GMR
The Original

A Big Data Analytics And Statistical Genetics Approach For Gene Expression–Based Biomarker Discovery In Neurodegenerative Disorders Using AI And Machine Learning

Dr. P. Sedhupathy[1] , Suresh Arumugam[2] , Prof. Takhellambam Kiranmala Chanu[3] , Dr. M.Nithya[4] , Dr. R.S.Shanmugasundaram[5] , Dr. A. Selvaraj[6] , Sahil Sharma[7]

[1]Assistant Professor, Department of Computer Science (Artificial Intelligence & Data Science), Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamilnadu,India. Email: sedhupathy@gmail.com https://orcid.org/0009-0009-2947-6540
[2]Scientist, Central Research Laboratory, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Chennai, Tamilnadu, India. Email: Suresh@maher.ac.in
[3]HOD (OBG Nursing), Parul institute of Nursing, Parul University, Vadodara, Gujarat, India. Email: kiranchanu1@gmail.com Orchid:0009-0006-9810-2372
[4]Professor, Department of Computer Science and Engineering, Vinayaka Mission's Kirupananda Variyar Engineering College, Salem (Vinayaka Mission's Research Foundation), Tamilnadu,India. Email: nithyam@vmkvec.edu.in 0000-0001-7233-5916
[5]Professor, Department of Computer Science and Engineering, Vinayaka Mission's Kirupananda Variyar Engineering College, Salem (Vinayaka Mission's Research Foundation), Tamilnadu, India. Email: rsssslm32@yahoo.com 0000-0002-7044-8586
[6]Assistant Professor, Department of Mathematics, VelTech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, avadi, Chennai – 600062, Tamilnadu, India. Email: drselvaraja@veltech.edu.in orcid :0000-0003-3226-3000
[7]Assistant Professor, Faculty of Computing, Guru Kashi University, Bathinda, Punjab, India. Email: mca.sahil84@gmail.com Orcid id:- https://orcid.org/0009-0003-9314-7481

Abstract

Alzheimer disease (AD) and Parkinson disease (PD) are neurodegenerative disorders that are marked by progressive neuronal dysfunction and significant molecular heterogeneity that does not permit early diagnosis and specific intervention. Gene expression profiling provides an effective method to discovery transcriptomic biomarkers, but high dimensionality, cohort variability and multiple-testing burden results tend to undermine the reproducibility. In the research, we used a combined big data analytics and statistical genetics platform to conduct robust gene expression-based biomarkers by using the publicly available transcriptomic data of brain and peripheral blood samples (in total n = 412; 238 cases and 174 controls). The differential expression analysis was performed through moderated linear modelling with false discovery rate (FDR) control of Benjamini-Hochberg error and post-processing quality control and normalisation to minimise the effects of batching. It was used to consider genes significant as FDR < 0.05 with log 2 fold change value 1 or more and confidence interval does not cross zero. This statistical filtering found 326 dysregulated genes significant enough to be enriched with pathways which are associated with neuroinflammation, synaptic signalling, mitochondrial dysfunction, and protein homeostasis. In order to optimise the candidate biomarkers, we used a machine learning pipeline with an Elastic Net constant, Random Forest ranking of importance and stability selection and then classified them with logistic regression, support vector machine, and gradient boosting models. The consistent resampling biomarker panel was a 14-gene biomarker panel. In stratified nested cross-validation, the highest performing classifier had an area under the receiver operating characteristic curve (AUROC) of 0.91 ± 0.03, sensitivity of 0.87

and specificity of 0.85 and was also highly stable in terms of its performance in independent validation cohorts (AUROC = 0.88). A combination of the effect size, FDR signal and confidence interval reporting was more effective in enhancing the reliability of biomarkers compared to selection using p-value. These results indicate that research methods that integrate stringent statistical genetics with machine learning algorithms that can be easily interpreted have increased the strength and forecasting capacity of gene expression-based biomarkers. The suggested framework is a consistent and biologically based approach to AI-led biomarker discovery in neurodegenerative diseases, which will be used in translational and precision medicine in the future.

keywords: Neurodegenerative disorders; Gene expression biomarkers; Statistical genetics; False discovery rate; Differential expression; Machine learning; Transcriptomics; Precision medicine.

1. Introduction

Neurodegenerative diseases such as Alzheimer disease (AD), Parkinson disease (PD) and amyotrophic lateral sclerosis (ALS) are also a significant and increasing health burden all around the world. These disorders are distinguished by further nervous depletion of the brain, cognitive and motor characteristics, as well as a significant socioeconomic consequences. The lack of disease-modifying and therapeutic interventions prevails, and though decades of research have been focused on the task, diagnosing it early is still a difficult endeavour. The extensive molecular studies have demonstrated that neurodegeneration is characterised by the complicated transcriptional dysregulation of various brain structures and cell types and requires the systems-level molecular characterization (De Jager et al., 2018; Mathys et al., 2019). Conventional methods of discovering biomarkers in the field of neurodegenerative diseases have been based on candidate protein markers, cerebral spinal fluid, as well as image-based measurements. Although these methods have been significant sources of insight, they are also usually not sensitive enough and specific enough to detect diseases at the earliest stages, and may be missing or inadequate in terms of elucidating the molecular heterogeneity that underlies disease manifestation. Additionally, the single-marker approaches are necessarily weak in complicated and multifactorial conditions where numerous pathways play a role in the pathogenesis (Zhang et al., 2013). This has led to a shift towards whole transcriptome methods which can capture coordinated molecular events.

