STEM Fellowship Big Data Challenge 2023

Exploring the relationship between Protein Expression in Cerebrospinal fluid and Parkinson's Disease Progression using Machine Learning

Carlos Pariona[1,2] and Fiorella Ojeda[1,2]

[1]Pontificia Universidad Catolica del Peru
[2]Universidad Peruana Cayetano Heredia

June 26, 2023

Abstract

Parkinson's disease (PD) is a chronic neurodegenerative condition characterized by the degeneration of nerve cells in the brain, leading to motor symptoms such as tremors and rigidity. While the exact cause of PD is not fully understood, studies have shown the involvement of specific proteins, including alpha-synuclein. This paper explores the predictive potential of proteins in PD using data derived from mass spectrometry readings of cerebrospinal fluid (CSF) samples. The goal is to identify proteins that may have relevance in early detection and personalized treatments for PD. The study utilizes machine learning techniques, including regression and classification models, to establish the relationship between normalized protein expression (NPX) and PD progression. The results indicate that there is no significant relationship between NPX and PD progression based on the analyzed approaches. However, limitations in the study, such as the absence of complete clinical information and the need for imputation, must be considered. Further research is required to gain a more comprehensive understanding of the complex molecular mechanisms associated with PD.

Keywords
Parkinson's disease, protein expression, MDS-UPDRS machine learning

1 Introduction

Parkinson's disease (PD) is a chronic neurodegenerative condition that affects approximately 6.5 million people worldwide [1], ranking as the second most prevalent neurodegenerative disease [2]. It is characterized by the progressive degeneration of nerve cells in a specific region of the brain known as the substantia nigra, which leads to insufficient production of the neurotransmitter dopamine, responsible for the control of movement and muscular coordination. This dopamine deficiency results in the emergence of distinctive motor symptoms such as tremors, rigidity, and difficulty with locomotion [2].

Although the exact cause of PD is not yet fully understood, various scientific studies have suggested that certain proteins play a fundamental role in the development and progression of this disease [2] [3]. The discovery of proteins associated with Parkinson's has opened new perspectives in understanding the underlying mechanisms of this pathology, such as alpha-synuclein, which is found in nerve cells and its abnormality leads to the formation of Lewy bodies, believed to contribute to the characteristic neuronal degeneration of the disease [3]. Alongside conventional diagnostic and clinical approaches, research has explored the potential of peptides as a promising class of bioactive compounds with therapeutic applications in PD [3].

Early and accurate diagnosis of PD is crucial for proper disease management. In recent years, significant advances have been made in the use of artificial intelligence (AI) and machine learning (ML), allowing for the detection of subtle patterns in complex data [4] [5]. This facilitates a

more precise and objective assessment of symptoms and disease progression, aiding doctors in early and differential diagnosis, as well as monitoring treatment response over time.

The Unified Parkinson's Disease Rating Scale (UPDRS) is a standardized assessment tool that provides physicians with the ability to accurately evaluate and quantify both motor and non-motor symptoms in Parkinson's patients [6]. The scale consists of several domains, including the evaluation of motor function, activities of daily living, motor complications, as well as non-motor symptoms such as depression, anxiety, and quality of life. Each domain is assessed through a series of items and assigned a score, enabling an objective measurement of symptom severity and tracking of disease progression over time [6].

Despite advances in PD research, there is still significant uncertainty regarding the complete set of proteins involved in this disease. Therefore, it is necessary to thoroughly investigate proteins that may have relevant predictive value in relation to PD. Our approach is based on a data core consisting of protein abundance values derived from mass spectrometry readings of cerebrospinal fluid (CSF) samples collected from 248 patients over several months of visits. The goal is to explore the predictive potential of proteins in PD, which could provide new insights for early detection and the development of personalized treatments for this debilitating disease.

2 Materials & Methods

All analyses were performed using Python 3.9 and related scientific libraries.

