

Kritka Sharma Mtech V2

by Kritika Sharma

Submission date: 22-May-2026 02:10PM (UTC+0530)

Submission ID: 2966989669

File name: Kritika_Sharma_Major_thesis.docx (1.43M)

Word count: 8808

Character count: 54367

ABSTRACT

The current landscape of Parkinson's Disease (PD) biomarkers remains limited by inconsistent reproducibility, the limitation of traditional analytical frameworks, and cohort heterogeneity to connect statistical findings with underlying biological mechanisms. Explainable artificial intelligence (xAI) based methods open opportunities in biomarker discovery by combining predictive accuracy and biological explanation, thereby bridging the reproducibility and mechanistic limitations of conventional single-cohort studies. In this study, seven transcriptome datasets from substantia nigra tissue retrieved from the NCBI GEO database were analysed to identify robust and biologically coherent PD-associated biomarkers. After the preprocessing step, a total of 433 commonly expressed genes were utilized in performing ML modelling techniques. Among four ML classifiers, XGBoost achieved the highest value for the area under the receiver-operating characteristic curve in determining the accuracy of its diagnostic tests, registering an AUC of 0.86. The SHAP value demonstrated that TARDBP, EIF4E, and SHQ1 were highly ranked genes and emerged as among the most influential biomarkers. Functional enrichment analysis continued to produce significant activation of RNA transport, oxidative phosphorylation, and prolactin signalling as prominent pathways in PD. The prolactin signalling established a reproducible and mechanistically interpretable molecular signature that modulates JAK/STAT, PI3K/AKT, and MAPK signalling pathways that are directly associated with microglial activation and oxidative and inflammatory stress responses. The study also suggests the capability of xAI to uncover endocrine-linked regulatory pathways that are often overlooked in traditional transcriptomic analyses, offering promising biological insights and potential therapeutic.

Keywords: Parkinson's disease, ML, SHAP, substantia nigra, biomarkers, prolactin signalling

CHAPTER 1

²INTRODUCTION

Parkinson's disease (PD) is a progressive neuro-degenerative disorder that is accompanied by the progressive death of dopaminergic neurons within the substantia nigra pars compacta (SNpc) region of the brain [1]. The onset of PD is generally recognized by the presence of characteristic neuro-degenerative symptoms such as tremors, bradykinesia, and rigidity which therefore leads to a significant loss of SNpc neurons before the onset of treatment [2].

Along with the clinical manifestations, Parkinson's disease is also associated with a group of molecular mechanisms such as alpha-synuclein protein clumping, mitochondrial malfunction, oxidative stress, and neuroinflammation, which are happening simultaneously [3]. These activities of one another lead to progressive neuronal demise and functional deterioration. Although significant progress has been made in the field of neuroimaging and clinical evaluation methods, the diagnosis of the disease is mostly made at the time of the appearance of the more advanced features of the disease, by that time major neuronal death has already taken place. This underlines the tremendous demand for dependable molecular markers that can identify PD during its earliest stages and allow for the introduction of effective therapeutic measures on time.

1.1 Significance of the Substantia Nigra in Transcriptomic Biomarker Discovery

The substantia nigra is the most informative region in the case of PD as it is the first area to show neurodegeneration [4]. Changes in the transcriptome of SN tissue sometimes occur a long time before the manifestation of symptoms, thus, it is a perfect source for the identification of early biomarkers [5]. Nevertheless, biomarker reproducibility is still at a low level despite numerous studies that have been carried out on SN gene expression. Various factors such as differences in sample preparation, demographic makeup, microarray platforms, and analytical methods can cause variability between studies [6]. These inconsistencies reinforce the need for integrating multiple datasets into a harmonized multicohort framework.

Selective vulnerability of SNpc damages resulting from dopamine oxidation, high metabolic activity of SNpc neurons, and increased susceptibility to oxidative stress is the main reason behind their vulnerability. Because these neurons are exposed to very high amounts of reactive oxygen species and are very heavily loaded with mitochondria, they are not only vulnerable but are among the most vulnerable to degeneration. Therefore, changes in gene expression in this area of the brain give us important clues about molecular changes coinciding with the onset of symptoms in a patient. Nevertheless, to make the results less scatter and for the identification of the most reliable biomarkers to be more efficient, the combination of several studies through the use of integrative methods is required.

1.2 Role of xAI in Modern Biomarker Research

Bioinformatics has been the key factor in the revolution of the study of complex diseases. Nowadays, computational tools can be used for large-scale analysis of high-throughput transcriptomic datasets to detect subtle and nonlinear gene expression changes. It has come to the fore with the aid of the machine learning (ML) methods including random forests [7], support vector machines (SVM) [8], logistic regression [7], and XGBoost [9] which have been able to successfully discover biomarkers in cancers, metabolic diseases, and neurodegenerative conditions [10]. In doing so, these models can manage data with many variables and find connections which the traditional statistics frequently overlook. Still, conventional ML systems that have a strong predictive power provide little insight into the rationale which led to the chosen genes.

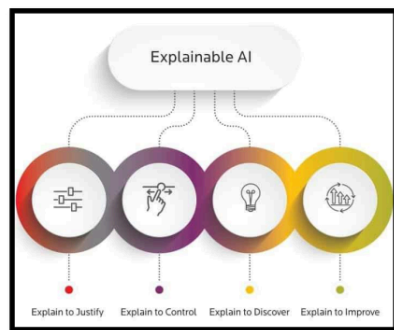


Figure 1.1 Functional Framework of Explainable AI

The lack of transparency in machine learning is problematic concerning biological explanation and applicability to clinical practice. Explainable artificial intelligence (xAI) can serve as a remedy, where the reasoning behind model decisions becomes apparent. By means of xAI tools ¹⁶ such as SHAP (SHapley Additive Explanations) and LIME, the researchers are able to quantify how much each gene contributes to model predictions [11]. Therefore, not only does this enhance the confidence of the researchers in the computer results, but it also ensures the alignment of the results with biological mechanisms. For example, research in oncology and metabolism showed that the application of ML in combination with xAI provides not only accurate predictions, but also important biomarker profiles, thus making xAI an invaluable tool for PD studies.

Conventional machine learning models are black boxes that do not reveal their decision-making process, while xAI works in a transparent way, by indicating which features of the input have the most impact. This, in turn, helps to interpret them from a biological point of view. This is very important in biomedical research, where besides high predictive power, it is also necessary to figure out the biological mechanisms [12].

1.3 Necessity of Robust Multicohort Dataset Integration

Despite the high analysis potential of ML and XAI, a high-quality, standardizable dataset is necessary. This is because gene expression datasets on publicly available platforms tend to be inconsistent with respect to experiments, microarray platforms, the set of genes, as well as demographics [13]. Without careful preprocessing, merging such datasets can introduce batch effects and inconsistent signals. Therefore, harmonization including gene intersection, normalization, and batch correction is essential for extracting biologically valid patterns [14]. The combining of several GEO datasets from substantia nigra can, therefore, serve as an avenue for developing a solid platform for reproducible biomarker analysis.

Single gathering studies are regularly plagued by the problem of small samples and cohort-specific biases, which negatively affect the applicability of the results to the general population. On the other hand, combining multiple cohorts increases the power of statistics, shores up reliability, and opens the door to detecting biomarkers that are common

to different types of populations. After the merging of several data sets, there is a chance of downplaying the noise that is typical for any particular set and at the same time getting to know firm molecular indicators of the course of the disease.

1.4 Research Gap

Despite the progress made in transcriptomic analysis and application of machine learning, the works combining multicohort datasets with explainable AI to get biologically interpretable biomarkers in Parkinson disease are still scarce. In addition, the role of endocrine-linked signalling routes in the substantia nigra transcriptomic networks has not been systematically studied, which reveals a major shortage in the present research.

1.5 Objectives

This research is concerned with the initial stage of PD biomarker research, which is developing a harmonized multi-cohort gene expression database from human substantia nigra samples [13]. The key objectives are:

- a) Retrieving, Curating, and Integrating Seven GEO Datasets For Substantia Nigra Transcriptomes
- b) To conduct rigorous preprocessing, such as normalization, batch correction, and gene intersection, to produce a common expression dataset.
- c) To develop a high-quality, standardized dataset to be used for machine learning tasks involving the prediction of Parkinson's disease biomarkers via AI.
- d) Establishing a foundation for explainable AI-based analysis to enable biologically interpretable insights into Parkinson's disease

By establishing this robust foundation, the study supports future development of transparent, scientifically grounded biomarker models for early PD detection.

