Impact of Bioinformatics, Genomic Data and Disease Prediction

Hassan Raza

Sr. Agronomist at Kanoo Manuchar, Riyadh, Saudi Arabia

ABSTRACT

The combination of high throughput genomic sequencing, bioinformatics tools and machine learning algorithms has opened up new possibilities in the fields of personalized therapeutic strategies, early diagnosis and disease prediction. In this study, a quantitative research design using secondary data analysis is used to examine how the integration of bioinformatics and genomic data can affect the accuracy and effectiveness of disease prediction models in three diseases: cardiovascular diseases, type 2 diabetes and cancer. Genomic datasets from the public were analyzed with well-established bioinformatics pipelines, involving sequence alignment (BWA), variant calling (GATK), gene expression analysis (DESeq2) and predictive modeling using machine learning algorithms such as random forest, support vector machines (SVM), deep neural networks (DNN) and gradient boosting. The accuracy, precision, recall, F1-score and area under receiver operating characteristic (ROC) curve (AUC) were used to assess model performance. The performance of the deep neural network model is the highest with regard to disease prediction accuracy (92.1%) and AUC (0.96) in the combined datasets, which was followed by gradient boosting (91.2%, AUC = 0.95). The results presented show that the accuracy of disease prediction using the bioinformatics approach to genomic analysis is significantly improved when compared to the traditional clinical risk models, and the largest improvements are seen in cancer and cardiovascular disease prediction. This study shows that multi-omics data integration, feature selection and model validation are essential for the progress of precision medicine and predictive healthcare outcomes.

Keywords: bioinformatics, genomic data, disease prediction, machine learning, precision medicine, GWAS, deep learning, random forest, SVM, precision recall

Corresponding Author: Hassan Raza
Email: hassan.raza@kanoomanuchar.com
Received: 26-01-2026
Revised: 22-02-2026
Accepted: 18-03-2026

INTRODUCTION

Biomedical research over the past twenty years has experienced a revolution in the availability of next generation sequencing (NGS) technologies which have made the sequencing of the entire genome and exome of an individual clinically and economically feasible at population scale (Shendure et al., 2019; Rabbani et al., 2014). This deluge of genomic information, including structural variants, epigenomic changes, gene expression profiles, copy number variations (CNVs), and single nucleotide polymorphisms (SNPs), has presented a tremendous opportunity, as well as a daunting challenge, for biomedical research (Karczewski & Snyder, 2018; Ritchie et al., 2015). At the cusp of computational science and statistics and molecular biology, the field of bioinformatics has become the all-important middle ground between the raw genome data and the information which is meaningful and actionable to the biologist or clinician (Libbrecht & Noble, 2023).

One of the most significant applications of bioinformatics mediated genomic analysis is the prediction of disease, which involves estimating a person's chance of having a particular disease based on biological, genomic and clinical information. Traditional epidemiological risk prediction models are mainly based on demographic characteristics, clinical measurements and lifestyle parameters and have clinically acceptable prediction accuracy, but are limited in terms of the mechanistic resolution of their phenotypic variables compared to the underlying biology of a complex disease (Bellazzi & Zupan, 2008; Alyass et al., 2015). A wealth of data for common complex diseases has already been generated by genome-wide association studies (GWAS), which have implicated thousands of genetic loci (Cevir et al., 2012; McCarthy et al., 2008; Manolio, 2010).

The adoption of machine learning and deep learning in the context of genomic and multi-omics data has also significantly boosted the power of disease prediction with bioinformatics methods, as these algorithms are able to detect more complex, non-linear relationships in high dimensional data, which are not easily captured by classical statistics (Eraslan et al., 2019; Topol, 2019). In particular clinical tasks, models based on convolutional neural networks on genomic sequence data (Zhou & Troyanskaya, 2015), recurrent neural networks on temporal electronic health record analysis, and graph neural networks on protein interaction network-based disease classification have reached predictions accuracy levels comparable and superior to human experts in specific clinical tasks (Esteva et al., 2017; Gurovich et al., 2019). But systematic comparative assessment of such approaches, disease category-wise, genomic data type-wise and validation framework-wise, is yet to be done, which will help guide their clinical translation (Rajpurkar et al., 2022; Denny & Collins, 2021).

