INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Parkinson's Disease Interstage Prognostic Biomarkers for Early Detection through Hybrid Machine Learning Model

Archana Chandrashekhar Magare[1]*, Maulika Patel[2]

1 PhD Scholar, Gujarat Technological University, Ahmedabad, Gujarat, India
2 Professor & Head, Department of Computer Engineering,, G H Patel College of Engg. & Tech, CVM University, Vallabh Vidyanagar, Gujarat, India

OPEN ACCESS

Received: 12-07-2024
Accepted: 04-10-2024
Published: 29-10-2024

Citation: Magare AC, Patel M (2024) Parkinson's Disease Interstage Prognostic Biomarkers for Early Detection through Hybrid Machine Learning Model. Indian Journal of Science and Technology 17(40): 4209-4220. https://doi.org/10.17485/IJST/v17i40.2254

* **Corresponding author**.

archanamagare@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: Parkinson's Disease (PD) is a neurodegenerative condition that takes 15 to 20 years to exhibit the symptoms. However, during the initial stages of the disease, there are microlevel changes in the body that are typically undetected. The pathogenesis of Parkinson's disease requires understanding of these microlevel changes. **Objectives:** The objective of the present study is to develop a hybrid ensembled machine learning pipeline model for identifying inter-stage Parkinson's Disease (PD) biomarkers for understanding disease progression as well as etiology. **Methods:** The proposed work was carried out on the dataset GSE202667 from Gene Expression Omnibus, containing time-resolved RNA signatures of CD4+ T cells at various stages. Differentiating genes were identified in different interstage groups. Two types of unsupervised learning methods- distance-based (K-means, Agglomerative, Density-Based Spatial Clustering of Application with Noise) and probability-based (Hidden Markov Model, Gaussian Mixture Model Latent Dirichlet Allocation) were applied. The best algorithms were selected and applied to optimize clusters. Enrichment analysis was conducted on the top 10 PD biomarkers in each category. **Findings:** The top 10 PD biomarkers in each category are identified and gene set enrichment analysis resulted into their enrichment in three KEGG Pathways and depletion in two GO molecular functions**.** These biomarkers' depletion is observed in 215 Reactome pathways and enrichment in 18 Reactome pathways. Twelve GO-Cellular components had an enrichment whereas 111 GO-Cellular components had a depletion in the gene set. A total of 25 GO-Biological process components were enriched, while 339 GO-Biological process components were depleted. **Novelty:** The proposed hybrid ensembled machine learning pipeline model works as a tool to identify Parkinson's Disease biomarkers from omics data. The model contributes to identify implicit patterns in omics data in order to unveil Parkinson's disease progress mechanism through biomarkers discovery.

Keywords: Gaussian Mixture Model clustering; Early Detection; k-means clustering; Machine learning; Parkinson's Disease Interstage Biomarkers

1 Introduction

Parkinson's Disease (PD) is a neurodegenerative disease that impacts on dopamine-making substantia nigra area of the brain. The progression of the disease develops slowly over several years. The progression symptoms vary from person to person due to the disease's diverse nature. The main characteristic of Parkinson's Disease is there is neuronal death due to dopamine loss. There are evidences that the autoimmune system is responsible for the dopaminergic cell death [(1)]. CD4+ T cells are T helper cells. They act as a conductor for immune response. Peripheral CD4+ T cells play key roles in neuroinflammation. The dysregulated peripheral CD4+ T cells are also correlated with cognitive decline in Parkinson's Disease [(2)]. The T cells play a significant role in the pathogenesis of Parkinson's Disease [(2)].

The existing studies suggest that the t-cells interaction with peripheral blood and other body fluid impacts on the central nervous system along with dysfunction of the immune system [(3)].

Previous studies are for identifying various types of Parkinson's Disease biomarkers through machine learning, but the time-resolved RNA signature along with the CD4+ T-cells activation and the activation effect at various Parkinson's Disease stages is not been explored to the fullest. The studies have been carried out to identify significant genes with altered expressions in different time frames [(2)].

The RNA signatures along with CD4+ T cells activation at different Parkinson's Disease stages and its interstage effect in order to identify differentiating biomarkers along with biological significance needs to be explored. Such RNA Signature analysis will help to understand the minute level alterations between different Parkinson's Disease stages. Various studies have identified the significant RNA signatures in different time frames but interstage RNA signatures and their implicit cohorts need to be uncovered.

These cohorts will help in recognizing differentiating features during various stages of Parkinson's disease. Unsupervised learning algorithms aim to identify implicit cohorts in the input data.

The reason behind the selection of an ensemble of clustering algorithms of different categories and then constructing the proposed framework based on combined approaches is to get the complementary strengths of both distance-based clustering algorithms and probability-based clustering algorithms. Distance-based clustering algorithms are suitable for large datasets as they are computationally efficient and scalable. Moreover, these algorithms are versatile as they have the ability to incorporate various distance metrics. These methods can handle arbitrary shape clusters.

Probability-based clustering methods handle overlapping clusters in complex input data. These clustering methods can refine the clusters and capture subtler patterns.

The proposed validation-based hybrid ensembled unsupervised learning approach is used to identify preliminary Parkinson's Disease interstage biomarkers. The performance of each unsupervised learning method is evaluated through the Davies-Bouldin Index (DBI), Calinski- Harabasz Index (CHI), and Silhouette Score. Initially, two different categories of unsupervised learning methods-distance based learning methods and probability-based learning methods are applied independently to identify primary groups and then the optimization of clusters is performed by applying a hybrid clustering pipeline through a validated set of methods in each category. This resulted in identifying a set of interstage PD biomarkers. The biological significance of the biomarkers is explored through gene set enrichment analysis.