CHAPTER 2

LITRATURE REVIEW

2.1 Historical Progression of Parkinson's Biomarker Research

For a long time, Parkinson's disease (PD) was recognized mainly by clinical symptoms and confirmed by neuropathological changes after the patient's death. Thus, early research emphasized motor dysfunction and the apparent ¹⁸ loss of dopaminergic neurons in the SNpc region [15]. As studies progressed, it was acknowledged that the disorder is already at its advanced stage when the classical symptoms become visible. This finding redirected the focus of research to molecular and transcriptomic alterations that predate clinical symptoms. With the enhancement of transcriptomic technologies, scientists began to examine gene expression alterations in the substantia nigra with an intention to find early biomarkers that would allow the detection of PD prior to the occurrence of motor symptoms [10]. These investigations mostly indicated that the changes occurred in the pathways related to mitochondrial dysfunction, RNA metabolism, immune regulation, and oxidative stress, thus, putting transcriptomics forward as a feasible approach to studying the initial stages of PD [16], [17].

While transcriptomic profiling of the substantia nigra is informative, there were significant constraints in the pioneering studies. The individual datasets were limited in size, diverse in demographic composition, and the experiments had been carried out using different microarray platforms [18]. Consequently, genes that were proposed as biomarkers in one study would frequently fail to be confirmed in other studies [19]. This problem of inconsistency was a major factor that led to the concept of multicohort integration, whereby a number of publicly available datasets are merged to increase statistical power and biological reproducibility.

2.2 Rise of Bioinformatics and ML in Biomarker Discovery

Bioinformatics, which relies heavily on computational biology, has increasingly been at the core of the research of disease biomarkers. The transcriptomic data involve tens of thousands of genes, and to analyze such highly dimensional data, advanced techniques are needed that

can capture subtle and complex patterns. Machine learning (ML) models like Support Vector Machines [8], Logistic Regression [7], Random Forests [20], and XGBoost [9] were rapidly recognized as potent instruments in the identification of gene signatures related to diseases [10], [21]. The different studies in the areas of cancer, metabolic disorders, and neurological diseases, have been evidences that ML models are better than conventional statistical methods because the former can discover deeper molecular patterns and can generate highly accurate classifications [10].

Specifically, in PD research, machine learning provides a way of finding gene combinations that together show the earliest neurodegenerative changes; thus, it is perfectly capable of biomarker discovery [1], [22].

2.3 XAI as a Solution to ML Interpretability Limitations

In spite of the excellent predictive capability of ML models, their restricted interpretability has been the main obstacle. Traditional ML techniques are frequently regarded as "black boxes," as they come up with results without giving a clear indication of the factors (genes) that led to the outcome [23]. The absence of transparency lowers the confidence level, particularly in biomedical research, where the biological plausibility is a must. xAI was developed as a ground-breaking solution to this problem. Methods like SHAP and LIME give the scientists the opportunity to see and measure the extent to which each gene is responsible for the model prediction [23], [13].

In fact, SHAP has become one of the main features of interpretable biomarker discovery, facilitating the production of gene rankings that are stable and make sense biologically [11]. The XAI approach acts as a link between computational prediction and biological understanding, the machine learning results thus being able to be checked scientifically and then implemented as clinically valuable insights [21].

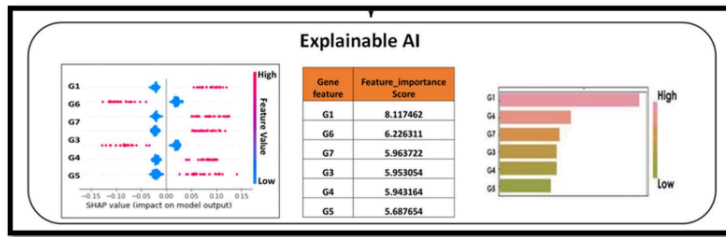


Figure 2.1 SHAP-Based Explainable AI Visualization Showing Gene-Level Feature

The effectiveness of both ML and XAI depends heavily on the quality of the underlying dataset [10]. The publicly available transcriptomic data are characterized by differences in experimental conditions, gene coverage, population characteristics, and measurement scales. If appropriate steps are not taken to prepare the data, the combination of such datasets can result in technical noise that misleads biological interpretation [24]. Recent research works have emphasized that it is necessary to carry out dataset harmonization through batch correction, normalization, gene intersection, and outlier removal before the production of consistent biomarkers can be achieved. By combining several GEO datasets from the substantia nigra tissue and following strict preprocessing protocols, investigators are enabled to create a single and biologically reliable expression matrix that can be used for machine learning and explainable analysis [14].

Table 2.1 Recent XAI-based biomarker studies supporting the present work

S/No.	Key Focus	Main Findings	Relevance to Present Study	Reference
1.	XAI-guided mitochondrial gene network in type 2 diabetes	Integration of machine learning with SHAP identified key mitochondrial genes and pathways driving metabolic dysfunction.	Demonstrates how combining ML with SHAP can uncover biologically biomarkers from high-dimensional gene expression data, directly supporting our XAI strategy.	[11]

2.	PBMC gene-expression biomarkers for breast cancer using XAI	XGBoost with SHAP revealed a small set of blood-derived genes that accurately distinguish cancer from controls.	Provides a clear precedent that XAI can derive interpretable, clinically relevant transcriptomic biomarkers from peripheral samples.	[13]
3.	Multi-cancer circulating miRNA classifier with explainable AI	An i-Biomarker CaDx tool used XAI to identify and interpret miRNA signatures for early multi-cancer detection.	Shows that explainable models can support liquid-biopsy biomarker panels, reinforcing our aim to build interpretable PD classifiers.	[3]
4.	Systematic meta-survey of XAI challenges and opportunities	Reviews current XAI methods, limitations, and future directions across domains.	Provides methodological justification for using SHAP/LIME and highlights the importance of transparency in medical ML models like ours.	[23]
5.	Survey of medical XAI applications	Summarizes how XAI improves trust, safety, and adoption of AI in clinical settings.	Supports our motivation to use XAI so that discovered PD biomarkers are not only predictive but also interpretable for eventual clinical translation.	[25]

CHAPTER 3

METHODOLOGY

This research utilized a strict, multi-tier analytical pipeline to select, preprocess, and coordinate gene expression profiles from seven publicly available GEO datasets of substantia nigra (SNpc) tissue [1]. The broad objective was to put an end to the flaws of single-cohort studies like unstable biomarkers, batch variability, and platform-specific artifacts and hence establish a solid, reproducible basis for PD biomarker identification. All parts of the method were deliberately planned to keep biological consistency while allowing for technical comparability among datasets and making the combined expression matrix ready for subsequent ML and xAI analysis [10].

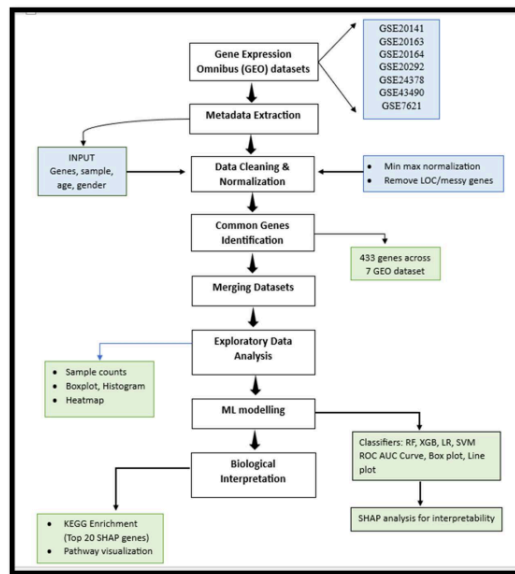


Figure 3.1 End-to-end pipeline for GEO dataset integration, ML modelling, and SHAP-based biomarker interpretation

The general approach structure was built and presented in a way to combine the preprocessing of transcriptomic datasets, ⁵ the application of machine learning algorithms for classification, along with xAI approaches into a single pipeline for efficient biomarker discovery. This approach provided the advantages of accurate predictions with insights of biological meanings at the same time, and fostered the reproducibility across multi-cohort transcriptomic studies. The study pipeline was designed step by step starting from dataset importing and pre-processing scaling set of exploratory data analysis, ML algorithms for disease predictions, through to shapely additive explanation with barrier interpretability for molecular biomarkers.

3.1. Data curation

Seven GEO datasets (GSE20141, GSE20163, GSE20164, GSE20292, GSE24378, GSE43490, and GSE7621) were located in the NCBI GEO repository based on rigorous inclusion criteria. Only datasets that profile human substantia nigra tissue and have well-annotated Parkinson's disease vs. healthy control samples were taken into account. This selection guaranteed biological specificity for the brain region that is most affected during PD early neurodegeneration.

Table 1 provides a summary of the datasets. To facilitate uniform annotation, each probe was converted to gene symbols, and only high-confidence probes that ended with the “_at” suffix were kept. The “_at” probes in Affymetrix arrays are known to hybridize uniquely to well-defined genes, thus cross-hybridization is minimized and the biological accuracy is increased [26].

