Identification of Parkinson's Biomarkers:
Using Univariate Feature Selection and Comparative Machine Learning

Thabsheer Machingal
thabsheerjm1@gmail.com

***Abstract*—Diagnosing Parkinson's Disease early is a major challenge because physical symptoms often appear only after significant neurodegeneration has occurred . This study used machine learning to identify hidden patterns in patient blood samples that could serve as early warning signs. We combined standard statistical tests with three different models(Logistic Regression, SVM, Random Forest) to find a reliable set of biomarkers, genes that consistently present in patient samples. Our "consensus" approach identified three key genes, *CEACAM4*, *IFI27*, and *ITCH*. that distinguished patients from healthy individuals with over 85% accuracy. Importantly, biological analysis linked these genes to known problems in Parkinson's, such as inflammation and cellular energy failure. These findings suggest that using multiple models together can uncover robust biological signals, leading to early detection.**

***Index Terms*—Bioinformatics, Computational Biology, Machine Learning**

I. Introduction

Bioinformatics combines biology, computer science, and statistics to analyze complex biological data, such as gene expression profiles. In the context of Parkinson's Disease (PD), identifying reliable blood-based biomarkers is critical for early diagnosis, as current clinical methods rely heavily on observing motor symptoms after significant neurodegeneration has already occurred. By applying computational methods to transcriptomic data, we can uncover hidden genetic signatures that signal the disease in its earliest stages.

The remainder of this paper is organized as follows: Section II reviews related work in transcriptomic analysis for PD. Section III details our methodology, integrating univariate statistical filtering with comparative machine learning. Section IV presents the experimental results and the identification of consensus biomarkers, followed by the conclusion and future directions in Section V.

II. Related Work

The application of machine learning in transcriptomic profiling has been a promising strategy for identifying non-invasive biomarkers for Parkinson's Disease(PD). Recent studies have demonstrated that computational approaches can effectively decode complex gene expression patterns that traditional univariate statistics often miss. For instance, [1] conducted a meta-analysis of blood-based transcriptomics, applying classification algorithms like Support Vector Machines and Random Forest to distinguish PD patients from healthy controls. Their work highlighted the critical need for robust feature selection to prevent overfitting in high-dimensional datasets, a challenges directly addressed in our study through multi-model consensus approach. Similarly, [2] utilized early microarray data to develop a predictive spline model, successfully identifying a set of biomarkers including *VDR* and *HIP2* that correlated with disease risk.

More recently, [3] leveraged large scale RNA sequencing data to build a multi-omics classifier. Their findings confirmed that integrating transcriptomic signatures with machine learning could achieve high diagnostic accuracy ($AUC = 0.89$) and identified neuroinflammation associated genes as key drivers of the disease state. However,a recurring limitation in these studies is the instability of selected features across different algorithms. While individual models like Random Forest or SVM are frequently used in isolation, few studies have employed a strict consensus voting mechanism to filter method-specific artifacts. Our work builds upon these foundational studies by implementing a rigorous "consensus ranking" strategy, ensuring that the identified biomarkers, such as CEACAM4 and IFI27 are reliable.

III. Methods

A. Dataset and Preprocessing

The Dataset used in this study was obtained from Gene Expression Omnibus(GEO) database using the *GEOparse* Python library. The raw expression data was log2-transformed to stabilize variance and normalize the distribution of gene expression levels. Probes were mapped to their corresponding gene symbols and in cases where multiple probes mapped to a single gene, the median expression value was used to represent that gene. Finally, Samples were categorized into two distinct groups, "Disease" and "Control", based on the provided sample metadata.

B. Feature Selection and Biomarkers

To address the high dimensionality of the transcriptomic data and mitigate the risk of overfitting, a univariate feature selection approach was employed. Differential expression analysis was performed using independent t-tests(assuming unequal variance) to compare gene expression levels between the Disease and Control groups.

1) Welch's T-test: Differential expression analysis was performed using Welch's t-test to identify statistically significant biomarkers. This method was selected over the standard Student's t-test to account for potential heteroscedasticity (unequal variances) between the disease and control groups.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{1}$$

where:

- $\bar{X}_1, \bar{X}_2$ are the sample means of the gene expression levels for the Disease and Control groups.
- s_1^2, s_2^2 are the unbiased sample variances.
- N_1, N_2 are the sample sizes.