Thus, the proposed work aims to find the differentiating RNA expression biomarkers between various stages of Parkinson's Disease (PD) through various machine learning techniques with the help of input data GSE202667 [2] which is RNA expression data collected through activation of peripheral CD4+ T cells. The comparative analysis of each group's unsupervised learning method is carried out in order to find the best clustering technique for identifying implicit clusters in the input data based on various evaluation metrics. The enrichment analysis is carried out to get the biological significance of the Parkinson's Disease interstage biomarkers.

1.1 Parkinson's Disease and Different Immune Cells

Guang Yang et al. in [4], identified endoplasmic Reticulum stress Parkinson's Disease biomarkers using Random Forest along with LASSO logistic regression. The correlation between PD diagnostic biomarkers and immune cell penetration is explored through Spearman's rank correlation method. Two genes- ABCA7 and TBL2 are identified as diagnostic biomarkers. Plasma cells, NK cells, and resting CD4 memory cells exhibited a positive correlation with identified diagnostic PD biomarkers.

Yanchen Chen et al. in [5], addressed the immuno-pathogenesis of Parkinson's Disease in order to find inflammatory PD biomarkers. Immune-related differential expressed genes were identified. Eight genes- AVP, CBL, CD3E, GALR3, RELA, OGFR, and SEMA6B were recognized as dysregulated genes. CBL is identified as a potential diagnostic biomarker.

Wu Jiecong et al. in [6], explored therapeutic target PD genes using weighted genes co-expression network analysis along with Random Forest, Support Vector Machine recursive feature elimination, and LASSO Regression. Two hub genes- SV2C and DENR for PD are identified as diagnostic PD biomarkers.

Guanghao Xin et al. in [7], identified four immune-related Parkinson's Disease hub genes- dld, dlk1, iars, and ttd19 using LASSO and mSVM-RFE algorithms. Gene Set Enrichment Analysis (GSEA) and Gene Set Variation Analysis (GSVA) were performed to identify significant genes.

Y Huang et al. in [8], used weighted gene co-expression network analysis along with LASSO Regression, Support Vector Machine, and Random Forest to screen out immune-related genes in Parkinson's Disease. A total of 218 immune-related genes were identified and ANK1 was identified as the hub gene.

Shaonan Hu et al. in [9], explored the genetic connection between Parkinson's Disease and periodontitis via immunology. Inflammatory response-related genes from both disease samples were clustered using k-means clustering. The LASSO regression was used for feature selection. Five genes were found to be associated with both diseases, indicating immunological involvement.

1.2 Parkinson's Disease stages and progress biomarkers

John Michael Templeton et al. in [10], used a Decision Tree classification algorithm to compare tablet-based assessment and self-reported metrics for various [Hoehn and Yahr (H&Y) Stages 1–5] stages of Parkinson's disease. The decision Tree algorithm is further used to compare perceived and sensor-based neurocognitive abilities of individuals with PD. The at advanced stages of PD, the individuals have elevated perception as compared to the initial stage. The sensor score is lower at the beginning in comparison with the advanced stages of PD.

M. B. Makarious et al. in [11], addressed the use of a multi-modality approach for early detection of Parkinson's Disease. The RNA and DNA sequencing data were preprocessed separately and extremely randomized tree classifiers were used on combined modality data in order to remove redundant features. The model with the Adaboost classifier algorithm was the best-performing classifier with the highest accuracy which was further optimized using hyber parameter tuning and cross-validation.

Ouiem Bchir in [12], developed a classifier for automatically detecting speech disorder in Parkinson's Disease. The classifier categorized auditory information and detected articulatory impairments. In order to manage the high dimensional data, relevant feature weighting was performed using the Gaussian Mixture Model (GMM).

Tran et al. in [13], worked on automating the diagnostics of Parkinson's Disease. The study used retinal fundus imaging to predict the prevalence and incidents of Parkinson's Disease through deep learning.

Birkenbihl et al. in [14], analyzed multi-modal data in order to identify patients' subtypes and disease progression using machine learning.

Zhang et al. in [15], detected the network biomarkers from gene co-expression network data using Gaussian mixture model clustering which can detect clusters of any shape such as strip or oval.

2 Methodology

2.1 Dataset Used

This work used the dataset GSE202667 [2], which contains time-resolved RNA markers of CD4+ T cells in Parkinson's disease. The dataset is downloaded from the Gene Expression Omnibus (GEO)database (https://www.ncbi.nlm.nih.gov/geo/). The samples are collected from peripheral blood samples for subsequent T-cell isolation [2]. The samples are from 5 Parkinson's Disease patients of stages 1, 2.5, and 4 respectively, and five healthy donors [2]. The dataset GSE202667 [2] has a total of 60 samples (PD of different stages=30 and Healthy Control=30) of Parkinson's Disease stage 1,2.5, 4 and healthy controls along with the CD4+ T Cells activation status in the 24hrs time frame. This activation status is collected at 0,2,4,8,12, 24 hours [2]. We selected 35 samples of PD Stage 1, Stage 2.5, Stage 4, and PD Stage 0 that are healthy controls along with CD4+ activation for the span of 24 hrs as given in the dataset.

2.2 Evaluation Metrics

The performance of various clustering algorithms is evaluated using the following evaluation metrics.

i. Davies-Bouldin Index (DBI) [16] method examines the quality of the clusters. It measures how well the clusters are separated from each other and how compact each cluster is. Thus, it focuses on measuring inter-cluster differences and intra-cluster similarity. Lower DBI values indicate better clustering.

ii. Calinski- Harabasz Index (CHI) [17] cluster evaluation method evaluates the Variance Ratio Criterion which is the ratio of the sum of inter-cluster distance and intra-cluster dispersion for all clusters. Higher CHI values indicate better clustering

iii. Silhouette Score [18] is used to evaluate the cluster quality. It measures how each data point is similar to its own cluster as compared to other clusters. The Silhouette score is calculated by finding each data point's average distance to its own cluster and the minimum average distance to other clusters.