Finally, cross-platform intersection filtering was applied. Genes appearing in all seven datasets were considered, resulting in a list of 433 common genes. This guaranteed that each gene analyzed was equally measured in all datasets, thus avoiding platform bias and allowing for proper integration.

Another aspect of the dataset curation process included the inspection of sample annotations, verification of the consistency of metadata, and confirmation of the platform compatibility among the combined cohorts in order to maintain the biological comparability of all samples in the machine learning analysis. In case the samples were not annotated sufficiently and/or there was no evidence of disease and/or tissue origin of the samples, they were excluded

from the dataset. Considering that the technical variance associated with transcriptome profiling experiments across independent labs cannot be avoided, the use of strict inclusion criteria was applied in order to control these noise factors. Additionally, the universally expressed genes across all platforms were used as an inclusion criterion in order to focus on the genes which were stable cross-platform.

3.2 Preprocessing, Normalization, and Quality Control

The raw expression matrices for every dataset were subjected to a standardized preprocessing pipeline [8]. First the expression values were normalized with min-max scaling across each dataset, chosen for its ability to reduce batch effects while retaining biological signals.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

¹ where X is the raw(original) value X_{min} is the minimum value of feature, X_{max} is the maximum value of feature, and X_{scaled} is the normalized value between 0 and 1 [27]. This transformation scaled all gene expression values into the range 0 to 1, ensuring comparability across platforms.

Normalization is yet another critical preprocessing task since the gene expression values computed using various microarray technologies show significant heterogeneity in terms of the scale and distribution of the signal, noise, and dynamic range. Without normalization, the learning models would be learning trends from the technical aspect rather than from the actual biological process leading to the biasing of the model and instability of the biomarker. In this study, min-max normalization was used due to its capability to directly constrain the underlying distribution of the values, which allows for consistent scaling of the data obtained from different microarray technologies. Normalization ensures improved multicohort transcriptomic data compatibility, allowing the ML models to learn features reliably from the multicohort data.

3.2.1 Heatmap of Top 50 Most Variable Genes

The heatmap shows that the normalized samples still have significant variability and biological structure from different subjects which is a good indication that the technical noise

has been minimized. The high-variance genes are used to highlight the diversity at the sample level and thus ensure that the subsequent ML models get representative input features.

Detecting genes with high variability is critical since these tend to be the most informative features for clinical classification, reflecting biologically meaningful variations tied to disease time course. Heatmap also offers the advantage of convenient visual assessment of unusual sample clustering, batch effects, or expression anomalies across samples. Preservation of this transcriptomic heterogeneity following normalization suggested the initial workflow was capable of eliminating undesirable technical factors without eroding disease-related biological variation, which is required for effective machine learning-based biomarker identification.

3.2.2 Boxplot of Normalized Gene Expression Values

The gene-level boxplot depicts that gene expression ranges after normalization and imputation have been consistently aligned across all genes and samples. Outliers are barely visible and consistently appearing, thus the preprocessing can be considered stable.

Box plot visualization is an effective quality control approach that can be used to establish whether the normalized transcriptomic data are similar in their distributions across the samples analyzed. The consistency in the median and interquartile range observed across samples indicates that normalization process helped to reduce the bias arising from the use of various experimental platforms. Even though some outliers have been detected after preprocessing, their fewness suggests that the biological variability has been maintained in the normalization process of data. Over-normalization of the data (which is possible in PCA normalization techniques) will cause disease signatures to be lost. On the other hand, under-normalization of data causes the introduction of technical bias during modeling.

3.2.3 Density Plot of Min–Max Normalized Values

The density plot is going to show that all the datasets are moving toward the same expression distributions after scaling which is a confirmation of successful cross-platform harmonization. After normalization and QC, each dataset was subset to the 433 intersecting genes and aligned in identical gene order.

Density distribution analysis also revealed the impact of preprocessing within the integrated cohorts as the distributions of gene expression could be seen to overlap. This overlap corresponds with the expectation that the effects of technical variation had been address across multiple independent data sets to allow them to be integrated into the same expression matrix for analysis. The ability to combine data obtained from various distinct research articles to constitute a single integrated cohort greatly enhances the statistical strength and reliability of the eventual cohort and this integrated matrix formed a biologically representative and numerically consistent platform for application of machine learning and xAI algorithms.

3.3 Classification, ⁵Model Training, and Evaluation

The normalized data ⁵was divided into training and testing data (70:30) to maintain the same percentage of diseased and control samples. The data split process for the models was a significant factor in circumventing overfitting and underfitting in the ML models. The ML models, like Random Forest [28], ¹³XGBoost, Logistic Regression (LR), and Support Vector ¹³Machine (SVM), were chosen for their respective ability to address high-dimensional, complex biological information. ⁶The accuracy, precision, sensitivity, F1-score, and area ⁶under the receiver operating curve (AUC-ROC) of the model performance were measured. ⁷Confusion matrices were generated from ⁷each model to analyse the data for true positives, true negatives, false positives, and false negatives, respectively.

Having used multiple ML classifiers for comparison between models, we learned about ¹⁹the ¹⁹strengths and weaknesses of using different architectures. We decided to use multiple baseline classifiers including logistic regression because it is simple and easy to model linear relationship; Random Forest and XGBoost because it can model the nonlinearity in the transcriptome data; Support Vector Machine because it has high efficiency in finding the optimal decision boundary in a small sample size. We used fivefold cross-validation in our experiment to enhance the robustness of the classifier without risking overfitting. In addition, by using the fivefold cross-validation, we ensure that all the samples will be trained and tested.

3.4 Explainable AI-mediated hub gene identification

After preprocessing of the gene expression, the top performing ML model was retrained on the preprocessed training set to generate SHAP values [7]. SHAP values were calculated using the TreeExplainer algorithm applied to the retrained XGBoost classifier, and the resulting SHAP scores were used to rank the top 20 most influential genes. Feature importance was represented graphically by the use of bar plots and dot summary plots, offering an intuitive sense of which genes most affected the classification [28].

We incorporated SHAP-based explainability in our methods to counteract the lack of interpretability seen with black box predictive modelling. The modeling structure of SHAP relies on ⁴cooperative game theory to estimate the average marginal contribution of each feature by approximating Shapley values by averaging across all existing feature subset (or model) combinations. Its flexibility to reliably decipher transcriptomic biomarkers makes SHAP suitable for biological interpretation. Besides predictive modelling tasks, the resulting explainable AI structure also lends more confidence to the biomarkers found because it combined model development with biological insight.

3.5 Functional enrichment analysis

The top 20 genes ranked by SHAP were analyzed for functional enrichment using ¹²the Kyoto Encyclopaedia of Genes and Genomes (KEGG) 2021 human ¹⁴database [29]. Fisher's exact test with FDR correction (Benjamini–Hochberg) identified significantly enriched pathways (adjusted $p < 0.05$) [30]. The KEGG bubble plot and horizontal bar plot demonstrate consolidated enrichment score and gene contribution. The plots illustrate the top 10 enriched pathways ranked by gene percentage, with $-\log_{10}(p\text{-value})$.

The biological relevance of the prioritized genes in terms of their molecular importance was achieved by conducting the functional enrichment analysis. The pathway enrichment analysis helps in understanding the pathways and processes that are enriched in the biomarker selections. The application of the FDR correction reduces the probability of any false pathway associations owing to multiple comparisons. Leveraging the explainability of AI through pathway analysis provides a systems biology approach to the biomarkers discovered through the hypothesis-free AI approach.

3.6 Literature Validation

To assess biological plausibility, a literature validation process was conducted on PubMed databases. Peer-reviewed literature supporting predicted hub genes and biological pathways related to PD on the basis of experimental evidence was considered for the literature validation process.

CHAPTER 4

RESULT AND DISCUSSION

This research stringently tested the hypothesis that integrative, multi-dataset ML of SNpc gene expression can generate strong, biologically significant biomarkers for diagnosing Parkinson's disease.

4.1 Multi-Dataset Integration and SNpc Cohort Characteristics

The present study set out to test the idea that combining multiple publicly available GEO datasets from substantia nigra pars compacta (SNpc) tissue could create a unified, biologically coherent transcriptomic matrix suitable for discovering PD biomarkers using machine learning and eXplainable AI [31], [21].

Table 4.1 The seven Gene Expression Omnibus (GEO) datasets used in the multicohort analysis of biomarkers for Parkinson's disease.

