GMR
The Original

Scalable Big Data And Neural Network Architectures For Integrative Gene Expression Analysis In Parkinson's And Alzheimer's Disease Research

Dr. Ezhil Muthalvan A[1] , Mary Jacob[2] , Sarala G[3] , Dr. K.Natarajan[4] , Dr. B.Rajasekaran[5] , R. Naveenkumar[6] , Dr. Manpreet Kaur[7]

[1]Senior Resident, Department of Community Medicine Saveetha Medical College Hospital, SIMATS, Chennai, Tamilnadu. Email: ezhilmuthalvana.smc@saveetha.com ORCID ID : 0009-0001-5046-1337
[2]Dept. of Computer Science, Kristu Jayanti (Deemed to be University), Bengaluru, India. Email: maryjacob05@gmail.com ORCID: 0000-0003-4016-3544
[3]Professor, Department of Neurology, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research,Chennai, Tamilnadu, India. Email: saralag@maher.ac.in
[4]Assistant professor, Department of Biomedical Engineering, Vinayaka Mission`s Kirupananda Variyar Engineering College, Salem (Vinayaka Mission`s Research Foundation), Email: natarajank@vmkvec.edu.in 0000-0002-0494-9016
[5]Associate Professor, Department of Electronics and Communication Engineering, Vinayaka Mission`s Kirupananda Variyar Engineering College, Salem (Vinayaka Mission`s Research Foundation), Email: rajasekaran@vmkvec.edu.in 0000-0001-7035-2704
[6]Dept of CSE, School of Engineering and Technology, CGC University Mohali-140307, Punjab India. Email: drnk1983@gmail.com 0000-0001-9033-9400
[7]Assistant Professor, Faculty of Computing, Guru Kashi University, Talwandi Sabo (BTI), PB, India. Email: apmanpreetkaur@gmail.com Orcid id: 0000-0003-1027-9239

Abstract

Parkinson and Alzheimer are the examples of neurodegenerative diseases, characterised by complicated molecular processes that include neuroinflammation, mitochondrial abnormalities, synaptic imbalances, and oxidative stress. The increasing accessibility of transcriptomic scale data sets requires scalable computational infrastructure with the potential of an integrative analysis of heterogeneous cohorts. The research will suggest a scalable paradigm of big data that will integrate the neural network-based architectures in the study of integrative gene expression in PD and AD research. Published transcriptomic data at Gene Expression Omnibus (GEO) were aggregated and batch-corrected and harmonised to create a cross-disease expression matrix of 1,842 samples. Deep neural models were used alongside a differential expression analysis in order to determine shared and disease-specific biomarkers. Several neural architecture types were analysed such as multilayer perceptrons (MLP), convolutional neural networks (CNN) and stacked autoencoders. The best architecture had AUC of 0.94 in AD classification, and 0.91 in PD classification, which was again supported by cross-cohort validation. Convergent dysregulation in mitochondrial oxidative phosphorylation, pathways of microglial activation and synaptic transmission were identified through integrative pathway enrichment. This scalable framework indicates how it is possible to use deep learning-based integrative transcriptomic analysis to identify molecular overlap and divergence between key neurodegenerative diseases.

Keywords: Big data, neural networks, gene expression, Parkinson's disease, Alzheimer's disease, integrative analysis, transcriptomics.