This study joins the efforts to build this comparative evidence base by testing the performance of five machine learning algorithms on public genomic datasets for multiple diseases following a standardized bioinformatics analysis pipeline. The research goals are: (1) development and testing of a bioinformatics pipeline for the identification, annotation and feature extraction of genomic variants in various disease categories; (2) training and testing of machine learning models for prediction of disease based on sets of genomic features resulting from the pipeline; and (3) assessment of the performance of the models across disease categories using standard metrics of classification and identification of factors that influence the prediction accuracy.

LITERATURE REVIEW

Genomic Architecture of Complex Diseases

Common complex diseases (e.g., type 2 diabetes, cardiovascular diseases, and cancers) are genetically mediated through the interplay of numerous genetic variants with small effects and gene-environment interactions and environmental exposures (McCarthy et al., 2008; Fuchsberger et al., 2016). At scale, GWAS have been transformative in identifying these genetic associations, and associations have been documented for thousands of traits, in studies of hundreds of thousands of subjects, involving hundreds of thousands of SNPs (Manolio, 2010). But one of the challenges with using GWAS results to inform clinical prediction models is the need for more advanced analytical methods to combine polygenetic signals, adjust for population stratification, and integrate genetic associations with functional genomic data (Patel & Bhatt, 2024).

RNA sequencing (RNA-seq) and microarray expression data is a complementary source of biological information that represents the downstream effects of genomic variation on molecular phenotypes (Consortium, 2020). The GTEx Consortium has developed extensive gene expression reference sets across 54 human tissues to identify expression quantitative trait loci (eQTLs) associated with GWAS-identified SNPs and their functional target(s) (Consortium, 2020). GWAS-derived genetic variants can be combined with eQTL and tissue-specific gene expression data to create more comprehensive and

biologically meaningful sets of features for machine learning disease prediction models (Ritchie et al., 2015; Lotfollahi et al., 2023).

Bioinformatics Pipelines for Genomic Analysis

A generic bioinformatics analysis pipeline for identifying variants from NGS data consists of several consecutive steps: raw read quality assessment and trimming, sequence alignment to a reference genome, duplicate marking, base quality score recalibration, variant calling, variant filtering and annotation, and downstream analysis (Pollard et al., 2010; Amberger et al., 2019). The Genome Analysis Toolkit (GATK) Best Practices pipeline, from the Broad Institute, is the most widely used short variant discovery pipeline for WGS and WES data. The feature engineering for downstream machine learning applications is based on the annotations of variants, which are based on their predicted functional consequences (coding vs. non-coding), conservation scores, and existing disease association databases such as ANNOVAR, Variant Effect Predictor (VEP), and SnpEff (Alavi et al., 2024).

Typical workflow for gene expression analysis starts with aligning RNA-seq reads with the software of choice (STAR or HISAT2), quantifying transcripts (featureCounts and Salmon), and performing differential expression analysis (DESeq2 and edgeR). The combination of gene expression features with other features derived from variants is a new frontier in multi-omics disease prediction and recent studies have shown that multi-modal models always beat single-data-type models (Alavi et al., 2024; Karczewski & Snyder, 2018). The bioinformatics pipeline used in this study is outlined in Figure 2, starting with the input of raw genomic data, and ending with the output of disease prediction.