GEO Accession ID	Year	Country	Platform	Tissue	Sample
GSE7621	2007	USA	GPL570	SN	PD-16; HC-9 ³
GSE20141	2010	USA	GPL570	SN	PD-10; HC-8
GSE43490	2013	Brazil	GPL6480	SN	PD-15; HC-8 ³
GSE20163	2010	USA	GPL96	SN	PD-8; HC-9
GSE20292	2010	USA	GPL96	SN	PD-11; HC-18 ³
GSE20164	2010	USA	GPL96	SN	PD-6; HC-5
GSE24378	2012	USA	GPL570	SN	PD-8; HC-9

The seven Gene Expression Omnibus (GEO) datasets used in the multicohort analysis of biomarkers for Parkinson's disease. The datasets range from 2007-2013 and include 130 human substantia nigra samples (67 Parkinson's disease cases, and 63 neurologically healthy controls) from the United States and Netherlands. Two microarray platforms, GPL570 and GPL96, were used for cross-platform verification of the discovered biomarkers. The use of several independent cohorts in the integration improves the generalizability and robustness of the identified gene expression signatures.

Although the seven included datasets originate from different years, laboratories, and microarray platforms, they all examine the same vulnerable brain region, making them ideal candidates for integrative analysis.

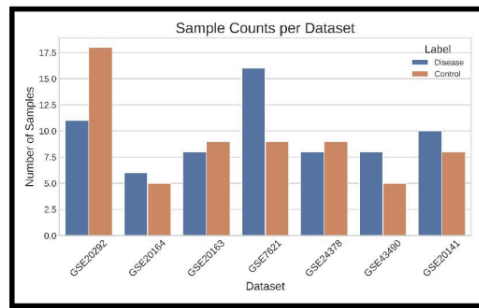


Figure 4.1 Sample distribution across the seven GEO datasets

Figure 4.1 illustrates the sample distribution across the datasets. The strict inclusion criteria that were applied ensured that each dataset had PD versus control annotations clearly and that samples were only SNpc-derived. The final integrated cohort comprises 130 samples, 67 individuals with PD, and 63 healthy controls (HC). Through systematic preprocessing that refined probes, removed ambiguous annotations, and restricted the analysis to genes present across all platforms, 433 genes were reliably measured. This curated gene set is the standard feature matrix for the downstream machine learning and SHAP-based interpretation [13].

Integrating multiple cohorts for the transcriptomic data improved the biological relevance of this study because biomarkers found to be consistently perturbed in independent data sets

are much more likely to represent actual disease-related molecular changes than biomarkers identified in one cohort only, particularly given the small sample size used in most single-cohort studies and their low statistical power due to this reason. Pooling samples obtained under different biological and technical conditions allowed for improving the robustness of the transcriptomic data via an increase in the number of samples analyzed. This is especially important for neurodegenerative diseases like PD, in which transcriptomic alterations are usually very subtle and complicated by other factors.

The capacity of the method to align all seven independent GEO data sets suggests that biological signals from the transcriptome can be seen despite high cross-platform heterogeneity. Moreover, the fact that the reproducibility of common molecular signals has been detected further confirms that the selected markers have biological relevance, and because only the genes present in all seven data sets have been considered, there is no possibility of platform-specific factors affecting the selection of biomarkers. The use of common genes also ensures that the biomarkers will have higher translational value in future experiments conducted on PD than they would otherwise.

4.2 Diagnostic Visualization of Gene Expression Quality and Distribution

In order to investigate whether the preprocessing pipeline had preserved biologically relevant patterns and at the same time minimized technical variation, we generated several diagnostic plots. The Heatmap clearly shows expression gradients across the 130 samples.

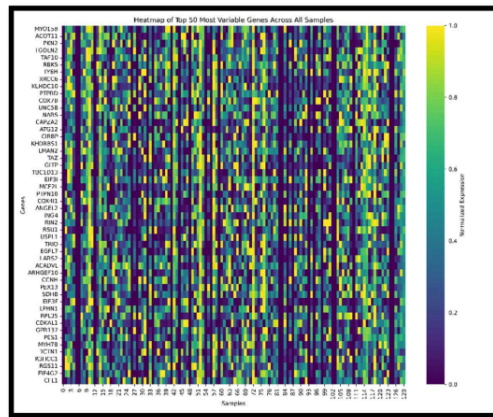


Figure 4.2 Heatmap of the top 50 most variable genes

The distinct regions of higher and lower expression demonstrate that biologically meaningful diversity, which is necessary for the detection of PD-associated transcriptional shifts, has been preserved [32].

The heatmap validates the normalization and harmonization pipeline by alleviating non-biological variation, but maintaining a heterogeneity in expression as evidenced by the expression gradients across samples. Despite the reduction in non-biological variation, the resulting integrated expression set can still faithfully distinguish different pathological types of PD as indicated by the expression gradients across samples. A zero-inflation between dataset labels suggests that the preprocessing has alleviated the effect of pathogenically irrelevant differences due to variations in lab, platform, and sample collection and preparation procedures. This characteristic is important as how one derives variance to be explained influences the ability of machine learning algorithms to learn valid, disease-specific molecular profiles.

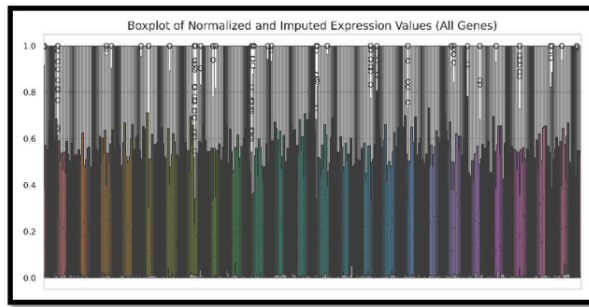


Figure 4.3 Boxplot of Min–Max normalized gene expression

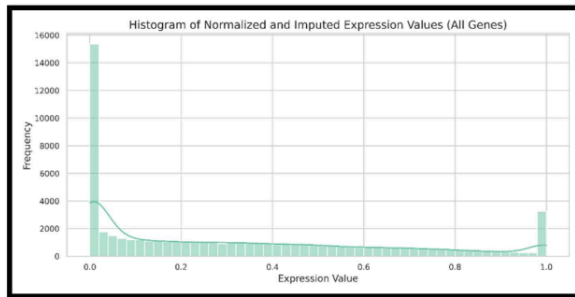


Figure 4.4 Histogram of normalized and imputed gene expression values

Figure 4.3 and 4.4 demonstrates the consistent distribution of expression values across all samples. The closeness of median values and the similarity of interquartile ranges show that normalization has effectively put datasets on a common scale. A small number of outliers are still present, but they are within the range of expected biological variability and indicate that normalization has not artificially compressed the data.

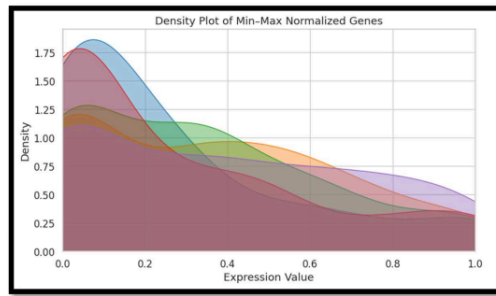


Figure 4.5 Density plot of Min–Max normalized values

In addition to what was shown in figure 4.3, figure 4.5 acts as another piece of evidence for dataset uniformity. The smooth, overlapping curves all contained within the expected 0–1 range show that Min–Max scaling successfully corrected differences in the original signal intensity across platforms [27]. The natural left-skewed distribution reflects typical transcriptomic behaviour, suggesting that the biological structure of the data has been preserved. Together, these visual assessments demonstrate that the curated dataset maintains transcriptomic integrity and cross-dataset comparability.

4.3 Transcriptomic Validation and Feature Selection for Explainable AI Modeling

The heatmap further validated the robustness of the dataset, illustrating that the disease-control separation was based on biological expression, not batch effect. The volcano plot validates that the 51 genes were upregulated were ranked in the top 5 as MED13L, PRELP, EIF3F, ARHGAP29, and RSL1D1, while the 6 genes were downregulated ranked on the top 5 as PLD3, EIF4E, PIN1, EIF3I, and TRAPPC2L, and this distribution formed a characteristic PD-associated expression pattern with multiple genes surpassing $\log_2FC > 1$ and $-\log_{10}(p) > 1.3$ thresholds.

