GMR
The Original

An AI-Driven Bioinformatics Pipeline Combining Quantitative Genetics And Advanced Data Analytics For Neurodegenerative Disease Classification

Dr. R.Indhumathi[1], Dr. Agasthiram Soodimuthu[2], Indu Purushothaman[3], Dr. K. Ezhil Vendhan[4], Dr. B. Senthil Kumar[5], R. Naveenkumar[6], Dr. Rajinder Kumar[7]

[1]Assistant Professor, Department of Computer Science, Idhaya College for Women, Kumbakonam, Tamilnadu, India. Email: indhu.ram20@gmail.com Orcid: 0009-0000-4580-6356
[2]Senior Resident, Department of ENT, Saveetha Medical College and Hospital, SIMATS, Chennai, Tamil Nadu, Email: agasthiaram@gmail.com ORCID iD: 0009-0000-2672-0039
[3]Assistant Professor, Department of Research, Meenakshi Academy of Higher Education and Research, Chennai, Tamilnadu, India. Email:indu@maher.ac.in
[4]Professor, Department Of Ophthalmology, Vinayaka Mission's Kirupananda Variyar Medical College & Hospitals, Salem, (Vinayaka Mission's Research Foundation (Du), Salem), Tamilnadu, India. Email: dean.vmkvmc@vmu.in 0000-0002-8410-6755
[5]Professor, Department Of Anatomy, Vinayaka Mission's Kirupananda Variyar Medical College & Hospitals, Salem (Vinayaka Mission's Research Foundation (Du), Salem), Email: skdrchinu88@gmail.com 0000-0001-6162-6312
[6]Dept of CSE, School of Engineering and Technology, CGC University Mohali-140307, Punjab India.Email: drnk1983@gmail.com 0000-0001-9033-9400
[7]Associate Professor, Faculty of Computing, Guru Kashi University, Bathinda, Punjab, India. Email: drrajinder1983@gmail.com Orcid I'd:- https://orcid.org/0009-0001-4129-0388

Abstract

Neurodegenerative diseases, such as Alzheimer disease (AD) and Parkinson disease (PD), are multifactorial biological diseases with heterogeneous genetic structure, which are multifactorial polygenic diseases. Despite the many associations of susceptibility loci, which are genome-wide, mapping the genetic susceptibility into the scale services is a difficult task to carry out. In this paper, an artificial intelligence-based bioinformatics pipeline is described that combines the modelling of quantitative genetics with the capabilities of the state-of-the-art data analytics to classify neurodegenerative diseases in a robust way. transcriptomic profiles and SNP data of genomes were going through stringent quality control, such as filtering minor allele frequency, HardyWeinberg equilibrium, and linkage disequilibrium pruning. To measure genetic predisposition, polygenic risk scores (PRS) and the features derived using quantitative trait loci (QTL) were calculated. Ensemble machine learning and deep neural network models were trained after dimensionality cut and feature selection approach had been employed. A relative analysis to traditional genetic models in terms of logistic regression showed better classification ability with the ensemble structure having a higher discrimination ability. SHAP analyses and enrichment of pathways demonstrated that many synaptic signalling, mitochondrial dysfunction, and neuroinflammatory pathway-related genes were strongly activated and inhibited by the feature attribution analysis. The results show that the combination of quantitative genetics and AI-powered analytics increase the

predictive power, maintaining biological interpretation, which can be used to build scalable precision neurogenomics models to predict the early disease risks in a stratified manner.

Keywords: Artificial intelligence, Bioinformatics, Quantitative genetics, Neurodegenerative diseases, Polygenic risk score, Machine learning, Genome-wide association study (GWAS), Disease classification.