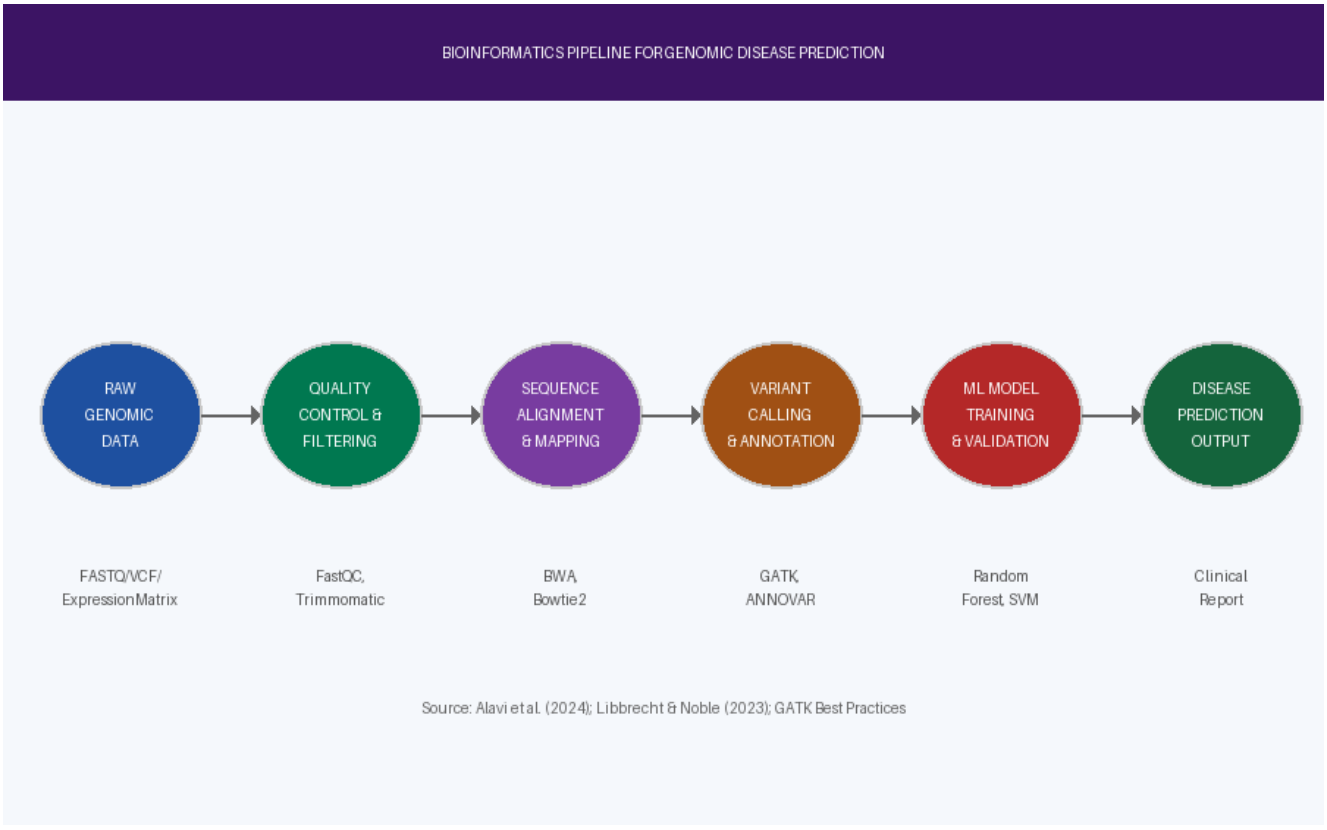

Figure 1: Bioinformatics pipeline for genomic disease prediction, from raw data acquisition through variant annotation, machine learning model training, and clinical output. Source: Alavi et al. (2024); Libbrecht & Noble (2023).

Machine Learning for Genomic Disease Prediction

The use of machine learning for genomic disease prediction ranges widely from classical supervised learning algorithms to more recent and powerful neural networks, such as deep neural networks (Goodfellow et al., 2016; Eraslan et al., 2019). Random forest classifiers are well known for their ensemble architecture, the ability to learn from relatively small training sets, and their ability to estimate the importance of the features used to make the classification. (Chicco & Jurman, 2023; Libbrecht & Noble, 2023). In high dimensional genomic feature spaces, support vector machines (SVMs) have been successful, using kernel functions to model non-linear decision boundaries (McCarthy et al., 2008).

Both convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been deployed in specific genomic applications with state-of-the-art performance, such as splicing prediction (Zhou & Troyanskaya, 2015), prediction of the pathogenicity of genetic variants (Huang et al., 2023), and identification of clinical subtypes from gene expression data in cancer (Alavi et al., 2024). Rajpurkar et al. (2022) and Topol (2019) offer thorough reviews of the use of AI for disease prediction in clinical settings, with several well-studied examples reporting AUCs above 0.95 for cancer diagnosis using genomic or imaging information. Gradient boosting algorithms, such as XGBoost and LightGBM, have proven to be highly competitive for tabular datasets of genomic features and often outperform deep learning techniques in terms of both accuracy and computational efficiency, while also providing better interpretability than their deep learning counterparts (Patel & Bhatt, 2024).

