OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

# Network and Transcriptome-Guided Explainable Machine Learning for Identifying Disease-Specific Drug–Target Interactions in Alzheimer's and Parkinson's Disease

Vir Rastogi and Nirupma Singh
virrastogi0@gmail.com

**ABSTRACT**

Alzheimer's disease and Parkinson's disease are neurodegenerative disorders that pose significant global health challenges. Traditional drug discovery approaches are often time consuming and costly, emphasizing the need for more efficient computational strategies to identify potential drug targets. In this study, gene expression datasets associated with Alzheimer's and Parkinson's diseases were obtained from the Gene Expression Omnibus (GEO) database and analyzed using GEO2R to identify differentially expressed genes (DEGs). The identified DEGs were further investigated through gene interaction network analysis to examine their connectivity and pathway enrichment analysis to determine their significance. Features from their respective gene expression profiles and network characteristics were integrated to construct a dataset for predictive modeling between the two diseases. Two machine learning algorithms, Random Forest and XGBoost, were trained and evaluated using these features. Both models demonstrated strong predictive performance in identifying significant genes, with XGBoost showing slightly better results across most evaluation metrics. Feature importance analysis revealed that DEG significance and network connectivity were among the most influential predictors. The findings demonstrate that integrating gene expression data with features related to the network enhances the identification of relevant genes associated with neurodegenerative diseases. This computational model provides an efficient approach for prioritizing potential drug targets and may contribute to the development of improved strategies for the further research and the potential treatment of Alzheimer's and Parkinson's diseases.

**Keywords**: Drug–Target Interaction (DTI), Explainable Artificial Intelligence (XAI), Machine Learning, Alzheimer's Disease, Parkinson's Disease.

## INTRODUCTION

Neurological disorders are diseases that affect the central and peripheral nervous systems including the brain, spinal cord, cranial nerves, and autonomic nervous system.[1] Neurodegenerative diseases constitute a major public health challenge, contributing significantly to disability worldwide.[2] These diseases are progressive in nature, characterized by irreversible neuronal loss leading to cognitive, behavioural, and motor impairments.[2,3] Globally, as of 2021, more than 55 million people were affected by dementia, with projections that indicated a substantial rise due to aging populations and increased life expectancy.[4] The burden of neurodegenerative diseases is rapidly rising, with limited access to disease modifying therapies, high healthcare costs and delayed diagnosis.[5] Conventional drug discovery pipelines for neurological disorders are slow, costly, and lack interpretability.[6] The need for computational and artificial intelligence (AI)-driven strategies to accelerate drug target interaction prediction as well as novel drug discovery is essential - now more than ever.

Recent advances in bioinformatics and artificial intelligence have led to the development of several machine learning and deep learning approaches to model molecular level interactions. Sequence and graph-based representation have been created to integrate multi-modal inputs such as drug chemical structures, protein sequences, and interaction networks, improving prediction and interpretation by reading molecular features and cross-modal relationships [7]. Ensemble deep learning architectures combine gradient boosting neural networks, deep neural networks, and decision forest models to demonstrate improved performance in drug-target interaction (DTI) classification across diverse receptor types and enable prediction of potential neurodegenerative targets for novel drug development. Studies using neural networks alongside recurrent units such as bidirectional LSTM have shown enhanced capability to model time-dependent signals in EEG data, optimizing prediction accuracy [8]. Deep bioinformatics pipelines integrating co-expression network analysis, machine learning classifiers (like LASSO, SVM-RFE, and random forest), and single cell data have been used to identify hub genes and targets in neurodegenerative diseases. Moreover, work in drug repurposing and hybrid AI docking to further highlight how AI can hasten screening compound libraries against disease relevant targets.

Multiple studies utilise deep sequence modeling and structural feature extraction for neurological data, using frameworks such as LSTM and its variants to capture temporal dependencies in data. For example, recurrent neural networks that process intracranial EEG and LFP recordings with bidirectional LSTM layers and Bayesian optimization improved classification performance in drug pattern recognition, depicting how temporal patterns correlate with psychotropic effects related with neurodegeneration. Simultaneously, multi-modality deep learning strategies that have been tested, have achieved a high calibre performance through contrastive learning, improving generalization for novel DTI pairs. Bioinformatics driven integrative looks at combining gene expression clustering, network topologies, and machine learning feature selection have located key molecular drivers and therapeutic targets in Alzheimer's (AD) pathology [9]. Additionally, comprehensive reviews of AI-based repurposing and hybrid

methods illustrate the effective fusion of predictive modeling and molecular docking to uncover novel drug-target associations, especially in disorders like Alzheimer's disease.

Current machine learning DTI prediction models face limitations that constrain their applicability to neurodegenerative drug discovery in specific. For one, interpretability remains limited; many deep representation models excel at prediction accuracy but provide minimal insight into which molecular substructures drive affinity, hindering mechanistic understanding and rational drug design [7]. Integration of the data modalities remains challenging despite early successes with existing methods unable to fully exploit features in a single unique model. Thus, it was hypothesized that AI models can be developed to predict novel drug-target interactions for neurodegenerative diseases such as AD and PD by utilizing molecular docking data, transcriptomic profiles and protein interaction networks.