2.3 Proposed Methodology

The aim of the proposed work was to identify the interstage Parkinson's disease biomarkers so differentially expressed transcriptomes are identified individually within the pair (PDStage4 Vs PDStage0, PDStage1 Vs PDStage4, PDStage2.5 Vs PDStage1, PDStage0 Vs PDStage2.5) through LIMMA package in R [19]. The pairwise comparison will help to understand the expression alterations that drive the progression of the disease as certain genes may be upregulated or downregulated at specific stages. LIMMA package is used as it uses the empirical Bayes method and linear modeling which gives the optimized results for large datasets. The proposed methodology is a validation-based hybrid ensembled optimized clustering pipeline approach that identifies interstage differentiating Parkinson's Disease biomarkers. Figure 1 shows the methodology for identification of interstage Parkinson's Disease Biomarkers. The input RNA expression values are preprocessed using log2 transformation and quantile normalization in order to stabilize the variance across the range of RNA expression values along with making the expression values consistent across all samples. Then the experimental design is defined through a model matrix where each row represents input sample categories and the column represents different variables in the experiment. Next, the linear regression is used to fit the linear model to preprocessed RNA expression data. The top-ranked differentially expressed RNA expressions are selected for further analysis based on t-statistics and p value<=0.05. Then for each group of DEGs, standardization that is z-score normalizationwas performed so that each feature has a mean of 0 and a standard deviation of 1. This was performed because the features in the input data with different ranges can disproportionately influence the machine learning model. Standardization ensures that each feature in the input data set contributes equally to the model. Then distance-based clustering methods means, Agglomerative and DBSCAN clustering and Probability-based clustering methods- Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), and Latent Dirichlet Allocation (LDA) were applied to the standardized data. Davies-Bouldin Index, Calinski-Harabasz Index, and Silhouette Score were used to evaluate the performance of k-means, Agglomerative, HMM, GMM, and LDA methods. DBI and CHI were not calculated for DBSCAN since these methods assume that all points are assigned to clusters but in DBSCAN, noise points are labeled as '-1', indicating that they do not belong to any cluster. Figures 2, 3, 4, 5, 6, 7, 8 and 9 show the results of cluster assignments of distance-based clustering methods. Table 1 shows the evaluation metrics scores of k-means, DBSCAN, and Agglomerative clustering. DBSCAN algorithm has the highest Silhouette score as compared to the other two methods. Based on the average rank of the other two scores, k-means clusters are concise and compact as shown in Figures 2, 3, 4, 5, 6, 7, 8 and 9. The goal of this study was to obtain clusters that are set of inter-stage PD biomarkers that exhibit distinct dispersion and separation, allowing for clear identification of interstage biomarkers. To achieve this, the proposed methodology selected k-means clustering for further analysis, as it outperformed Agglomerative clustering in terms of CHI score, although it did not perform as well in terms of DBI score in some cases.