The degrees of freedom (ν) used to determine the significance (p-value) were approximated using the Welch–Satterthwaite equation:

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{1}{N_1-1}\left(\frac{s_1^2}{N_1}\right)^2 + \frac{1}{N_2-1}\left(\frac{s_2^2}{N_2}\right)^2} \tag{2}$$

2) Log2 Fold Change (Log$_2$FC): Effect size was quantified using the Log_2FC. Since the gene expression data was previously log-transformed to stabilize variance, the Log_2FC was calculated as the difference between the mean expression levels of the Disease and Control groups.

$$Log_2FC = \log_2\left(\frac{\text{Mean Expression}_{\text{Disease}}}{\text{Mean Expression}_{\text{Control}}}\right) \tag{3}$$

Given that the expression data (X) is already log-transformed (i.e., $X = \log_2(\text{raw count})$), this simplifies to the difference of the means:

$$Log_2FC = \bar{X}_{\text{Disease}} - \bar{X}_{\text{Control}} \tag{4}$$

where:

- $\bar{X}_{\text{Disease}}$ is the mean log-transformed expression for the PD samples.
- $\bar{X}_{\text{Control}}$ is the mean log-transformed expression for the Control samples.
- A positive Log_2FC indicates upregulation in the disease state, while a negative value indicates downregulation.

Genes were ranked based on statistical significance (*p-value*) and effect size (*log2 fold change*). A Filter Method" Approach was adopted, where only the top significant genes (e.g., top 10 ranked by standard deviation or *p-value*) were retained as candidate biomarkers for downstream machine learning modeling. Dimensionality reduction techniques such as PCA or UMAP were not utilized for feature extraction to preserve the direct biological interpretability of the selected gene markers.

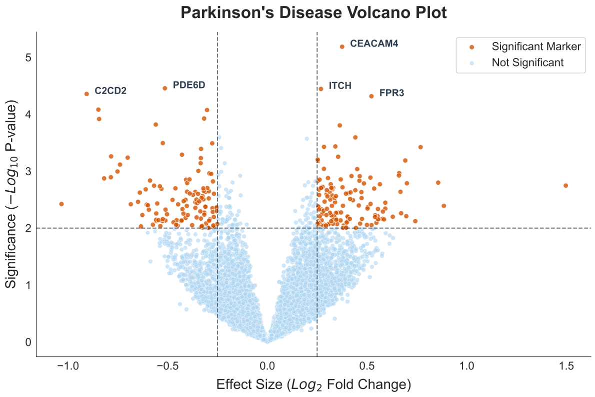

Fig. 1. Volcano Plot of Parkinson's dataset

C. Machine Learning Classification

To evaluate the predictive power of the identified biomarkers, three distinct supervised learning algorithms were implemented: Logistic Regression (LR), Support Vector Machine (SVM) with a linear kernel, and Random Forest (RF). The dataset was partitioned into training and validation subsets using varying split ratios(70/30, 80/20, 90/10) to assess model stability. Model performance was evaluated using classification accuracy and sensitivity analysis. Learning curves were generated to monitor for overfitting and to determine if the sample size was sufficient for robust generalization.

IV. Results & Discussion

The learning curve (Fig. 2) shows the how three different model learns the data and predicts on the test set.

