Unsupervised Machine Learning Using Cerebrospinal Fluid Proteomics for Understanding Parkinson's Disease Progression

Lubna Mahmoud Abu Zohair[1], **Hind Zantout**[1], **Marta Vallejo**[2], **MD Azher Uddin**[1]
School of Mathematical and Computer Science, Heriot-Watt University
[1]Dubai, United Arab Emirates
[2]Edinburgh, United Kingdom
la2015@hw.ac.uk

Abstract

This study explores the potential of advanced, context-aware machine learning algorithms, such as autoencoders, to represent longitudinal cerebrospinal fluid proteomic data, enabling the objective discovery of two patient strata with significance.

Background, Problem, and Objective

Parkinson's disease (PD) is the second most common progressive neurodegenerative characterized by defective dopaminergic neurons in the brain, causing motor and nonmotor symptoms, such as tremors, bradykinesia, mood changes and postural instability (Kouli et al. 2018). The exact cause and full clinical progression of PD remain unclear and vary among patients due to the abnormal and heterogeneous nature of the disease's progression (Balestrino and Schapira 2020). As a result, current treatments can manage motor symptoms but do not halt neurodegeneration or the disability progression associated with the disease, threating patient quality of life (Balestrino and Schapira 2020; Kouli et al. 2018). PD management and monitoring heavily rely on clinical scores, like the Unified Parkinson's Disease Rating Scale (UPDRS), to assess disease severity and progression based on observable symptoms. However, these scores are subjective, affected by clinicians' judgments and patients' self-reported symptoms, making them prone to errors and variability (Martínez-Martín et al. 2015). In addition, they cannot capture changes in the brain's underlying biology which occur during the course of the disease, and are crucial for understanding the full trajectory. Hence, there is a growing need for more objective understanding of the disease progression trajectory. Cerebrospinal Fluid (CSF) has emerged as a promising source of biomarkers that capture the brain's internal environment that could complement existing PD clinical scores (Evers et al. 2019). Proteins, peptides, and their abundance in CSF are closely linked to the brain's neural, structural, and fluid state, making them attractive candidates for providing a more reliable reflection of disease progression by monitoring changes in these fluids over time (Bader et al. 2020). With the current advancements in artificial intelligence models, analyzing high dimensional data, such as proteomic data, became possible, and proved to outperform traditional statistical analysis methods (Stahl 2024). However, the potential of deep learning models, such as autoencoders, in creating predictive features from longitudinal proteomic data to identify patient strata has yet to be fully explored. In summary, this is the aim of this research

Data and Methods

In this study, 24 records of PD patients with complete 6-year proteomics visit data were selected from the Accelerating Medicines Partnership (AMP) for Parkinson's Disease Program dataset on Kaggle (Kirsch et al. 2023). This dataset, which contains features such as patient IDs, visit months, protein information, and peptide abundances, was restructured into both tabular and graph forms. The proposed graph structure represents each patient's longitudinal data as a graph, where visit months are nodes. Node features include peptide and peptide abundance, while edge features represent the correlation strength between nodes. Feature encoding was applied using Long Short-Term Memory (LSTM) and Graph autoencoders (Nguyen et al. 2021; Pan et al. 2018). And those deep learning approaches were compared with traditional dimensionality reduction techniques, like Kernel Principal Component Analysis (Kernel PCA), and t-Distributed Stochastic Neighbor Embedding (t-SNE). The original features and their representation were hierarchically clustered, with silhouette scores being used as evaluation metric. The significance of overall peptide abundance variability over the years, and the differences in peptide type abundances between the identified groups, was confirmed through the Kruskal-Wallis H test and Permutational Multivariate Analysis of Variance (PERMANOVA), respectively (Anderson 2017). A p-value threshold of 0.05 was set to reject the null hypothesis, which stated that there were no significant differences in abundance temporal variability and group differences.