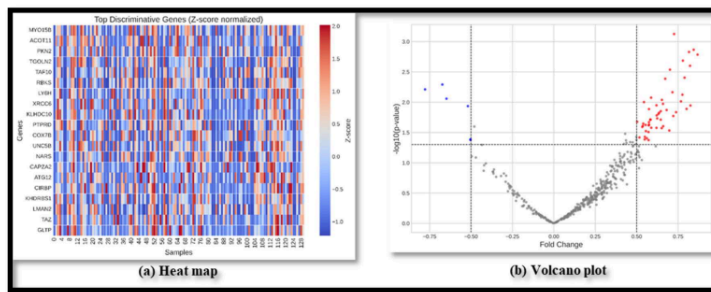


Figure 4.6 Heatmap of top 20 genes and (b) Volcano plot of 433 common genes

This taken from the heatmap is evidence that Parkinson's disease-related molecular changes identified are sufficiently unique to distinguish diseased versus healthy samples at the gene expression level. Such separation can lend credence to the hypothesis that the integrated SNpc transcriptomic matrix may contain biological information to support the use of predictive machine learning models. Really other gene dysregulation is cluster reveals the possibility of other coordinated molecular pathways proposed to be directly involved in neurodegeneration, mitochondria malfunction and anomalies in RNA metabolism. The volcano plot also shows the consistent presence of upregulated and downregulated transcriptomic variation throughout the integrated cohort.

4.4 ML Model Performance

The integrated multi-cohort datasets, after preprocessing, served as the input for downstream ML analyses.

Machine Learning Details

Original data shape: (130, 435)

Training features shape: (91, 435)

Testing features shape: (39, 435)

Training labels shape: (91,)

Testing labels shape: (39,)

Number of gene features (after potential encoding): 433

Following rigorous preprocessing, including outlier removal, min-max normalization, and cross-platform harmonization, four ML classifiers were built and tested with five-fold cross-validation [33].

Table 4.2 Summary of ML classifier performance

ML Technique	¹⁷ Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Logistic Regression	73.08	71.43	76.92	74.07	85.21
Random Forest	65.38	66.67	61.54	64	74.26
SVM	57.69	57.14	61.54	59.26	65.09
XG Boost	76.92	73.33	84.62	78.57	86.39

The above Table 4.2 summarizes the cross-validated performance of the four classifiers against the 433 gene signature. XGBoost produced the best AUC (86.39%) with balanced recall and precision (71.43%, 76.92%), whereas XGBoost provided the best accuracy (76.92%) and recall (84.62%). This complementary profiling indicates that tree-based and linear models capture distinct yet convergent transcriptomic signatures relevant to PD.

Furthermore, the results of machine learning classifiers indicate that the integrated transcriptomic dataset contains quite a large amount of biologically meaningful information that can be used for distinguishing between patients with PD and healthy individuals with high accuracy. All considered machine learning models showed different levels of classification accuracy, but XGBoost turned out to have the highest value of this metric, which indicates its potential usefulness for modeling non-linear relationships in the transcriptomic space related to neurodegeneration (as well as other complex systems). The highest recall of the XGBoost classifier implies that it shows higher sensitivity to detecting biomarkers associated with PD, which is useful for biomarker identification tasks.

4.5 Model Discrimination and Calibration

Figure 4.7 (a, b, c) showed ROC-AUC plot, model wise metric distribution, and comparative performance across metrics for each dataset to evaluate the discriminatory power of the model.

Logistic regression performed consistently in terms of recall and AUC. On the other hand, XGBoost produced high F1 score, which implies that the model strikes the perfect balance between precision and recall. From the combined analysis, we deduce that the XGBoost classifier is more suitable for risk stratification than boosting algorithms, which are better suited for balancing in reality. XGBoost classifier's confusion matrix shows 9 true negatives (controls classified correctly), 4 false positives, 2 false negatives, and 11 true positives (PD classified correctly), indicating 84.6% sensitivity and 69.2% specificity.

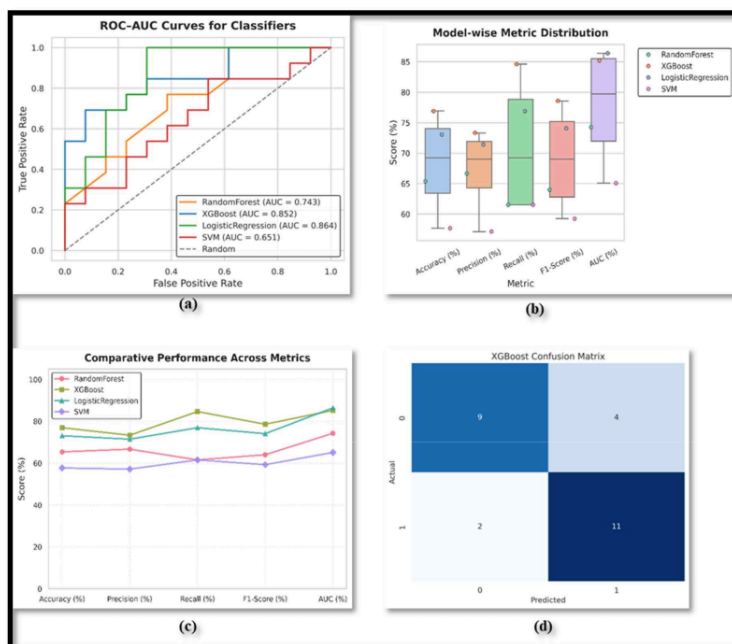


Figure 4.7 (a) ROC–AUC curves, (b) Boxplot of model Performance Distribution, (c) Comparative performance trend Analysis, and (d) XGBoost Confusion matrix

The ROC-AUC analysis shows that the tested classifiers had a good ability to discriminate between disease samples and controls. The high AUC scores of XGBoost and Logistic Regression suggest that changes in the transcriptome in the substantia nigra hold dependable predictive information related to neurodegenerative pathology. The high sensitivity in the XGBoost model indicates that the classifier really "understood" the disease-associated molecular features and So did not suffer from many false-negative predictions, a feature that is quite valuable Mostly for early biomarker identification and diagnostic screening purposes.

In fact the confusion matrix represent the clinical value of the classification model by showing how it was able to identify nearly all PD cases correctly whilst also not compromising too much the specificity towards the controls. Only few false positives and false negatives were to be expected, and the overall classification trade-off is indicative of Really the combined transcriptomic features have genuine predictive power despite the issue of biological heterogeneity that is frequently a cause of concern in Parkinson's disease cohorts. Taken together these results suggest that interpretable machine learning methods could play an important part in creating transcriptomics based diagnostic strategies for neurodegenerative diseases.

4.6 SHAP analysis and biomarker validation

In order to determine the influence of different genes on our model's predictions, we utilized SHAP explainability technique. The mechanism behind our XGBoost classifier was made explicit using the explainability of AI through SHAP analysis, with TARDBP/TDP-43, EIF4E, and SHQ1 being some of the most important biomarkers, as shown in Figure 4.8. These genes have been linked to RNA metabolism and neurodegenerative processes in the past. TARDBP/TDP-43 protein aggregation is common in both ALS and frontotemporal dementia, EIF4E is important for cap-dependent translation, and SHQ1 plays a role in ribonucleoprotein assembly [34]. These SHAP-identified genes align with established PD mechanisms involving mitochondrial dysfunction, impaired RNA metabolism, prolactin signalling, and stress-response pathways.

The bar plot, horizontally, represents the top 20 genes on the basis of their mean absolute SHAP values. The x-axis represents the mean (|SHAP value|) ranging from 0.0 to 0.6, indicating each gene's average effect on the model output size. TARDBP has the highest bar (~0.52), followed by EIF4E (~0.41) and SHQ1 (~0.38). For TARDBP, increased expression (pink) corresponds to very positive SHAP values, whereas reduced expression (blue) accumulates around zero or negative effects.

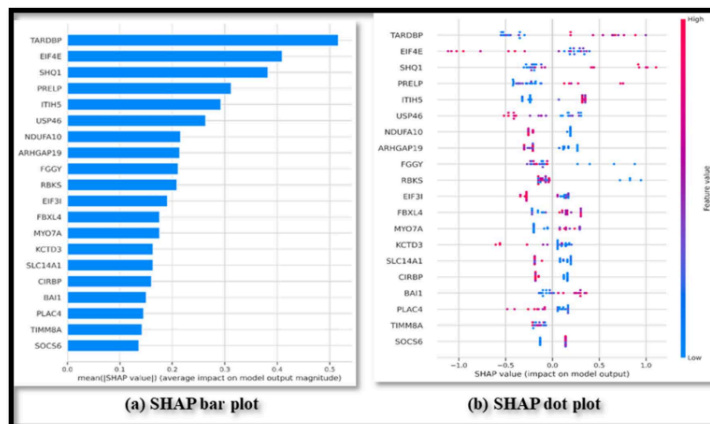


Figure 4.8 (a) SHAP bar, and (b) SHAP dot plot

The SHAP analysis revealed a clear molecular basis of how individual transcriptomic features were driving Parkinson's disease classification. Differently from traditional gene importance methods, SHAP measures not only the direction but also the size of gene impact on disease prediction, which allows the biological understanding of machine learning results. The discovery of TARDBP as the most strongly influencing biomarker is Mostly meaningful since TARDBP is the gene for TDP-43, a protein which is heavily linked to RNA deregulation, protein clumping, and brain disease pathology. The high SHAP value of TARDBP indicates that the transcriptome changes related to RNA processing might be one of the key molecular factors in the development of Parkinson's disease.