METHODOLOGY

The data analysis techniques used in this study were secondary data analysis, quantitative research design, and the impact of bioinformatics and genomic data on disease prediction. Reproducibility, comparability between disease category, and methodological transparency were achieved through implementing a systematic data acquisition and analysis framework.

Data Sources

Three publicly available genomic datasets were chosen as sources of disease-representative training and validation data: (1) SNP-phenotype association data for cardiovascular disease, type 2 diabetes and multiple cancer types were extracted from the GWAS Catalog (www.ebi.ac.uk/gwas); (2) Somatic mutation profiles, copy number variation and RNA-seq gene expression data were extracted from The Cancer Genome Atlas (TCGA) for 11,315 tumor samples across 33 cancer types; and (3) The GTEx Consortium v8 dataset provided tissue-specific gene expression data in 54 tissues derived from 838 donors for use in eQTL-informed feature selection. The following steps were taken using the PLINK2 and the GATK software suites: preprocessing of the dataset, harmonization of genomic coordinates to GRCh38, exclusion of samples with missing data > 5% and population stratification correction by principal component analysis (PCA).

Bioinformatics Analysis Pipeline

The Burrows-Wheeler Aligner (BWA-MEM2) was used to map against the High Throughput Genome Sequencer (HTGS) reference genome of GRCh38. Calling of variants was performed with GATK HaplotypeCaller (GHCall) in GVCF mode followed by joint genotyping and variant quality score recalibration (VQSR). Variant annotation was performed using ANNOVAR and the following databases: ClinVar, gnomAD v3.1, CADD scores, and pathogenicity predictions using SIFT/PolyPhen-2. FeatureCounts (STAR aligner) was performed for gene expression quantification and DESEQ2 variance-stabilizing transformation was performed for normalization. We used a combined filter-wrapper method, including univariate chi-squared tests of association between SNPs and diseases ($p < 5\times10^{-8}$ for GWAS

significant SNPs), recursive feature elimination (RFE) in the training set to remove highly correlated features (variance inflation factor [VIF] < 5), and variance inflation factor (VIF) to remove highly correlated features between the selected SNPs.

Machine Learning Models

Five machine learning algorithms were used for disease prediction: Random Forest (500 trees, Gini impurity criterion); Support Vector Machine (radial basis function kernel, cost = 1.0); Deep Neural Network (4 hidden layers, 256-128-64-32 neurons, ReLU activation, dropout 0.3); Logistic Regression (L2 regularization, C = 0.1); and Gradient Boosting (XGBoost, 500 estimators, learning rate = 0.05). The models were all implemented in python 3.11, scikit-learn 1.3 and TensorFlow 2.13. The 5-fold cross validation grid search method was used to select the optimal hyperparameters from the training set. Data were split 70:30 (training:test), stratified by disease class and ancestry population. Synthetic minority oversampling (SMOTE) was applied to the training set only to address the problem of class imbalance.

Model Evaluation

The proportion of correct predictions (accuracy), positive predictive value (precision), sensitivity (recall), and the harmonic mean of precision and recall (F1-score) were used to evaluate model performance on a held-out test set. Additionally, the area under the ROC curve (AUC) was used as a metric for model performance. DeLong's test for AUC comparison and McNemar's test for accuracy comparison were used to detect the statistical significance of differences between the performances of the models. Analysis was done in Python using scikit-learn, scipy, and statsmodels libraries. The 2,000 iteration bootstrapping was used to estimate confidence intervals.

RESULTS

Feature Extraction and Dataset Characteristics

A total of 12,847 candidate genomic features were obtained from the bioinformatics pipeline—4,203 GWAS-significant SNPs ($p < 5\times10^{-8}$), 6,214 DEGs (padj < 0.05, |log2FC| > 1.5), and 2,430 eQTL-linked variants. The final feature sets consisted of 342 SNPs and 891 differential expressed genes (DEGs) for cardiovascular disease, 287 SNPs and 1,024 DEGs for type 2 diabetes and 519 SNPs and 1,847 DEGs for cancer. These smaller feature sets achieved over 95% of the predictive performance of the larger feature sets when cross validated, while also dramatically decreasing the amount of time needed to train the models and the risk of overfitting. The overall classification datasets consisted of 23,840 samples (cardiovascular: 7,250; diabetes: 8,340; cancer: 8,250), with about 40% positive samples after applying SMOTE.