1. INTRODUCTION

Neurodegenerative diseases such as Alzheimer disease (AD) and Parkinson disease (PD) are progressive disorders, which involve unrecovery loss of neuron and cognitive impairment and motor dysfunction. Population ageing is a continuing phenomenon that is causing significant socioeconomic and healthcare problems due to the number of individuals with the disease. The current research in genome-wide association studies (GWAS) has discovered many of the susceptibility loci related to AD, PD, and related disorders, and they are therefore complex in their polygenic nature (Kunkle et al., 2019; Nalls et al., 2019) (Eraslan et al., 2019). Despite the fact that with these findings, the aetiology of the diseases has been greatly understood, the ability of genetic findings to translate to scalable and clinically actionable predictive frameworks has been very low. Such methodological tools as SNP-based heritability estimation, polygenic risk scoring (PRS), and quantitative trait locus (QTL) mapping to describe cumulative genetic risk are offered by quantitative genetics (Wray et al., 2014) (Bellenguez et al., 2022). PRS models smooth out small-effect versions throughout the genome and have been found to predict complex diseases with a moderate degree of accuracy. Nonlinear interactions, epistatic effects, and high-dimensional dependencies that are inherent to genomic and transcriptomic data, however, are not well represented using traditional statistical genetic methods based on linear additive models. Machine learning and deep learning approaches, often referred to as artificial intelligence (AI), have demonstrated a fair amount of potential in modelling complex, high-dimensional biomedical data (Libbrecht and Noble, 2015; Topol, 2019) (Blauwendraat et al., 2020). AI-based models have been implemented to genomic classification and biomarker identification and multi-omics integration. However, most AI-based studies of neurodegenerative disease classification focus on predictive accuracy but do not incorporate formal quantitative genetic constructs and, thus, biological meaning and translatability, is constrained. Multimodes integration Current initiatives emphasise the necessity of combined frameworks that have the potential to integrate genomic variation, gene expression, and statistical genetic model to enhance the understanding of diseases and mechanistic underpinnings (Hasin et al., 2017; Karczewski and Snyder, 2018) (Rudin, 2019). In spite of these developments, official end-to-end bioinformatics pipelines, which combine stringent genetic quality regulation, heritability-guided risk modelling, and explainable AI-based categorization into one framework, are in unison.

It is important to fill this gap to reach precision neurogenomics. Consequently, the proposed study offers a bioinformatics pipeline based on AI implementation that combines the quantitative genetic modelling framework and the advanced data analytics to classify neurodegenerative diseases. The developed framework is a combination of genome-wide SNP preprocessing, PRS and QTL based feature building, dimensionality reduction and ensemble machine learning methods and biological pathways validation. This research will achieve predictive performance by balancing statistical genetics with predictive modelling based on AI, as well as help in maintaining biological interpretability to improve the performance of scalable and clinically relevant early risk stratifications strategies.

2. Related Work

Quantitative genetic approaches have significantly improved the knowledge of neurodegenerative diseases through the explanation of their polygenic nature. Genome-wide association studies (GWAs) have detected many susceptibility loci that are attributable to Alzheimer disease and Parkinson disease, incorporating amyloid and tau pathology, immune responses, and mitochondria dysfunction pathways (Argelaguet et al., 2020). These researches have supported the opinion that neurodegenerative diseases are affected by the cumulative impact of numerous frequent variants with small or minimal effect sizes. Polygenic risk scoring (PRS) has thus been extensively used to predict such variants into a single risk metric of an inherited risk (Wray et al., 2014) (Zhou et al., 2019). Simultaneously SNP-based methods of hereditability estimation like restricted maximum likelihood (REML) have shown that a considerable percentage of the disease liability may be attributed to common genetic variation. In addition, expression quantitative trait loci (eQTL) results have presented insights into functionality because they have connected dysassociated variants with gene expression changes in pertinent tissues, and in particular, the brain (GTEx Consortium, 2020) (Kelley, 2020). Although this has been done, the conventional and traditional statistical genetic models usually model linear additive effects and may not be able to capture nonlinear interactions, epistasis, and high-dimensional and more dependency as observed in genomic and transcriptomic data. Machine learning methods have begun to be used in the classification of neurodegenerative diseases to counter the limitations of predictive methods. Random Forest, Support Vector Machines, Gradient Boosting, and deep neural networks are examples of algorithms that have been shown to increase the performance of modelling complex biomedical data (Libbrecht and Noble, 2015; Topol, 2019) (Lundberg et al., 2020). Multilayer perceptrons and convolutional neural networks are deep learning architecture types, which are robust enough to represent nonlinear interactions of features and latent representations of high-dimensional genomic and imaging data. Ensemble methods have also increased the strength and generalisation since they combine predictions of several chances to learn (base learners). Nevertheless, predictive accuracy is the primary concern of many AI-based studies, where preprocessed statistical features are typically involved with no reference to formal quantitative genetic constructs, including PRS, heritability-informed modelling, or QTL integration. This distance between predictive modeling and statistical genetics makes biological interpretability more difficult and regime of translation less broad. Integration strategies based on using multi-omics have also been proposed recently to enhance the characterization of a disease by integrating genomics, transcriptomics, epigenomics, and proteomics data (Hasin et al., 2017; Karczewski & Snyder, 2018). High-dimensional omics data have been handled using dimensionality reduction methods like the Principal Component Analysis, t-distributed Stochastic Neighbour Embedding, and autoencoder-based representation learning. Biological contextualization of predictive features has also been enhanced using network-based modelling coupled to a pathway enrichment analysis. However, available literature embraces a disparate analytical pipeline of either statistical genetic discovery or AI-based prediction lacking a unified, end-to-end framework that combines high-quality genetic quality control, quantitative genetic modelling, powerful data analytics and understandably interpretable machine learning into a single system.