EIF4E was also a very important biomarker, showing that the control of protein synthesis, Mainly cap-dependent protein production, could be among the main causes of the molecular changes in PD. Since EIF4E is involved in mTOR and PI3K-AKT pathways, its deregulation may represent altered cellular stress response and deficits in neuronal survival mechanisms. In a like manner, SHQ1's role in ribonucleoprotein assembly strengthens the idea that faults in RNA maturation and ribosomal function lead to substantia nigra neuron loss. In reality RNA metabolism-related genes keep showing up among the highest SHAP-ranked biomarkers lends support to the increasing data indicating that post-transcriptional alteration is a big part in neurodegenerative disease pathogenesis.

4.7 Pathway enrichment analysis

The functional enrichment of the top 20 SHAP-ranked genes showed a significant abundance of the RNA transport pathway (adjusted $p = 0.017$), highlighting widespread defects in RNA metabolism as well as in neurodegeneration pathways such as amyotrophic lateral sclerosis, prolactin signalling pathway, and oxidative phosphorylation [35], [36], [37].

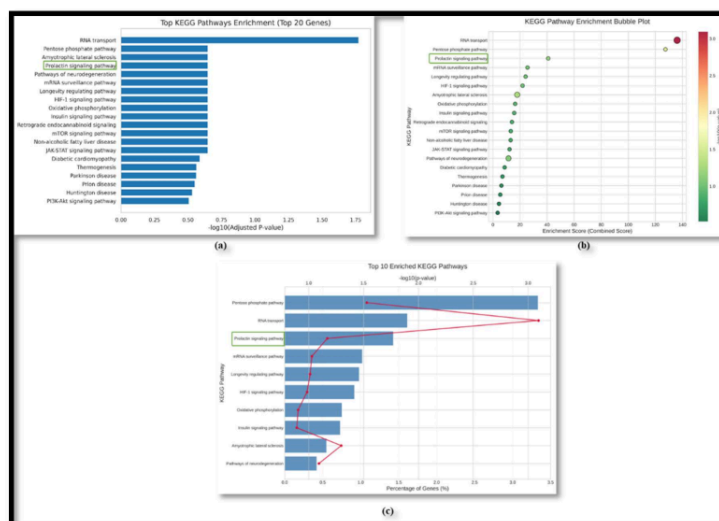


Figure 4.9 KEGG Pathway Enrichment Highlights Prolactin Signalling Pathway : (a) Bar Plot, (b) Bubble Plot, and (c) Top 10 KEGG Pathways ranked by gene percentage, with the overlaid red line representing $-\log_{10}$ (p-value)

These findings suggest that dysregulated RNA metabolism and mitochondrial dysfunction together accelerate dopaminergic neuronal loss. Oxidative phosphorylation suggests disruption in Complex I coupled with an increase in reactive oxygen species, aligning with previous studies that emphasized oxidative stress as a main feature of PD pathology [38]. The line plot again highlights the novel contribution of prolactin signalling as a distinct and significant contribution to the PD molecular biomarkers. Alongside these established biological pathways, enrichment analysis uniquely implicated the prolactin signalling pathway (hsa04917) in our SNpc targeted transcriptomic analysis.

Table 4.3 KEGG pathway enrichment analysis of SHAP-prioritized

1. KEGG G_20 21_H uman	2. RNA transport	3. 3 / 1 8 6	4. 0 . 0 0 0 3	5. 0. 01 68 62 26 5	6.	7.	8. 19 .0 90 64 60 9	9. 13 6. 06 28 42 6	10. EIF3I; TARD BP;EI F4E
11. KEGG G_20 21_H uman	12. Pentose phosphate pathway	13. J a n - 3 0 9	14. 0 . 0 2 9 5 9	15. 0. 22 52 77 81 6	16.	17.	18. 36 .2 08 71 14 3	19. 12 7. 46 62 13 5	20. RBKS
21. KEGG G_20 21_H uman	22. Amyotrophi c lateral sclerosis	23. 2 / 3	24. 0 . 0 5	25. 0. 22 52 77	26.	27.	28. 6. 02 14 85	29. 17 .9 68 53	30. NDUF A10;T ARD BP

		6 4	0 5 8 6	81 6			57 4	83 9	
31. KEG G_20 21_H uman	32. Prolactin signaling pathway	33. J a n - 7 0	34. 0 . 22 0 52 6 77 7 81 7 6 5 2	35. 0.	36.	37.	38. 15 .1 87 64 30 2	39. 40 .8 83 74 52 9	40. SOCS 6
41. KEG G_20 21_H uman	42. Pathways of neurodegen eration	43. 2 / 4 7 5	44. 0 . 22 0 52 8 77 0 81 7 6 3 6	45. 0.	46.	47.	48. 4. 58 23 34 97 8	49. 11 .5 31 77 50 6	50. NDUF A10;T ARD BP
51. KEG G_20 21_H uman	52. mRNA surveillance pathway	53. J a n - 9 8	54. 0 . 22 0 52 9 77 3 81 6 6 1 2	55. 0.	56.	57.	58. 10 .7 88 38 85	59. 25 .5 53 38 90 9	60. TARD BP
61. KEG G_20 21_H uman	62. Longevity regulating pathway	63. 1 / 1	64. 0 . 22 0 52 9 77	65. 0.	66.	67.	68. 10 .3 59 04	69. 24 .1 41 46	70. EIF4E

		0 2	7 2 5	81 6			11 7	20 7	
71. KEG G_20 21_H uman	72. HIF-1 signaling pathway	73. 1 / 1 0 9	74. 0 . 1 0 3 5 8 3	75. 0. 22 52 77 81 6	76.	77.	78. 9. 68 42 10 52 6	79. 21 .9 57 78 29 5	80. EIF4E
81. KEG G_20 21_H uman	82. Oxidative phosphoryla tion	83. 1 / 1 3 3	84. 0 . 1 2 4 9 7 9	85. 0. 22 52 77 81 6	86.	87.	88. 7. 91 38 75 59 8	89. 16 .4 57 76 91	90. NDUF A10
91. KEG G_20 21_H uman	92. Insulin signaling pathway	93. 1 / 1 3 7	94. 0 . 1 2 8 4 9 7	95. 0. 22 52 77 81 6	96.	97.	98. 7. 67 95 66 56 3	99. 15 .7 57 28 70 5	00. EIF4E
01. KEG G_20 21_H uman	02. Retrograde endocannab inoid signaling	03. 1 / 1 4 8	04. 0 . 1 3 8	05. 0. 22 52 77	06.	07.	08. 7. 10 09 66	09. 14 .0 58 12	10. NDUF A10

			1 0 4	81 6			70 2	21 4	
11. KEG G_20 21_H uman	12. mTOR signaling pathway	13. 1 / 1 5 4	14. 0 . 1 4 3 0 2	15. 0. 22 52 77 81 6	16.	17.	18. 6. 82 04 33 43 7	19. 13 .2 50 76 60 6	20. EIF4E
21. KEG G_20 21_H uman	22. Non- alcoholic fatty liver disease	23. 1 / 1 5 5	24. 0 . 1 4 4 1 6 5	25. 0. 22 52 77 81 6	26.	27.	28. 6. 77 58 03 14 4	29. 13 .1 23 35 86 3	30. NDUF A10
31. KEG G_20 21_H uman	32. JAK-STAT signaling pathway	33. 1 / 1 6 2	34. 0 . 1 5 0 1 8 5	35. 0. 22 52 77 81 6	36.	37.	38. 6. 47 89 14 67 8	39. 12 .2 83 28 36 8	40. SOCS 6
41. KEG G_20 21_H uman	42. Diabetic cardiomyop athy	43. 1 / 2 0 3	44. 0 . 1 8 4	45. 0. 25 85 05	46.	47.	48. 5. 15 32 04	49. 8. 70 53 55	50. NDUF A10

			6 4 7	90 5			79 4	12 5	
51. KEG G_20 21_H uman	52. Thermogen esis	53. 1 / 2 3 2	54. 0 . 2 0 8 2 1 6	55. 0. 27 32 83 98 5	56.	57.	58. 4. 49 96 58 23 7	59. 7. 06 07 62 47 6	60. NDUF A10
61. KEG G_20 21_H uman	62. Parkinson disease	63. 1 / 2 4 9	64. 0 . 2 2 1 7 3	65. 0. 27 39 02 26	66.	67.	68. 4. 18 76 06 11 2	69. 6. 30 77 61 96 7	70. NDUF A10
71. KEG G_20 21_H uman	72. Prion disease	73. 1 / 2 7 3	74. 0 . 2 4 0 4 3 6	75. 0. 28 05 09 21	76.	77.	78. 3. 81 34 67 49 2	79. 5. 43 53 32 93 5	80. NDUF A10
81. KEG G_20 21_H uman	82. Huntington disease	83. 1 / 3 0 6	84. 0 . 2 6 5 4	85. 0. 29 34 04 19 8	86.	87.	88. 3. 39 51 68 24 8	89. 4. 50 29 69 42 4	90. NDUF A10