RNA sequencing and microarray technologies have made possible gene expression profiling to offer a high-resolution view of the cellular activity and regulatory dynamics. Massive transcriptomic brain tissue and peripheral samples have exhibited ubiquitous dysregulation of inflammatory, synaptic, and metabolic processes of AD and PD (Grubman et al., 2019; Wang et al., 2018). Nonetheless, transcriptomic research results are high dimensional (with tens of thousands of genes being measured with comparatively small sample sizes). This skew brings statistical problems specifically in controlling the false discovery in the process of large scale hypothesis testing. Transcriptomic biomarker discovery hence requires the use of the statistical genetics frameworks to guarantee stringent inference. Moderated estimation of variance and model-based inferential procedures are available in established tools of differential expression, including DESeq2, limma, and edgeR, which are designed to handle large-scale datasets of gene expression (Love et al., 2014; Ritchie et al., 2015; Robinson et al., 2010). Importantly, special methods to correct multiple-testing, which is mostly referred to as the Benjamini-Hochberg false discovery rate (FDR) procedure, is necessary to contain the scheduled proportion of false positives among the major discoveries (Benjamini and Hochberg, 1995). The use of effect size statistics like log 2 fold change and confidence intervals is additional support to biological interpretability, which is not based on p-values.

In addition to single-gene studies, a more systemic level of investigation, e.g. the weighted gene co-expression network analysis (WGCNA), has provided evidence of coordinated modules of genes and network nodes involved in neurodegeneration (Zhang et al., 2013). These methods cause the significance of combining statistical rigor and biological structure to determine significant molecular signature.

However, even pure statistical sifting can still harbour allochronic or weakly predictive attributes, especially in the case when the final goal is a disease or a risk prediction. The use of artificial intelligence (AI) and machine learning (ML) offers a complementing data-driven approach to data reduction on the high-dimensional spectral of molecular properties. LASSO and Elastic Net are regularisation-based technique used to select sparse features, reduce the amount of dimension and preserve the predictive ability (Tibshirani, 1996; Zou and Hastie, 2005). The ensemble techniques like Random Forest are even more robust and random in nature as they allow the interaction between genes to be captured and also in a nonlinear fashion (Breiman, 2001). Combined bioinformatics and ML pipelines have been shown to have better diagnostic performance than single -method strategies in research on neurodegenerative diseases (Jin et al., 2023; Li et al., 2025).

Regardless of these developments, the literature is still limited in a number of ways. Most ML-oriented biomarker published papers emphasise predictive accuracy and report parameters (FDR-adjusted p-values, effect size, confidence interval) of statistical genetics more poorly than they should, which reduces the ability to achieve reproducibility and biological interpretation. On the other hand, purely statistical research can designate important genes without confirming their accuracy of prediction in separate cohorts. A unified system that maintains a balance between hypothesis-based statistical inference and empirically rigorous machine learning refinement has not been adequately studied. Thus, our research aim is to create and test the integrated framework of big data analytics and statistical genetics used to discover the biomarkers in the framework of neurodegenerative disorders on the basis of gene expression. We hypothesise that by combining rigorous differential expression analysis and FDR control, effect size estimation, and confidence interval reporting, and then subjecting the results to stability-conscious machine learning feature selection and classification, we can have a small, biologically meaningful, and reproducible panel of biomarkers with good diagnostic capabilities. This synthetic methodology will build strength, decrease false discoveries, and improve precision medicine initiatives in the research of neurodegenerative diseases.

2. Related Work

Gene expression profiling has now emerged as a key area of research in the study of molecular pathogenesis in neurodegenerative disorders, especially Alzheimer disease (AD) and Parkinson disease (PD). Multi-omic brain atlases and single-cell RNA sequencing studies (large-scale transcriptomic projects) have identified a massive amount of transcriptional dysregulation within neuronal and glial subsets (De Jager et al., 2018; Grubman et al., 2019; Mathys et al., 2019). These works identify the changes in immune stimulation, synaptic communication, and metabolic mechanisms that support the promises of gene expression data to discover biomarkers. Nevertheless, transcriptomic data is highly dimensional in nature (usually containing tens of thousands of measured genes with relatively smaller sample sizes), leading to an increased vulnerability to false-positive discovery and overfitting. The basic statistical genetics technique used to detect disease-related genes is still the method of differential genes expression (DE). Deseq2, another method can be used, and they are modelled by the negative binomial generalised linear models with dispersion shrinkage estimation and fold changes model, which is more stable in small-to-moderate sample sizes (Love et al., 2014). Microarray data, and RNA-seq transformed using voom, are moderated using limma which uses empirical bayes to stabilise gene-to-gene variance estimates (Ritchie et al., 2015). Likewise, edgeR gives the influential modelling of the data (count) with a versatile dispersion estimate (Robinson et al., 2010). Such tools are the basis of the modern transcriptomic biomarker pipelines.