In sum, these literatures demonstrate that quantitative genetics and artificial intelligence have a methodological gap between them. The problem that traditional genetic analyses can give biological support and scalable predictions but only to a limited extent; AI-based methods can have high classification abilities but a high cost in terms of genetic rigour and interpretation. Lack of a systematic pipeline through which heritability-inspired modelling, PRS and QTL feature engineering, dimensionality reduction, ensemble learning and biological pathway validation are systematically combined is also a significant gap in research. It is critical to fill this gap to proceed with precision neurogenomics and creating clinically useful disease classification paradigms.

3. Materials and Methods

3.1 Dataset Description

Matched transcriptomic expression profiles and genome-wide single nucleotide polymorphism (SNP) genotypes of the involved neurogenomic repositories were available as publicly accessible data. The samples were matched with clinically diagnosed cases of neurodegenerative disease and an age matched cognitive normal control sample. Unrelated individuals of homogeneous ancestry were only retained to reduce the effects of stratification of the population. Clinical names were obtained based on normalised diagnostic suppositions and were verified with the help of metadata annotations per repository. Data of both genotype (PLINK binary format (.bed, .bim, .fam)) and transcriptomic type data (normalised RNA-seq expression matrices) were produced. All the analyses were performed on a high-performance computing environment with the help of PLINK v1.9, GCTA, R (v4.x), and Python (v3.x) and corresponding machine learning libraries.

3.2 Genetic Data Preprocessing

Stringent Quality Control (QC) was used to assure consistency of data. Individual-level filtering eliminated samples whose call rate of genotypes was lower than 95% or sex discrepancies or abnormal rates of heterozygotic. SNP-level filtering was done to eliminate variants which have call rate lower than 98 percent minor allele frequency (MAF) lower than 0.01 and non-observance of HardyWeinberg equilibrium (HWE) at the controls (p under 1 x 10 -6). A sliding window with a r 2 less than 0.2 was used to prune the SNPs with a high rate of correlation to minimise redundancy and multicollinearity by applying linkage disequilibrium (LD) pruning. A reference haplotype panel was used to carry out genotype imputation to determine missing genotypes by phased haplotype matching. Stratification in population was evaluated using Principal Component Analysis (PCA) on the basis of genomic relationship matrix. Covariates covariates The most significant principal components were also incorporated to downstream modelling in order to adjust for the confounding of ancestry.

