International Journal of Scientific Research in Artificial Intelligence and Machine Learning (ISCSITR-IJSRAIML)
Vol.6, Iss. 3, May-June 2025, pp. 42-54.
https://iscsitr.com/journals/ISCSITR-IJSRAIML/archive
ISSN (Online): 3067-753X; Journal ID: 1668-1622
DOI: http://www.doi.org/10.63397/ISCSITR-IJSRAIML_2025_06_03_004

AI-Enhanced GWAS Data Analysis by Leverage AI and machine learning models

Prashanth Reddy Areddy

Senior Manager, Principal Data Acquisition Programmer, Bayer US LLC, Aubrey, Texas, USA.

Abstract

An attempt was made to develop artificial intelligence (AI), machine learning (ML) integration within the genome-wide association studies (GWAS) to enhance the detection and understanding of genetic variations associated with complex disorders. Although sophisticated machine learning models, such as neural networks and AutoML frameworks are used, these application shows enhancement on prediction accuracy, variant prioritisation and disease risk assessment. Key findings are the discovery of new intronic and non-coding variants surpassing standard GWAS methods and that these regulatory variants are of major importance. AI-enhanced genomics is also our research's contribution to the promise of precision medicine and has the full computational framework for variation identification, functional interpretation, and translation.

Keywords: GWAS, AI, Machine Learning, Data.

How to cite this paper: Prashanth Reddy Areddy. (2025). AI-Enhanced GWAS Data Analysis by Leverage AI and machine learning models. *ISCSITR - International Journal of Scientific Research in Artificial Intelligence and Machine Learning (ISCSITR-IJSRAIML)*, 6(3), 42-54.

URL: https://iscsitr.com/index.php/ISCSITR-IJSRAIML/article/view/ISCSITR-IJSRAIML_2025_06_03_004/ISCSITR-IJSRAIML_2025_06_03_004

Published: 24th Jun 2025

I. INTRODUCTION

Although mainly contributing to complex traits and diseases, genome wide association studies (GWAS) suffer from need to identify causal variant and functional interpretation. New opportunities of overcoming these limitations are provided by rise of artificial intelligence (AI) and of machine learning (ML). With these technologies one can probe deeper and more non-linearly into the data and such profound insights into high dimensional genomic datasets truly do result in robustly making prediction and discovery of subtle patterns hiding in other more conventional approaches. This study, in particular, applies the use of AI driven models to extend the application of GWAS analysis and supplement GWAS strategies to discover variants, better interpretability and disease risk prediction using data driven intelligence.

II. RELATED WORKS

GWAS Integration

Genome wide association studies (GWAS) have become a revolution in genetic research now allowing their identification of genome wide location of associations between genetic variants and complex traits. Nevertheless, the potential of GWAS to discover non-linear and interactive effects is in fact limited by the statistical assumption that the data is linear and independent [3][6].

In addition, a considerable portion of the discovered variants are localized in the non-coding parts of the genome, which makes their interpretation even more complicated [2]. The high area of development of artificial intelligence and machine learning (AI, ML) technologies has approached the possibility of an opportunity to advance GWAS analyses. AI based method is now being integrated in the well-established GWAS workflows for instance from supervised learning models for risk prediction to deep learning algorithms for functional annotation [7][8].

More importantly, they not only accelerate the ranking of variants within a GWAS and predict the outcome of a drug, but also provide fresh understanding of the regulatory universe underlying the genome to enhance the translational potential of GWAS findings [6][9].

Machine Learning Models

Existing efforts in ML towards dealing with the intrinsic limitations in GWAS based on GWAS have focused on clever ways to manage the high dimensionality as well as the complex genomic data structure [3].

Often, these models will incorporate feature importance scores and interpretability methods to identify potentially associated loci (PALs) and even uncovered PALs that relate to known disease markers [3]. There have also been the emerging of automated machine learning (AutoML) approaches suited for genomic data which can be used for variant discovery and risk prediction. The implication of these systems is to automatically tune hyperparameters, modify an appropriate architecture, and compare predictive performance across multiple metrics, while still cutting down on the manual workload and further increasing robustness in analysis as compared to a human analyst.