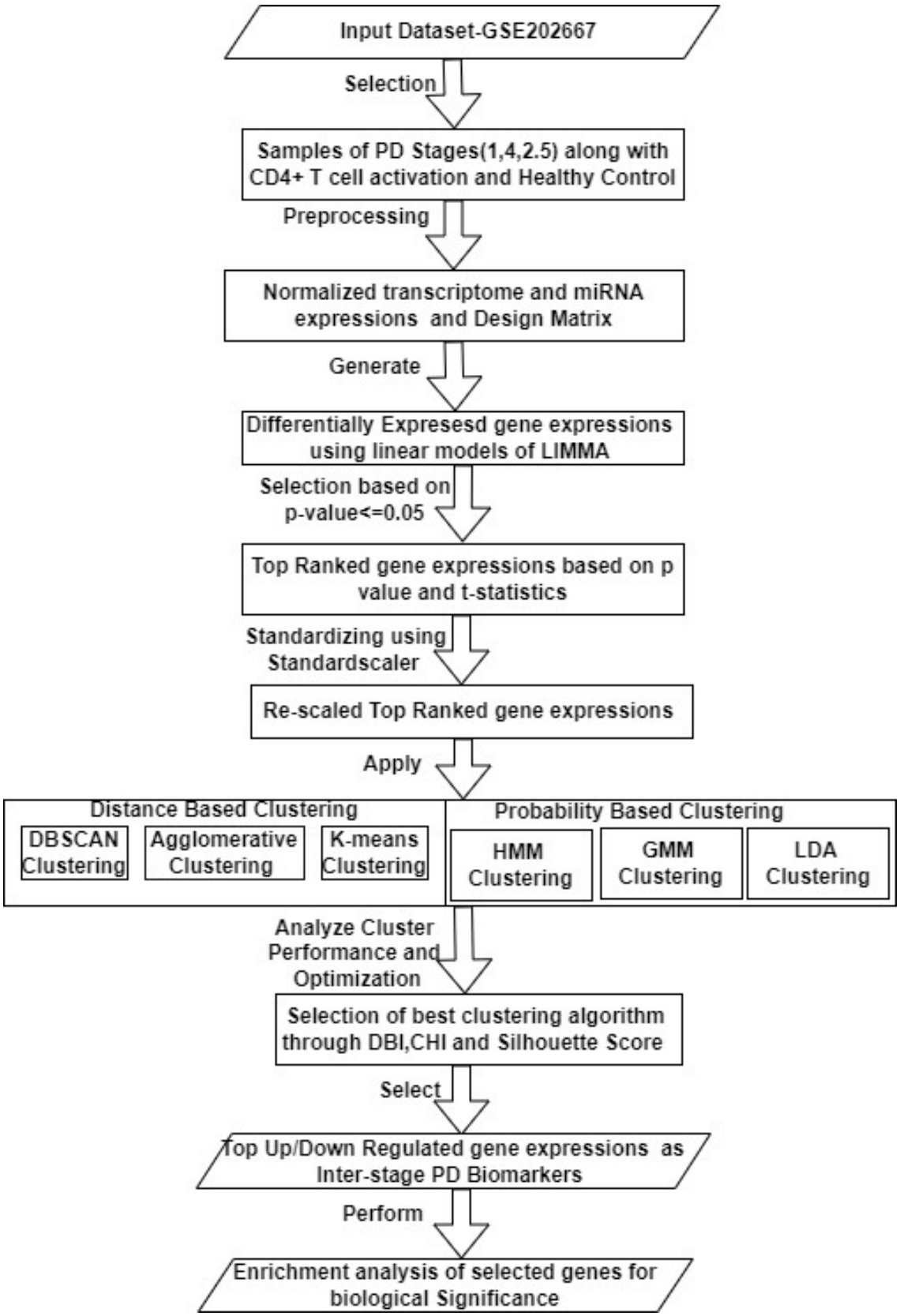

Fig 1. Machine learning-based predictive model for interstage Parkinson's Disease Biomarker Identification

Table 1. Evaluation Metrics Scores

Dataset	Algorithms	Evaluation Metrics	Score
PD Stage1Vs Stage4	K-Means	Silhouette	0.4746
	Agglomerative		0.4908
	DBSCAN		0.7382
	KMeans	Davies-Bouldin Index	0.9824
	Agglomerative		0.8004
	KMeans	Calinski-Harabasz Index	11571.9211
	Agglomerative		9747.4383
PDStage4VsStage0(HC)	KMeans	Silhouette	0.4513
	Agglomerative		0.4509
	DBSCAN		0.6903
	KMeans	Davies-Bouldin Index	1.1426
	Agglomerative		1.1589
	KMeans	Calinski-Harabasz Index	4891.6207
	Agglomerative		4669.4331
PDStage2.5 VsStage1	KMeans	Silhouette	0.4770
	Agglomerative		0.5060
	DBSCAN		0.4010
	KMeans	Davies-Bouldin Index	0.9311
	Agglomerative		0.8136
	KMeans	Calinski-Harabasz Index	4798.53044
	Agglomerative		3777.7416
PDStage0(HC) Vs PDStage2.5	KMeans	Silhouette	0.5869
	Agglomerative		0.5822
	DBSCAN		0.62311
	KMeans	Davies-Bouldin Index	0.9920
	Agglomerative		1.0668
	KMeans	Calinski-Harabasz Index	1199.0013
	Agglomerative		1133.2815

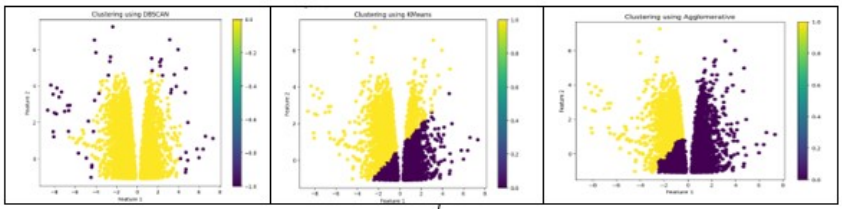

Fig 2. Cluster assignments of DBSCAN, K-means, and Agglomerative for PD Stage1 Vs Stage4

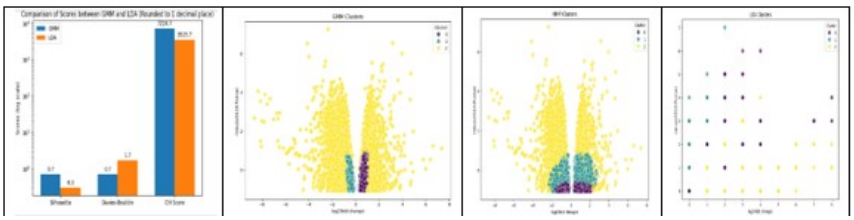

Fig 3. Cluster assignments of GMM, LDA and HMM for PD Stage1 Vs Stage4

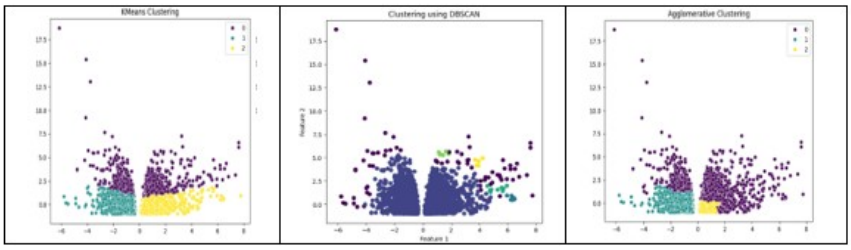

Fig 4. Cluster assignments of K-means, Agglomerative and DBSCAN for PD Stage4 Vs Stage0