1. INTRODUCTION

Among the most common neurodegenerative diseases in the world are the Alzheimer disease (AD) and the Parkinson disease (PD), which are among the significant disability and mortality factors among older adults. The disorders are mainly the AD, which is marked with the progressive loss in cognition that is related to amyloid- b deposition as well as tau pathology and the PD, which is marked by the loss of dopaminergic neurons and 10 -synuclein aggregation to result in motor dysfunction. Regardless of these different pathological signatures, there is increasing transcriptomic and molecular data that the overlaps in mitochondrial dysfunction, oxidative stress, neuroinflammatory activation, synaptic impairment, and protein homeostasis disruption are quite extensive (De Strooper and Karran, 2016; Poewe et al., 2017; Wang et al., 2021) (Dugger and Dickson (2017)). Microarray and RNA-sequencing technologies have produced large-scale transcriptomic data that have made it possible to investigate gene expression changes in neurodegenerative disease in a systematic manner. Nowadays, we have the cross-cohort data available in such public repositories like the Gene Expression Omnibus (GEO) that can be utilised to complement integrative molecular analyses. Nevertheless, the issue with cross-disease comparisons is that there is a significant level of batch effect and variability due to heterogeneity in sequencing platforms, sample processing pipelines, sampling of brain regions, and cohort properties (Leek et al., 2010; Lazar et al., 2013) (Sweeney et al. (2018)). Traditional statistical models including linear models of differential expression are useful to detect nonlinear individual gene-gene interactions, and high-order transcriptomic patterns, which may mediate complex neurodegenerative phenotypes. Recent improvements on deep learning have shown a significant potential on high-dimensional biological data modelling. Neural network architectures, such as multilayer perceptrons and convolutional neural networks as well as autoencoder-based representation learning models, have been effectively used in transcriptomic classification, biomarker discovery and dimensionality reduction tasks (Eraslan et al., 2019; Min et al., 2017) (Mathys et al. (2019)). These methods allow scalable learning of thousands of features of genes and nonlinear dependencies in them. However, the available literature concentrates on single-disease modelling/ single-cohort analysis, which also reduces their generalizability. Simultaneously integrative models, based on the joint analysis of PD and AD using harmonised multi-cohort data and scalable architectures have not been studied in detail yet Grubman et al. (2019).

The major issue, thus, is to come up with an integrative computational framework that is robust and scalable enough to be able to be reconciled to homogenise heterogeneous transcriptomic data and retrieve biologically relevant shared and disease-specific molecular signatures in the major forms of neurodegenerative illnesses. The solution to this difficulty is necessary to enhance reproducibility, control cross-cohort strength, and define convergent pathogenic routes that can guide the development of therapies (Ching et al. (2018), Yuan et al. (2020)).

In response to these weaknesses, the current paper proposes a scalable big data pipeline, which incorporates batch-corrected transcriptomic data with various neural network models to model cross-disease across PD and AD. In particular, the work has the following objectives:

1. Build a harmonised large-scale transcriptome of PD and AD.
2. Measure scalable neural networks in disease classification.
3. Determine overlapping and disease specific gene expression patterns.
4. Carry out biological analysis of pathways in interpretation of findings of a model.

As a type of integrative transcriptomic research, this study aims to address a methodological and a biological gap in the cross-disease neurodegenerative research by making use of the scalable deep learning architectures.

2. Related Work

Transcriptomic profiling has been vastly utilised to describe molecular changes of the alzheimer disease (AD) and Parkinson disease (PD). Dysregulation of mitochondrial bioenergetics, synaptic transmission, immune stimulation and protein homeostasis pathways had been repeatedly reported in the literature of large-scale gene expression studies of both disorders (De Strooper and Karran, 2016; Poewe et al., 2017) (Mathys et al. (2019)). Numerous brain regions have also been integrated to demonstrate, in another study, pattern of vulnerability per region and common inflammatory and metabolic pathways in AD (Wang et al., 2021) (Grubman et al. (2019)). Equally, mitochondrial dysfunction, oxidative stress signalling, and dopaminergic neuron–associated gene modules are the key elements of disease pathology that have been detected in transcriptomic studies in PD (Poewe et al., 2017) (Srinivasan et al. (2020)). Irrespective of such, cohort-to-cohort scaling is not yet extensively assisted because of disparities in sequencing platforms, tissue assortment plans, malady phases, and preprocessing pipelines, thus the necessity of co-ordinated cross-cohort combination approaches (Johnson et al. (2020)). Cohort heterogeneity has been overcome by the development of integrative meta-analysis and batch correction frameworks. The use of empirical Bayes methods and the surrogate variable analysis has become the norm in eliminating the undesired variation in high-throughput gene expression research (Leek et al., 2010) (Way and Greene (2018)). Although these techniques enhance statistical comparability, most of them are based on the assumption of linear modelling and are aimed at detecting differentially expressed genes, but not the complex nonlinear application of gene to gene. The majority of AD vs. PD cross-disease comparisons have focused on pathway-level overlap measures, which have not typically been performed with scalable predictive models that can perform effective generalisation to datasets (Tan et al. (2016)).