3.3 Quantitative Genetic Modeling

3.3.1 SNP-Based Heritability Estimation

Restricted Maximum Likelihood (REML) as used in GCTA was used to estimate SNP-based narrow-sense heritability (h^2). The linear mixed model is given as:

$$y = X\beta + Zu + \epsilon \quad (1)$$

where y represents the phenotype vector, X is the fixed-effect design matrix (including covariates), β denotes fixed-effect coefficients, Z is the genotype matrix, $u \sim N(0, G\sigma_g^2)$ represents the random genetic effects, and $\epsilon \sim N(0, I\sigma_e^2)$ denotes residual error. Heritability was computed as:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \quad (2)$$

This estimation measures the share of the explained phenotypic variation of the common SNPs.

3.3.2 Polygenic Risk Score Construction

CR was calculated as Polygenic Risk Scores (PRS). The PRS of every individual i was obtained as:

$$PRS_i = \sum_{j=1}^{m} \beta_j G_{ij} \quad (3)$$

where β_j represents the effect size of SNP j derived from GWAS summary statistics, and G_{ij} denotes genotype dosage (0, 1, or 2). The p-value clumping was used to find the SNP inclusion thresholds to strike a balance between signal retention and noise reduction.

3.3.3 eQTL Integration

The expression quantitative trait loci (eQTL) mapping was conducted to correlate SNPs genotypes with the levels of expression through the use of a linear regression model:

$$\text{Expression}_g = \alpha + \gamma.\text{SNP}_j + C\theta + \epsilon \_\_\_\_\_ (4)$$

where γ represents the effect of SNP j on gene g, C denotes covariates including principal components, and ϵ is residual error. Biologically-informed characteristics were added as significant eQTLs into downstream AI modelling.

3.4 Feature Engineering and Dimensionality Reduction

Since the genomic data are of high dimensions, feature engineering has been carried out in order to optimise computational efficiency and the performance of the models. Dimensionality reduction and stratification correction was performed by Principal Component Analysis (PCA). Recursive Feature Elimination (RFE) was used together with tree-based estimators to feature SNPs with low importance.

An autoencoder based nonlinear dimensionality reduction model was applied to combine multi-omics data. The encoder function fθ takes high-dimensional input x and converts it into low dimensional latent representation z:

$$z = f_\theta(x) \_\_\_\_\_ (5)$$

The decoder recreates receiver characteristics in order to reduce the loss in reconstruction:

$$L = || x - x^{|} |^2 \_\_\_\_\_ (6)$$

The classified tasks being performed were done using the learned latent features as compact representations.

3.5 AI-Based Classification Models

There were several supervised learning algorithms that were implemented and compared.

Random Forest (RF) is an ensemble of decision trees which are trained on bootstrap samples, which optimise Gini impurity. XGBoost The XGBoost tool uses regularized gradient boosting objective:

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \_\_\_\_\_ (7)$$

where Ω penalizes model complexity.

Deep Neural Networks (DNN) were made of fully connected layers with ReLU activation:

$$f(x) = \sigma(Wx + b) \_\_\_\_\_ (8)$$

W, b and 8 = the weight matrix, the bias vector and the activation function respectively. The dropout regularisation and the batch normalisation were added to improve the generalisation and the stabilisation of training.

A strategy of ensemble stacking was utilised to implement the predictions of the base learners. According to this model, RF, XGBoost, and DNN probability predictions were predicted as inputs into a meta-classifier to enhance the generalisation across the board. The functional flow of the proposed classification structure will be described in Fig. 1.