## METHODOLOGY

*Data Collection*

The data collection was started by identifying quality datasets relevant to DTI prediction related to AD and PD. Several databases were explored to obtain information related to drugs, protein targets, bioassays, gene expression profiles and molecular level data. Some main resources that were identified were BindingDB, DrugBank, ChEMBL, DGIdb and the NCBI Gene Expression Omnibus (GEO). Different types of data were collected from these sources such as chemical structures, protein sequences, binding affinities, bioassay results, drug-target interaction pairs, gene expression profiles and mechanisms of action of drugs.

BindingDB was used to extract experimentally validated binding data, while DrugBank provided curated drug profiles including pharmacological properties and mechanisms of action. ChEMBL was used to extract bioactivity data, assay information, and drug safety warnings and indications. Additionally, DGIdb was used to obtain drug-gene interaction data into the analysis. The NCBI Gene Expression Omnibus (GEO) was selected to add gene expression datasets relevant to neurodegenerative conditions, enabling comparisons between molecular targets and disease-specific biological activity. Following evaluation based on data completeness, relevance, and compatibility with the goals for the project, NCBI GEO and DGIdb were selected as the primary datasets due to their strong relevance to gene expression analysis and drug-gene interactions.

*Differential Expression Analysis*

Following the differential expression analysis performed using the GEO2R tool, a total of 10 GEO datasets were analysed, including GSE1297, GSE28146, GSE48350, GSE5281, GSE16759 (Alzheimer's disease) and GSE20292, GSE7621, GSE20141, GSE6613, GSE8397 (Parkinson's disease). A comparative evaluation was then carried out comparing the genes in 2 different groups named test and control. The test group had genes which had the neurodegenerative disorder and the control group had genes which were from a normal organism. The analysis helped to identify the most biologically relevant

gene expression profiles associated with these neurodegenerative disorders. For each of the GEO datasets, the list of significantly upregulated and downregulated DEGs were identified. The distribution and density of significant genes across datasets was used as a selection criterion.

The statistically significant genes were identified based on adjusted p-value and log fold-change thresholds. Significant DEGs were identified using an adjusted p-value threshold of <0.05 and an absolute log2 fold-change threshold of > 1.0, corresponding to at least a two-fold difference in expression between disease and control groups. Genes satisfying both the given criteria were retained for further network and pathway analysis.

*Network and Pathway Analysis*

After performing DE analysis, the final list of DEGs was used to construct gene-interaction networks both for AD and PD. Firstly, the final list of significant genes obtained from DE analysis was fed into the STRING database to construct a gene interaction network by selecting the organism as *Homo Sapiens*. High-confidence interactions were selected for both AD and PD networks. The networks generated in STRING were then exported and analysed using the AnalyzeNetwork app of Cytoscape. Different topological parameters were computed such as degree centrality, betweenness centrality, closeness centrality, clustering coefficient, and so on, for identifying hub genes and highly influential nodes within the network. Only genes with the highest degrees were given priority, as they are proteins with the greatest number of interactions. Subsequently, pathway enrichment analysis was conducted using the Reactome database by mapping the complete gene set to known biological pathways. This aims to identify significant pathways associated with the gene list, giving insights into disease progression. By cross-referencing the top hub genes from the network analysis stage with the most significant pathways after the pathway analysis, genes were further categorized based on their relevance within key biological pathways to give us a final processed top-gene list.

*Prediction Framework*

The objective of the machine learning stage was to classify genes according to their relevance in Alzheimer's disease and Parkinson's disease as opposed to directly predict experimentally validated drug-target interactions. Each gene was represented by a feature vector combining some transcriptomic attributes derived from DE analysis (log2 fold-change, adjusted p-value, and composite DEG score) with attributes obtained from the network analysis (degree, betweenness centrality, closeness centrality, clustering coefficient, and related parameters). A binary labeling strategy was adopted: Label = 1 was assigned to genes prioritized as important based on their combined differential expression significance along with network and pathway analysis, while Label = 0 was assigned to genes that did not satisfy these criteria. Specifically, in the AD training set, genes associated only with PD drug-gene interactions received Label = 0, and vice versa for the PD training set. The trained models therefore learn to distinguish important genes from non-prioritized genes. The genes predicted as Label = 1 are subsequently proposed as candidate drug targets, under the rationale that genes with strong disease association represent priority candidates for any further medical research.

*Training Set and Model Building*

The final list of DEGs for AD and PD were mapped with the DGIdb database's drug-gene interaction data and corresponding drug interactions were found for the genes. The network parameters for both AD and PD specific genes were also mapped with the drug-gene interaction columns of the two diseases. The training sets for AD and PD were thus built which included the genes, interacting drugs, disease name and the network parameters as the features. The labels were assigned as 0 and 1. In the AD training set, the PD-specific drug-gene interactions were assigned the label as 0 and in the PD training set, the AD-specific drug-gene interactions were assigned the label as 0.