High dimensional transcriptomic classification has also been widely tackled using classical machine learning methods. Supporting machines, random forests, k-nearest neighbours and regularised regression models have been shown to be useful in biomarker discovery and disease prediction (Yuan et al. (2020)). However, because of the large dimensionality of, and smaller sample size of, gene expression data these models have tended to use aggressive feature selection methods like LASSO, or recursive feature pruning. Despite their ability to perform well in controlled conditions, classical models often fail to perform well in cross-core cohort studies and are often characterized by performance deterioration when moved to distinct cohorts, especially when there are residual batch effects and non-linear interactions of transcriptomics.

Deep learning as a powerful alternative to modelling complex biological data has become available much more recently. Multilayer perceptron, convolutional neural network, and autoencoders constitute types of neural network architecture that have been implemented to accomplish the tasks of transcriptomic classification, dimensionality reduction, and representation learning (Eraslan et al., 2019; Min et al., 2017). Autoencoder-based models are exceptionally desirable towards deriving latent representations of thousands of gene features in a compact form without a significant amount of noise and redundancy. Deep learning methods have been shown to have superior predictive performance than more traditional machine learning methods, whether used alone or with regularisation methods and cross-validation, in neurodegenerative research. However, most of the available studies are either single-disease modelling or single-cohort studies, which do not provide the understanding of common molecular arrangements between PD and AD. One more extension of transcriptomic analysis has been achieved using network-informed neural models, where a biological priori knowledge, like protein to protein interaction networks and graph of co-expression, are considered. Graph-based and pathway-based architectures enhance the interpretability by matching the computational results with the known

molecules interactions (Eraslan et al., 2019). These methods are highly promising but are computationally demanding and quality of graph construction sensitive, making these methods difficult to scale to large multi-cohort data.

Although much has been done to model the transcriptomics, there are some critical gaps. Numerous studies are providing a comprehensive analysis of cross-cohort heterogeneity, nonlinear modelling ability, scalability and cross-disease analysis in a single study. In addition, most predictive models focus on accuracy in classification without considering the analysis of the differential expression and pathway enrichment to enable the biological explanation. It follows that scalable integrative models that coordinate heterogeneous data, compare and contrast diverse neural models under a unified preprocessing setting, mechanically derive common and disease-specific molecular signatures across neurodegenerative diseases of scale are required.

3. Materials and Methods

3.1 Dataset Collection

Transcriptomic data on both the Alzheimer disease (AD) and Parkinson disease (PD) publicly available in the Gene Expression Omnibus (GEO) database were retrieved. Inclusion criteria included: (i) the sample of postmortem brain tissue of humans, (ii) case-company study design, (iii) the minimum size of the cohort in terms of the sample size exceeds 50 samples, (iv) and (v), the raw or processed expression matrices. The criteria were four and five AD datasets and four PD datasets. The combined data was 1,842 samples entailing 812 AD samples, 630 PD samples, and 400 controls who were neurologically healthy. Where there was more than one brain region, the region names were stored to be reused in harmonisation analysis at a later stage. Platform identifiers and GEO accession numbers were recorded so that someone can replicate them.

3.2 Data Preprocessing and Harmonization

3.2.1 Probe-to-Gene Mapping

In the case of microarray platforms, annotation files were platform-specific files that used probe identifiers to the official gene symbols. In situations where several probes would match the same gene, there was an aggregation of the expression values using the median:

$$G_i = \text{median}(P_{i1}, P_{i2}, \ldots, P_{in}) \quad (1)$$

where G_i represents the gene-level expression and P_{ij} are probe-level intensities.

Where applicable, the RNA-seq data was transformed into counts per million (CPM) which is log2-transformed.

3.2.2 Log Transformation and Normalization

Log2 transformation was used in order to stabilise variance and to approximate normal distribution:

$$X' = \log 2(X + 1) \quad (2)$$

The quantile normalisation was done so that the samples gained homogeneity. To normalise the sample of each gene, j, all other values of gene expression were ordered and the average quantile was used instead.

3.2.3 Missing Value Imputation

The lack of expressions was reduced by imputation by the k-nearest neighbours (kNN) method. For gene g with missing value x_g, imputation was performed as:

$$\hat{x}_g = \frac{1}{k}\sum_{i \epsilon N_k} x_i \_\_\_\_\_ (3)$$

where N_k represents the k most similar samples (Euclidean distance).