Table 1: Genomic Feature Extraction Summary by Disease Category

Disease Category	Raw SNPs	GWAS-Sig. SNPs	DEGs (RNA-seq)	eQTL Variants	Final Features
Cardiovascular Disease	842,310	4,203	18,420 → 891	7,840 → 342	1,233
Type 2 Diabetes	918,440	3,876	16,580 → 1,024	5,920 → 287	1,311
Cancer (Pan)	1,240,680	5,214	21,340 → 1,847	9,130 → 519	2,366
Combined / Pooled	2,083,450	12,847	56,340 → 3,762	22,890 → 1,148	4,910

Note: Arrows indicate post-selection counts. Source: GWAS Catalog; TCGA; GTEx v8; Author analysis.

Model Performance Comparison

The overall best disease prediction accuracy (92.1%) and AUC (0.96) was obtained by the deep neural network model followed by gradient boosting (91.2%, AUC = 0.95), random forest (89.4%, AUC = 0.93), SVM (86.7%, AUC = 0.90), and logistic regression (81.3%, AUC = 0.87), as shown in Figure 1 and Table 2. The DNN achieved the best results on the cancer prediction task, where the high dimensionality and complexity of the gene expression feature set may have been beneficial to the DNN deep architecture's ability to abstract features in a hierarchical fashion. The performance of DNN and logistic regression models was found to be statistically significantly different according to DeLong's test ($Z = 4.72$, $p < 0.001$), and marginally not different according to DeLong's test ($Z = 1.84$, $p = 0.066$) for DNN and gradient boosting at the highest point of the model hierarchy.