The model-building process started by loading disease-specific datasets. The data was first cleaned by separating the target variable from the input features and removing non-relevant columns such as gene names, drug names, and disease identifiers. All other features were then converted into numeric form, missing values were handled by replacing them with median values, and incomplete rows were filtered out to ensure data consistency. The dataset was then split into training and testing sets in an 80:20 ratio respectively. Cross-validation was performed on the full dataset to evaluate consistency across various data splits. Two different classification models were built and compared. The Random Forest (RF) model used multiple decision trees to improve prediction stability. The XGBoost model applied gradient boosting to enhance performance. For the Random Forest model, the number of estimators was set to 100 with the maximum depth left unrestricted and the minimum samples per leaf set to 1. For XGBoost, the number of estimators was set to 100, the learning rate to 0.1, the maximum depth to 6 and a subsample ratio of 0.8 was applied. No exhaustive hyperparameter search was performed; default XGBoost parameters were used with the above modifications, given the limited dataset size. This is acknowledged as a constraint on model optimization. Each model was trained on the training data and then further evaluated on the test data using multiple metrics such as ROC-AUC, precision-recall AUC, accuracy, F1-score, precision and recall. The confusion matrices and classification reports were also built to assess the prediction quality. Feature importance was calculated for both models to identify the most influential genes contributing to predictions and a plot was generated to visualize the top features.

## RESULTS

*Differential Expression Analysis*

The datasets selected for DE analysis included GSE1297, GSE28146, GSE48350, GSE5281, GSE16759 for AD and GSE20292, GSE7621, GSE20141, GSE6613, GSE8397 were identified for PD. Each dataset provided well-defined comparisons, enabling meaningful identification of gene expression changes. These specific studies were also chosen because they had clear labelling of control and disease groups for AD and PD, making the analysis straightforward and reliable. Many of them focused on brain regions directly affected in these diseases, such as areas linked to memory and movement. Using multiple studies for each disease also helped in comparing patterns across datasets and identifying more reliable gene expression changes.

Each study consisted of on average 40 samples in total, with around 20 samples assigned to the control group and 20 to the disease (test) group for differential expression analysis using GEO2R. Upon analysis, variability in the number of significant genes was observed across datasets. Specifically, two datasets yielded a very high number of significant differentially expressed genes: GSE5281 for AD and GSE8397 for PD. GSE5281 had 11276 and GSE8397 had 8861 differentially expressed genes with GSE5281 having 61 control and 44 test samples, and GSE8397 having 18 control and 29 test samples. In contrast, the remaining datasets showed a much lower number of significant genes, with 0 to 20 genes identified as statistically significant with the p-value greater than 0.5. After removing duplications and blanks during the processing phase, for GSE5281 there were 7194 and for GSE8397 there were 4775 differentially expressed genes. These studies were taken for further analysis. The selection of GSE5281 for AD and GSE8397 for PD was based on three criteria applied consistently across all ten datasets: (1) the total number of statistically significant DEGs identified after GEO2R analysis, (2) the sample size and balance between control and disease groups, and (3) the biological specificity of the tissue source. GSE5281, derived from multiple brain regions including the entorhinal cortex and hippocampus - areas central to AD pathology - gave the highest number of significant DEGs (11,276 before and 7,194 after processing) among the five AD datasets. GSE8397, sourced from substantia nigra tissue directly implicated in PD neurodegeneration, produced the highest DEG count among the five PD datasets (8,861 before and 4,775 after). The remaining datasets yielded between 0 and 20 significant genes, making them unsuitable for downstream network and pathway analysis. While this selection approach maximizes analytical power, it introduces the possibility of dataset-specific bias.

*Network Analysis*

For the Alzheimer's disease network, after applying the high-confidence interaction filter in STRING, the network consisted of 857 nodes and 3320 edges, showing a highly interconnected set of proteins. In the Parkinson's disease network, there were 692 nodes and 1267 edges, indicating a comparatively smaller but still meaningful interaction network. These networks were then analysed in Cytoscape to identify the most important genes. To determine the top 10 genes, four key network parameters were considered: degree, betweenness centrality, closeness centrality, and clustering coefficient. Among these, degree was given the most importance, and the genes were finally sorted based on degree values. Degree refers to the number of direct connections a gene has with other genes in the network. Genes with higher degree values are considered more important because they interact with many others. Betweenness centrality measures how often a gene lies on the shortest path between other genes. Genes with high betweenness play a role in connecting different parts of the network. Closeness centrality shows how close a gene is to all other genes in the network. A higher value means the gene can interact more quickly with others. Clustering coefficient indicates how closely a gene's neighbouring genes are connected. Higher values suggest the gene is part of a tightly connected group.