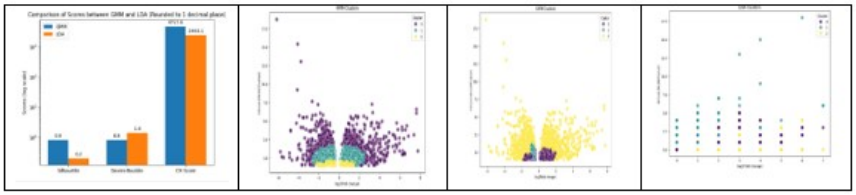

Fig 5. Cluster assignments of GMM, HMM and LDA for PD Stage4 Vs Stage0

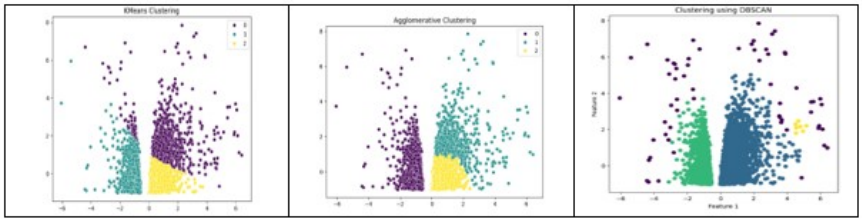

Fig 6. Cluster assignments of K-means, Agglomerative and DBSCAN for PD Stage2.5 Vs Stage1

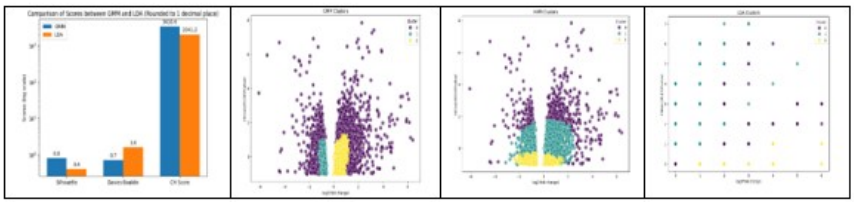

Fig 7. Cluster assignments of GMM, HMM and LDA for PD Stage 2.5 Vs Stage1

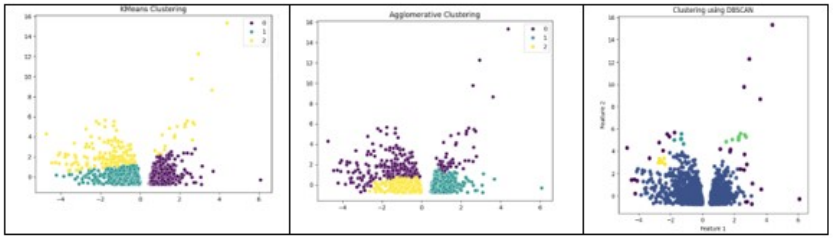

Fig 8. Cluster assignments of K-means, Agglomerative and DBSCAN for PD Stage0(HC) Vs Stage2.5

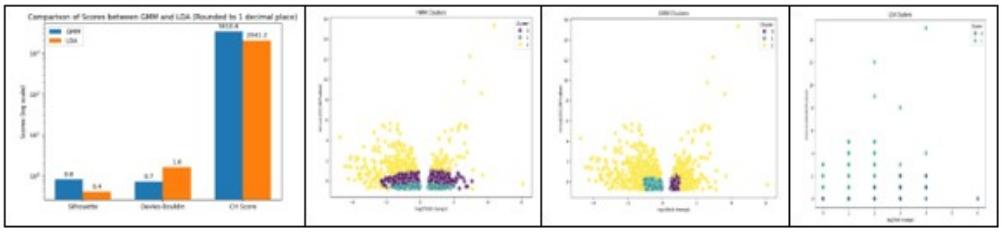

Fig 9. Cluster assignments of GMM, HMM and LDA for PD Stage 0(HC)Vs Stage2.5

Figure 2 to Figure 9 shows cluster assignments of distance-based clustering methods and probabilistic clustering methods for interstage groups. It is observed that in traditional distance-based clustering, k means has outperformed and in the probabilistic clustering model GMM clustering has shown well-defined clusters as compared to the other methods. So, in order to get the optimal set of interstage Parkinson's Disease biomarkers the combined hybrid clustering model is proposed. The pseudocode for the hybrid ensembled clustering model is as follows.

Input: Dataset of PDInterstageTopDEGs

Output: Optimal set of PD interstage biomarkers

1. Preprocess data by converting all columns into numeric and dropping the NA values.
2. Determine an optimal number of clusters by finding inertia and silhouette scores for the 'k' cluster ranging from 2 to 10.
3. Determine optimal components for GMM by finding the BIC (Bayesian Information Criterion) score and AIC (Akaike Information Criterion) score.
4. Fit the k-means with the optimal number of clusters on 'Input Data' and assign labels as k_means_labels.
5. Fit the GMM with optimal k components on 'Input Data' and assign labels as gmm_labels.
6. Combine cluster assignments and Relabel the clusters to match between k-means and GMM algorithms.
7. Refine the clusters using Agglomerative clustering.
8. Visualize the final cluster results.

Figure 10 to Figure 13 shows well-defined clusters of PD interstage biomarkers based on a validation-based hybrid ensembled clustering model.