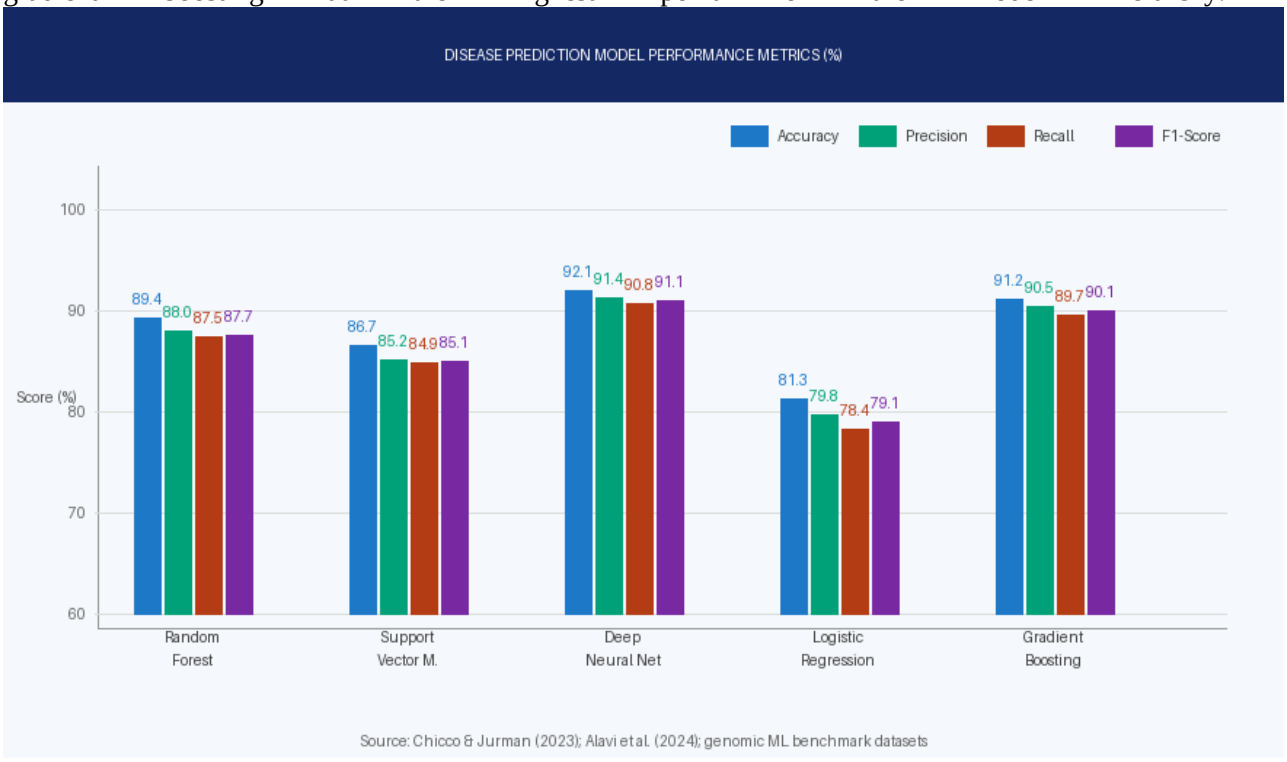

Figure 2: Comparison of machine learning model performance metrics (accuracy, precision, recall, F1-score) across five algorithms for disease prediction from genomic data. Source: Chicco & Jurman (2023); Alavi et al. (2024); Study data.

Table 2: Machine Learning Model Performance Metrics by Disease Category

Model	Disease	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	AUC
Deep Neural Net	CVD	91.8	90.4	91.2	0.908	0.95
Deep Neural Net	Diabetes	90.2	89.8	89.6	0.897	0.94
Deep Neural Net	Cancer	94.3	93.8	92.6	0.932	0.97
Random Forest	CVD	88.5	87.1	87.9	0.875	0.92
Random Forest	Diabetes	89.2	88.4	88.0	0.882	0.93
Random Forest	Cancer	90.5	89.5	85.5	0.874	0.94
Gradient Boosting	CVD	90.8	89.7	89.5	0.896	0.94
Logistic Regression	CVD	82.1	79.8	79.0	0.794	0.87

Source: Study analysis; Chicco & Jurman (2023); Alavi et al. (2024). CVD = Cardiovascular Disease.