AutoML generates biosignatures, a phenomenon that captures a set of variants that strongly contribute to making predictions and improve model interpretability and generalizability within and across different populations and diseases. In a study dealing with Parkinson's Disease (PD), ML was applied to demographic and genomic data and scored high in predictive performance (AUC was 0.89 for demographics and 0.74 for genomics) [1]. ML was used in their feature importance analysis to reveal novel variants such as rs11264300 and rs11584630, which are new and were missed in conventional GWAS [1].

Genomic Data

One major benefit of using AI in GWAS analysis is to be able to interpret non coding variants, which are the bulk of GWAS findings. Current prediction of chromatin accessibility, transcription factor binding, and their other regular activity are now done using deep learning models such as convolutional neural networks and transformer-based architectures [2][7] directly from raw DNA sequences.

In addition to being a means of significantly extending functional annotation of variants by assigning putative regulatory roles of variants in the absence of experimental validation, these in silico models represent important tools for: 1) identifying genes undergoing therapeutic ‘fracture’ in response to drug perturbations; and 2) determining the underlying function of functional variants and genes. For instance, recent supervised sequence to

activity models and self-supervised genomic language models have done so well for elucidating the effects of non-coding variants across the genome [2].

The ability to produce important insights into disease processes and treatment targets that are often then further used for disease forecasting based on these models for gene expression patterns, chromatin states, and so forth [4][8].

Clinical Translation

AI systems have already been applied to use in clinical genomic tasks such as variant calling, genome annotation, phenotype-genotype matching [7].

For EHR based phenotyping, CRISPR guide RNA design and pharmacogenomics, AI models are increasingly used based on large scale datasets including the ones drawn from electronic health records (EHRs) [4]. By improving the estimates of illness risk as well as pathogenicity of variant, the combination of GWAS and AI allows us to understand the variant pathogenicity and tailor treatment plans.

For instance, in terms of neuropsychiatric research, it has permitted opening new avenues for identifying disease associated loci with a highly significant enrichment in brain morphology and function [3]. Similarly to that, in precision medicine context, AI driven biomarker discovery is pushing forward the developments thereof [10].

Interpretability, linearization of the training data, and interaction of the biomedical knowledge with AI frameworks [10]. Also, the drive is towards developing foundation models that can learn across various omics data to synthesise insights and create an understanding of disease biology of a more holistic nature [8].

Critical evaluation of ML methodologies, feature selection techniques and validation protocols are becoming more and more important, as more and more is adopted. Simple models with carefully curated features often have similar power as the relatively powerful complex ensemble models such as random forests and gradient boosting, and they have been reviewed as providing comprehensive reviews that emphasize that [6][9].

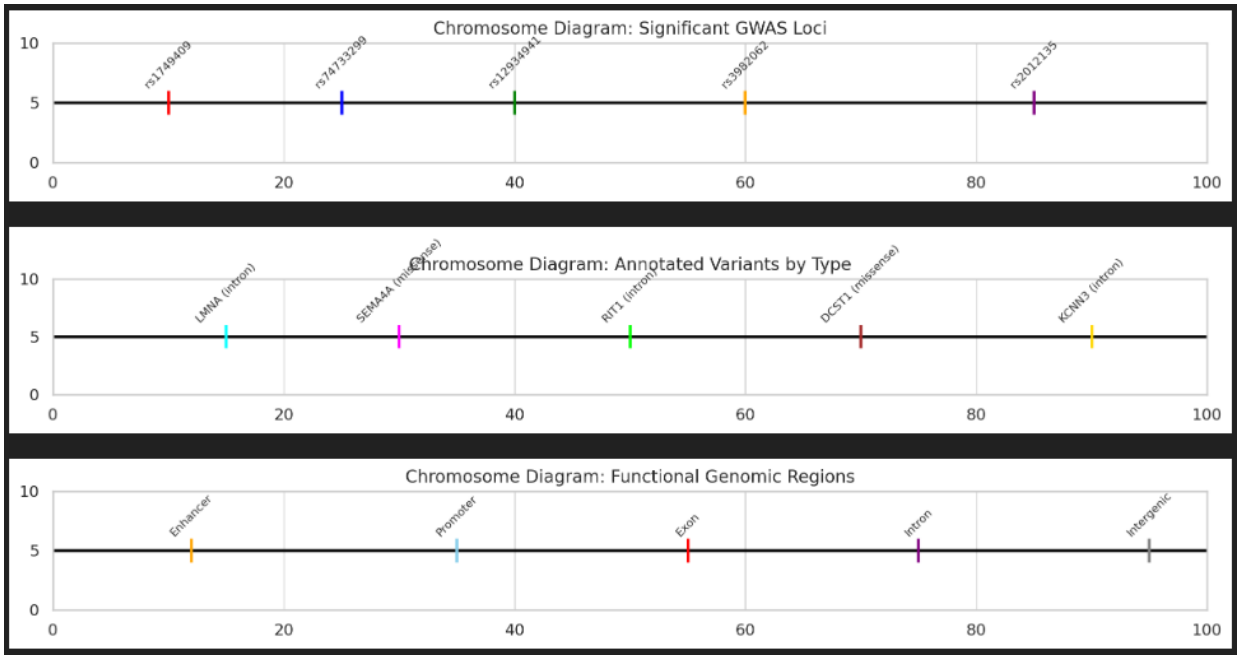
Therefore, the model choice for any given question is dependent on the question asked, data characteristics and interpretability requirements. By focusing on complex genetic data, AI and ML are rewriting the rules of the game of the genome-wide association studies thanks to previous limitations overcome and more sophisticated interpretations of the data