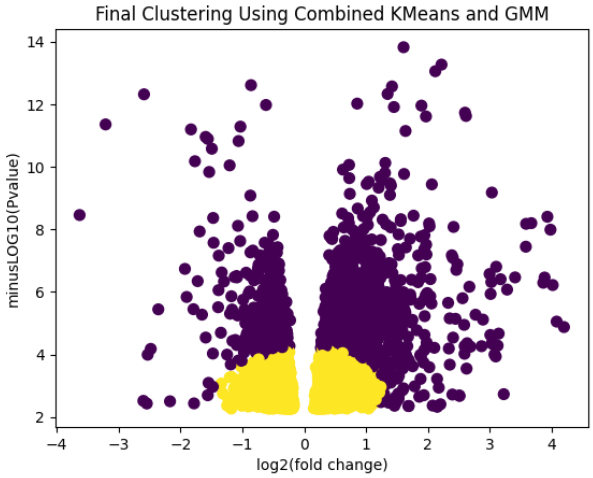

Fig 10. Final Clustering using hybrid clustering model for PDStage2.5 Vs Stage1

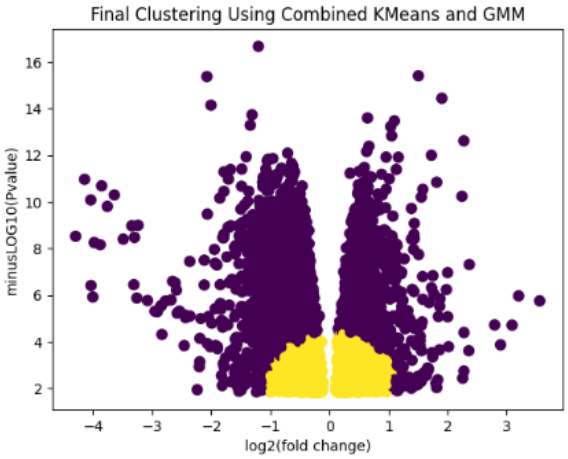

Fig 11. Final Clustering using hybrid clustering model for PDStage1 Vs Stage4

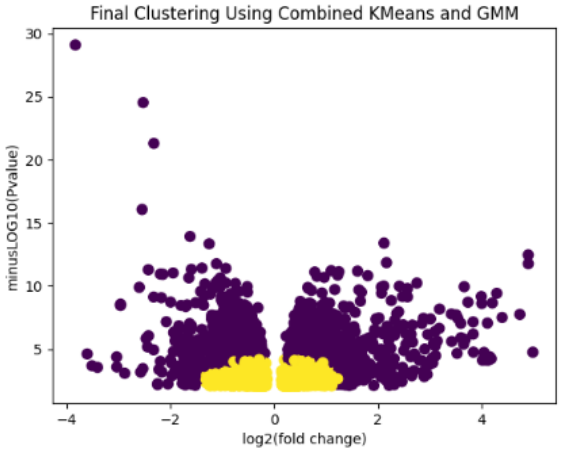

Fig 12. Final Clustering using hybrid clustering model for PDStage4 Vs Stage0(HC)

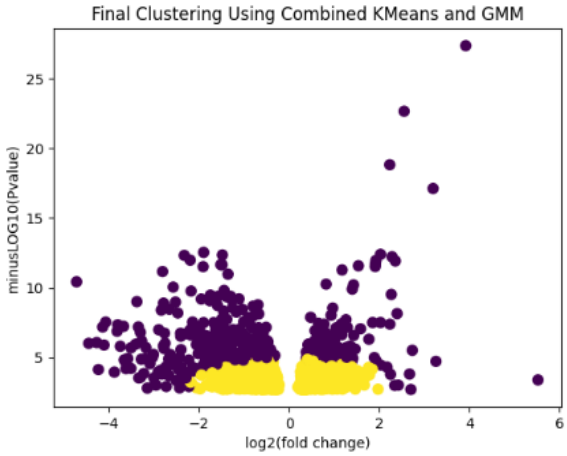

Fig 13. Final Clustering using hybrid clustering model for PDStage0(HC) Vs Stage2.5

3 Results and Discussion

The proposed approach identified the set of PD interstage biomarkers. The biological significance of these biomarkers reveals in which pathways or gene ontology components these RNAs are enriched or depleted. So top selected PD interstage biomarkers were analyzed to get enrichment analysis of these genes in biological processes and pathways. Figure 14 shows a Volcano plot of gene set enrichment analysis using GeneTrail3.2 [20]. As shown in Figure 14, Gene Set Enrichment Analysis was performed using GeneTrail3.2 [20] where the selected genes were shown depletion in 215 Reactome pathways and enrichment in 18 Reactome pathways. The top genes were found to be enriched in seven KEGG pathways and deficient in 46 KEGG pathways. Twelve GO-Cellular components had an enrichment whereas 111 GO-Cellular components had a depletion in the gene set. 25 GO-Biological process components were enriched, while 339 GO-Biological process components were depleted. 52 GO-Molecular Functions showed depletion and 15 GO-Molecular Functions showed enrichment in the input gene collection. These sets of genes are depleted in two GO Molecular Function components that is Molecular Function and protein binding. The enrichment has been shown in one of the Wiki pathways known as Vitamin D Receptor Pathways.

ists the up and downregulated Parkinson's Disease Inter-Stage Biomarkers. The up-regulated genes are overexpressed between the condition and down-regulated genes are under-expressed between two groups. The enrichment analysis of these up-regulated and down-regulated genes was performed, which showed that these genes enriched in three KEGG pathways-Adrenergic signaling in cardiomyocytes, Dilated cardiomyopathy (DCM), and Oxytocin signaling pathway. This enrichment indicates autonomic dysfunction [21], adrenergic dysfunction, and cardiac manifestations in PD.

Along with performing the enrichment analysis, the literature is also explored to find associations of these interstage biomarkers with neurodegenerative diseases or any other diseases as mentioned in Table 2.

APOC1 is identified as downregulated interstage PD biomarkers between PD stage 4 Vs stage 1 in our study, in literature [22] it is identified as dysregulated and overlapped risk genes between Alzheimer's Disease and PD.

CST6 has been found as a CSF diagnostic biomarker for neuropsychiatric systemic lupus erythematosus [23]. It is recognized as a PD Interstage biomarker that is downregulated in Stage 4 compared to Stage 1.

FAM9C is recognized as an up-regulated interstage PD biomarker between PD stage 4 and healthy control in the proposed work. It is analyzed as an altered expression in cblC and cblG groups of fibroblasts [24].

SNHG5 is identified as up regulated PD biomarker in stages 2.5 and Stage 0, in previous studies, it is one of the significant long non-coding RNAs for regulating programmed cell death which causes initiation of neurodegenerative diseases such as Alzheimer's Disease and Parkinson's Disease [25].