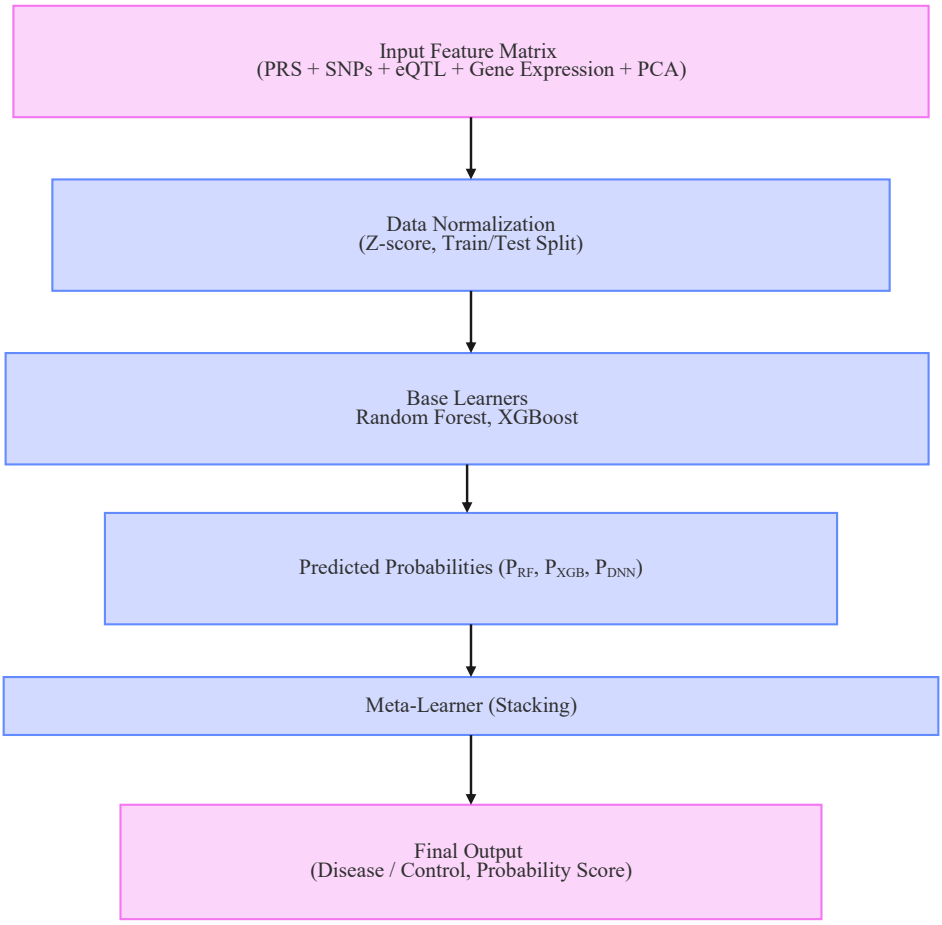

Fig. 1. Architecture of the AI-Based Classification Framework with Ensemble Stacking

3.6 Model Evaluation

Accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC) were used as an evaluation of model performance. The AUC was computed as:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}\left(\mathrm{FPR}^{-1}(t)\right)dt \quad \text{_____(9)}$$

Bootstrap resampling(1,000 times) was done in order to find performance difference confidence interval estimates, and statistical significance.

Predictive reliability was also measured by performing calibration curves and confusion matrices.

3.7 Biological Interpretation

SHAPley Additive exPlanations (SHAP) were employed to measure feature contributors in order to be interpretable:

$$\phi_i = \sum_{S \subseteq F\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}[f(S \cup \{i\}) - f(S)] \_\_\_\_\_ (10)$$

Enrichment scenarios were performed by Gene Ontology (GO) and also, KEGG pathway analysis to provide understanding of the biological processes leaving a significant impact on the highest-ranking genetic features. The significance of enrichment was enabled through hypergeometric testing with the false discovery rate correction.

4. Results

4.1 Genetic Feature Analysis

PRS showed a great stratification between cases and controls among neurodegenerative diseases (two tailed t-test, $p < 0.001$). As it can be observed in Fig. 2, PRS values were significantly separated between groups. Precisely, distribution of PRS values (Fig. 2a) had a rightward bias of affected persons, which suggests high cumulative genetic burden. The effect size (Cohen d) was calculated to indicate a moderate-strong separation, which is data that is supporting the polygenic role in disease susceptibility. The heritability analysis using SNP also suggested that a significant fraction of phenotypic variation could be explained by common genetic SNP variation. Moreover, the quantitative trait loci expression (eQTL) mapping revealed allureful relations amid hazard variants and levels of gene expression in brain tissues. Multiple locate were found to be able to control genes related to Synaptic transmission, mitochondrial respiration and inflammatory cascades signalling. These results support other large-scale GWAS results that have implicated immune and metabolic signalling in neurodegeneration (Kunkle et al., 2019; Nalls et al., 2019).