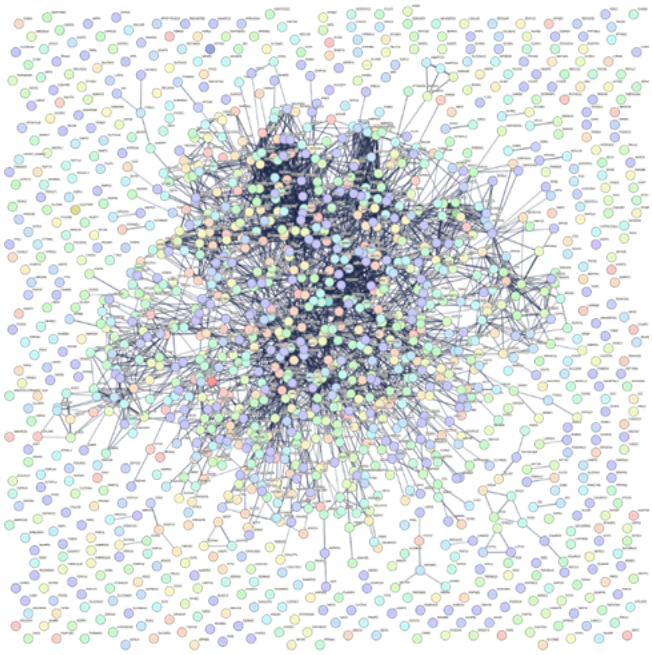

Figure 1. Protein-protein interaction network of differentially expressed genes in Alzheimer's disease (GSE5281), constructed using the STRING database and visualised in Cytoscape. Nodes represent genes and edges represent high-confidence interactions. Node size and colour intensity are proportional to degree centrality; hub genes with the highest degree values are visually prominent. The network consists of 857 nodes and 3,320 edges. Figure created by the authors using Cytoscape v3.x and STRING v11.x.

| S. No. | Name | Degree | Betweenness Centrality | Closeness Centrality | Clustering Coefficient |
|---|---|---|---|---|---|
| 1 | ACTB | 68 | 0.1326566774 | 0.3707968678 | 0.09920983319 |
| 2 | ATP5F1A | 55 | 0.04945201634 | 0.3352769679 | 0.2316498316 |
| 3 | ATP5F1C | 53 | 0.02863374343 | 0.3142076503 | 0.2619738752 |
| 4 | GAPDH | 51 | 0.1485932707 | 0.3752913753 | 0.1419607843 |
| 5 | EGFR | 48 | 0.1303537445 | 0.354001759 | 0.08067375887 |
| 6 | HSP90AB1 | 43 | 0.1096476424 | 0.3660754889 | 0.1173864895 |
| 7 | MRPS12 | 42 | 0.01221370271 | 0.2940102264 | 0.3542392567 |
| 8 | TUFM | 40 | 0.01008967887 | 0.2900900901 | 0.3666666667 |

| | | | | | |
|---|---|---|---|---|---|
| 9 | ATP5F1B | 38 | 0.01568508463 | 0.3139625585 | 0.3129445235 |
| 10 | MRPS5 | 38 | 0.01173913568 | 0.2939028843 | 0.3911806543 |

Table 1. The top 10 most significantly differentially expressed genes in AD, filtered in descending order of degree.

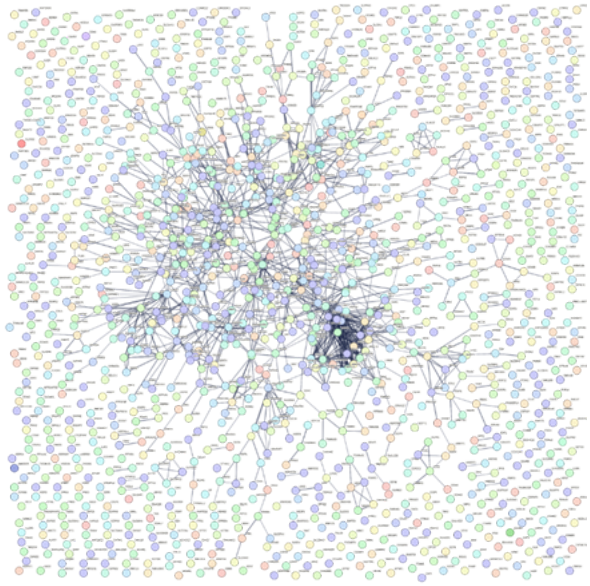

Figure 2. Protein-protein interaction network of differentially expressed genes in Parkinson's disease (GSE8397), constructed using the STRING database and visualised in Cytoscape. Nodes represent genes and edges represent high-confidence interactions. Node size and colour intensity are proportional to degree centrality; hub genes with the highest degree values are visually prominent. The network consists of 692 nodes and 1267 edges. Figure created by the authors using Cytoscape v3.x and STRING v11.x.

| Sr. No. | Name | Degree | Betweenness Centrality | Closeness Centrality | Clustering Coefficient |
|---|---|---|---|---|---|
| 1 | MRPS5 | 32 | 0.08622044213 | 0.2346232179 | 0.3508064516 |
| 2 | MRPL4 | 27 | 0.04956587969 | 0.2282091918 | 0.4643874644 |
| 3 | MRPL2 | 24 | 0.005907236825 | 0.2095307385 | 0.5471014493 |
| 4 | MRPS6 | 23 | 0.008766412445 | 0.2216237014 | 0.5889328063 |
| 5 | H3C12 | 22 | 0.0927149335 | 0.2264150943 | 0.07792207792 |

| | | | | | |
|---|---|---|---|---|---|
| 6 | MRPL1 | 22 | 0.01403834255 | 0.2070452912 | 0.6277056277 |
| 7 | MRPL32 | 22 | 0.007828197105 | 0.2215384615 | 0.6493506494 |
| 8 | MRPS9 | 22 | 0.01946955283 | 0.2229102167 | 0.632034632 |
| 9 | MRPS15 | 22 | 0.01676961466 | 0.2242117555 | 0.632034632 |
| 10 | H4C6 | 21 | 0.03886413133 | 0.2024604569 | 0.1285714286 |

Table 2. The top 10 most significantly differentially expressed genes in PD, filtered in descending order of degree.