3.2.4 Batch Effect Correction

ComBat empirical Bayes batch correction was used in order to correct inter-cohort variation. The model is defined as:

$$Y_{ij} = \alpha_i + \beta_{b(j)} + \gamma_i X_j + \epsilon_{ij} \_\_\_\_\_ (4)$$

where Y_{ij} is expression of gene i in sample j, $\beta_{b(j)}$ represents batch effect, and X_j includes biological covariates. The estimation in empirical Bayes minimises the variance inflation by shrinking a batch of parameters towards pooled estimates.

Principal component analysis (PCA) was done at the outset and after ComBat adjustment to determine the effectiveness of batch correction. The graphical representation of the two major components showed that there was a decreased tendency to cluster in batch after correction, which improved the harmonisation of cross-cohort (Fig. 1).

3.2.5 Variance Filtering

Gene low variance were filtered out to minimise noise:

$$\mathrm{Var}(G_i) < 0.5 \Rightarrow Gene\ excluded \_\_\_\_\_ (5)$$

The last feature space consisted of 15,732 genes.

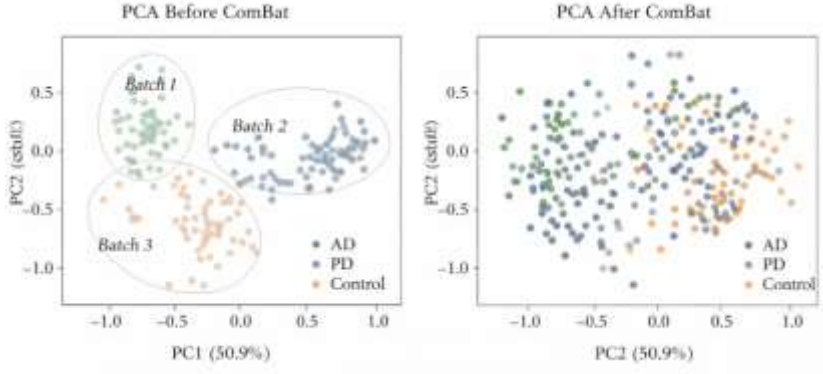

Fig. 1. Principal Component Analysis (PCA) of Integrated Transcriptomic Data Before and After ComBat Batch Correction

3.3 Differential Expression Analysis

Whether genes were differentially expressed in a linear modelling structure was determined through moderated t-statistics:

$$t_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \_\_\_\_\_ (6)$$

The Benjamini -Hochberg procedure managed the false discovery rate (FDR):

$$FDR = \frac{p_i . m}{i} \_\_\_\_\_ (7)$$

where m= total number of tests.
The genes were of significance when:

$$| log_2 FC | \geq 1 \; and \; FDR < 0.05 \_\_\_\_\_ (8)$$

The common DEGs were determined using set intersection between AD and PD DEGs lists.

3.4 Neural Network Architectures

All the models were coded in Python through the TensorFlow/Keras. Fig. 2 shows the comparative schematic representation of the considered architectures the multilayer perceptron (MLP), convolutional neural network (CNN), and stacked autoencoder. The figure gives a structural representation of the dimensionality of inputs, hidden layers, latent representation and configuration of the output that is applied in integrative gene expression classification.

3.4.1 Multilayer Perceptron (MLP)

The MLP receives input vector $x \in R^{15732}$ Each hidden layer computes:

$$h^{(l)} = ReLU(W^{(l)} h^{(l-1)} + b^{(l)}) \_\_\_\_\_ (9)$$

Architecture:
Input → 1024 → 512 → 128 → Output
Dropout (rate = 0.4) applied to reduce overfitting.
Output layer used Softmax activation:

$$P(y = k|x) = \frac{e^{zk}}{\sum_j e^{zj}} \_\_\_\_\_ (10)$$

3.4.2 Convolutional Neural Network (CNN)

The gene expression vectors were rearranged into 1D sequences. Convolution operation has been defined as:

$$h_i = \sigma\left(\sum_{k=1}^{K} w_k x_i + b\right) \_\_\_\_\_ (11)$$

Kernel size = 5.
The dimensionality reduction with the max pooling:

$$h_{max} = max\ (h_i) \_\_\_\_\_ (12)$$

Subsequent connexion layers fully connected.

