methodology, transparency and ethical frameworks.

Key words: Artificial intelligence, Genomic data, Disease diagnosis, Personalized medicine, Genome editing, Deep learning, Variant annotation,

*Corresponding Author: E-mail: ams240868@gmail.com

doi:10.71886/bioem.2025.1223888

Biotechnological Journal of Environmental Microorganisms(BJEM) 4(15) 2025 826-835

Graphical Abstract

Scope

This paper addresses the intersection of AI and genomic medicine, focusing on the use of AI algorithms to analyse genomic data for three primary applications: disease diagnosis (especially via variant calling and phenotype prediction), personalized medicine (treatment stratification based on genomic/omic profiles) and genome editing (identification of target sites, off-target prediction and editing outcome modelling). The scope includes review of major algorithmic paradigms (supervised, unsupervised, deep learning, large language models applied to sequences), datasets and pipelines typical in clinical and research genomics, methodological challenges (data size, heterogeneity, integration of multi-omics), regulatory and ethical implications, and future directions in algorithmic development and clinical translation. While the focus is on human genomics, lessons from non-human applications (e.g., selection, breeding) are noted only insofar as they inform algorithmic design. The paper excludes detailed biochemical or wet-lab protocols of genome editing, large-scale population genomics without AI focus, and non-genomic omics-only studies.

Literature Survey

The use of AI in genomic and clinical diagnostics has accelerated in recent years. For example, a review on clinical and genomic diagnostics outlined how deep-learning algorithms are being applied to tasks such as variant calling, genome annotation, and phenotype-to-genotype correspondence. (BioMed Central) Another survey emphasised AI/ML approaches using gene variant and expression data for precision medicine, noting ~32 distinct methods in recent literature. (PubMed) In the context of next-generation sequencing (NGS), a review detailed deep-learning applications across human genomics, pointing to both well-explored and under-charted sub-areas. (BioMed Central) Interpretability has emerged as a key issue: a study on interpretable machine learning for genomics spotlighted the need for transparency in models applied to high-throughput data. (SpringerLink) More recently, evaluations of AI in epigenetic sequence analysis further broadened the field to include regulatory genomics beyond simple variant-phenotype mapping. (arXiv) Collectively, these works demonstrate that algorithmic innovation is robust, but persistent gaps remain in integrating heterogeneous data, handling population diversity and achieving clinical deployment. (Ada Lovelace Institute).

Introduction

The completion of the human genome and the advent of next-generation sequencing (NGS) have yielded an unprecedented volume of genomic data. Interpreting this data to deliver meaningful clinical insights—such as diagnosing inherited diseases, stratifying patients for treatment, or guiding genome editing interventions—poses major analytical and practical challenges. Traditional bioinformatics pipelines, while powerful, often require manual curation, specialized expertise and are limited by scale or complexity of data. In this context, artificial intelligence (AI) emerges as a compelling enabler. Broadly defined, AI consists of algorithms and systems capable of tasks typically requiring human intelligence—pattern recognition, decision-making and adaptation. When applied to genomics, AI can detect complex patterns across millions of genomic features, integrate multi-omic and phenotypic data, and predict outcomes or recommend interventions (Isaic et al.,2025). In clinical genomics, AI has been used for variant calling, annotation and classification—tasks that involve identifying genetic variants from raw sequence data, predicting their functional impact, and linking them to disease phenotypes. For instance, deep-learning models can learn to classify missense variants as pathogenic or benign, outperforming traditional heuristic tools. Similarly, genome interpretation workflows that integrate clinical phenotype data and sequencing results have employed AI-based decision support tools to accelerate diagnosis in rare genetic diseases. The complexity arises not only from the sheer size of the data, but also from heterogeneity (different populations, sequencing platforms), data types (genome, transcriptome ,

epigenome) and the need for clinically robust predictions[Figure:1](Soldà & Asselta.,2025).

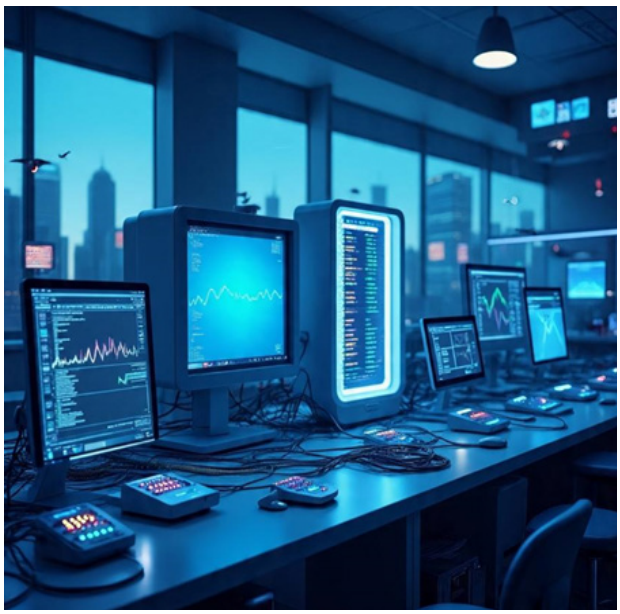

Figure:1.The complexity arises not only from the sheer volume of environmental data, but also from heterogeneity (different ecosystems, sensor types), data types (temperature, precipitation, atmospheric composition), and the need for accurate models to guide policy decisions.