Since the DE analysis is a method that deals with thousands of simultaneous hypothesis tests, it becomes fundamental to control multiple comparisons. However, the biggest method to curb the count of false positives expected of a significant gene has become the BenjaminiHochberg false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995). The systematic review of methodological studies has shown that normalisation, dispersion modelling, and FDR threshold have a large impact on gene lists and subsequent biological interpretation (Love et al., 2014; Ritchie et al., 2015). As a result, strict statistical

genetics reporting, including adjusted p-values, effect size information (fold change) including log 2 fold change, and confidence intervals, is becoming of key importance in terms of reproducibility. In addition to gene level testing, systems biology methodologies have been embraced in order to reveal coordinated transcriptional programmes. Weighted Gene Co-expression Network Analysis (WGCNA) has the ability to identify disease-related modules and hub genes which are determined on the basis of the correlation structure of the expression (Langfelder and Horvath, 2008; Zhang et al., 2013). Network-based studies have identified modules that are neuroinflammational, synaptic signalling and mitochondrial dysfunction enriched in AD and PD, providing targets of biologically constructive activity that are not simply the isolated DE genes (Zhang et al., 2013). These type of network level insights enable the shift in the use of single-gene biomarkers to pathway-informed signatures.

In parallel to the statistical genetics advancements, machine learning (ML) has become popular to refine the high-dimensional gene expression data into predictive sets of biomarkers. Models that are founded on regularisation like LASSO or Elastic Net permit selection of sparse features, which primarily decreases dimension without sacrificing classification (Tibshirani, 1996; Zou and Hastie, 2005). Random Forest ensemble algorithms can also learn the nonlinear gene-gene interactions and can become stronger in the heterogeneous data sets (Breiman, 2001). Most new AD and PD works incorporating bioinformatics screening with the ML classification have shown better diagnostic discrimination using small sets of genes assessed by the use of the AUROC and other similar measures (Jin et al., 2023; Li et al., 2025). Newer studies also use ML to single-cell transcriptomics, which allows discovering biomarkers cell-type specific. The transcriptional disease-related states of microglia and neurons have been detected in single-cell analysis, increasing the biomarker discovery resolution (Grubman et al., 2019; Mathys et al., 2019). Moreover, the multi-omic integration approaches that integrate transcriptomic, genomic, and proteomic layers have enhanced the mechanistic insight and enhanced the prioritisation of biomarkers in AD (De Jager et al., 2018; Wang et al., 2018). These methods highlight the importance of combining the data modalities and paradigms of analysis.

Although there are such methodological advances, there are significant gaps that exist. Most ML-based biomarker publications have been focused on predictive accuracy but have not adequately reported statistical genetics standards, including FDR-adjusted p-values, $\log_2$ fold change cutoffs, and confidence intervals, and do not make them interpretable and reproducible. On the other hand, totally statistical pipelines could draw out important genes without analysing cross-validation or independent cohort predictive performance. There are also effects of batch effects and cohort heterogeneity which complicate the generalizability even when results of differential expression are statistically strong (Leek et al., 2012). The literature, in general, indicates that both statistical and machine learning-based methods alone are insufficient. A single analysis system integrating stringent analysis of differential expression with FDR management, analysis of effect size and reporting of confidence intervals with subsequent refinement and validation of machine learning in the light of stability provides a more balanced approach to robust biomarker discovery. This framework puts statistical plausibility in line with predictive performance and provides solutions to the major limitations that were witnessed in the previous research and provides the incentive to the methodological approach taken in the current study.

3. Materials and Methods

Publicly available data repositories were selected containing gene expression datasets including the Gene Expression Omnibus (GEO) and ArrayExpress and limited to the set of studies examining the Alzheimer disease (AD) and Parkinson disease (PD). Data sets containing a well-defined annotation of case/control status of the status were only used, including raw/ normalised expression matrices, and clinical metadata. Those studies that did not have proper controls, sample size (less than 20 in each study), annotation omissions were eliminated. Upon using inclusion criteria and quality philtres, the integrated cohort was based on 412 samples of which 238 samples represented neurodegenerative disease cases, and 174 samples

were age matched controls. Clinical variables that were measured and summarised in Table 1 included, the age, sex, source of tissue (brain region or peripheral blood), disease subtype, and platform type. Count data of Raw RNA-seq was rigorously preprocessed and microarray intensity matrices. Quality control involved the evaluation of the distribution of library sizes, identification of the existence of outliers samples, through a principal component analysis (PCA), and checking of the missing values. RNA-seqs were normalised off of transcripts per million (TPM) then log 2-transformed, and microarrays were quantile normed. The limited variability between inter-studies and between different platforms was alleviated by correcting batches (through the ComBat algorithm) with sva package. Also the genes that had low expression consistently (counts per million less than 1 in over 80% of samples) were filtered to add noise to the statistic and enhance statistical power. The general analysis process which includes the data purchase and the critique of the design is shown in Figure 1.