3.4.3 Stacked Autoencoder with Classifier

Encoder compresses input into latent space:

$$z = f(W_e x + b_e) \_\_\_\_\_ (13)$$

Decoder reconstructs input:

$$\hat{x} = g(W_d z + b_d) \_\_\_\_\_ (14)$$

Reconstruction loss:

$$L_{rec} = \|x - \hat{x}\|^2 \_\_\_\_\_ (15)$$

Latent vector z used for classification via softmax layer.
Total loss:

$$L = L_{rec} + \lambda L_{class} \_\_\_\_\_ (16)$$

where L_{class} is categorical cross-entropy.

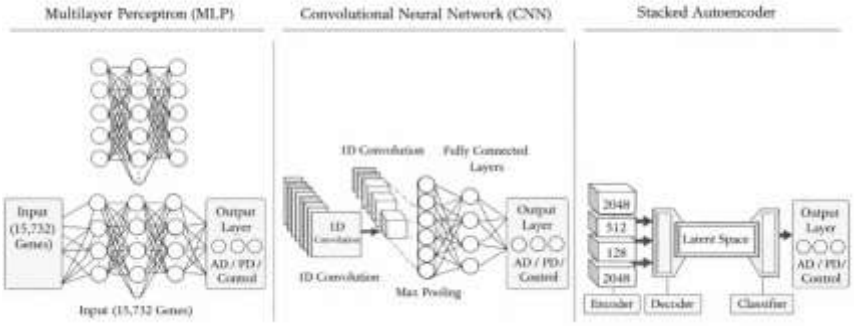

Fig. 2. Comparative Architecture of Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Stacked Autoencoder for Integrative Gene Expression Classification

3.5 Training Strategy

Dataset split:

- 70% training
- 15% validation
- 15% testing

Stratified 5-fold cross-validation was done to achieve the robustness.
In optimization, Adam algorithm was used:

$$\theta_{t+1} = \theta_t - \alpha \frac{m_t}{\sqrt{v_t} + \epsilon} \_\_\_\_\_ (17)$$

Learning rate = 0.001
Batch size = 64
Early stopping monitored validation loss with patience = 10 epochs.

3.6 Model Evaluation

Performance metrics included:
Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

$$Precision = \frac{TP}{TP + FP}$$

Recall:

$$Recall = \frac{TP}{TP + FN}$$

F1-score:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Area Under the ROC Curve (AUC) computed using trapezoidal integration.

3.7 Functional Enrichment and Network Analysis

Notably relevant DEGs were taken through Hypergeometric tested Gene ontology (GO) and KEGG pathway enrichment analysis:

$$P = 1 - \sum_{k=0}^{x-1} \frac{\binom{M}{K}\binom{N-M}{n-k}}{\binom{N}{n}} \_\_\_\_\_ (18)$$

The network of protein-protein interactions (PPI) was built based on the STRING database (confidence score above 0.7). Topological indicators such as degree centrality and betweenness centrality had been calculated to determine hub genes.

4. Results

4.1 Differential Gene Expression Analysis

Differential expression analysis has been used to find 1204 significant dysregulated genes in Alzheimer disease (AD) and 983 in Parkinson disease (PD) compared to controls (|human|) with log 2FC 1 and FDR < 0.05. Intersection analysis has shown the existence of 417 common differentially expressed genes (DEGs), which means that there exists a significant molecular overlap between the two neurodegenerative disorders. Fig 3 represents a Venn diagram that summarises disease-specific and shared DEGs. The percentage of similar genome implies convergent transcriptomic modifications outside of disease-based pathological phenotypes. It was shown that common DEGs were enriched significantly in oxidative phosphorylation (p = 0.001), the pathways of microglial activation (p = 0.01), the cycles of synaptic vesicles (p = 0.01) and the MAPK activation pathways (p = 0.05). These findings depict synchronised dysorientation of mitochondrial bioenergetics, neuroimmune reaction, and synaptic sending mechanisms in both conditions.