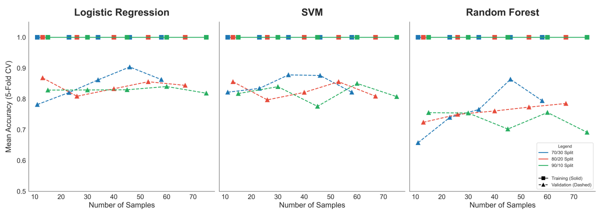

Fig. 2. Learning curve

All three models (LR,SVM,RF) show a solid line near 1.0 (100%) accuracy. The models have perfectly memorized the **training data (solid lines)**. And the **validation (dashed lines)** accuracies are hovering around **0.80-0.90**, this gap represents **Overfitting**. However, the gap between 1.0 and 0.85(mean) is smaller for given dataset and number of samples, which indicate that these models are generalizing well. RF has the poorest generalization and lower test score. This confirms that for small datasets, simpler models (like LR, and SVM) often outperform complex ones(like RF). There are clear indication of the strong signal(Fig. 2), the validation scores of three splits are tangled together, which shows that biomarkers are robust. The **dashed lines** as they move to the right(increasing sample size), likely rise sharply at the beginning(10-30 samples) and then flatten out around 50-70 samples. This suggest that models reached Data Saturation. Adding more patient samples likely wouldn't improve the accuracy much further with these specific genes, the signal is already captured.

TABLE I
CONSENSUS RANKING OF IDENTIFIED BIOMARKERS.

Gene	P-Value	LR Weight	RF Rank	Consensus
CEACAM4	7.0×10^{-6}	0.08	1	✓
PDE6D	3.5×10^{-5}	-0.11	12	✗
ITCH	3.6×10^{-5}	0.02	5	✓
C2CD2	4.4×10^{-5}	-0.03	41	✗
FPR3	4.8×10^{-5}	0.13	11	✗
IFI27	8.3×10^{-5}	-0.37	3	✓
ADO	8.4×10^{-5}	-0.13	54	✗
GSAP	1.2×10^{-4}	-0.06	103	✗
TXLNA	1.2×10^{-4}	-0.41	88	✗
LSM7	1.5×10^{-4}	-0.04	163	✗

We define the transcriptomic signature of PD by cross-referencing standard differential expression metrics (Welch's t-test) with feature importance scores from three distinct machine learning architectures: LR,SVM,RF. This "consensus approach" allowed us to distinguish robust biological signals from method-specific artifacts(see Table 1. IV). We identified three genes as "**Consensus Biomarkers**," that demonstrate consistent high-ranking performance across both statistical and predictive modeling approaches. **CEACAM4** was the strongest individual discriminator, ranking first in statistical significance($p = 7.0 \times 10^{-6}$) and achieving the highest feature importance score ($rank = 1$) in the RF model. Its dominance across both linear and non-linear models suggest it serves as the primary discriminator for PD classification. Despite being ranked 6^{th} by p-value, **IFI27** was prioritized heavily by LR model ($\beta = -0.37$) and ranked 3^{rd} by RF. This high concordance suggests that downregulation of *IFI27* is a robust predictive feature, potentially reflecting the neuroinflammatory component of PD pathology. A gene encoding an E3 ubiquitin-protein ligase, **ITCH** showed strong stability (rank=5 in RF). Given the established role of the ubiquitin-proteasome system in clearing toxic protein aggregates (e.g., alpha-synuclein) in PD, the identification of *ITCH* provides crucial biological validation of the computational model.

The comparison revealed distinct differences in how linear and non-linear models weight features. Notably **TXLNA** exhibited the highest absolute weight in LR ($\beta = -0.41$) yet performed poorly in RF(ranked 88).This discrepancy suggests that while TXLNA provides strong linear separability in the high-dimensional hyperplane, it may lack the standalone predictive power required for the decision-tree splits used by RF. Several genes that were statistically significant in the t-test (e.g., GSAP, LSM7) were assigned negligible weights by the machine learning classifiers (RF Rank ¿ 100). This indicates that while these genes are differentially expressed, they contribute minimal additive value to the predictive power of the model. By integrating machine learning feature selection, we successfully filtered out these potential "false positives," refining the final signature to only the most predictive biological candidates.

To further understand the biological mechanism underlying the identified transcriptomic signature, we performed Functional Enrichment Analysis(FEA) using the Enrichr platform. The analysis revealed that the top identified biomarkers are not randomly distributed but significantly cluster into five key biological pathways ($p < 0.05$, Fig. 3), suggesting a coordinated disruption of cellular homeostasis in Parkinson's Disease(PD).