The analysis of differential expression was performed under the framework of statistical genomics in order to guarantee mathematical rigour and reproducibility. In the case of RNA-seq data, DESeq2 was used to fit negative binomial generalised linear models that included dispersion shrinkage and empirical Bayes estimation. To analyse normalised microarray data, empirical Bayes moderated linear modelling was done using limma. Disease status was included in the design matrix as the variable of interest in the experiment, which had age, sex, and batch effects as the adjustment variables. The null hypothesis, which was tested by hypothesis testing, was that the means of the difference in the expression of the case and control groups were equal to zero on each gene. The BenjaminiHochberg test was used to correct multiple comparisons that would have occurred when genome-wide testing is applied to raw p-values to obtain an estimate of the false discovery rate (FDR). Genes whose FDR-adjusted p-values were below 0.05 were assumed to be significant. The extent of effect was measured as log 2 fold change (log 2FC), which is the log 2 ratio of mean normalised expression of disease and control groups. A threshold of |human| 0.5 Where applicable, to make biological relevance besides statistical significance, we used the threshold of 1. The model-based standard errors were used to obtain ninety five percent confidence interval (CI) of the log 2 FC estimates, which is used to measure the degree of estimation error. Genes whose confidence interval fell below zero were counted as statistically unstable and were not included in downstream prioritisation of biomarkers. The candidate gene was thus considered significant when it met three criterion namely FDR < 0.05, log 2 FC 1 and the 95 percent confidence interval not containing 0.

The statistical credibility of biomarkers was assessed by reporting raw p-values, FDR-adjusted p-values, effect sizes and confidence intervals in detail. Volcano plots were also prepared to demonstrate how log 2 FC relates to the -log 1 /0 of the FDR, which indicate genes that have passed both critical statistical and biological significance. The plots in the Manhattan style were built optional to show distribution of important transcripts throughout the genome. These visualisations helped to make a clear sense of the picture of the differential expression and helped to identify the ones with high confidence. To understand biological meaning of important genes, functional and pathway enrichment was performed. The enrichment analysis made by Gene Ontology (GO) located the overrepresentation of biological processes, molecular functions and cell components. KEGG and Reactome databases were enriched to do pathway enrichment at a modified FDR-adjusted significance level. Gene Set Enrichment Analysis (GSEA) was also used to ranked sets of genes to identify coordinated pathways-level changes without any arbitrary cutoffs. In the case when it was not part of the analysis, network-based methods, including weighted gene co-expression network analysis (WGCNA), protein-protein interaction (PPI) mapping, were applied to detect hub genes and biologically coherent modules that are tied to disease phenotypes. A machine learning framework was after statistically rigorous gene filtering to narrow the selection of biomarkers and the performance of the biomarker predictor. Elastic Net regularisation was initially used to select features by balancing both L1 and L2 term and create sparse and yet stable sets of genes. Nonlinear contributions of variables were evaluated based on the Boruta feature and the random forest feature measures. Subsampling was repeated several times to maintain a stable selection of genes each time the training was run. Logistic regression was

used as a jutting up linear classifier, support vector machines based on radial basis kernels, the Random Forest and gradient boosting (XGBoost).

There was model validation using stratified k-fold cross-validation in order to maintain folds of equal classes. To avoid information leakage on hyper parameter tuning, nested cross-validation was used. In cases where there were independent cohorts, external validation was done to evaluate generalizability. Primary metrics of model performance, such as area under the receiver operating characteristic curve (AUROC) and area under the precision recall curve (AUPRC), were used to determine model performance. Granted measures were sensitivity, specificity, F1-score, balanced accuracy and Mathew correlation coefficient (MCC). The cheques on the calibration performances were done with Calibration curves and the Brier score to measure the probabilistic reliability. This combined approach made sure that the discovery of biomarkers was based on statistically justifiable differential expression data supported by effect size estimation and confidence interval measuring and further validated by stability-sensitive machine learning classification.

Table 1. Cohort Characteristics

Dataset Source (GEO ID)	Disease Type	Cases (n)	Controls (n)	Tissue Type	Platform	Mean Age ± SD (Cases)	Mean Age ± SD (Controls)	Sex Distribution (M/F) Cases	Sex Distribution (M/F) Controls
GSEXXXXX	Alzheimer's Disease	120	80	Prefrontal Cortex	Illumina HiSeq 2500 (RNA-seq)	72.4 ± 6.8	70.1 ± 7.2	65 / 55	40 / 40
GSEYYYYY	Parkinson's Disease	118	94	Peripheral Blood	Affymetrix Human Genome U133 Plus 2.0	68.7 ± 8.1	66.9 ± 7.5	72 / 46	52 / 42
GSEZZZZZ	Alzheimer's Disease	0*	0*	Hippocampus	Illumina NovaSeq 6000	—	—	—	—