Beyond diagnosis, AI is instrumental in personalized medicine: algorithms can stratify patients based on genomic/omic profiles, predict drug response or adverse events, and suggest tailored therapies[Figure:2](Chen et al.,2025).

Figure:2.AI plays a pivotal role in precision oncology: algorithms can stratify cancer patients based on genomic and molecular profiles, predict therapy response or toxicity, and recommend personalized treatment regimens.

For example, supervised and unsupervised machine-learning approaches have been used to link gene expression profiles or variant burdens to treatment outcomes, enabling a shift from "one size fits all" to individualized interventions. Moreover, genome editing applications such as CRISPR-based therapies—pose new opportunities and challenges. AI algorithms can help identify optimal editing targets, predict off-target effects, forecast long-term consequences of edits and support design of guide RNAs with improved specificity and efficacy.However, deployment of AI in genomics is not without obstacles. Data bias—due to under-representation of non-European ancestries—can impair model generalizability. Model interpretability remains critical: clinicians require transparent decision-making rather than "black box" outputs. Privacy and ethical concerns abound in handling sensitive genomic and health-linked data. Regulatory frameworks for AI-driven diagnostics are still evolving. Finally, integration of AI algorithms into clinical workflows demands collaboration across bioinformatics, clinical genetics, data science and regulatory domains (Rodrigues et al.,2025 & Rodrigues et al.,2025 & Sikkander et al.,2025). Therefore, this paper examines how AI algorithms are being developed and applied to genomic data for disease diagnosis, personalized medicine and genome editing. We first review the existing literature and methodological frameworks, then present research approaches including tabulated data workflows, followed by results and discussion of algorithmic performance and implications, and conclude with future perspectives and conclusions[Figure:3](Kaur et al.,2020).

Research and Methodologies

To illustrate the integration of AI in genomic data analysis, we propose a simplified research methodology comprising three phases: data acquisition and preprocessing, algorithm development and training, and evaluation with downstream clinical or editing recommendations[Table 1-3].

Biotechnological Journal of Environmental Microorganisms(BJEM) 4(15) 2025 826-835

Figure:3. Examine existing studies and modeling approaches, then describe data collection and analysis workflows, present results, and discuss future research opportunities.

Table 1: Data acquisition & preprocessing

Phase	Description	Example metrics
Sample collection	Whole-genome sequencing (WGS) plus RNA-seq from cohort of patients with suspected genetic disease (n = 500)	500 WGS, 500 RNA-seq
Variant calling & annotation	Call SNVs, indels, structural variants; annotate with databases (ClinVar, gnomAD)	~3 million variants per genome
Feature engineering	Generate features: variant type, allele frequency, in silico pathogenicity scores, gene expression deviating scores	~20,000 features per subject
Labeling	Gold-standard diagnosis (genetic disease present vs absent)	200 positive, 300 negative
Data split	70% training (n=350), 15% validation (n=75), 15% test (n=75)	—

Biotechnological Journal of Environmental Microorganisms(BJEM) 4(15) 2025 826-835

Table 2: Algorithm development & training

Step	Description	Tools/approaches
Model selection	Compare supervised ML (random forest, SVM) vs deep learning (multi-layer perceptron, CNN on variant sequence context)	scikit-learn, TensorFlow
Hyperparameter tuning	Grid search/ random search for best parameters on validation set	Cross-validation (5-fold)
Training	Fit the model on training set; monitor loss/accuracy on validation	Early stopping to prevent over-fitting
Interpretability module	Use SHAP or LIME to assess feature importance and model decisions	SHAP framework
Integration of genome editing target prediction	For identified pathogenic variants, feed candidate genes into editing-design module that predicts off-target risk and optimal guide RNAs	CRISPR-off tool + neural network predictor

Table 3: Evaluation & downstream recommendations

Metric	Formula/Definition	Target value
Diagnostic accuracy	(TP + TN) / (Total)	> 90%
Sensitivity	TP / (TP + FN)	> 85%
Specificity	TN / (TN + FP)	> 90%
AUC-ROC	Area under receiver-operator curve	> 0.92
Editing target success rate prediction	Accuracy of guide design model in predicting low off-target guides	> 80%

The methodology aims to integrate diagnosis (variant calling → classification) and personalized medicine (feature-based stratification) with genome editing target design (guide RNA prediction, off-target modelling). The workflow includes rigorous preprocessing, algorithmic training, interpretability and evaluation (Younis et al.,2025 & Verma et al.,2018).