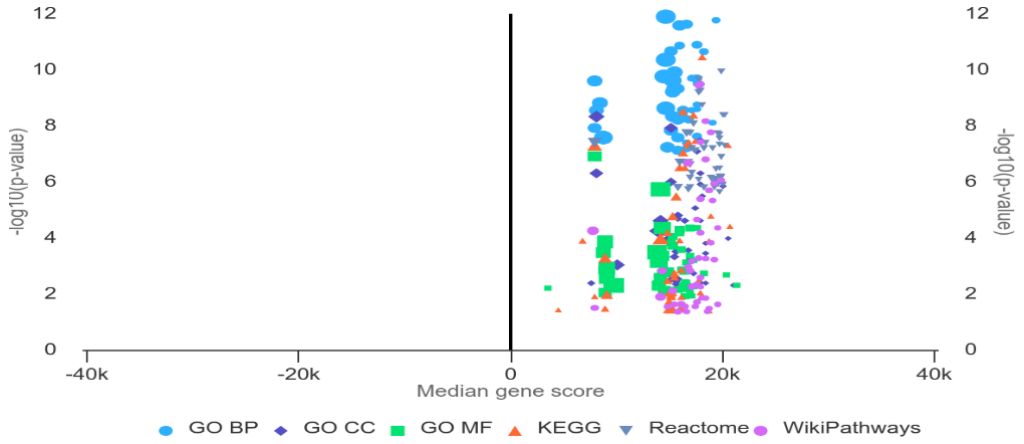

Fig 14. Volcano plot of gene set enrichment Analysis through GeneTrail

Table 2. Parkinson's Disease Inter-Stage Gene Biomarkers

Dataset	Gene Name	Regulation
PDStage4 Vs Stage 1	APOC1	Down
	CST6	Down
	lnc-LIX1-1	Down

Continued on next page

Table 2 continued

	THSD7A	Down
	KRT86	Down
	lnc-DYDC1-4	UP
	lnc-FCRL3-1	UP
	GOLGA8EP	UP
	lnc-FAM27D1.1-3	UP
	ALG1	UP
PDStage0(HC) Vs PDStage2.5	CACNG6	Down
	LINC00944	Down
	IL1RL1	Down
	ALDH1A2	Down
	EMR2	Down
	LOC403323	UP
	lnc-FAM75A6-2	UP
	LOC403323	UP
	SNHG5	UP
	ACAD10	UP
PD Stage4Vs Stage 0(HC)	ISM1	Down
	lnc-AP4S1-2	Down
	LINC01424	Down
	XLOC_l2_003856	Down
	LBH	Down
	ADAMTS9-AS1	UP
	CTXN2	UP
	FAM9C	UP
	LINC01556	UP
	lnc-TCERG1-2	UP
PD Stage 2.5 Vs Stage1	HERC2P7	Down
	lnc-LIX1-1	Down
	lnc-MBP-1	Down
	ZNF880	Down
	lnc-GMDS-3	Down
	lnc-LINGO2-2	UP
	LINGO2	UP
	KIR2DL2	UP
	GNAS	UP
	FSD1	UP

4 Conclusion

The proposed work focused on identifying the interstage Parkinson's disease time-resolved RNA signatures of CD4+ T cell biomarkers. This work proposed a novel hybrid ensembled pipeline of machine learning methods for the identification of Parkinson's Disease Inter-stage biomarkers. A total of 40 Parkinson's Disease interstage biomarkers are identified. These inter-stage PD biomarkers can be useful to understand Parkinson's disease progress and its etiology. It can be also helpful in understanding the role of each biomarker at different disease stages along with critical stages. Thus, the study can be useful for understanding transitional insights along with stage-specific understanding.

The literature was explored to find the association of these biomarkers with other diseases in which a total of four interstage biomarkers - APOC1, CST6, FAM9C**,** and SNHG5 are found to be associated with other diseases such as Alzheimer's disease, neuropsychiatric systemic lupus erythematosus, fibroblasts.

In the future, there is a need to explore the proposed approach on different biological experiment categories for gaining early insight into disease progression.

References

1) Tansey MG, Wallings RL, Houser MC, Herrick MK, Keating CE, Joers V. Inflammation and immune dysfunction in Parkinson disease. *Nature Reviews Immunology*. 2022;22(11):657–73. Available from: https://www.nature.com/articles/s41577-022-00684-6.pdf.