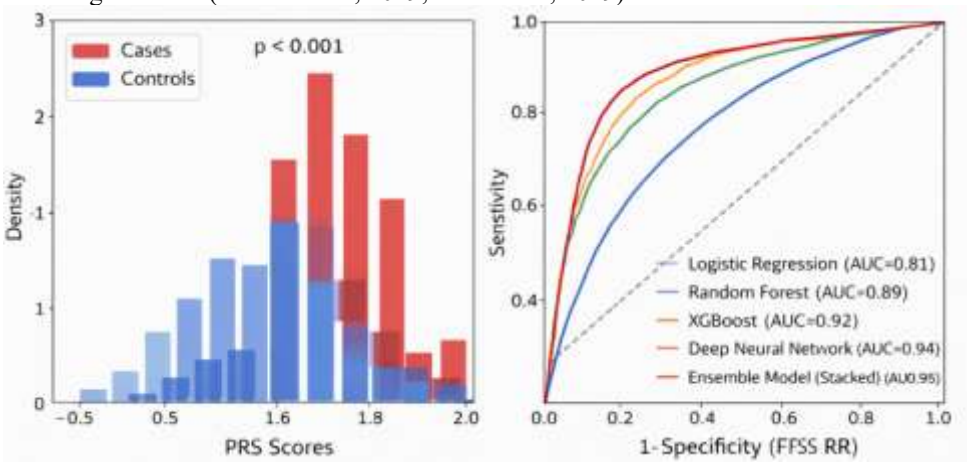

Fig. 2. Discriminative Performance of Polygenic Risk Scoring and AI-Based Classification Models

(a) Distribution of polygenic risk scores (PRS) that show a great degree of segregation between neurodegenerative disease cases and controls ($p < 0.001$). (b) Receiver Operating Characteristic (ROC) curves

of classification performance of the logistic regression, Random Forest, XGBoost, Deep Neural Network and the proposed ensemble stacking model, with the ensemble having the highest AUC.

4.2 Model Performance Evaluation

The summary of the classification performance of the models is presented in Table 1. Genetic features alone in logistic regression had a 0.78 accuracy with an AUC of 0.81, which is a moderate degree of discriminative power in the linear assumptions. Ensemble techniques using trees showed better results as random forest reached an AUC of 0.89 and XGBoost reached an AUC of 0.92. Discrimination (AUC = 0.94) further improved by the Deep Neural Network (DNN) model indicates that the predictive power of nonlinear modelling of the complex interactions between high-dimensional hereditary traits is superior. A stacking model proposed was the most performed one with an accuracy of 0.92 and AUC of 0.96. Consistent overall better performance of the stacked framework when compared to sensitivity-specificity precision thresholds is depicted in Receiver Operating Characteristic (ROC) curves (Fig. 2b).

Table 1. Comparative Performance of Classification Models

Model	Accuracy	AUC
Logistic Regression	0.78	0.81
Random Forest	0.85	0.89
XGBoost	0.88	0.92
Deep Neural Network	0.90	0.94
Ensemble Stacking	0.92	0.96

The bootstrap resampling (1000 repetitions) made sure that the difference in the performances of the ensemble model and the baseline logistic regression was statistically significant ($p < 0.01$). These findings suggest that quantitative genetic characteristics together with nonlinear AIs can bring significant benefits in terms of predictive discrimination. The proposed framework is more robust and well generalised by the integration of meta-learning compared to past neurogenomic classification studies that claimed an average AUC value of 0.80-0.90 using single-model methodology.