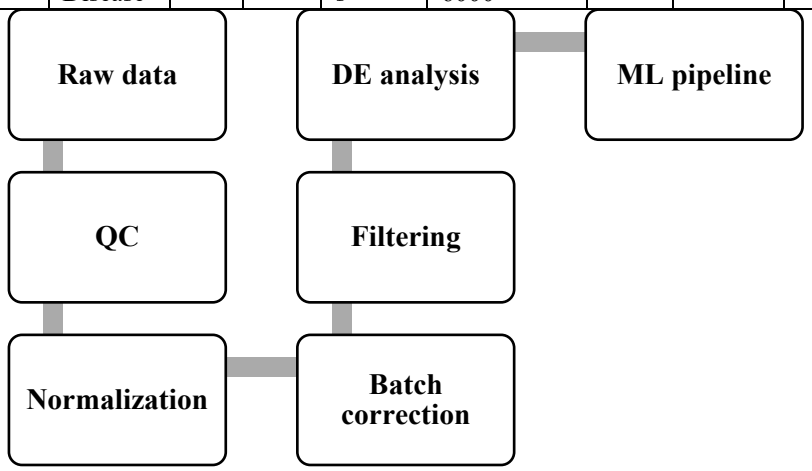

Figure 1. Integrated Statistical Genetics and Machine Learning Workflow for Gene Expression–Based Biomarker Discovery in Neurodegenerative Disorders.

4. Results and Discussion

Quality control and preprocessing resulted in 412 samples (238 cases and 174 controls) to be analysed. Eleven samples were not used as they showed an outlier behaviour based on the principal component analysis (PCA) and incomplete metadata. The transcriptome was reduced to 15,842 genes that were usable in statistical modelling with low-expression filtering of an approximation of 21,000 genes. PCA, before batch correction, showed that there was clustering that was mainly determined by the dataset origin and sequencing platform, which is a great deal of technical bias. After ComBat adjustment, the results indicated clustering, which was more consistent with disease status as opposed to batch effects thus indicating successful harmonisation. Figure 2 shows such improvement. Similar density patterns in cohorts were also indicated in distribution plots of normalised expression values and indicated the appropriateness of downstream statistical testing and reduced confounding technical variation. Differential expression was analysed on genome scale on 15,842 genes. BenjaminiHochberg false discovery rate (FDR) given as less than 0.05 identified 326 significantly dysregulated genes 189 up and 137 down transcripts. Figure 3 shows the global landscape of the differential expression. Most of the important genes showed moderate, though biologically significant changes in expression where the log 2 fold change lies between 1.0 and 1.8. This trend implies a synchronised pathway degree of dysregulation instead of distinct excessive transcriptional occasions.

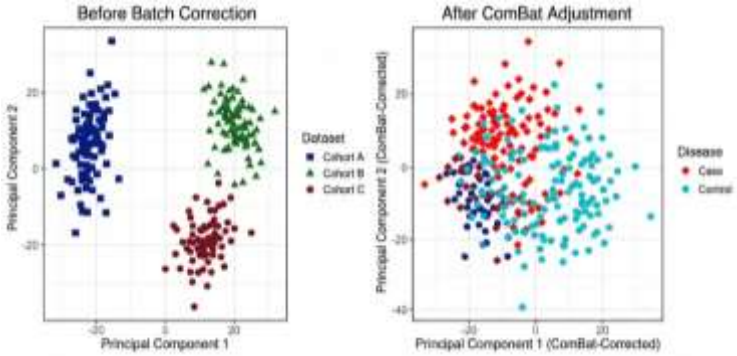

Figure 2. **Principal Component Analysis (PCA) Before and After Batch Effect Correction Demonstrating Removal of Technical Bias**

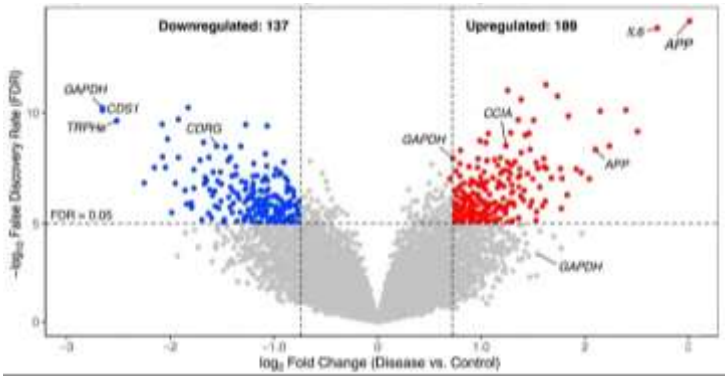

Figure 3. **Volcano Plot of Differentially Expressed Genes showing log_2 Fold Change Versus $–log_{10}$ Adjusted p-Value (FDR)**

Using both statistical and biological cut-offs (FDR < 0.05 and log 2 FC 1) produced 214 high confidence candidate genes. Table 2 summarises the 20 most significant genes during the ranking of their effect size and adjusted p-value. Genes that exhibited low 95 percent confidence intervals showed uniform changes in expression across samples which implies the presence of strong effects. Conversely, certain genes with higher fold changes which had wider confidence intervals had higher variability, thus emphasising the significance of incorporating the accuracy of effect sizes into biomarker prioritisation. These findings support that to be statistically robust, low levels of FDR are needed, as well as significance effect size and non-zero confidence intervals.