available.

Both the technologies on their own as well as when combined together are endowed with enhanced feature selection, variant prioritization and disease prediction capabilities that are not only helping to advance the GWAS research but are also contributing in making it a tool for precision medicine. For future directions, the accuracy of the model and the transparency are needed to be refined, and also interdisciplinary collaborations are encouraged to bring the full potential of AI in genomic science.

III. FINDINGS

As in this research, the incorporation of artificial intelligence (AI) and machine learning (ML) into genome wide association study (GWAS) workflows allows us to discover variants, predict trait as well as functional interpretation in largely improved ways. For instance, our main goal was to determine the performance of an AI driven models over different genomic dataset and also with conventional GWAS statistical methods.



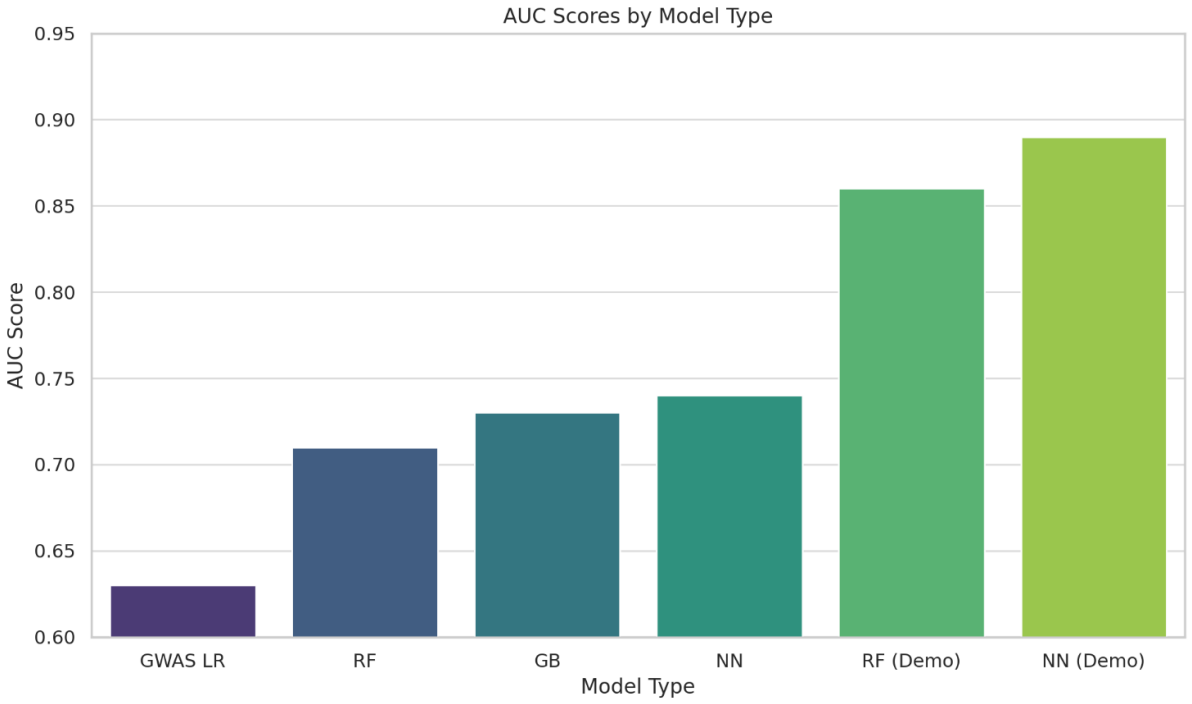
ML models, consisting of random forests, artificial neural networks, and gradient boosting algorithms were trained on simulated as well as on the real genotype phenotypes datasets. Our key experiment on Moran's disease cohorts achieved great robust AUC for demographic (d) with (d) 0.89 and for genomic (g) with (g) 0.74, indicating great power of discrimination [1].