Results

For all employed feature representation algorithms, one component/embedding dimension and two clusters were found to yield the best clustering performance, leading to the discovery of two distinct patient groups (or clusters) by the LSTM Autoencoder, as shown in the heatmap in Figure 1. The clustering had a silhouette score of 0.6901 and was found to be statistically significant, with a PER-MANOVA p-value of 0.0008 ($p < 0.05$), confirming that the differences in peptide abundance between the two groups were not due to chance. From Figure 2, further characteristics of the discovered patient groups were observed. Plots (A) and (B) illustrate the mean peptide abundance for 15 patients in group 1 (top plots) and 7 patients in group 2 (bottom plots), represented by the bold line, with standard deviation shaded areas for all patients in each group over time. They also show in plot (A) the motor (Part-III) UPDRS scale cutoffs (mild/moderate and moderate/severe) represented by dotted lines, while the middle plots display the HY scale from 1 to 5. Plots (C) present the peptide abundance values for all patients in the dataset who are taking Levodopa (top plot), compared to those who are not subject to any intervention (bottom plot). For Group 1 patients, peptide levels were found to change over time, and these changes were determined to be statistically significant, with a Kruskal-Wallis p-value of 0.0028. Conversely, for Group 2 patients, peptide levels showed a steady increase over time, without any significant changes, although a slight increase became noticeable starting from month 48. Regarding the health status of these patients, Cluster 2 patients were still in the early stages of Parkinson's disease (with mild symptoms), transitioning from HY1 (mild) to HY2 (more noticeable symptoms). However, some patients in Group 1 were already in a more advanced stage of the disease, reaching HY3 (severe symptoms). Levodopa medication is typically prescribed when patients begin to show uncontrolled symptoms that affect daily life activities (Salat & Tolosa 2013). This may explain the interesting peptide abundance pattern observed in Group 1, which is like the temporal pattern seen in patients on medication across the entire dataset. This suggests that the higher peptide levels in Group 1 may reflect the worsening of their disease condition, which is why medication is required to manage their symptoms.

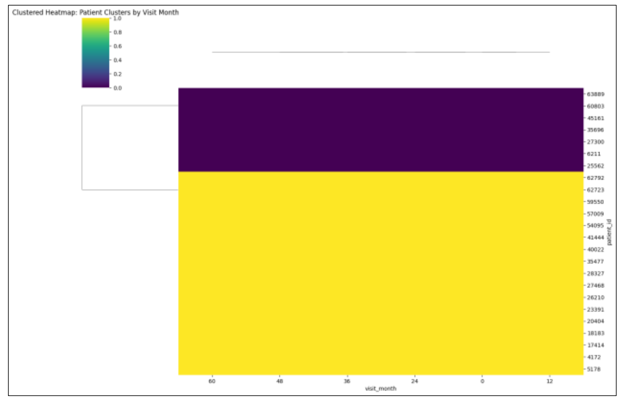

Figure 1: Discovered Patients Groups

Conclusion

These findings highlight how biomarkers, such as CSF biofluids, combined with deep learning models and clustering algorithms, have led to the discovery of two patient groups. They also demonstrate that peptide abundance patterns can serve as key biomarkers for understanding and tracking disease progression. However, clinician input, result reproducibility on larger PD patients' group, and the utility of this approach with other biomarkers, such as brain structural changes over time, as well as its extension to other neurodegenerative diseases, will be worth exploring.

(A) Mean Peptide Patterns by Months for patients in Group 1 (top) and Group 2 (bottom) with UPDRS-III Scale.

(B) Mean Peptide Patterns by Months for patients in Group 1 (top) and Group 2 (bottom) with Hoehn and Yahr scale (HY).

(C) All Patients Peptide Patterns by Months: on medication (Top) and off medication (bottom).