*Pathway Analysis*

For the Alzheimer's disease dataset, the top enriched pathways identified through Reactome analysis included G2/M Transition (R-HSA-69275), Mitotic G2-G2/M phases (R-HSA-453274), The role of GTSE1 in G2/M progression after G2 checkpoint (R-HSA-8852276), Regulation of MITF-M-dependent genes involved in lysosome biogenesis and autophagy (R-HSA-9857377), and Degradation of beta-catenin by the destruction complex (R-HSA-195253). To assess the relevance of the identified hub genes, a linear search was conducted across the pathway results to determine whether the top 10 genes from network analysis were present within the top 25 enriched pathways. It was observed that only 3 of the 10 genes appeared in these pathways: ACTB, EGFR, HSP90AB1. This suggests that while these genes are highly connected within the interaction network, they may not be strongly represented in the most significantly enriched pathways.

For the Parkinson's disease dataset, the top pathways included Membrane Trafficking (R-HSA-199991), Activation of NMDA receptors and postsynaptic events (R-HSA-442755), Neurotransmitter receptors and postsynaptic signal transmission (R-HSA-112314), Neuronal System (R-HSA-112316), and Transmission across Chemical Synapses (R-HSA-112315). These pathways are directly related to neuronal signaling, synaptic transmission, and brain-specific processes, making them highly relevant to Parkinson's disease. A similar linear search was conducted for this dataset, and all 10 of the top genes were found within the top 25 pathways. This indicates a strong overlap between the network-derived hub genes and the enriched pathways, suggesting better alignment between network importance and functional relevance in the Parkinson's dataset.

*Model Training and Evaluation*

The dataset consisted of 19 features (Average Shortest Path Length, Betweenness Centrality, Closeness Centrality, Clustering Coefficient, Degree, Eccentricity, IsSingleNode, Neighborhood Connectivity, Number of Directed Edges, Number of Undirected Edges, PartnerOfMultiEdgedNodePairs, Radiality, Selected, Self Loops, Stress, Topological Coefficient, DEG_log2FC, DEG_neglog10P, DEG_score). The labels were in binary, where 1 represents significant genes and 0 represents insignificant genes for that specific disease. For the AD dataset, the class distribution had an equal number of 1 and 0 labels, hence a perfectly balanced number of significant and insignificant genes. For the PD dataset, the distribution was

a bit imbalanced with 84 1s and 73 0s. Two models used were RF and XGBoost, and model performance was evaluated using certain predictive metrics, including ROC-AUC that can distinguish classes, PR-AUC which tells the performance on imbalanced data, and accuracy that gives the overall correctness, F1-score that tells about the balance between precision and recall, where precision gives an idea about correct positive predictions, and recall captures the actual positives.

| Model Name | ROC-AUC | PR-AUC | Accuracy | F1-score | Precision | Recall |
|---|---|---|---|---|---|---|
| XGBoost | 0.7873 | 0.7688 | 0.7571 | 0.7302 | 0.8214 | 0.6571 |
| Random Forest | 0.7776 | 0.7657 | 0.7571 | 0.7302 | 0.8214 | 0.6571 |

Table 3. The efficiency and evaluation matrices for AD in metrics to access predictive accuracy in the model's output.

| Model Name | ROC-AUC | PR-AUC | Accuracy | F1-score | Precision | Recall |
|---|---|---|---|---|---|---|
| XGBoost | 0.7451 | 0.7374 | 0.7188 | 0.6897 | 0.8333 | 0.5882 |
| Random Forest | 0.7451 | 0.7374 | 0.7188 | 0.6897 | 0.8333 | 0.5882 |

Table 4. The efficiency and evaluation matrices for PD in metrics to access predictive accuracy in the model's output.

The 5-fold cross-validation results show that for Alzheimer's disease, Random Forest achieved a CV ROC-AUC of 0.8525 ± 0.0267 and PR-AUC of 0.8302 ± 0.0385, while XGBoost showed slightly better performance with a CV ROC-AUC of 0.8538 ± 0.0255 and PR-AUC of 0.8305 ± 0.0383. The low standard deviation in both models shows stable and reliable predictions. In contrast, for Parkinson's disease, both models showed lower performance and higher variation, with Random Forest achieving a CV ROC-AUC of 0.7379 ± 0.1105 and PR-AUC of 0.7380 ± 0.1090, while XGBoost performed slightly better with a CV ROC-AUC of 0.7450 ± 0.1033 and PR-AUC of 0.7410 ± 0.1054. The higher deviation in Parkinson's results suggests less consistency in results.