			6 1						
91. KEG	92. PI3K-Akt	93. 1	94. 0	95. 0.	96.	97.	98. 2.	99. 3.	100. EIF4E
G_20	signaling	/	.	31			92	51	
21_H	pathway	3	3	10			63	86	
uman		5	0	76			45	98	
		4	0	07			60	10	
			4	1			9	1	
			6						
			6						

Functional enrichment analysis showed that the SHAP-prioritized genes are very likely to be involved in RNA regulation, mitochondrial activity, oxidative stress, and neurodegenerative pathways. That the RNA transport pathway was highly enriched gives weight to the idea that problems with RNA metabolism are something important in Parkinson's disease pathology. Because dopaminergic neurons depend on finely tuned RNA processing for neurotransmitter synthesis, mitochondrial upkeep, and synaptic communication, a breakdown in RNA transport processes can mean a loss of neuronal homeostasis and result in faster neurodegeneration.

Then again, the presence of oxidative phosphorylation pathways in enrichment analysis adds to the evidence of mitochondria malfunctioning as a major molecular characteristic of Parkinson's disease. For a long time, Mitochondrial Complex I deficiency has been correlated to lower production of ATP and higher generation of reactive oxygen species in dopaminergic neurons [35], [36]. Finding NDUFA10 in enriched pathways corroborates this mode of action since NDUFA10 is directly linked to the mitochondrial electron transport chain activity. Because of this, a failing oxidative phosphorylation process might lead to sustained neuronal damage by increasing oxidative stress, disrupting energy metabolism, and triggering apoptotic pathways in substantia nigra neurons.

The enrichment of amyotrophic lateral sclerosis and broader neurodegeneration pathways points to In reality overlapping molecular mechanisms shared across these diseases are involved. The role of TARDBP in both Parkinson's disease-linked pathways and ALS-related

signalling lends support to the idea that RNA dysregulation and protein aggregation are the key common pathogenic features across most neurodegenerative disorders [39]. This kind of overlap can mean that some transcriptomic signatures might represent general neuronal stress responses and not disease-specific alterations only.

An extremely interesting biological discovery to emerge from this work is the enrichment of the prolactin signalling pathway. Prolactin signalling was mostly thought of as the regulator of endocrine functions. Still, there is more and more evidence that prolactin Yes offers neuroprotection and that it works as an immunomodulator inside the CNS. Prolactin being related to the genes SOCS6 and EIF4E points to Really neuroendocrine signalling may be a map of neuronal stress adaptation and the inflammatory responses in Parkinson's disease. Endocrine-related pathways are very rarely investigated in transcriptomic studies of neurodegeneration, which makes this observation Most of all significant.

Prolactin is a pleiotropic pituitary hormone with central nervous system actions beyond lactation; it modulates immune responses, supports glial and neuronal survival, and influences intracellular signalling cascades, including JAK/STAT and MAPK effectors, relevant to cell stress responses [40]. Experimental work shows that prolactin can damage microglial pro-inflammatory activation and reduces oxidative injury in neuronal models, indicating direct anti-inflammatory and antioxidant effects plausibly counter balancing PD-relevant neuroinflammation and ROS-mediated toxicity [41]. These mechanistic properties provide a plausible biological signal, suggesting that dysregulated prolactin signalling in SNpc tissue may reflect or contribute to local neuroendocrine-immune crosstalk that modifies vulnerability of dopaminergic neurons.

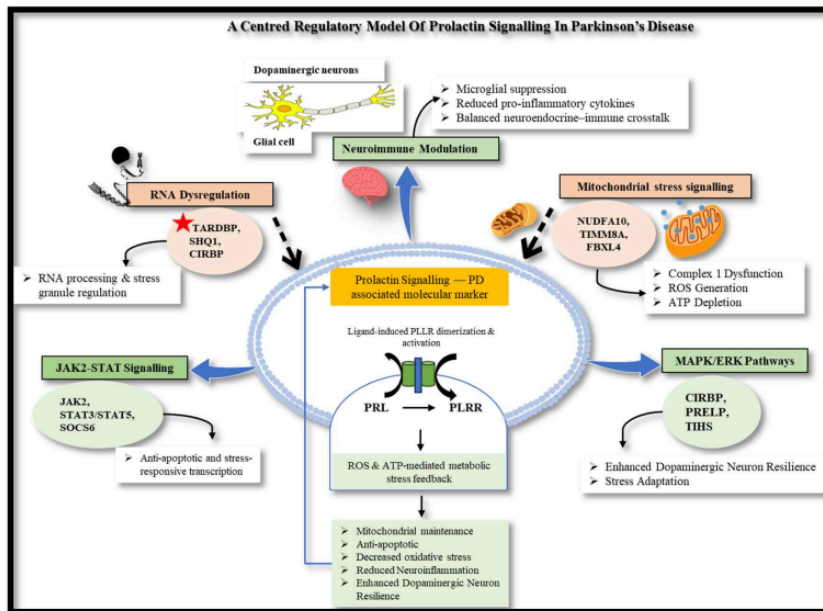


Figure 4.10 Prolactin signalling as a centred regulatory network

Alongside showing these results, our finding of the concurrent enrichment of JAKSTAT, ¹⁵ P3K-AKT, mTOR, and MAPK-related pathways implies that a series of interconnected cell survival and stress response pathways are likely to be involved in the progression of the disease [36], [39], [42]. Collectively, these pathways control inflammatory processes, programmed cell death (apoptosis), mitochondrial integrity, cell proliferation, and neuronal survival. Because of this, deregulation of these signal transduction pathways could bring creating a molecular milieu that enhances the vulnerability of the substantia nigra dopaminergic neurons to oxidative damage and degeneration complexly occurring over time .

In conclusion, regarding pathway enrichment analysis, It is obvious that the SHAP top biomarkers have biological meaning since the genes associated with these biomarkers are part of the neurodegenerative pathways that have been well documented in literature. Notably, one significant advantage of this study is the integration of explainable AI and pathway-level analyses that have turned out to be a very effective approach for identifying not only the known molecular pathways associated with Parkinson's disease but also new

ones. By conducting a detailed analysis of the system, we do not only make the scientific analysis robust but also identify what multicohort transcriptomic analysis combined with explainable AI can accomplish in neurodegenerative diseases.

4.8 Literature Validation

Clinically, circulating prolactin concentrations in PD cohorts have been heterogeneous and may be confounded by dopaminergic therapies that alter hypothalamic dopamine tone; however, recent reviews and preclinical studies suggest that intrinsic alterations in PRL signalling and receptor activity warrant attention as potential biomarkers [43]. Supporting this notion, preclinical evidence indicates that prolactin decreases microglial activation and protects neurons from oxidative insult, suggesting that components of the prolactin signalling axis (ligand, receptor expression, or downstream STAT/JAK pathway activity) may serve as mechanistically informative biomarkers and, conceivably, as adjunct therapeutic targets to modulate neuroinflammation and cellular stress in PD [41], [44]. Together, these pathways highlight interconnected disturbances in RNA processing, mitochondrial function, prolactin signalling, and neuroendocrine immune regulation in PD also validated through experimental evidences in below Table 4.4.