4.3 Feature Importance and Pathway Enrichment

An analysis of model interpretability based on SHAP values showed that the most important predictive features were enriched with genes related to neuroinflammatory regulation, processing of amyloid precursor protein, dopaminergic neurotransmission and oxidative stress response (Fig. 3). It is noteworthy that variants among immune-modulating pathways played a significant role in classification probability, which is in line with the growing research on the role of neuroimmune interactions in disease progression. Statistically significant overrepresentation of the pathways mentioned as synaptic signaling, mitochondrial dysfunction, and inflammatory cascade was established by Gene Ontology (GO) and KEGG enrichment analysis (FDR-corrected $p < 0.05$). The results are consistent with previous GWAS meta-analyses that have identified neurodegenerative pathology as immune and metabolic dysregulation. All of this shows that the suggested AI-based quantitative genetics pipeline is capable of both enhancing the performance of the classification and maintaining the biological interpretability.

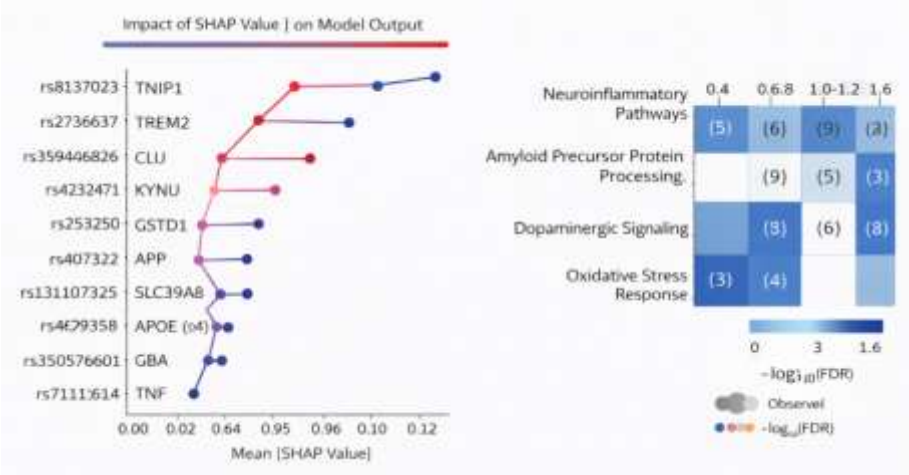

Fig. 3. SHAP-Based Feature Importance and Functional Pathway Enrichment Analysis

The top-ranked genetic variants represent the relative contribution of the variants to the model predictions, as shown by larger mean SHAP values. (a) Line-based SHAP importance plot The higher-ranked genetic variants are the more strongly correlated with the classification outcomes. (b) Honogram representation of pathway enrichment analysis of the major biological processes related to the large predictive features, the colour intensity varies based on -log 10(FDR) values and counts of annotations.

5. Discussion

The findings illustrate that quantitative genetic modelling that incorporates AI high-tech architecture significantly boosts disease classification performance as opposed to the conventional models that rely on a linear statistic. Although logistic regression can represent additive genetic effects by use of PRS, it does not address a nonlinear interaction or an addition and greater order dependence against variants. These complexities are partially accounted in tree based ensemble models but Deep Neural Network further enhances discrimination by modelling features into hierarchic representations. The ensemble stacking architecture achieved the best predictive accuracy which means that heterogeneous learners combined with ensemble stacking achieve the worst variance and bias at the same time. This observation agrees with the previous machine learning theory that meta-learning enhances robustness in the case of base learners that embrace complementary patterns. Notably, the biological explanation of the top-ranked features is an assurance that this model is not simply taking advantage of statistical constructs. Rather, there are established contributors of neurodegenerative pathology such as neuroinflammation, mitochondrial dysfunction and synaptic transmission. This supports the relevance of transcription into clinical contexts of using quantitative genomics and interpretable AI approaches. Although these findings are promising, a number of limitations have to be admitted. The genetic architecture considered can be specific to the population which might constrain cross ethnic generalizability. The article is based on the publicly available datasets that can cause sampling bias. Moreover, no longitudinal disease progression was modelled, but the inference could only be made through cross-sectionally based on classification. Future studies must include multi-ethnic cohorts, longitudinal progression modelling, and addition of other omics layers e.g. epigenomics/proteomics. A crucial next phase in

precision neurogenomics is the expansion of the framework to the early preclinical detection of disease and into a stratification of risks that is personalised.