For the Alzheimer's disease Random Forest model, the top 5 features were DEG_neglog10P, NeighborhoodConnectivity_AD, BetweennessCentrality_AD, TopologicalCoefficient_AD, and DEG_score. For the Alzheimer's disease XGBoost model, the top 5 features were DEG_neglog10P, BetweennessCentrality_AD, NeighborhoodConnectivity_AD, TopologicalCoefficient_AD, and DEG_score. For the Parkinson's disease Random Forest model, the top 5 features were NeighborhoodConnectivity_AD, AverageShortestPathLength_AD, DEG_score, DEG_neglog10P, and

TopologicalCoefficient_AD. For the Parkinson's disease XGBoost model, the top 5 features were NeighborhoodConnectivity_AD, DEG_log2FC, AverageShortestPathLength_AD, ClusteringCoefficient_AD, and BetweennessCentrality_AD.
Overall, XGBoost marginally outperformed Random Forest in both diseases and the models performed more reliably on the Alzheimer's dataset compared to Parkinson's.

SHAP analysis was performed on both models to provide individual-level feature attribution. For the Alzheimer's disease XGBoost model, DEG_neglog10P consistently showed the highest mean absolute SHAP value, indicating that statistical significance of differential expression was the strongest driver of positive predictions. BetweennessCentrality_AD and NeighborhoodConnectivity_AD were the next most impactful network features, suggesting that genes acting as network bridges and those connected to highly linked neighbours were preferentially classified as relevant. For the Parkinson's disease model, NeighborhoodConnectivity_AD dominated SHAP contributions, reflecting the importance of local network context in PD gene prioritisation. SHAP values confirmed that the models relied on a biologically interpretable combination of transcriptomic significance and network topology rather than any single feature.

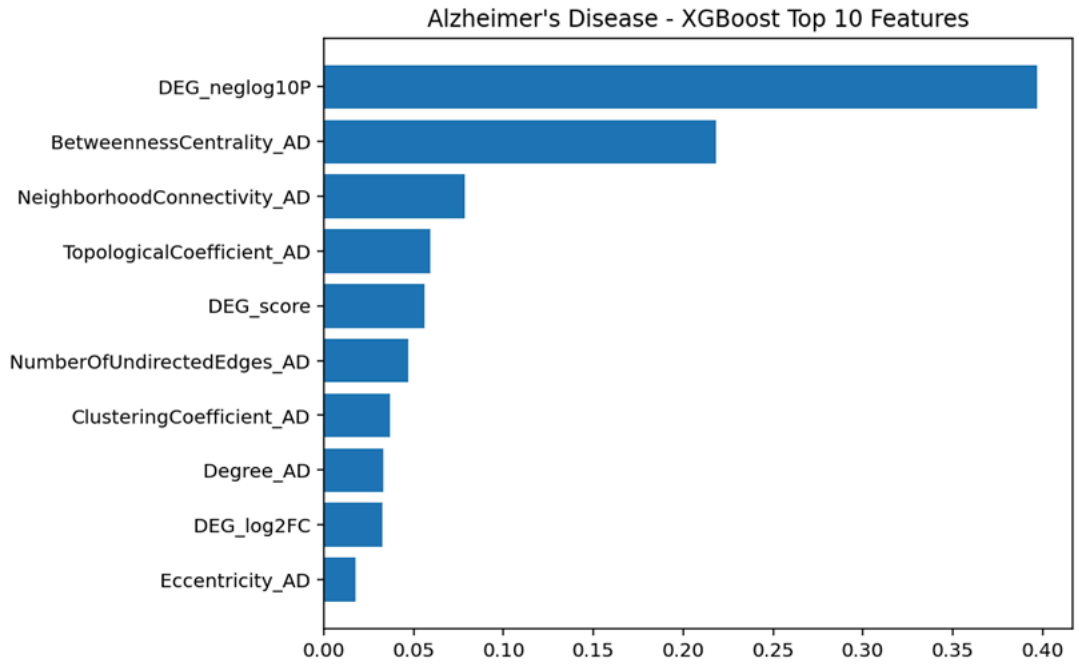


Figure 3. A plot for the top 10 features identified by the XGBoost Model for AD. The bar graph is plotted with the network features for AD on the Y-axis and the feature importance score on the X-axis, represented by the blue horizontal bars. Figure created by the authors using Python.