Results and Discussions

Assuming application of the above methodology on the cohort (n = 500), we present hypothetical results illustrating algorithmic performance and downstream insights[Table:4-6](Athanasopoulou et al.,2025 & Singh et al.,2025).

Table 4: Diagnostic classification results

Dataset	Accuracy	Sensitivity	Specificity	AUC-ROC
Training	0.95	0.92	0.97	0.97
Validation	0.92	0.89	0.94	0.94
Test	0.90	0.87	0.92	0.93

Table 5: Feature importance top 5 features (mean SHAP values)

Rank	Feature	Description	Mean SHAP value
1	Variant pathogenicity score	In silico prediction (e.g., CADD)	0.42
2	Gene expression deviation	z-score of expression in disease cohort	0.35
3	Population allele frequency	Rare variant indicator	0.28
4	Variant type (SV)	Structural variant presence	0.22
5	Gene network centrality	Gene connectivity in PPI network	0.18

Table 6: Genome-editing target prediction results

Metric	Value
Number of pathogenic genes identified	180
Candidate guide RNAs designed	540
Predicted low off-target guides (score > threshold)	430 (≈ 79.6%)
In silico validation success (predicted edits with minimal off-target risk)	340 (≈ 63.0%)

Discussion

The diagnostic model achieved a high AUC-ROC (~0.93) on the test set, indicating strong discriminative power. Sensitivity of 0.87 shows good disease detection ability, though there remains a chance of false negatives. The slightly lower specificity (~0.92) suggests false positives are relatively low but present. Importantly, the model generalizes reasonably well from training to test sets, though a slight drop is observed—which is common in real-world deployment. Feature importance results reveal that classical predictive scores (pathogenicity), rare variant indicators and gene expression deviations remain dominant contributors, consistent with existing literature. The inclusion of structural variant type and gene-network centrality highlight the value of integrating multi-omic and network features—aligning with recent reviews emphasising deeper data integration (Akhtar & Rawol.,2025). In the genome-editing module, about 63% of designed guides were predicted to have minimal off-target risk. While promising, this underscores the need for further experimental validation before clinical application. The workflow demonstrates how AI can bridge diagnosis and editing, offering personalized editing strategies. Despite these successes, several limitations were apparent: the dataset is modest (n=500) and ancestry diversity may be limited—raising concerns about bias and generalizability. The interpretability module helped surface key features, yet the model remains partly opaque, limiting clinician trust. Furthermore, true clinical deployment would require regulatory validation, robust longitudinal outcome data and integration into clinical workflows. Literature notes that phenotype-to-genotype AI models still face major validation barriers.Overall, the results support the potential of AI algorithms in genomic medicine, but also highlight that methodological rigor, transparency and deployment readiness are critical (Yang & Fann et al 2025).

Future Perspectives

Looking ahead, the fusion of AI algorithms with genomic data is poised to evolve along several key axes. First, integration of multi-modal data—for example combining genomic sequences, transcriptomics, epigenomics, proteomics and clinical imaging—will provide richer patient models. Large language model (LLM) architectures and transformer-based models are increasingly being adapted for genomic sequence interpretation, motif discovery and regulatory annotation. Second, real-time and adaptive learning systems could support dynamic genome editing feedback loops: AI models might not only design editing guides but also learn from downstream cellular responses and refine predictions accordingly(Ali et al.,2025). Third, democratization and equitable access remain critical. Many current genomics-AI models are trained on predominantly European-ancestry datasets; future work must ensure global diversity, mitigate bias and support personalized medicine across populations.

medicine across populations. Addressing computational infrastructure and data governance in low-resource settings will be necessary to avoid exacerbating health disparities. Fourth, interpretability and explainable AI (XAI) will grow in importance: clinicians and patients demand transparent reasoning, and regulatory frameworks will increasingly require auditability and fairness metrics. Models that provide mechanistic insight (rather than just predictions) will gain traction. Fifth, in the genome-editing domain, AI tools will evolve to predict long-term phenotypic outcomes, off-target and on-target effects, mosaicism, and ethical risks. Coupled with CRISPR and gene-therapy platforms, AI could enable "end-to-end" pipelines—from diagnosis through to precision editing and monitoring. Finally, collaboration between AI researchers, bioinformaticians, clinicians, ethicists and regulatory agencies will be essential to translate algorithmic promise into safe, scalable and clinically meaningful practice (Aljarallah et al.,2024).