2) Diener C, Hart M, Kehl T, Becker-Dorison A, Tänzer T, Schub D, et al. Time-resolved RNA signatures of CD4+ T cells in Parkinson's disease. . *Cell Death Discovery*. 2023;9:18–18. Available from: https://www.nature.com/articles/s41420-023-01333-0.pdf.
3) Wang P, Yao L, Luo M, Zhou W, Xu JX, Yan Z, et al. Single-cell transcriptome and TCR profiling reveal activated and expanded T cell populations in Parkinson's disease. . *Cell Discovery*. 2021;7:52–52. Available from: https://www.nature.com/articles/s41421-021-00280-3.
4) Yang G, Zhang B, Xu CY, Wu JW, Zhang Y, Yu Y, et al. Utilizing Machine Learning to Identify Biomarkers of Endoplasmic Reticulum Stress and Analyze Immune Cell Infiltration in Parkinson's Disease. . *Molecular Neurobiology*. 2024;23:1–8. Available from: https://link.springer.com/article/10.1007/s12035-024-03948-5.
5) Chen Y, Tu Y, Yan G, Chen JX, Niu S, Liao C, et al. Integrated Bioinformatics Analysis for Revealing CBL is a Potential Diagnosing Biomarker and Related Immune Infiltration in Parkinson's Disease. *International Journal of General Medicine*. 2024;p. 2371–86. Available from: https://www.tandfonline.com/doi/full/10.2147/IJGM.S456942.
6) Wu J, Wu W, Jiang P, Xu Y, Yu M. Identification of SV2C and DENR as Key Biomarkers for Parkinson's Disease Based on Bioinformatics, Machine Learning, and Experimental Verification. *Journal of Molecular Neuroscience*. 2024;74(1). Available from: https://link.springer.com/article/10.1007/s12031-023-02182-3.
7) Xin G, Niu J, Tian Q, Fu Y, Chen L, Yi T, et al. Identification of potential immune-related hub genes in Parkinson's disease based on machine learning and development and validation of a diagnostic classification model. *PloS One*. 2023;18(12). Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0294984.
8) Huang Y, Wang A, Wang F, Xu Y, Zhang W, Shi F, et al. Screening of key immune-related gene in Parkinson's disease based on WGCNA and machine learning. Zhong nan da xue xue bao. *Journal of Central South University Medical Sciences*. 2024;49(2):207–226. Available from: https://europepmc.org/article/pmc/pmc11103055.
9) Hu S, Li S, Ning W, Huang X, Liu X, Deng Y, et al. Identifying crosstalk genetic biomarkers linking a neurodegenerative disease, Parkinson's disease, and periodontitis using integrated bioinformatics analyses. . *Frontiers in Aging Neuroscience*. 2022;14:1032401–1032401. Available from: https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2022.1032401/full.
10) Templeton JM, Poellabauer C, Schneider S. Classification of Parkinson's disease and its stages using machine learning. . *Scientific reports*. 2022;12:14036–14036. Available from: https://www.nature.com/articles/s41598-022-18015-z.
11) Makarious MB, Leonard HL, Vitale D, Iwaki H, Sargent L, Dadu A, et al. Multi-modality machine learning predicting Parkinson's disease. . *npj Parkinson's Disease*. 2022;8:35–35. Available from: https://www.nature.com/articles/s41531-022-00288-w.
12) Bchir O. Parkinson's disease classification using Gaussian mixture models with relevance feature weights on vocal feature sets. *International Journal of Advanced Computer Science and Applications*. 2020;11(4). Available from: https://thesai.org/Downloads/Volume11No4/Paper_56Parkinsons_Disease_Classification_using_Gaussian_Mixture.pdf.
13) Tran C, Shen K, Liu K. Deep learning predicts prevalent and incident Parkinson's disease from UK Biobank fundus imaging. *Sci Rep*. 2024;14:3637–3637. Available from: https://doi.org/10.1038/s41598-024-54251-1.
14) Birkenbihl C, Ahmad A, Massat NJ. Artificial intelligence-based clustering and characterization of Parkinson's disease trajectories. *Sci Rep*. 2023;13:2897–2897. Available from: https://doi.org/10.1038/s41598-023-30038-8.
15) Zhang H, Zhu Z, Li H, He S. Network Biomarker Detection From Gene Co-Expression Network Using Gaussian Mixture Model Clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2023. Available from: https://doi.org/10.1109/TCBB.2023.3297388.
16) Ros F, Riad R, Guillaume S. PDBI: A partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing*. 2023;528:178–99. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0925231223000528.
17) Ashari IF, Nugroho ED, Baraku R, Yanda IN, Liwardana R. Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and rand-index evaluation on k-means algorithm for classifying flood-affected areas in Jakarta. *Journal of Applied Informatics and Computing*. 2023;7(1):95–103. Available from: https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/4947.
18) Lenssen L, Schubert E. Medoid Silhouette clustering with automatic cluster number selection. . *Information Systems*. 2024;120:102290–102290. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0306437923001266.
19) Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43. Available from: https://doi.org/10.1093/nar/gkv007.
20) Gerstner N, Kehl T, Lenhof K, Müller A, Mayer C, Eckhart L, et al. GeneTrail 3: advanced high-throughput enrichment analysis. *Nucleic Acids Research*. 2020;48(W1):W515–W520. Available from: https://dx.doi.org/10.1093/nar/gkaa306.
21) Palma JA, Thijs RD. Treatment of autonomic dysfunction in Parkinson disease and other synucleinopathies. *Movement Disorders*. 2024;p. 81–92. Available from: https://doi.org/10.1002/mds.27344.
22) Torshizi AD, Truong DT, Hou L, Smets B, Whelan CD, Li S. Proteogenomic network analysis reveals dysregulated mechanisms and potential mediators in Parkinson's disease. *Nature Communications*. 2024;15:6430–6430. Available from: https://dx.doi.org/10.1038/s41467-024-50718-x.
23) Ni J, CChen, Wang S, Liu X, Tan L, Lu L, et al. Novel CSF biomarkers for diagnosis and integrated analysis of neuropsychiatric systemic lupus erythematosus: based on antibody profiling. *Arthritis Research & Therapy*. 2023;25(1):165–165. Available from: https://dx.doi.org/10.1186/s13075-023-03146-z.
24) Rashka C, Hergalant S, Dreumont N, Oussalah A, Camadro JM, Marchand V, et al. Analysis of fibroblasts from patients with cblC and cblG genetic defects of cobalamin metabolism reveals global dysregulation of alternative splicing. *Human Molecular Genetics*. 2020;29(12):1969–1985. Available from: https://dx.doi.org/10.1093/hmg/ddaa027. doi:10.1093/hmg/ddaa027.
25) Zhao N, Wang J, Huang S, Zhang J, Bao J, Ni H, et al. The landscape of programmed cell death-related lncRNAs in Alzheimer's disease and Parkinson's disease. *Apoptosis*. 2024;29(9-10):1584–1599. Available from: https://dx.doi.org/10.1007/s10495-024-01984-z.