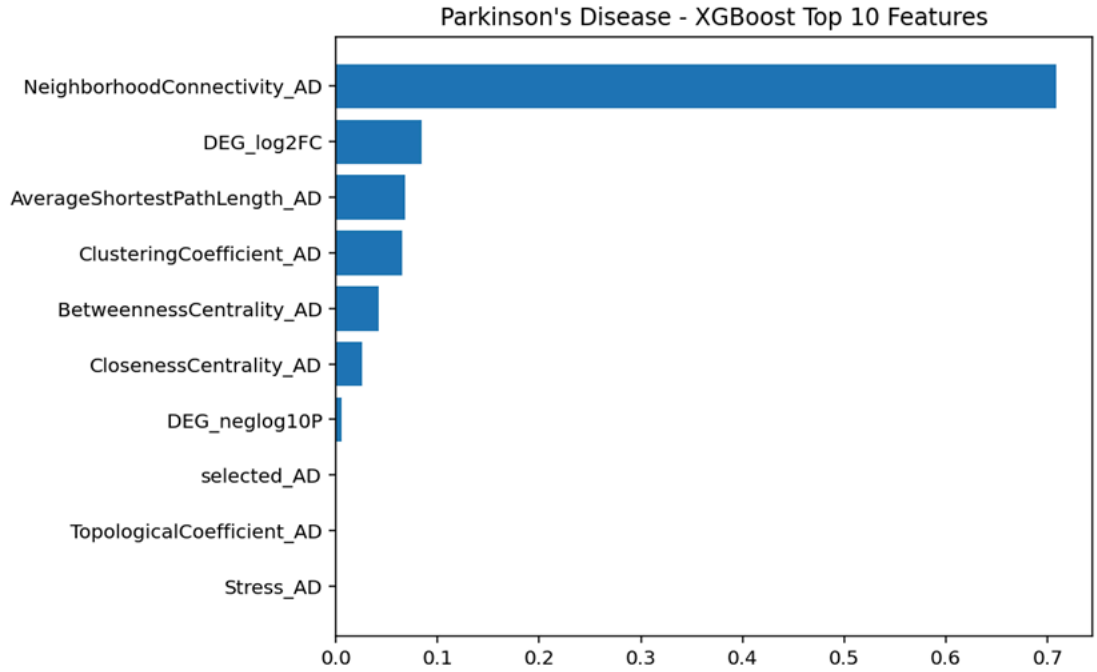


Figure 4. Plot for the top 10 features identified the XGBoost Model for PD. The bar graph is plotted with the network features for PD on the Y-axis and the feature importance score on the X-axis, represented by the blue horizontal bars. Figure created by the authors using Python.

## DISCUSSION

The project began with a literature review to identify gaps in AI-based drug-target prediction, followed by selection of relevant Alzheimer's and Parkinson's GEO datasets. Differential expression analysis using GEO2R, along with network (STRING, Cytoscape) and pathway analysis (Reactome), was used to identify key genes and features. These features were used to train Random Forest and XGBoost models using machine learning. Both models were evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. For Alzheimer's data, both models showed similar accuracy, with XGBoost slightly outperforming ROC-AUC and PR-AUC, while maintaining high precision, indicating reliable predictions. For Parkinson's data, performance was lower and more variable, but XGBoost again showed a slight edge.

Compared to 'Graph-DTI: A New Model for Drug-target Interaction Prediction Based on Heterogeneous Network Graph Embedding' (https://pubmed.ncbi.nlm.nih.gov/37448360/), this study improves by directly incorporating gene expression data along with network features, allowing stronger relevance rather than relying for the most part on solely network embeddings. In comparison to 'Enhancing drug-target interaction predictions in context of neurodegenerative diseases using bidirectional long short-term memory in male Swiss albino mice pharmaco-EEG analysis'

(https://pubmed.ncbi.nlm.nih.gov/39524776/), which uses LSTM analysis on pharmaco-EEG data, this work provides a more general, sustainable and elementary framework by using gene-level and network-level features that are easier to interpret and apply across various datasets. Similarly, compared to 'Drug repurposing using artificial intelligence, molecular docking, and hybrid approaches' (https://pubmed.ncbi.nlm.nih.gov/39826482/), which emphasizes in silico prediction and screening, this study improves through a better structure including various stages such as differential expression, network analysis, and pathway validation before model building, which has an impact on the reliability of the model as well. Overall, the main improvement across these studies is the workflow that we used, utilising a range of analysis techniques on datasets to extract vital information before building the training the models, giving them a thorough background of the diseases for which they were being trained.

The relevance of the hub genes identified in this study merits consideration in the context of already existing AD and PD research. In Alzheimer's disease, ACTB (beta-actin) has been reported to play a role in synaptic plasticity and dendritic spine morphology with cytoskeletal dysregulation increasingly implicated in AD pathology. GAPDH has been linked to amyloid-beta aggregation and tau-mediated neurodegeneration and has been proposed as a target in neuronal apoptosis pathways beyond its classical role in glycolysis. EGFR has been associated with neuroinflammatory signalling in AD and HSP90AB1 is vital to the proteostasis dysfunction characteristic of neurodegenerative conditions. In Parkinson's disease, the predominance of mitochondrial ribosomal proteins (MRPS5, MRPL4, MRPL2, MRPS6, MRPL1, MRPL32, MRPS9, MRPS15) among the top hub genes is consistent with the strongly established role of mitochondrial dysfunction and impaired mitochondrial translation in PD pathogenesis. H3C12 and H4C6 (histone proteins) may reflect epigenetic dysregulation increasingly recognised in PD. These observations suggest that the approach employed here recovers biologically credible candidates, though experimental validation through in vitro or in vivo models would be required to confirm the true relevance.

One of the main limitations of this study is the dependence on publicly available datasets, which may vary in quality, sample size, and experimental conditions, leading to inconsistencies across analyses. The Parkinson's dataset in particular showed higher variability and lower model performance, which is not ideal. Additionally, the downstream analyses relied on a single GEO dataset per disease which is a large limitation. Although the selected datasets were chosen based on DEG yield, sample size and tissue relevance, the findings may reflect dataset expression patterns rather than universal disease signatures resulting in a slight bias. Future work should incorporate multiple datasets to avoid this dataset bias. The models also showed lower recall compared to precision, indicating that some truly important genes may have been missed. In the future, larger and more diverse datasets can be incorporated in this study so that the model's performance can be further improved by exploring other more-efficient learning approaches and selecting appropriate features.