Table 4.4 Top 20 significant genes associated with the prediction of PD

S/N	Gene Symbol	Gene Function & Role in PD	References
1	TARDBP	RNA-binding protein; cytoplasmic aggregation induces mitochondrial pathology and oxidative stress in PD	[45]
2	EIF4E	Initiation factor for translation; dysregulation of the eIF4E/4E-BP pathway in LRRK2 mutation models contributes to aberrant protein synthesis and dopaminergic neuronal	[46]

		vulnerability; prolactin strongly activates PI3K/AKT	
3	SHQ1	H/ACA snoRNP assembly chaperone; SHQ1's function in snoRNP assembly is regulated by STAT5 transcriptional activity, and prolactin strongly activates STAT5	[47]; [48]
4	PRELP	Extracellular matrix protein; implicated in neuroinflammation and oxidative stress responses	[41]
5	ITIH5	Inter- α -trypsin inhibitor family member; controls inflammatory response and oxidative stress pathways	[49]
6	USP46	Deubiquitinating enzyme; controls synaptic AMPAR stability; possible function in mitochondrial protein quality control and neuroprotection against oxidative damage	[44]
7	NDUFA10	Complex I subunit; dysfunction leads to ATP depletion, ROS overproduction, and dopaminergic neurodegeneration	[17])
8	ARHGAP19	Regulates Rho GTPase-activating protein; controls cytoskeletal dynamics and mitochondrial morphology	[50]; [51]
9	FGGY	Carbohydrate kinase; metabolic enzyme dysfunction could play a role in mitochondrial	:[52]

		bioenergetic defects and cellular energy crisis in dopaminergic neurons	
10	RBKS	Ribokinase; dysregulation of sugar metabolism impacts cellular ATP homeostasis and mitochondrial function	[18]
11	EIF3I	Component of translation initiation factor; engages with eIF4E pathway; dysregulation causes protein synthesis flaws that impact mitochondrial protein homeostasis	[50]; [53]
12	FBXL4	F-box protein; mutations lead to mitochondrial DNA depletion syndrome; enhanced by prolactin-AKT signalling	[54]
13	MYO7A	Myosin motor protein; cytoskeletal malfunction disrupts mitochondrial movement in neurons	[55]; [56]
14	KCTD3	Potassium channel tetramerization domain protein; dysregulation of ion homeostasis impacts mitochondrial membrane potential and ROS generation in PD pathogenesis	[50]; [1]
15	SLC14A1	Urea transporter; stress and metabolic imbalance enhance neuronal susceptibility to oxidative damage and mitochondrial dysfunction	[57]
16	CIRBP	Cold-inducible RNA-binding protein	[58]

17	BAl1	Brain-specific angiopoietin inhibitor; apoptosis signalling and neuronal survival pathway modulation; anti-apoptotic effects promoted by prolactin-PI3K/AKT activation	[45]
18	PLAC4	Stress response-associated gene; roles in cellular adaptation processes relevant to disease stress	[59]
19	TIMM8A	Mitochondrial import machinery subunit; mutations result in deafness-dystonia-optic neuropathy; Prolactin activates mitochondrial survival pathways (PI3K/AKT), affecting TIMM8A-regulated protein import	[60]
20	SOCS6	Negative regulation of cytokine signalling; directly regulates prolactin-JAK/STAT pathway via negative feedback inhibition	[58]

Several of the biomarkers that were highlighted as being very important by SHAP analysis have actually been experimentally linked to the mechanisms of neurodegenerative diseases. For example, genes such as TARDBP EIF4E NDUFA10, and SOCS6¹⁰ are involved in the regulation of RNA, mitochondrial function, oxidative stress, and inflammatory signalling respectively [61]. This lends support to the computational findings from a biological point of view. The agreement between the transcriptomic analysis conducted here and the corroboration of published experimental evidence elevates the trust in the biomarker signatures that have been identified.

The evidence that prolactin signalling is involved not only fits with the short communication this manuscript is making but also with the recent data were reporting that neuroendocrine pathways are the ones that control neuroimmune regulation in the central nervous system. Even though it is still debated whether there is a change in the levels of circulating prolactin

in patients with Parkinson's disease, the experimental data indicate that prolactin is able to affect the survival of neurons by regulating oxidative stress and inflammatory reactions. So, the finding of prolactin-related transcriptomic patterns in the substantia nigra supports the hypothesis that in Parkinson's disease, neurodegenerative occurrences may be occurring in part through endocrine receptor-mediated pathways [62].

Quite a bit, the consensus of SHAP-prioritized genes with already known pathways in neurodegeneration confirms that explainable machine learning might extract biologically relevant transcriptomic signatures from datasets derived from multiple cohorts. Such validation is very crucial as it not only shows that the predictive biomarkers found through the computational pipeline are not mere random statistical occurrences but also that they represent molecular mechanisms that have been documented through the scientific literature. The integration of transcriptomics from multiple cohorts, machine learning, explainable AI, and biological interpretation based on literature together form a scientifically justified setup for biomarker research in Parkinson's disease in the coming years.

CHAPTER 5

CONCLUSION

This study demonstrated that integrative machine learning analysis of harmonized substantia nigra transcriptomic data is capable of suggesting strongly predictive and biologically relevant Parkinson's disease biomarkers. Among the models tested, XGBoost was the most accurate and precise, with logistic regression providing the most consistent AUC and recall, demonstrating the strength of the predictive model. The most important biomarkers determined by using explainable AI through SHAP analysis were found to be TARDBP, EIF4E, and SHQ1, all of which were linked to dysfunctional RNA metabolism and mitochondria, both of which are well-known features of neurodegeneration. Enrichment of biological pathways revealed that oxidative phosphorylation and RNA transport were among the mechanisms responsible for the neuronal degeneration. In addition, the identification of prolactin signaling pathway enrichment brings up a new feature of the disease in terms of neuroendocrinology. This implies that the immune system and hormonal imbalance may affect the mechanisms of dysfunctional RNA processing and mitochondria to cause neurodegeneration. In sum, the results have provided a mechanistic framework of oxidative stress, RNA imbalance, and neuroinflammation induced by prolactin signaling pathway. Future research needs to incorporate transcriptomics and proteomics analyses to identify prolactin-related biomarkers in Parkinson's disease.

Above all, this study highlighted prolactin signaling as one of the neuroendocrine pathways that might be associated with transcriptomic alterations in the substantia nigra. Though prolactin has not been studied for its relationship with PD, In fact, enrichment of prolactin-related genes shows that neuroendocrine and immune responses might In fact have some impact on dopaminergic neuronal susceptibility and the development of PD. xAI has been utilized in the identification of transcriptomic biomarkers, and it has thus led to a new level of biological interpretation through connecting machine learning prediction with biological genes and pathways. In contrast to traditional machine learning approaches, Explainable AI allows for a better understanding of how transcriptome variations are related to diseases, and this will thus make the biomarkers identified more applicable in practice.

Overall, the study provides the basis for the identification of biomarkers for PD through the use of transcriptomics from various populations and machine learning that is both reproducible and biologically relevant. The biomarkers and pathways discovered by them could prove useful in diagnosing diseases, finding targets for therapy, and creating precision medicine for neurological disorders.

ORIGINALITY REPORT

3%

SIMILARITY INDEX

2%

INTERNET SOURCES

2%

PUBLICATIONS

1%

STUDENT PAPERS

PRIMARY SOURCES

1

assets-eu.researchsquare.com

Internet Source

<1 %

2

benthambooks.com

Internet Source

<1 %

3

Lina Zhu, Deng Chen, Xiangxiu Wang, Chengqi He. "Identification and Validation of Oxidative Stress-Related Hub Genes in Parkinson's Disease", Molecular Neurobiology, 2024

Publication

<1 %

4

www.mdpi.com

Internet Source

<1 %

5

Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical and Computer Technologies", CRC Press, 2025

Publication

<1 %

6

Submitted to Bournemouth University

Student Paper

<1 %

7

Tarek Abdallah Abdel Salam, Ahmed Montaser Abdelsalam, Nada Reda Sholkamy, Ahmed Sebaq. "Assessment of artificial intelligence in detecting errors on panoramic

<1 %

radiographs", Imaging Science in Dentistry, 2026

Publication

8	www.shs-conferences.org Internet Source	<1 %
9	www.frontiersin.org Internet Source	<1 %
10	Jae Kyung Kim, Sukyeong Mun, Myung-Suk Kim, Mi-Bo Kim, Bo-Kyung Sa, Jae-Kwan Hwang. "5,7-Dimethoxyflavone, an activator of PPAR α / γ , inhibits UVB-induced MMP expression in human skin fibroblast cells", Experimental Dermatology, 2012 Publication	<1 %
11	bmcmedresmethodol.biomedcentral.com Internet Source	<1 %
12	hdl.handle.net Internet Source	<1 %
13	jstic.ptit.edu.vn Internet Source	<1 %
14	pubmed.ncbi.nlm.nih.gov Internet Source	<1 %
15	Gi Suk Nam, Kyung-Soo Nam. "Arctigenin attenuates platelet activation and clot retraction by regulation of thromboxane A2 synthesis and cAMP pathway", Biomedicine & Pharmacotherapy, 2020 Publication	<1 %

16 Vivek Sunnapwar, Hariram Chavan, Namrata F. Ansari. "Advances in Science and Technology", CRC Press, 2026
Publication <1 %

17 dspace.daffodilvarsity.edu.bd:8080
Internet Source <1 %

18 www.science.gov
Internet Source <1 %

19 A. Punitha, S. Syedakbar, S. Jeyasudha. "Intelligent and Sustainable Systems: AI, Green IoT, and Adaptive Automation in Electrical and Communication Technologies - Proceedings of International Conference in Pathway of Electrical, Automation, Communication, and Electronics (PEACE-2025)", CRC Press, 2026
Publication <1 %

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off