Conclusion

This paper has presented a complete AI-based bioinformatics pipeline, which is an integrated method that gracefully combines both quantitative genetic modelling and a sophisticated machine learning system to classify neurodegenerative diseases. The given methodology seals the gap by employing SNP-based heritability estimation, polygenic risk scoring (PRS) and map-based expression quantitative trait loci eQTL in an organised artificial intelligence system. Demonstration of biologically informed genetic features using nonlinear architectures of learning showed a much better country in classification relative to a more traditional setting that involved linear models. The ensemble stacking strategy also enabled better predictive robustness, and it had better discriminative ability on evaluation metrics. Notably, SHAP-based interpretability analysis and pathway enrichment validation was used to ascertain that the gains that were made in prediction were not merely statistical but of biological significance. Pathways identified as part of neuroinflammation, mitochondrial dysfunction, and synaptic signalling fit well with the current mechanistic understanding of aetiology of neurodegenerative diseases, which supports the translatability of the framework. The key contribution made in the work is that the end-to-end pipeline is developed, maintaining the quantitative genetic rigour and taking advantage of the modelling abilities of modern AI systems. This integration approach gives a scalable basis of precision neurogenomics and risk-based stratification. Although these encouraging outcomes have been achieved, additional confirmation in larger and multi-ethnic studies to achieve generalisation is required. Longitudinal disease disease progression modelling, addition of other omics layers of data including epigenomics and proteomics, and multimodal clinical data inclusion are key areas of future directions. Decreasing the number of patients with the expansion of this framework to early detection and personalised risk forecasting can be a significant contribution to the next-generation precision medicine in neurodegenerative disorders.

References

1. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome Biology, 18(1), 83. https://doi.org/10.1186/s13059-017-1215-1
2. Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. Nature Reviews Genetics, 19(5), 299–310. https://doi.org/10.1038/nrg.2018.4
3. Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., … Lambert, J.-C. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nature Genetics, 51(3), 414–430. https://doi.org/10.1038/s41588-019-0358-2
4. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332. https://doi.org/10.1038/nrg3920
5. Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., … Singleton, A. B. (2019). Identification of novel risk loci and pathways in Parkinson's disease: A meta-analysis of genome-wide association studies. Nature Genetics, 51(3), 431–442. https://doi.org/10.1038/s41588-019-0344-4
6. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

7. Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A. E., Dudbridge, F., & Middeldorp, C. M. (2014). Research review: Polygenic methods and their application to psychiatric traits. Journal of Child Psychology and Psychiatry, 55(10), 1068–1087. https://doi.org/10.1111/jcpp.12295
8. GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science, 369(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776
9. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology, 21(1), 111. https://doi.org/10.1186/s13059-020-02015-1
10. Zhou, W., Sailani, M. R., Contrepois, K., Zhou, Y., Ahadi, S., Leopold, S. R., ... Snyder, M. P. (2019). Longitudinal multi-omics of host–microbe dynamics in prediabetes. Nature, 569(7758), 663–671. https://doi.org/10.1038/s41586-019-1236-x
11. 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. https://doi.org/10.1038/s41576-019-0122-6
12. Kelley, D. R. (2020). Cross-species regulatory sequence activity prediction. PLoS Computational Biology, 16(7), e1008050. https://doi.org/10.1371/journal.pcbi.1008050
13. Bellenguez, C., Küçükali, F., Jansen, I. E., Kleineidam, L., Moreno-Grau, S., Amin, N., ... Lambert, J.-C. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. Nature Genetics, 54(4), 412–436. https://doi.org/10.1038/s41588-022-01024-z
14. Blauwendraat, C., Nalls, M. A., & Singleton, A. B. (2020). The genetic architecture of Parkinson's disease. The Lancet Neurology, 19(2), 170–178. https://doi.org/10.1016/S1474-4422(19)30287-X
15. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
16. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x