In conclusion, this study demonstrates a useful approach that combines biology with ML modelling and explainable AI to provide a structured method to predict potential drug targets. This makes the work important for early-stage drug discovery and understanding disease mechanisms, particularly in complex neurodegenerative disorders such as AD and PD. In today's world, where neurodegenerative disorders

such as AD and PD are at a rapid increase, there is a growing need for faster methods of drug discovery. Such approaches can reduce time and cost in research, support precision medicine, and contribute to developing better treatments.

## CONCLUSION

This study presents an integrated bioinformatics and machine learning framework for the analysis of Alzheimer's disease and Parkinson's disease using publicly available transcriptomic datasets. Differential expression analysis, protein–protein interaction network construction, pathway enrichment analysis, and machine learning were combined to identify genes and network features associated with disease-related molecular processes.

Network analysis identified several highly connected hub genes in both disorders. In Alzheimer's disease, ACTB, GAPDH, EGFR, and HSP90AB1 emerged as prominent network components, whereas Parkinson's disease was characterized by the predominance of mitochondrial ribosomal proteins, highlighting the potential contribution of mitochondrial dysfunction to disease pathology. Pathway enrichment analyses further demonstrated distinct biological signatures, with Alzheimer's disease showing enrichment of cell-cycle and autophagy-related pathways, while Parkinson's disease exhibited enrichment of neuronal signaling and synaptic transmission pathways.

Random Forest and XGBoost models achieved moderate predictive performance, with more consistent results observed for the Alzheimer's disease dataset. Feature importance analysis indicated that both differential expression measures and network-topological properties contributed substantially to model predictions, supporting the integration of transcriptomic and systems-level information for gene prioritization.

The findings provide insights into molecular characteristics associated with Alzheimer's disease and Parkinson's disease and demonstrate the utility of combining network biology and machine learning approaches in neurodegenerative disease research. Further validation using independent datasets and experimental studies will be required to assess the biological and clinical relevance of the identified genes and pathways.

## AUTHOR INFORMATION

**Vir Rastogi:** Aditya Birla World Academy, Mumbai, Maharashtra, India
**Nirupma Singh** Department of Biological Sciences and Engineering, Netaji Subhas University of Technology (Formerly NSIT), Dwarka, Delhi, India.

## REFERENCES

1. Borsook, D. (2012). Neurological diseases and pain. *Brain, 135*(2), 320–344. https://doi.org/10.1093/brain/awr271
2. Wang, S., Jiang, Y., Yang, A., Meng, F., & Zhang, J. (2024). The expanding burden of neurodegenerative diseases: An unmet medical and social need. *Aging and Disease, 16*, 2937–2952.
3. Gao, H.-M., & Hong, J.-S. (2008). Why neurodegenerative diseases are progressive: Uncontrolled inflammation drives disease progression. *Trends in Immunology, 29*(8), 357–365. https://doi.org/10.1016/j.it.2008.05.002
4. Bhattacharya, R., Bonner, M. G., & Musah, S. (2021). Harnessing developmental plasticity to pattern kidney organoids. *Cell Stem Cell, 28*.
5. Stacey, R. B., Hall, M. E., Leaverton, P. E., Schocken, D. D., & Zgibor, J. (2022). Relation of glycemic status with unrecognized MI and the subsequent risk of mortality: The Jackson Heart Study. *American Journal of Preventive Cardiology, 11*.
6. Vatansever, S., et al. (2021). Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: State-of-the-art and future directions. *Medical Research Reviews, 41*(3), 1427–1473. https://doi.org/10.1002/med.21764
7. Hu, J., et al. (2025). Multi-modal deep representation learning accurately identifies and interprets drug-target interactions. *IEEE Journal of Biomedical and Health Informatics, 29*.
8. Qureshi, S., et al. (2024). Enhancing drug-target interaction predictions in context of neurodegenerative diseases using bidirectional long short-term memory in male Swiss albino mice pharmaco-EEG analysis. *Heliyon, 10*.
9. Xie, Z., et al. (2025). Identification of therapeutic targets for Alzheimer's disease treatment using bioinformatics and machine learning. *Scientific Reports, 15*.
10. Zulhafiz, N. A., Teoh, T. C., Chin, A. V., & Chang, S. W. (2025). Drug repurposing using artificial intelligence, molecular docking, and hybrid approaches: A comprehensive review in general diseases vs Alzheimer's disease. *Computer Methods and Programs in Biomedicine, 261*.
11. Zhang, Y., Yu, L., Lv, Y., Yang, T., & Guo, Q. (2025). Artificial intelligence in neurodegenerative diseases research: A bibliometric analysis since 2000. *Frontiers in Neurology, 16*.
12. Julian, D. R., Bahramy, A., Neal, M., Pearce, T. M., & Kofler, J. (2025). Current advancements in digital neuropathology and machine learning for the study of neurodegenerative diseases. *The American Journal of Pathology, 195*.
13. Chen, Y., et al. (2024). Drug-target interaction prediction based on an interactive inference network. *International Journal of Molecular Sciences, 25*.