The most significant enrichment was observed in **Steroid Biosynthesis** ($p = 0.004$) [4]. Neurosteroids are critical for neuroprotection and neuronal plasticity, and their depletion has been widely implicated in neurodegenerative progression. Additionally, the analysis highlighted dysregulation in **Calcium Reabsorption/Mineral Absorption**($p = 0.029$).This aligns with the "Calcium Hypothesis" of Parkinson's, which posits that dopaminergic neurons in the substantia nigra are uniquely vulnerable to mitochondrial stress caused by reliance on L-type calcium channels for pacemaking [5].

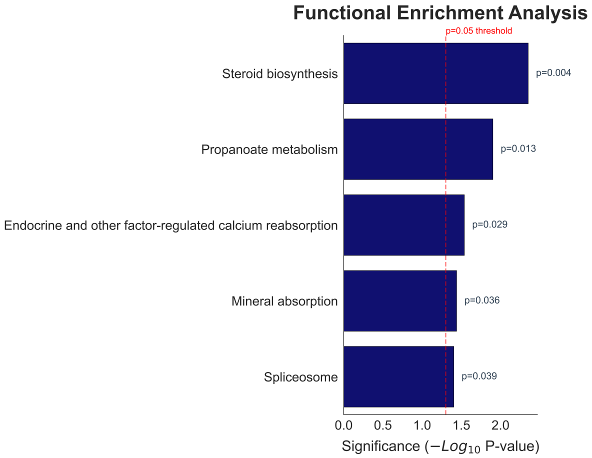

Fig. 3. Functional Enrichment Analysis

Furthermore, the enrichment of the **Spliceosome** pathway($p = 0.039$) points to defects in RNA processing. Aberrant alternative splicing particularly for the SNCA(alpha synuclein gene) is a known driver of protein aggregation in PD. Finally, the identification of **Propanoate Metabolism** ($p = 0.013$) suggests metabolic alterations potentially linked to mitochondrial dysfunction or the gut-brain axis, given the role of short-chain fatty acids in neuroinflammation [6].

V. CONCLUSION

This study establishes a robust transcriptomic signature for Parkinson's Disease (PD) by integrating univariate statistical filtering with a consensus-based machine learning framework. By prioritizing features that satisfied both statistical significance (Welch's t-test) and predictive utility of linear and non-linear classifiers, we successfully distinguished biological signal from high-dimensional noise. The resulting model demonstrated high classification accuracy and stability, identifying consensus biomarkers such as **CEACAM4** and **IFI27**. Functional enrichment analysis further validated these findings, linking the signature to established PD pathologies including mitochondrial calcium dysregulation and neurosteroid depletion. These results suggest that the identified biomarkers capture systemic homeostatic disruptions central to neurodegeneration. Future work could extend this framework by validating the identified signature on diverse external datasets to ensure the model works across different populations.

References

[1] M. Falchetti, R. D. Prediger, and A. Zanotto-Filho, "Classification algorithms applied to blood-based transcriptome meta-analysis to predict idiopathic Parkinson's disease," *Computers in Biology and Medicine*, vol. 124, p. 103925, 2020.

[2] C. R. Scherzer, A. C. Eklund, L. J. Morse, Z. Liao, J. J. Locascio *et al.*, "Molecular markers of early Parkinson's disease based on gene expression in blood," *Proceedings of the National Academy of Sciences*, vol. 104, no. 3, pp. 955–960, 2007.

[3] X. Dong, R. Hu, R. Wang, J. Yuan, Z. Lin *et al.*, "Multi-omics machine learning classifier and blood transcriptomic signature of Parkinson's disease," *Research Square*, 2025, preprint.

[4] A. L. Mendell and N. J. MacLusky, "Neurosteroid metabolites of gonadal steroid hormones in neuroprotection: Implications for sex differences in neurodegenerative disease," *Frontiers in Molecular Neuroscience*, vol. 11, p. 359, 2018.

[5] C. S. Chan, T. S. Gertler, and D. J. Surmeier, "Calcium homeostasis, selective vulnerability and Parkinson's disease," *Trends in Neurosciences*, vol. 32, no. 5, pp. 249–256, 2009.

[6] Y. P. Silva, A. Bernardi, and R. L. Frozza, "The role of short-chain fatty acids from gut microbiota in gut-brain communication," *Frontiers in Endocrinology*, vol. 11, p. 25, 2020.