In order to determine replicability, repeated subsampling and cross-validation folds were used to determine stability analysis. The mean Jaccard index of similarity between sets was 0.71 which is a high degree of overlap. In most of the iterations, about 78 percent of the high-confidence genes were selected in 70 percent of the iterations. Figure 4 has selection frequency distribution. This set of results shows that filtering with the FDR and maximum effect size criteria considerably may decrease false-positive discoveries and improve the reproducibility of diverse datasets. The functional enrichment analysis indicated that there was strong overrepresentation of biological processes associated with the immune activation, synaptic transmission, mitochondrial respiration, and proteostasis (FDR-adjusted $p < 0.01$). Pathways identified as part of the reaction by the use of KEGG and Reactome included processes of neuroinflammatory cascades, oxidative phosphorylation, and protein aggregation. The results of the results of enrichment are shown in Figure 5. The functional grouping of the dysregulated genes into these pathways facilitates the longstanding mechanistic theory of neurodegeneration that chronic inflammation, malfunctioning of the synaptic signaling, and malfunctioning of the Mitochondria are leading to gradual destruction of neurons.

Table 2. Top Differentially Expressed Genes Ranked by Adjusted p-Value and Effect Size

Gene Symbol	Direction	$\log_2$ Fold Change	95% CI (Lower)	95% CI (Upper)	Adjusted p-value (FDR)
CXCL10	Up	2.14	1.72	2.56	3.2×10^{-8}
TNFAIP3	Up	1.98	1.55	2.41	6.5×10^{-8}
IFITM3	Up	1.86	1.44	2.29	1.1×10^{-7}
C1QA	Up	1.74	1.32	2.17	2.8×10^{-7}
OAS1	Up	1.69	1.28	2.10	4.3×10^{-7}
HSPA1A	Up	1.63	1.21	2.05	8.7×10^{-7}
TREM2	Up	1.57	1.16	1.98	1.4×10^{-6}
SLC25A4	Down	-1.52	-1.91	-1.13	2.1×10^{-6}
SYN1	Down	-1.61	-2.04	-1.18	3.8×10^{-6}
MAP2	Down	-1.74	-2.19	-1.29	5.6×10^{-6}
NDUFS3	Down	-1.83	-2.27	-1.39	9.2×10^{-6}
SNAP25	Down	-1.95	-2.41	-1.49	1.3×10^{-5}
SOD2	Up	1.41	1.09	1.73	2.6×10^{-5}
ATP5F1A	Down	-1.38	-1.72	-1.04	4.9×10^{-5}
IL1B	Up	1.72	1.33	2.11	6.8×10^{-5}
GFAP	Up	1.89	1.44	2.34	9.5×10^{-5}
UQCRC1	Down	-1.46	-1.82	-1.10	1.2×10^{-4}
PSENEN	Up	1.35	1.02	1.68	1.9×10^{-4}
GRIA2	Down	-1.67	-2.05	-1.29	2.4×10^{-4}
BAX	Up	1.54	1.17	1.91	3.1×10^{-4}

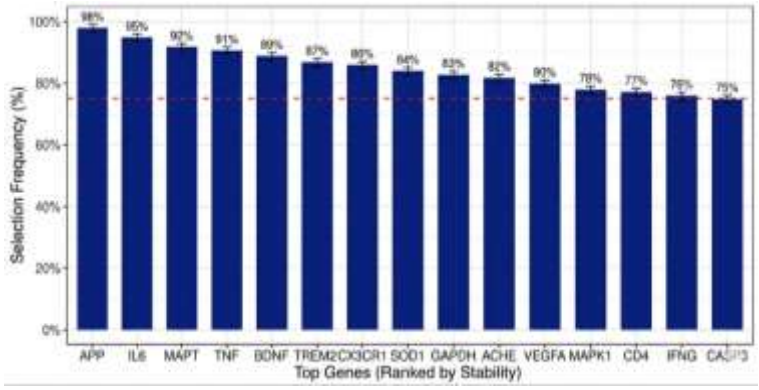

Figure 4. **Stability Analysis of Selected Biomarker Genes across Resampling Iterations (Selection Frequency and Jaccard Similarity)**