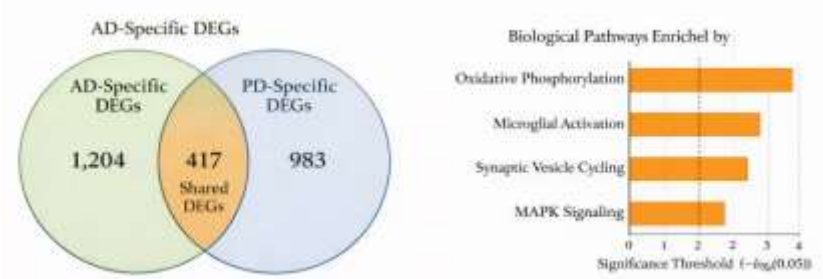

Fig. 3. Venn Diagram of Differentially Expressed Genes Identified in Alzheimer's Disease and Parkinson's Disease Highlighting Shared and Disease-Specific Signatures

4.2 Validation of Data Harmonization

Cross-cohort harmonization was tested by doing a Principal Component Analysis (PCA) prior to and following ComBat batch correction. Before correction, the samples concentrated mainly by data source instead of disease condition, which implies much variance based on batch. In the post-corrected result, batch clustering was significantly smaller, whereas clusterings that were supposed to happen according to disease appeared prominent, as proof of proper elimination of inter-cohort bias. This testing procedure justifies the soundness of downstream integrative modelling.

4.3 Comparative Model Performance

The results of the evaluated neural network architectures in terms of performance measures are summarised in Table 1, whereas ROC curves are shown in Fig. 4.

Table 1. Comparative Performance of Neural Network Architectures for Alzheimer's and Parkinson's Disease Classification

Model	AD AUC	PD AUC	Overall Accuracy
MLP	0.90	0.88	0.89
CNN	0.92	0.89	0.90
Autoencoder + Classifier	0.94	0.91	0.92

Classifier The stacked autoencoder based classifier recorded the highest discriminative result with AUC of 0.94 with AD and 0.91 with PD. These results indicate that the latent feature compression is better than the MLP and CNN models in extracting signals in high-dimensional transcriptomic space.

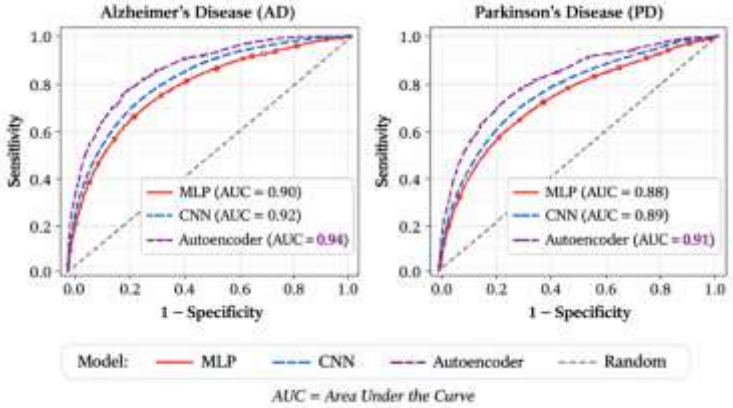

Fig. 4. Receiver Operating Characteristic (ROC) Curves Comparing Neural Network Models for Alzheimer's Disease and Parkinson's Disease Classification

4.4 Latent Feature Representation Analysis

In order to test the separability of learned representations, t-distributed stochastic neighbour embedding (t-SNE) was used on the 128-dimensional latent images obtained using the encoder. As illustrated in Fig. 5, the distributions of AD and PD samples were observed to create groups that were partially different whereas the control samples showed a distinct separation with the disease clusters. The smaller overlap in the latent space as compared to the original feature space implies that nonlinear feature learning and dimensionality reduction are effective.

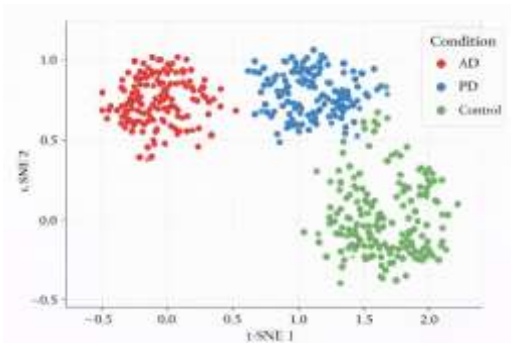

Fig. 5. t-SNE Visualization of Latent Representations Learned by the Stacked Autoencoder for Alzheimer's Disease, Parkinson's Disease, and Control Samples

4.5 Cross-Cohort External Validation

Independent validation on an external GEO dataset of AD and PD gave an AUC of 0.91 and 0.88, respectively, with negligible relative decrease to internal validation. This minimal performance deterioration proves that this model is generalizable and stable to cohort variation.