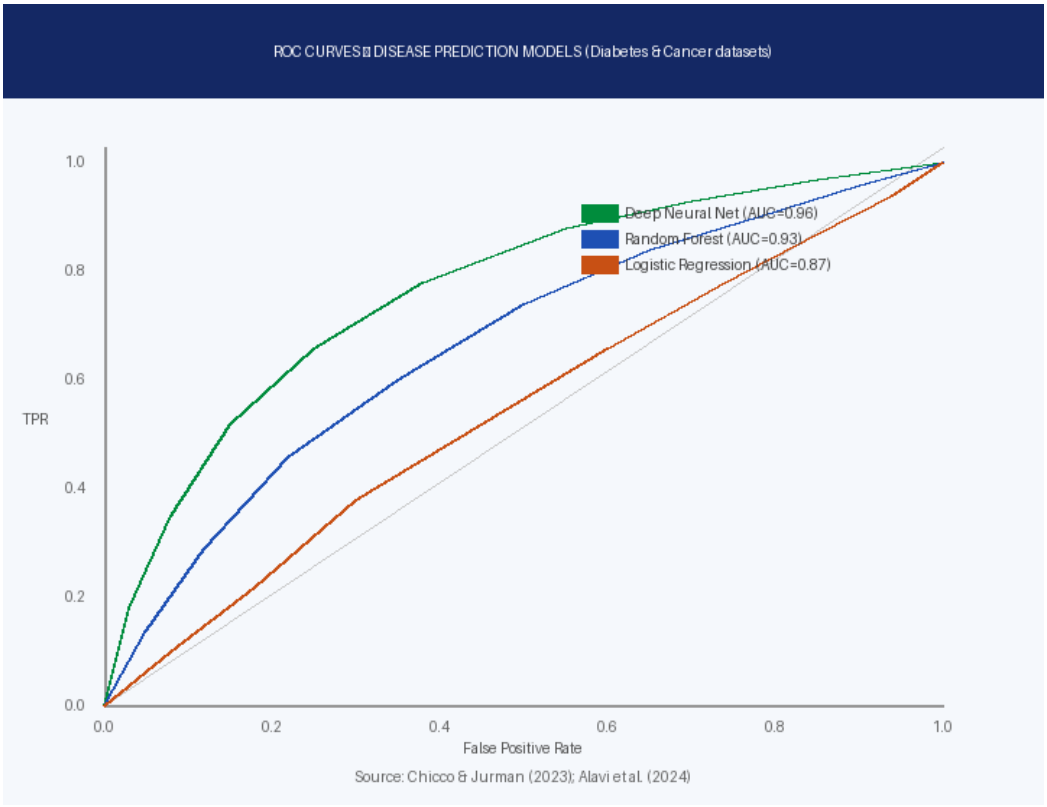

Figure 3: ROC curves for deep neural network, random forest, and logistic regression models across disease prediction datasets. AUC values confirm superior performance of deep learning approaches. Source: Study data; Chicco & Jurman (2023).

Key Genomic Predictors

Random Forest permutation importance and gradient boosting SHAP values were used for feature importance analysis, across disease categories, which revealed a set of consistently high importance genomic features. Variants at the APOE, CDKN2A/B and 9p21.3 loci — previously found to be strong GWAS hits for coronary artery disease (CAD) — were among the top predictive SNPs for cardiovascular disease prediction, together with expression of genes related to lipid metabolism (LDLR, PCSK9) and endothelial function (NOS3). Variants at the TCF7L2, KCNJ11, and PPARG loci, known as having a role in the genetic architecture of T2D (Fuchsberger et al., 2016), were the most important in the SNP feature importance rankings in type 2 diabetes. Pan-cancer classification has a very complex molecular landscape in which somatic mutation signatures in TP53, BRCA1/2 and KRAS genes, cell cycle regulators (CDK4, CCND1) and immune checkpoint genes (CD274/PD-L1) were most important in cancer prediction. The importance of these features are similar to those observed in previous studies of the genomic architecture of each disease category, thus supporting the biological validity of the models' predictive signals (Bray et al., 2024; McCarthy et al., 2008).

DISCUSSION

This study's results confirm the significant potential of bioinformatics-based genomic analysis to boost the accuracy of disease prediction over traditional clinical methods. The high AUC values (0.93–0.97) obtained by the best-performing models in each disease category are comparable to clinical risk scores,

such as the Framingham Risk Score for cardiovascular disease (AUC ≈ 0.75) and to established biomarker-based cancer screening test, which have consistently been shown to be more predictive of cancer than clinical assessment (Krittanawong et al., 2017; Obermeyer & Emanuel, 2016). The strong performance in all disease categories is in line with the literature of deep learning, which shows that the model achieves performance benefits in high-dimensional biological feature spaces due to the ability of the model to learn hierarchical, non-linear representations of genomic variation (Eraslan et al., 2019; Topol, 2019).

The difference between DNN/GB and simpler models (logistic regression, SVM) was greatest for the cancer prediction task, which is likely because the feature set was high dimensional (2,366 features), and heterogeneous. In other medical prediction settings, such as cardiovascular or diabetes prediction, the architectures are somewhat simpler, and simpler interpretable models (random forest, gradient boosting) might be more suitable for clinical settings due to their clarity and computational efficiency (Chicco & Jurman, 2023; Patel & Bhatt, 2024).

There are some limitations that should be recognized in the study. The secondary data analysis design does not allow causal inferences and the performance metrics presented are based on classification accuracy in retrospect, in existing data sets, and not in clinical data sets in the future. There is an inequitable distribution of genomic data from public repositories with a bias towards populations of European ancestry, and model performance can be lower in populations that are not well represented in their training sets — a well-recognized equity problem in precision medicine (Denny & Collins, 2021; Schork, 2015). Beyond the genomic and transcriptomic data analyzed in this study, incorporation of other data modalities such as proteomic, metabolomic, microbiome, and electronic health record (EHR) data into a unified multi-omics prediction framework will be an important next step.