2.1 Data description

In this study, we used two publicly available Kaggle datasets derived from the AMP PD Knowledge Platform, which can be accessed at the following link: https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction. The datasets contain information about the molecular characterization and longitudinal clinical profiling of 248 Parkinson's disease patients.

The data includes the patient ID, visit month, UniProt code of several proteins, normalized protein expression (NPX) values from mass spectrometry readings of cerebrospinal fluid (CSF) samples, clinical evaluation of the patient distributed in four columns based on the Unified Parkinson's Disease Rating Scale (MDS-UPDRS), and the clinical state on medication with Levodopa during the UPDRS assessment.

2.2 Data preprocessing

For our analysis, we worked with the described datasets containing protein and NPX information, as well as the clinical evaluation based on the MDS-UPDRS scale.

2.2.1 Proteins dataset feature engineering

As a first step, we applied One Hot Encoding to the protein column, which is related to the visit ID (patient ID + visit month). This created 227 columns, where each column corresponds to a protein expressed by a patient in their visit month. Next, we used the NPX column in logarithmic scale to assign protein expression values to each encoded column. Subsequently, we vertically grouped the data, resulting in a unique visit ID for each row, which is associated with the NPX values of each protein expressed by that patient in that visit month. Finally, we applied Robust Scaler as a scaling method to make the data more uniform and better distributed.

2.2.2 Clinical dataset feature engineering

Using the database with clinical information, we created a new column called 'updrs', which contains the sum of the four columns based on the MDS-UPDRS scale. This way, we obtained a single column with information about the PD progression through this scale. Due to the amount of missing data in the fourth part of the MDS-UPDRS scale, we had to apply imputation. The chosen strategy was k-Nearest Neighbors imputation, which fills in missing values using the mean value of the nearest neighbors found in the training set. This way, we do not lose half of the data as would happen if we removed the missing data.

2.3 Training dataframe generation

The preprocessing of these datasets allowed us to generate two strategies to determine the relationship between NPX and Parkinson's disease progression. The first strategy evaluates NPX against the MDS-UPDRS scale, which provides information about the symptomatic progression of Parkinson's disease. The second strategy evaluates NPX against the temporal evolution of the disease, i.e. the visit months of the patients. By merging the protein databases with the clinical data database, we noticed a large number of outliers in the protein database. Therefore, we

decided to impute these outliers by setting them equal to the median. Finally, two databases were generated, which were used for training the machine learning models.

The two chosen strategies to determine the relationship between normalized protein expression and Parkinson's disease progression were evaluated using regression and classification models due to the nature of our objective. Thus, for the classification models, the accuracy score was used as the evaluation metric because if we can predict the target with our model, we can determine which proteins are more important to measure the development of the disease. On the other hand, for the regression models, the r2 score was used as it provides information on how the target magnitudes can be explained by the differences in protein expression.

3 Results

3.1 NPX & MDS-UPDRS

As mentioned before, this first strategy required summing up the four parts of the MDS-UPDRS scale within the clinical database. Due to the amount of missing values in the column corresponding to the fourth part of the scale, imputation was necessary. Therefore, the imputed values became floating-point numbers, which prevents us from applying classification models. However, as mentioned earlier, our objective is to establish a relationship between proteins and the target, and since the target is an ordinal set, we can use regression models.

Regression Model	R2 Score
k-Nearest Neighbor	**29.00** %
Support Vector Machine	07.00 %
Decision Tree	00.00 %
Random Forest	24.00 %
Adaptive Boosting	14.00 %
Bootstrap Aggregating	13.00 %
Gradient Boosting	22.00 %
Histogram Gradient Boosting	26.00 %
Linear Regression	00.00 %
Multilayer Perceptron	11.00 %

Table 1: Performance metrics for ML regression models in NPX & MDS-UPDRS

The results obtained through regression are shown in Table 1. We observe that the best result corresponds to model k-Nearest Neighbor Regressor with an R2 score of 29.00%. This result can be interpreted as no relationship existing between the features and the target. This assumption was made during the exploratory data analysis based on the observed distribution between the proteins and the MDS-UPDRS scale, which seems to be a scattered cloud without any pattern (e.g., Fig 1).