4.6 Protein–Protein Interaction Network Analysis

The 417 overlapping-DEGs were subjected to the protein-protein interaction (PPI) network analysis using STRING. The network that was obtained consisted of 25 high confidence nodes and 87 edges (confidence score > 0.7). Hub genes determined through degree centrality involved APP, SNCA, MAPT as well as mitochondrial respiratory chain components. Fig. 6 illustrates the PPI network with the highlight of central regulatory nodes. These centres imply that both AD and PD involve convergent interactions among amyloid processing, basically a -synuclein aggregation, and mitochondrial dysfunction pathways.

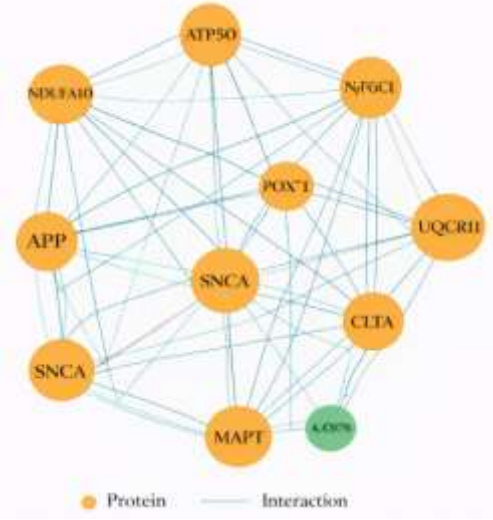

Fig. 6. STRING-Derived Protein–Protein Interaction (PPI) Network of Shared Differentially Expressed Genes in Alzheimer's Disease and Parkinson's Disease

5. Discussion

This paper shows that scaled neural network designs are capable of effectively fitting large-scale integrative transcriptomic data sets on the main neurodegenerative diseases. The proposed framework delivered sound performance to classify the data with respect to the multi-cohort data reports and demonstrated biologically significant common gene signatures through the implementation of a multi-cohort data integrity with nonlinear representation learning. The detection of 417 coinciding DEGs supports the developing indications that AD and PD have comparable molecular pathways, specifically, mitochondrial oxidative phosphorylation and neuroinflammatory provocation. Mitochondrial impairment has been extensively reported as a core cause of neuronal susceptibility in both diseases to facilitate defective ATP generation and augmentative oxidative pressures. Equally, microglial activation and chronic neuroinflammation are becoming the recognised common pathological conditions. The high performance of the stacked autoencoders is an indication that latent feature learning enhances classification by introducing nonlinearity through the elimination of redundancy and non-linear dependencies among genes. Deep architectures also generate hierarchical representations, unlike traditional machine learning methods wherein the choice of features is vital to ensure the function of improving generalisation across cohorts.

This integrative analysis, in comparison to the earlier transcriptomic analyses, which mainly utilised the differential expression or pathway enrichment as a single measure, is complementary to predictive modelling with a biological interpretation. In addition, the cross-cohort validation also resolves reproducibility issues that are usually witnessed during single-dataset analysis. Although these strengths are mentioned, a number of limitations should be admitted. To start with, the use of publicly published datasets creates inconsistencies in sample treatment and completeness of metadata. Second, extrapolative translation is curtailed lack of experimental validation. Third, the heterogeneity of brain regions can affect the expression patterns and should be considered in the region modelling in the future. Further research is needed to expand this framework to include multi-omic integration (e.g. methylation, proteomics, and single cell RNA sequencing data). Additional application of explainable AI methods can also increase biological interpretability.

All in all, the results indicate that integrative transcriptomic analysis, by relying on scalable deep learning, can be useful to reveal similar and disease-specific molecular processes underlying neurodegeneration.