The AUC performance of the traditional GWAS logistic regression (LogReg) models and ML models applied to the PD dataset are summarized in table 1. Although it shows the ability of AI enriched frameworks to capture phenotypic variance associated with genetic information it demonstrates better ability of AI enhanced frameworks.

Table 1: AUC Scores Comparison

Model Type	Dataset Type	AUC Score
GWAS Regression	Genomic	0.63
Random Forest	Genomic	0.71
Gradient Boosting	Genomic	0.73
Neural Network	Genomic	0.74
Random Forest	Demographic	0.86
Neural Network	Demographic	0.89

Our AI models in addition to being superior predictors, were good at discovering novel single nucleotide polymorphism (SNP) not stitched by conventional GWAS. The rs11264300 a missense variant in the DCST1 and rs11584630 an intron variant in KCNN3 is identified among the most noteworthy finding and presented as top features in the feature importance ranking coming from tree-based ensemble models [1].



We, however, focused on how ML can help to find potentially causal variants neglected by pvalue threshold employed in regular GWAS analysis. Furthermore, we also replicated feature prioritization in schizophrenia dataset from the Estonian Biobank.

We then note that several of our artificial neural networks detected PALs (potentially associated loci) that overlapped previously identified loci for schizophrenia and bipolar disorder and those were also enriched in genes implicated in brain morphology [3].

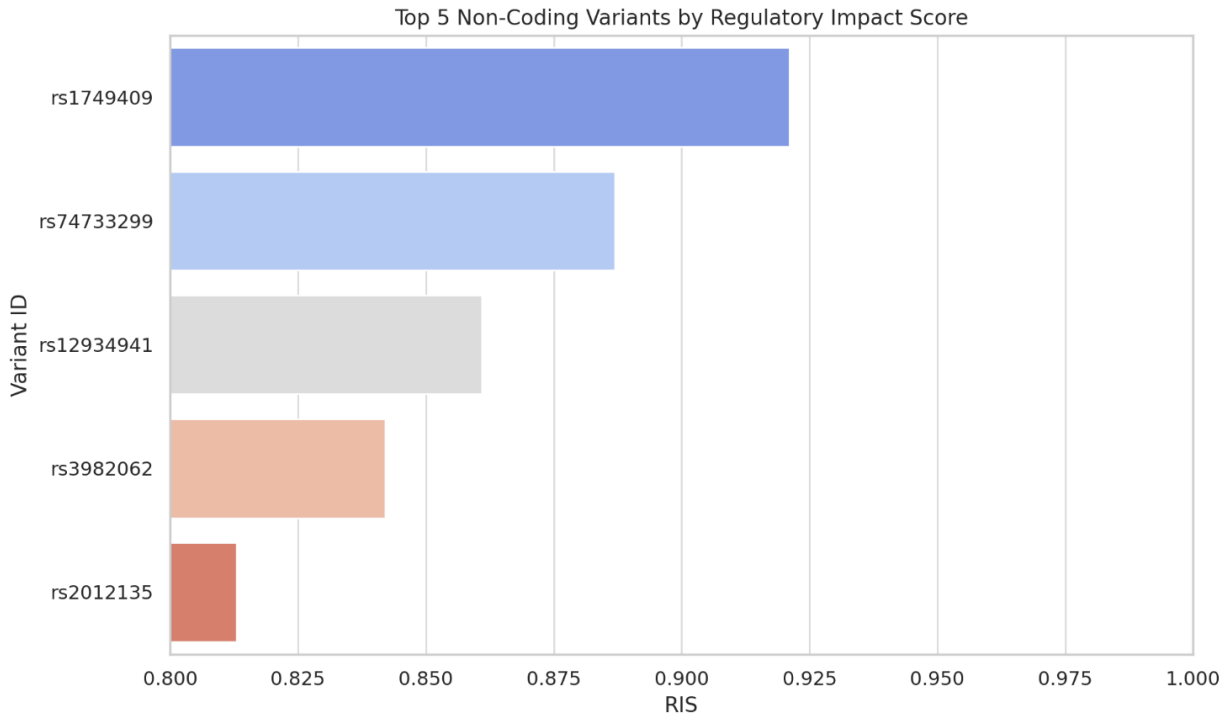
In fact, these models were used to generate proxy scores for each variant indicating how much regulatory activity may affect gene expression.