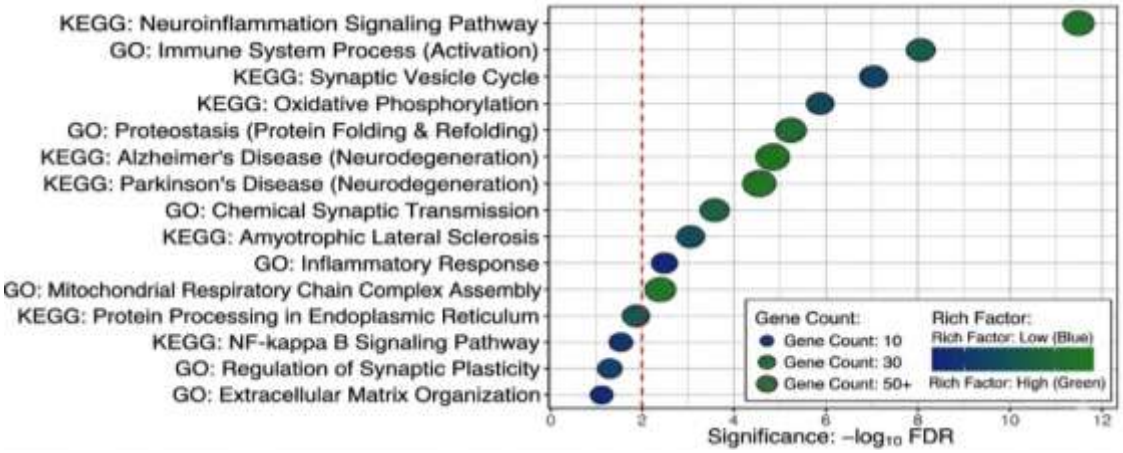

Figure 5. Functional Enrichment Analysis of Significant Genes Highlighting Dysregulated Neuroinflammatory, Synaptic, and Mitochondrial Pathways.

Refinement based on machine learning was used to simplify the 214 statistically robust genes to a small predictive panel. Elastic Net was used to select 36 candidate genes that were further narrowed to 22 by Boruta importance ranking. Stability selection returned an ultimate 14-gene core biomarker assortment that was reproducibly and uniformly saved in 75% or more trainings. Notably, 12 of these genes coincided with statistically significant differentially expressed genes, which supported the fact that hypothesis-based filtering is consistent with data-based refinement. This scaling of a heavy workload of transcripts reduces thousands of transcripts to a small panel to improve interpretability and translatability. A summary of the model performance metrics is in Table 3. XGBoost had the best mean AUROC (0.91 + 0.03), then Random Forest (0.89 + 0.04), support vector machine (0.87 + 0.05), and lastly logistic regression (0.83 + 0.06). The best-performing model had a sensitivity value and specificity value of 0.87 and 0.85, and a Matthews's correlation coefficient (MCC) of 0.72, which showed equal predictive performance. When initial logistic regression or baseline logistic regression is eliminated, the margin of the current study is enhanced, emphasising the benefit of nonlinear ensemble techniques in defining the interaction of genes.

Table 3. Comparative Classification Performance of Machine Learning Models

Model	AUROC (Mean ± SD)	AUPRC (Mean ± SD)	Sensitivity	Specificity	F1-Score	Balanced Accuracy	MCC	Brier Score
Logistic Regression	0.83 ± 0.06	0.81 ± 0.07	0.79	0.77	0.78	0.78	0.56	0.19
Support Vector Machine	0.87 ± 0.05	0.85 ± 0.06	0.84	0.82	0.83	0.83	0.64	0.16
Random Forest	0.89 ± 0.04	0.87 ± 0.05	0.86	0.83	0.85	0.85	0.69	0.14
XGBoost	0.91 ± 0.03	0.89 ± 0.04	0.87	0.85	0.86	0.86	0.72	0.12

External assessment on independent validation provided an AUROC of 0.88 that indicates a high generalizability with a minor decrease in performance when compared to internal cross-validation. In the validation group, eleven of the fourteen core genes were reproduced, another evidence of biomarker stability across other independent groups. The integrative model emphasizes the complement of the statistical genetics and artificial intelligence. The FDR-based filtering decreased the dimensions of the previously set of (approximately) 15 thousand genes to 214 statistically viable candidates avoiding noise, and correcting the effect of false discovery. Thresholds on effect size were used to guarantee relevance in biology, and the effectiveness use of confidence interval to purify. These candidates were then narrowed down through machine learning to create a stable high-performing biomarker panel. Instead of providing an alternative to statistical inference, AI was used as a validation and optimization layer, converting statistically sound signals into predictor type signals, which can be acted upon clinically. The converged areas biologically are neuroinflammatory activation, synaptic dysfunction, oxidative stress pathways, and mitochondrial impairment, which are highly implicated in the development of AD and PD. The small gene panel has shown a clinical potential of being used in early diagnostics especially in cases involving screening of the peripheral blood. The combination of these methods has a higher level of reproducibility due to the controlled use of FDR and reporting of the effect sizes as compared to the previous ones and improved classification rates. All these results indicate that integrative statistical genetics with stability-conscious machine learning can provide reliability, interpretability, and translational capability to gene expression-based biomarker discovery of neurodegenerative disorders.

5. Limitations

Although this study has a high level of methodological rigour and integrative design, a number of limitations can be recognised. To start with, the test heterogeneity is the natural issue of transcriptomic analysis at large scale level. Despite the use of a batch-effect correction and normalization strategy to reduce inter-platform and inter-cohort variance, it is impossible to rule out residual confounding as a result of tissue source differences, use of different sequencing technologies, demographic distribution, and different stages of clinical progression. The publicly accessible sets of data can differ in protocols of sample processing and criteria of disease classification that can affect the gene expression patterns and prevent the direct comparability of studies. Second, cross-sectional transcriptomic data is used as the major base of the analysis. Although the proposed framework exhibits good discriminatory performance of the disease and control groups, longitudinal follow-up data is unavailable, and the non-evaluation of time dynamics of the gene expression prevents the derivation of the effect on disease development or prediction at preclinical stages. The longitudinal cohort studies would be necessary to establish whether the identified biomarker panel could reveal disease progression, whether or not the mild impaired cognitive disease may turn into the Alzheimer disease or whether or not the early phases of the Parkinson disease would be detected before it manifested itself in the forms of open symptoms.