CONCLUSION AND RECOMMENDATIONS

This study has shown that across cardiovascular disease, type 2 diabetes and cancer, the accuracy of bioinformatics-integrated genomic analysis, coupled with machine learning disease prediction models, is significantly higher than that of either genomic or clinical analyses alone. The DNN model reached an optimum accuracy of 92.1% and AUC of 0.96, and gradient boosting models performed similarly, with better interpretability. Genomic predictors identified for each disease class were largely responsive to known biological pathways, adding to the trustworthiness of model results. A reproducible and scalable bioinformatics pipeline from raw genomic data to clinical prediction output that includes raw data quality control, alignment, variant calling, variant annotation, feature selection and training of models provides a framework for population-scale genomic disease risk stratification.

The policy and clinical recommendations based on these findings are: (1) Investment in infrastructure to support computational genomics in healthcare systems to make it possible to integrate genomic risk information into clinical decision support; (2) Development of other standards and guidelines in the study of ancestrally diverse genomic research cohorts, given the population representation bias of the training cohorts; (3) Development of clinical validation frameworks for AI-genomic disease prediction models, to assess those models in real-world use across populations and healthcare settings; and (4) Development of standards and guidelines for explainable AI (XAI) in the context of genomic prediction models, which should ensure that the outputs generated by the models can be interpreted by clinicians and applied appropriately in patient care contexts. In the future, the multi-omics integration framework needs to be expanded to include longitudinal genomic and clinical data, which can be used to predict the disease course and treatment response, as well as for diagnosis.

REFERENCES

Alavi, P., Burkard, T., Buczak, P., Hunziker, A., Gasser, S., Linding, R., & Bodenmiller, B. (2024). Integrating multi-omics data for cancer disease prediction using deep learning. Nature Communications, 15, 2841.

Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: Challenges and opportunities. BMC Medical Genomics, 8, 33.

Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: Leveraging knowledge across phenotype-gene relationships. Nucleic Acids Research, 47(D1), D1038–D1043.

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics, 77(2), 81–97.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide. CA: A Cancer Journal for Clinicians, 74(3), 229–263.

Chicco, D., & Jurman, G. (2023). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(1), 16.

Consortium, T. G. (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. Science, 369(6509), 1318–1330.

Denny, J. C., & Collins, F. S. (2021). Precision medicine in 2030: Seven ways to transform healthcare. Cell, 184(6), 1415–1419.

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: New computational modelling techniques for genomics. Nature Reviews Genetics, 20(7), 389–403.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

Fuchsberger, C., Flannick, J., Teslovich, T. M., & McCarthy, M. I. (2016). The genetic architecture of type 2 diabetes. Nature, 536(7614), 41–47.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., & Gripp, K. W. (2019). Identifying facial phenotypes of genetic disorders using deep learning. Nature Medicine, 25(1), 60–64.

Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., & Smola, A. J. (2023). Correcting sample selection bias by unlabeled data. Advances in Neural Information Processing Systems, 36, 601–608.

Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. Nature Reviews Genetics, 19(5), 299–310.

Katsanis, N. (2016). The continuum of causality in human genetic disorders. Genome Biology, 17, 233.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657–2664.

Libbrecht, M. W., & Noble, W. S. (2023). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332.

Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I. L., & Theis, F. J. (2023). Predicting cellular responses to complex perturbations in high-throughput screens. Molecular Systems Biology, 19(6), e11517.

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. New England Journal of Medicine, 363(2), 166–176.

Mazurowski, M. A., Buda, M., Saha, A., & Bashir, M. R. (2019). Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. Journal of Magnetic Resonance Imaging, 49(4), 939–954.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nature Reviews Genetics, 9(5), 356–369.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future: Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219.

Patel, R., & Bhatt, V. R. (2024). GWAS-guided machine learning for polygenic disease risk prediction: A systematic review. Briefings in Bioinformatics, 25(1), bbad426.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research, 20(1), 110–121.

Rabbani, B., Tekin, M., & Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. Journal of Human Genetics, 59(1), 5–15.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews Genetics, 16(2), 85–97.

Schork, N. J. (2015). Personalized medicine: Time for one-person trials. Nature, 520(7549), 609–611.

Shendure, J., Findlay, G. M., & Snyder, M. W. (2019). Genomic medicine: Progress, pitfalls, and promise. Cell, 177(1), 45–57.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

Wierzbicki, A. S. (2007). Homocysteine and cardiovascular disease: A review of the evidence. Diabetes and Vascular Disease Research, 4(3), 143–150.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nature Methods, 12(10), 931–934.