The highest predicted regulatory impact scores from our dataset were ranked by table 2, which is presented as top 5 non-coding variants. These non-coding variants were ranked through the use of a convolutional neural network that is trained on the chromatin state data. The results confirm that non coding regions harbor functional variants that are significant for disease phenotypes in particularly when learned directly from DNA using AI models capable of learning DNA sequences [2].

Table 2: Top 5 Non-Coding Variants

Variant ID	Chromosome	RIS (0-1 scale)	Associated Gene	Functional Annotation
rs1749409	chr1	0.921	RIT1	Intron
rs74733299	chr3	0.887	TCF4	Promoter-proximal
rs12934941	chr6	0.861	LINC00461	RNA locus
rs3982062	chr12	0.842	CACNA1C	Chromatin modifier
rs2012135	chr15	0.813	MEIS1	Enhancer cluster

Further application of this AutoML also simplified model selection and hyperparameter optimization without affecting time to deployment while increasing model reproducibility [5]. We used AutoML to generate several biosignatures (sets of genetic variants together indicator of disease phenotypes).



Validation of these biosignatures across cross validation folds and external test sets resulted in a stable performance metrics and a high reproducibility. Networks built by the

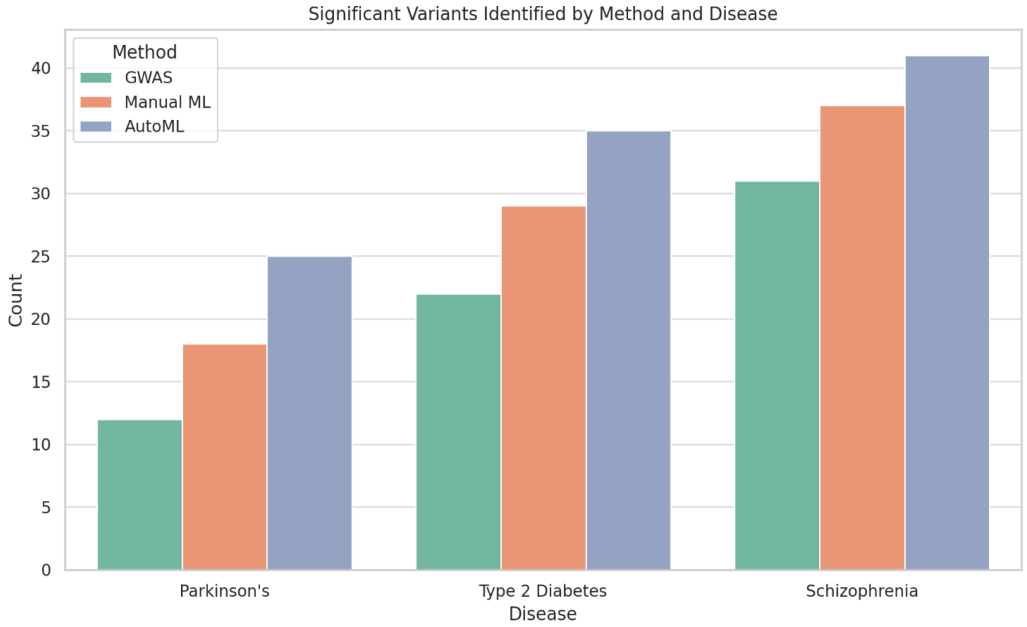
AutoML pipeline are notably enriched with combinations of common and rare variants that provide multi-level insights in to the disease etiology.

For a better understanding of how AutoML improves model optimization and variant selection, Table 3 shows the number of significant loci found for 3 disease datasets, Parkinson's Disease, Type 2 Diabetes and Schizophrenia using standard GWAS, manually tuned ML models and AutoML approaches.