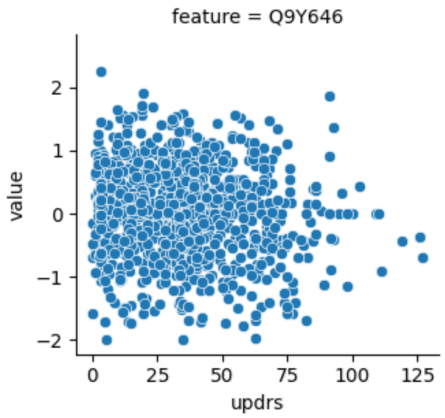

Figure 1: Q9Y646 protein distribution vs MDS-UPDRS

3.2 NPX & Visit Month

The second strategy aimed to find a relationship between normalized protein expression and time. For this purpose, no imputations were required since we had complete data, and we were able to apply classification models.

Classification Model	Accuracy
k-Nearest Neighbor	08.00 %
Support Vector Machine	12.00 %
Decision Tree	14.00 %
Random Forest	10.00 %
Adaptive Boosting	13.00 %
Bootstrap Aggregating	**16.00** %
Gradient Boosting	11.00 %
Histogram Gradient Boosting	07.00 %
Multilayer Perceptron	07.00 %

Table 2: Performance metrics for ML classification models in NPX & Visit Month

The obtained results are shown in Table 2. We observe that the best classification model is Bootstrap Aggregating Classifier, with an accuracy score of 16.00%. These results confirm the previous findings, concluding that there is no relationship between normalized protein expression and Parkinson's disease progression, at least based on the analyzed approaches.

4 Discussion

In recent years, in the field of neurodegenerative diseases, several genes and molecular mech-

anisms have been identified to have an association with neurodegenerative diseases and selective neuronal death. Intracellular pathways that increase the risk for the development of sporadic PD have been identified as well. However, despite association studies aimed at identifying genes that act as susceptibility factors, the results are inconclusive [7].

Similarly, it has been shown that the use of new sequencing techniques has allowed the identification of 90 independent risk variants for PD, located in several loci (physical sites in a genome), as well as more than 20 genes associated with PD. It should be noted that the association of some of these genes with this disease is still under debate due to contradictory results between different population-based studies. Therefore, it is emphasized that a high percentage of Parkinson's cases are attributed to the influence of risk genes rather than to the direct action of causal genes [8].

On the other hand, another study employed super-resolution fluorescence microscopy to directly visualize alpha-synuclein oligomers (a neuronal protein included in our database) in the brains of Parkinson's patients. It was discovered that the oligomers were more abundant and widely distributed compared to healthy individuals. These findings suggest that the oligomers could play a crucial role in the neurodegenerative processes of PD [9].

In another study, it was discovered that the presence of alpha-synuclein causes depolarization of mitochondrial membranes, altering their electrical potential. Additionally, a decrease in the phosphorylation capacity of mitochondria isolated from rat brains was observed, suggesting impairment in energy production. These findings are relevant for understanding the underlying mechanisms in Parkinson's disease, as they indicate that alpha-synuclein may have a detrimental impact on mitochondria, potentially contributing to the neuronal dysfunction observed in the disease [10].

Although alpha-synuclein has been extensively investigated in relation to PD, current studies do not yet provide sufficient evidence, along with other proteins, to definitively assert its involvement in the disease [11]. Further biological research is required to gain a more comprehensive understanding of the complex molecular mechanisms associated with the development and progression of Parkinson's.

Based on the available evidence and research findings, it can be stated that there is currently no conclusive relationship between protein expression and the progression of Parkinson's disease. While various genes, molecular mechanisms, and proteins have been investigated for their potential involvement, the results have been inconclusive and often contradictory. The complexity and multifactorial nature of Parkinson's disease suggest that it may be influenced by a combination of genetic and environmental factors, making it challenging to pinpoint a single protein or peptide responsible for the disease.