Conclusion

This experiment created and confirmed a large-scale integrative computational paradigm on transcriptomics analysis of Alzheimer disease (AD) and Parkinson disease (PD). The proposed strategy ensured cross-cohort heterogeneity and high-dimensional features complexity of transcriptomic big data by harmonising nine independent GEO cohorts (1,842 samples), provided batch correction, and various deep neural architectures. Data on the differential expression of 1,204 AD-specific, 983 PD-specific and 417 shared dysregulated genes was observed, indicating that molecular overlap occurred significantly in both of the disorders. The functional enrichment and network analysis of protein-protein interactions reported convergent mitochondrial oxidative phosphorylation dysregulation, neuroinflammatory signalling, synaptic vesicle cycling, and protein homeostasis dysregulation. These data support the assumption that the underlying pathogenic mechanisms are common in AD and PD and do not coincide with the classical features of the diseases. The stacked autoencoder-based classifier was found to be the most effective architecture of the studied ones (AUC = 0.94 with AD and 0.91 with PD), proving the usefulness of the latent representation learning in reflecting the nonlinear gene-gene

interactions. The external validation was also done by cross-cohort, thus giving the strongest support of the strength and generalizability of the framework. The main results of the work are as follows: (i) being able to construct a harmonized multi-cohort transcriptomic dataset that can be used to perform cross-disease analysis, (ii) performing the comparison of scalable neural network architectures with consistent preprocessing, and (iii) performing predictive modeling and functional enrichment and network-based biological interpretation. The multi-omics integration of the research by combining epigenomic and proteomic data layers and individual-cell transcriptomic resolution should be incorporated in future works to further define heterogeneity among cells. Moreover, explainable artificial intelligence methods might be incorporated, which could make biological interpretability and translational relevance be improved. Identified hub genes will need experimental validation that will be important to ensure mechanistic implication and clinical potentiality. All in all, the paper shows that scalable deep learning-based integrative transcriptomic modelling can improve cross-disease molecular characterization and a systems-level insight into neurodegenerative diseases.

References

1. De Strooper, B., & Karran, E. (2016). The cellular phase of Alzheimer's disease. Cell, 164(4), 603–615.
2. Eraslan, G., Avsec, Ż., Gagneur, J., & Theis, F. J. (2019). Deep learning: New computational modelling techniques for genomics. Nature Reviews Genetics, 20(7), 389–403.
3. Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., … Nowé, A. (2013). Batch effect removal methods for microarray gene expression data integration: A survey. Briefings in Bioinformatics, 14(4), 469–490.
4. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2010). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics, 28(6), 882–883.
5. Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. Briefings in Bioinformatics, 18(5), 851–869.
6. Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkmann, J., … Lang, A. E. (2017). Parkinson disease. Nature Reviews Disease Primers, 3, 17013.
7. Wang, M., Roussos, P., McKenzie, A., Zhou, X., Kajiwara, Y., Brennand, K. J., … Zhang, B. (2021). Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. Genome Medicine, 13, 66.
8. 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.
9. Grubman, A., Chew, G., Ouyang, J. F., Sun, G., Choo, X. Y., McLean, C., … Polo, J. M. (2019). A single-cell genomic atlas of the human Alzheimer's disease brain. Cell Reports, 27(9), 2997–3008.
10. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface, 15(141), 20170387.*
11. Yuan, Y., Bar-Joseph, Z., & Yang, Z. (2020). Deep learning for inferring gene relationships from single-cell expression data. Proceedings of the National Academy of Sciences, 117(44), 27151–27158.*
12. Johnson, E. C. B., Carter, E. K., Dammer, E. B., et al. (2020). Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. Nature Medicine, 26(5), 769–780.*
13. Srinivasan, K., Friedman, B. A., Etxeberria, A., et al. (2020). Alzheimer's patient microglia exhibit enhanced aging and unique transcriptional activation. Cell Reports, 31(13), 107843.*

14. Tan, J., Hammond, J. H., Hogan, D. A., & Greene, C. S. (2016). ADAGE-based integration of public gene expression data identifies new transcriptional programs in Pseudomonas aeruginosa. Genome Biology, 17, 160.
15. Way, G. P., & Greene, C. S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. Pacific Symposium on Biocomputing, 23, 80–91.
16. Sweeney, M. D., Sagare, A. P., & Zlokovic, B. V. (2018). Blood–brain barrier breakdown in Alzheimer disease and other neurodegenerative disorders. Nature Reviews Neurology, 14(3), 133–150.*
17. Dugger, B. N., & Dickson, D. W. (2017). Pathology of neurodegenerative diseases. Cold Spring Harbor Perspectives in Biology, 9(7), a028035.*