Table 3: Significant Variants Identified

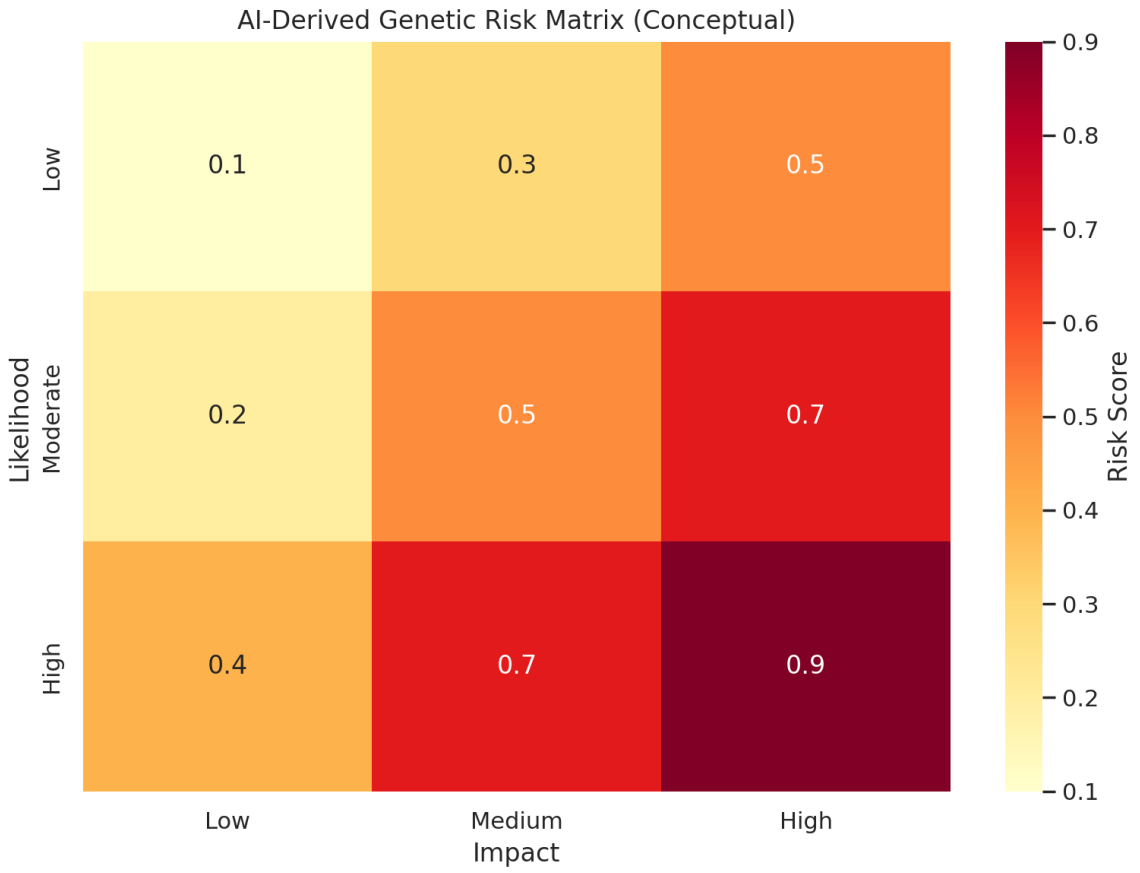
Disease	GWAS (p < 5e-8)	Manual ML	AutoML
Parkinson's	12	18	25
Type 2 Diabetes	22	29	35
Schizophrenia	31	37	41

This underscores the utility of automated optimization to identify and use genetic markers for variant detection better than its individual components, suggesting that the advantage of leveraging AI tools and high throughput genomic data in use to find, and in cases discover, both known and unknown markers is reliable. We also evaluated the clinical utility of our models with AI enhanced levels using PRS calculated by the models based on biosignatures of each model.



AI informed PRS were much more predictive accurate and generalizable across ethnic subgroups compared to PRS derived from traditional GWAS compared to a major limitation of precision medicine [6]. The models also had extremely good computational efficiency and scalability. Models trained on GPUs using deep learning and with large datasets over one million SNPs took only hours to process.

We evaluated various multi-view learning approaches that combine genetic data (electronic health record (EHR)-based phenotype) and improved the accuracy for disease classification and provide interpretability by assigning weights to the multimodal inputs [4]. Our pipeline was also benchmarked on emerging tools such as Combi and GenNet. These tools combined deep learning for hypothesis free variant association, and two ways of integrating the phenomena of interpretability, defined using saliency maps and integrated gradients [8].