Our chosen approach presents both advantages and limitations in its implementation. The use of samples collected by mass spectrometry readings of CSF, despite it not being an easily accessible biofluid, bestows certain benefits due to its relative metabolic simplicity and potential importance to central nervous system disorders [12]. In addition, it enables the quantification of a plethora of CSF-related proteins, which allows researchers to identify potential biomarkers associated with PD, and bestows a high level of sensitivity, specificity, and throughput, permitting the simultaneous measurement of numerous proteins in a single analysis [13].

Nevertheless, the lack of more clinical information, like the motor phenotype, the age of the patient, cognitive assessments or the presence of other diseases, poses a significant limitation to the study. These missing details hinder the ability to comprehend the influence of such related factors on the protein expression patterns and subsequent disease progression. The motor phenotype in particular has been proven to be correlated with the progression of PD [14]. Additionally, the imputation of UPDRS score aspects, while necessitated by the absence of complete clinical information, introduces an inherent loss of accuracy and reliability. This imputation approach, while attempting to compensate for the missing data, may inadvertently introduce biases or inaccuracies into the analysis, thereby compromising the fidelity of the predictive model. Consequently, the study's outcomes must be cautiously interpreted, considering the potential confounding effects of the data used and its processing.

Looking towards the future directions, we provide a compelling proposition to extend the sample collection method to encompass diseases characterized by a temporal evolution. By employing the proposed methodology to diseases that exhibit an established dynamic progression, like Alzheimer disease [15], a deeper understanding of the intricate molecular alterations associated with the temporal course of such conditions might be achieved. Moreover, it is recommended to incorporate the medication state, usually resulting from levodopa administration, as a crucial target variable within the analysis [16]. Especifically, by leveraging the levodopa-

protein affinities an extensive matrix can be constructed, enabling a comprehensive exploration of the intricate protein expression patterns in disease progression.

Conclusions

Machine Learning is often a useful tool for recognizing patterns that allow us to make inferences about the analyzed data. Nevertheless, in some cases, it is impossible to find a strong relationship within the data because it simply does not exist. The result of our analysis, after testing various Machine Learning algorithms and trying different approaches to data processing, indicates that there is no relationship between the protein expression of the set of cerebrospinal fluid proteins analyzed and the temporal and symptomatic evolution of Parkinson's disease. Although this conclusion may seem weak due to the ambiguous results that can be found in the literature. Ultimately, further research is required to refute or strengthen a relationship between the protein expression of certain proteins and the evolution of Parkinson's disease. All the same, it is necessary to highlight the importance of data science in carrying out these tasks, as it can allow us to find relationships that are not always visible to the naked eye.

Acknowledgements

We thank the STEM Fellowship team behind Big Data Challenge for giving us the opportunity to conduct research in the data science and the bioinformatics field, as well as for providing us with relevant tools, workshops and mentoring.

References

[1] E Dorsey, Todd Sherer, Michael S Okun, and Bastiaan R Bloem. The emerging evidence of the parkinson pandemic. *Journal of Parkinson's disease*, 8(s1):S3–S8, 2018.

[2] Werner Poewe, Klaus Seppi, Caroline M Tanner, Glenda M Halliday, Patrik Brundin, Jens Volkmann, Anette-Eleonore Schrag, and Anthony E Lang. Parkinson disease. *Nature reviews Disease primers*, 3(1):1–21, 2017.

[3] Maria Grazia Spillantini, Marie Luise Schmidt, Virginia M-Y Lee, John Q Trojanowski, Ross Jakes, and Michel Goedert. α-synuclein in lewy bodies. *Nature*, 388(6645):839–840, 1997.

[4] Shriniket Dixit, Khitij Bohre, Yashbir Singh, Yassine Himeur, Wathiq Mansoor, Shadi Atalla, and Kathiravan Srinivasan. A comprehensive review on ai-enabled models for parkinson's disease diagnosis. *Electronics*, 12(4):783, 2023.