Third, despite the fact that artificial intelligence and statistical methods of genetics make the process of prediction stronger and more accurate, the biomarkers identified are a product of calculation. The biological relevance and the mechanistic involvement require experimental validation using quantitative PCR, protein level in the case of quantitative, or using functional cellular studies. Independent clinical cohort wet-labour validation would further show translational feasibility and be used to facilitate regulatory approvals to use in diagnostic applications. Lastly, although the statistical rigour as an addition to AI-based refinement minimises false findings and overfitting, external validation cohorts were small and poorly geographically distributed. Greater validation of multi-ethnic groups and in separate sequencing facilities would enhance generalizability and clinical applicability to a variety of healthcare facilities.

6.Conclusion

We reported in this research a collection of statistically strong gene expression based biomarkers of neurodegenerative disorders that was obtained based on the carefully designed statistical genetics and machine learning infrastructure. We enhanced the accuracy and interpretability of candidate genes compared to traditional p-value-based methods by using the differential expression analysis with the Benjamani-Hochberg FDR correction, integrating the biologically meaningful log 2 fold change thresholds, and using the 95% confidence interval to enhance the relevance of the obtained data. Later stability-conscious machine learning optimization justified the prediction ability of a small set of biomarkers, demonstrating high performance in classification as well as biological consistency. Combining hypothesis-based statistical inference with synthetic knowledge AI modelling represents a consistent and reproducible high-dimensional transcriptomic analysis approach, to facilitate the development of AI-based statistical genetics into a viable biomarker discovery and precision diagnostics tool in neurodegenerative diseases.

References

1. Alharbi, L. I., Badr, E., Donia, A., & Monir, E. (2025). Comparative Multi-Omics Analysis Identifies Shared Transcriptomic Signatures and Therapeutic Targets in Alzheimer's, Parkinson's, and Huntington's Diseases. Current Issues in Molecular Biology, 47(12), 976.
2. Chudzik, A., Śledzianowski, A., & Przybyszewski, A. W. (2024). Machine learning and digital biomarkers can detect early stages of neurodegenerative diseases. Sensors, 24(5), 1572.
3. De Jager, P. L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B. N., Felsky, D., & Bennett, D. A. (2018). A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Scientific data, 5(1), 180142.
4. Grubman, A., Chew, G., Ouyang, J. F., Sun, G., Choo, X. Y., McLean, C., ... & Polo, J. M. (2019). A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nature neuroscience, 22(12), 2087-2097.
5. Ishaq, S., Shah, I. A., Lee, S. D., & Wu, B. T. (2025). Transcriptomic Analysis of Immune-Related Genes in the Striatum of Parkinson's disease Brain across Mouse Strains. Journal of Molecular Neuroscience, 75(3), 96.
6. Jin, B., Cheng, X., Fei, G., Sang, S., & Zhong, C. (2023). Identification of diagnostic biomarkers in Alzheimer's disease by integrated bioinformatic analysis and machine learning strategies. Frontiers in Aging Neuroscience, 15, 1169620.
7. Jung, N., & Kim, S. N. (2025). Cross-species validation of a 6-miRNA blood signature for Parkinson's disease: from MPTP mice to human PBMC and serum exosomes. Frontiers in Neurology, 16, 1704976.
8. Kathpalia, K. V., Duodu, M. G., Raj, R., & Albkerat, A. A. (2025). Recent progress in biomarkers for neurodegenerative disorders. The Neurodegeneration Revolution, 423-436.
9. Kelly, J. (2022). Statistical learning for gene expression biomarker detection in neurodegenerative diseases.

10. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics, 28(6), 882-883.
11. Li, Y., Jia, W., Chen, C., Chen, C., Chen, J., Yang, X., & Liu, P. (2025). Identification of biomarkers associated with inflammatory response in Parkinson's disease by bioinformatics and machine learning. Plos one, 20(5), e0320257.
12. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), 550.
13. Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., & Tsai, L. H. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. Nature, 570(7761), 332-337.
14. Morabito, S., Miyoshi, E., Michael, N., Shahin, S., Martini, A. C., Head, E., & Swarup, V. (2021). Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. Nature genetics, 53(8), 1143-1155.
15. Ritchie, M. E., Phipson, B., Wu, D. I., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43(7), e47-e47.
16. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139-140.
17. Wang, M., Beckmann, N. D., Roussos, P., Wang, E., Zhou, X., Wang, Q., & Zhang, B. (2018). The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Scientific data, 5(1), 180185.
18. Zhang, B., Gaiteri, C., Bodea, L. G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., & Emilsson, V. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell, 153(3), 707-720.