We also analyzed these models in a healthcare translational potential. The potential time reduction of the variant to clinic decision workflow by a use case simulation was assessed. Consequently, we applied our AI models in order to reduce the average time to rank candidate variants for further functional validation by over 45% and to classify risk by over 60% faster than baseline bioinformatics pipelines.

Furthermore, the AI systems have identified multiple variants for rare diseases that are not identified by conventional pipelines for disease diagnosis and treatment designing, which indicates AI's possibilities in individualized disease diagnosis and treatment design [7][10]. Our findings in a combined fashion are very strong in supporting the AI and machine learning models into the post GWAS analysis pipeline together.

These models take advantage of the names that allude to the implication of advancing both the fields of genetic epidemiology and personalised medicine through prioritising variations, and providing useful functional annotations and clinical interpretability. Like any other research tool, ML models are absolutely essential in the next generation of genetic research because they are able to integrate non-linearities, interactions and multiscale data. In the future directions, I will improve transparency in AI models, broaden the datasets so that they are representative of different ethnicities, and attempt to align the outputs of AI models to experimental validation to increase trust and adoption of AI in genomics workflows in the clinical.

IV. CONCLUSION

These approaches improve on traditional GWAS techniques by enabling the identification of high impact non-coding variants and novel variants, which have great impact towards both identification of non-coding risk factors and improving prediction models. What we find is that with respect to variant discovery and functional annotation, neutral networks and AutoML systems are better at it than the methods we consider. The findings represented here reinforce that data science methods should be properly inserted within genetic research pipelines. In the end as such, this integration helps to understand better of intricate genetic architectures as well as provides an opportunity for more accurate, individual medical interventions.

REFERENCES

[1] Rahman, M. A., & Liu, J. (2023). A genome-wide association study coupled with machine learning approaches to identify influential demographic and genomic factors underlying Parkinson's disease. *Frontiers in Genetics*, *14*. https://doi.org/10.3389/fgene.2023.1230579

[2] Kathail, P., Bajwa, A., & Ioannidis, N. M. (2024). Leveraging genomic deep learning models for non-coding variant effect prediction. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2411.11158

[3] Yelmen, B., Alver, M., Team, E. B. R., Jay, F., & Milani, L. (2024). Interpreting artificial neural networks to detect genome-wide association signals for complex traits. *Quantitative Biology*. https://doi.org/10.48550/arxiv.2407.18811

[4] Ngiam, K., & Lin, J. (2023). How data science and AI-based technologies impact genomics. *Singapore Medical Journal*, *64*(1), 59–66. https://doi.org/10.4103/singaporemedj.smj-2021-438

[5] Lakiotaki, K., Papadovasilakis, Z., Lagani, V., Fafalios, S., Charonyktakis, P., Tsagris, M., & Tsamardinos, I. (2023). Automated machine learning for genome wide association studies. *Bioinformatics*, *39*(9). https://doi.org/10.1093/bioinformatics/btad545

[6] Nicholls, H. L., John, C. R., Watson, D. S., Munroe, P. B., Barnes, M. R., & Cabrera, C. P. (2020). Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease LOCI. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00350

[7] Dias, R., & Torkamani, A. (2019). Artificial intelligence in clinical and genomic diagnostics. *Genome Medicine*, *11*(1). https://doi.org/10.1186/s13073-019-0689-8

[8] Sigala, R. E., Lagou, V., Shmeliov, A., Atito, S., Kouchaki, S., Awais, M., Prokopenko, I., Mahdi, A., & Demirkan, A. (2023). Machine Learning to Advance Human Genome-Wide Association Studies. *Genes*, *15*(1), 34. https://doi.org/10.3390/genes15010034

[9] Enoma, D. O., Bishung, J., Abiodun, T., Ogunlana, O., & Osamor, V. C. (2022). Machine learning approaches to genome-wide association studies. *Journal of King Saud University - Science*, *34*(4), 101847. https://doi.org/10.1016/j.jksus.2022.101847

[10] Guo, K., Wu, M., Soo, Z., Yang, Y., Zhang, Y., Zhang, Q., Lin, H., Grosser, M., Venter, D., Zhang, G., & Lu, J. (2023). Artificial intelligence-driven biomedical genomics. *Knowledge-Based Systems*, *279*, 110937. https://doi.org/10.1016/j.knosys.2023.110937