[5] Songyun Zhao, Li Zhang, Wei Ji, Yachen Shi, Guichuan Lai, Hao Chi, Weiyi Huang, and Chao Cheng. Machine learning-based characterization of cuprotosis-related biomarkers and immune infiltration in parkinson's disease. *Frontiers in Genetics*, 13, 2022.

[6] Haruko Tanji, Ann L Gruber-Baldini, Karen E Anderson, Ingrid Pretzer-Aboff, Stephen G Reich, Paul S Fishman, William J Weiner, and Lisa M Shulman. A comparative study of physical performance measures in parkinson's disease. *Movement Disorders*, 23(13):1897–1905, 2008.

[7] Margarita Gómez-Chavarín, M Carolina Torres-Ortiz, and Gabriel Perez-Soto. Interacción entre factores genéticosambientales y la epigénesis de la enfermedad de parkinson. *Archivos de Neurociencias*, 21(1):32–44, 2016.

[8] Daniel Macías García. *Biomarcadores séricos y riesgo vascular en la enfermedad de Parkinson esporádica y familiar*. PhD thesis, Universidad de Sevilla, 2022.

[9] Rosalind F Roberts, Richard Wade-Martins, and Javier Alegre-Abarrategui. Direct visualization of alpha-synuclein oligomers reveals previously undetected pathology in parkinson's disease brain. *Brain*, 138(6):1642–1657, 2015.

[10] Kalpita Banerjee, Maitrayee Sinha, Chi Le Lan Pham, Sirsendu Jana, Dalia Chanda, Roberto Cappai, and Sasanka Chakrabarti. α-synuclein induced membrane depolarization and loss of phosphorylation capacity of isolated rat brain mitochondria: Implications in parkinson's disease. *FEBS letters*, 584(8):1571–1576, 2010.

[11] Emily M Rocha, Briana De Miranda, and Laurie H Sanders. Alpha-synuclein: Pathology, mitochondrial dysfunction and neuroinflammation in parkinson's disease. *Neurobiology of disease*, 109:249–257, 2018.

[12] David S Wishart, Michael J Lewis, Joshua A Morrissey, Mitchel D Flegel, Kevin Jeroncic, Yeping Xiong, Dean Cheng, Roman Eisner, Bijaya Gautam, Dan Tzur,

et al. The human cerebrospinal fluid metabolome. *Journal of Chromatography B*, 871(2):164–173, 2008.

[13] Marcia Cristina T. dos Santos, Dieter Scheller, Claudia Schulte, Irene R Mesa, Peter Colman, Sarah R Bujac, Rosie Bell, Caroline Berteau, Luis Tosar Perez, Ingolf Lachmann, et al. Evaluation of cerebrospinal fluid proteins as potential biomarkers for early stage parkinson's disease diagnosis. *PLoS One*, 13(11):e0206536, 2018.

[14] Jennifer Michels, Hendrik van der Wurp, Elke Kalbe, Sarah Rehberg, Alexander Storch, Katharina Linse, Christine Schneider, Susanne Gräber, Daniela Berg, Judith Dams, et al. Long-term cognitive decline related to the motor phenotype in parkinson's disease. *Journal of Parkinson's disease*, 12(3):905–916, 2022.

[15] Bob Olsson, Ronald Lautner, Ulf Andreasson, Annika Öhrfelt, Erik Portelius, Maria Bjerke, Mikko Hölttä, Christoffer Rosén, Caroline Olsson, Gabrielle Strobel, et al. Csf and blood biomarkers for the diagnosis of alzheimer's disease: a systematic review and meta-analysis. *The Lancet Neurology*, 15(7):673–684, 2016.

[16] Parkinson Study Group. Levodopa and the progression of parkinson's disease. *New England Journal of Medicine*, 351(24):2498–2508, 2004.