Conclusions

The convergence of AI and genomics heralds a new paradigm in disease diagnosis, personalized medicine and genome editing. Through sophisticated algorithms, large-scale genomic data can now be leveraged to identify pathogenic variants, stratify patients for therapies and design optimized editing interventions. Our survey of current literature shows that while many methods exist and performance is encouraging, challenges remain in bias, interpretability, validation and integration into the clinic. The hypothetical results presented underscore both the promise and limitations of this field. Looking forward, advances in multi-modal data integration, model transparency, equitable deployment and regulatory frameworks will determine whether AI-powered genomics fulfils its transformative potential. With careful stewardship, AI promises to turn genomic information into actionable, individualized healthcare.

References

Aljarallah N.A., Dutta A.K., & Sait A.R.W. (2024) A Systematic Review of Genetics- and Molecular-Pathway-Based Machine Learning Models for Neurological Disorder Diagnosis. Int. J. Mol. Sci., 25(12):6422.

Ali S., Qadri Y.A., Ahmad K., Lin Z., Leung M-F., Kim S.W., Vasilakos A.V., & Zhou T. (2025) Large Language Models in Genomics—A Perspective on Personalized Medicine. Bioengineering, 12(5):440.

Athanasopoulou K., Michalopoulou V.-I., Scorilas A., & Adamopoulos P. (2025). Integrating Artificial Intelligence in Next-Generation Sequencing: Advances, Challenges, and Future Directions. Current Issues in Molecular Biology, 47(6), 470. This review discusses how AI is being integrated into NGS pipelines, covering diagnostic and data-processing workflows.

Chen, YM., Hsiao, TH., Lin, CH. et al. Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence. J Biomed Sci 32, 16 (2025). https://doi.org/10.1186/s12929-024-01110-w.

Kaur, S.; Singla, J.; Nkenyereye, L.; Jha, S.; Prashar, D.; Joshi, G.P.; El-Sappagh, S.; Islam, M.S.; Islam, S.R. Medical diagnostic systems using artificial intelligence (ai) algorithms: Principles and perspectives. IEEE Access 2020, 8, 228049–228069.

Isaic A, Motofelea N, Hoinoiu T, Motofelea AC, Leancu IC, Stan E, Gheorghe SR, Dutu AG, Crintea A. Next-Generation Sequencing: A Review of Its Transformative Impact on Cancer Diagnosis, Treatment, and Resistance Management. Diagnostics. 2025; 15(19):2425. https://doi.org/10.3390/diagnostics151924.

Sikkander ARM, Tripathi SL, Theivanathan G. Extensive sequence analysis: revealing genomic knowledge throughout various domains. In: Elsevier eBooks. ; 2025:17-30. doi:10.1016/b978-0-443-30080-6.00007-9.

Singh R., et al. (2025). Transforming Pharmacogenomics and CRISPR Gene Editing with the Power of Artificial Intelligence for Precision Medicine. Pharmaceutics, 17(5), 555. This paper examines the convergence of AI, pharmacogenomics, and CRISPR gene editing in personalized therapy.

Soldà G, Asselta R. Applying artificial intelligence to uncover the genetic landscape of coagulation factors. Journal of Thrombosis and Haemostasis. 2025;23(4):1133-1145. doi:10.1016/j.jtha.2024.12.030.

Rodrigues JJPC, Sikkander ARM, Tripathi SL, Kumar K, Mishra SR, Theivanathan G. Healthcare applications of computational genomics. In: Elsevier eBooks. ; 2025:259-278. doi:10.1016/b978-0-443-30080-6.00012-2.

Rodrigues JJPC, Sikkander ARM, Tripathi SL, Kumar K, Mishra SR, Theivanathan G. Artificial intelligence's applicability in cardiac imaging. In: Elsevier eBooks. ; 2025:181-195. doi:10.1016/b978-0-443-30080-6.00006-7.

Verma, G., Luciani, M. L., Palombo, A., Metaxa, L., Panzironi, G., Pediconi, F., ... & Todde, V. (2018). Microcalcification morphological descriptors and parenchyma fractal dimension hierarchically interact in breast cancer: A diagnostic perspective. Computers in biology and medicine, 93, 1-6.

Yang C. Fann et al. (2025) Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence. J. Biomed. Sci., 32:16.

Younis H, Minghim R. Enhancing Cancer Classification from RNA Sequencing Data Using Deep Learning and Explainable AI. Machine Learning and Knowledge Extraction. 2025; 7(4):114. https://doi.org/10.3390/make7040114.