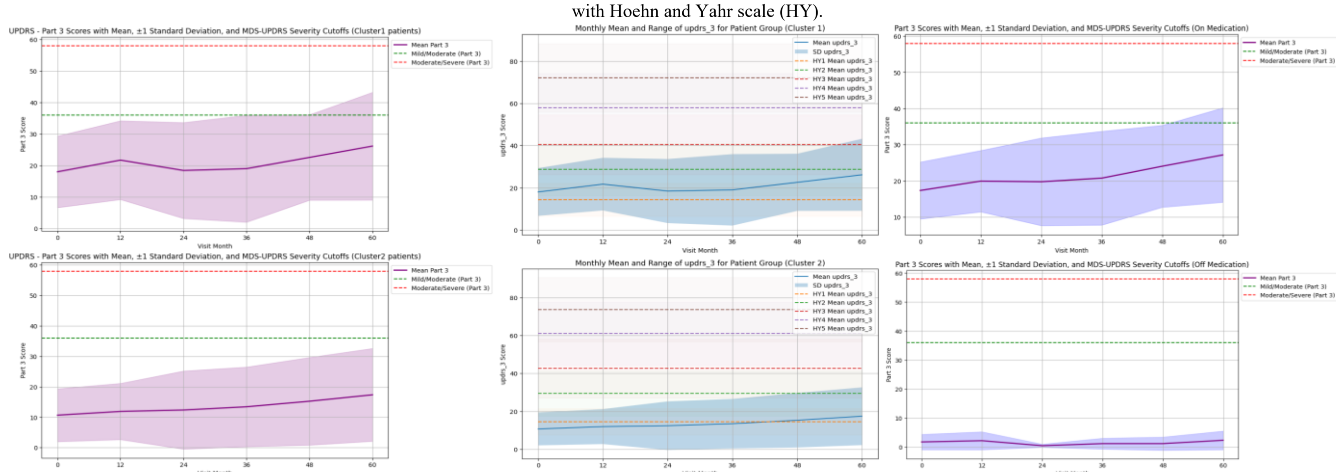

Figure 2: Peptide abundance temporal patterns for Group 1 and Group 2 patients, illustrating disease progression, motor scale cutoffs, and Levodopa treatment effects.

References

Ahmadi Rastegar, D., Ho, N., Halliday, G. M., & Dzamko, N. 2019. Parkinson's progression prediction using machine learning and serum cytokines. *Npj Parkinson's Disease, 5*(1), 14. doi.org/10.1038/s41531-019-0086-4

Anderson, M. J. 2017. Permutational Multivariate Analysis of Variance (PERMANOVA). In *Wiley StatsRef: Statistics Reference Online* (pp. 1–15). Wiley. doi.org/10.1002/9781118445112.stat07841

Bader, J. M., Geyer, P. E., Müller, J. B., Strauss, M. T., Koch, M., Leypoldt, F., Koertvelyessy, P., Bittner, D., Schipke, C. G., Incesoy, E. I., Peters, O., Deigendesch, N., Simons, M., Jensen, M. K., Zetterberg, H., & Mann, M. 2020. Proteome profiling in cerebrospinal fluid reveals novel biomarkers of Alzheimer's disease. *Molecular Systems Biology, 16*(6). doi.org/10.15252/msb.20199356

Balestrino, R., & Schapira, A. H. V. (2020). Parkinson disease. *European Journal of Neurology, 27*(1), 27–42. doi.org/10.1111/ene.14108

Díaz, S. A., Ciabis, V., Burgos, V., Belloso, W. H., & Risk, M. 2024. Predictive Modeling of Parkinson's Disease Progression Through Proteomic and Peptidomic Analysis (pp. 101–113). doi.org/10.1007/978-3-031-61960-1_10

Evers, L. J. W., Krijthe, J. H., Meinders, M. J., Bloem, B. R., & Heskes, T. M. 2019. Measuring Parkinson's disease over time: The real-world within-subject reliability of the MDS-UPDRS. *Movement Disorders, 34*(10), 1480–1487. doi.org/10.1002/mds.27790

Kouli, A., Torsney, K. M., & Kuan, W.-L. 2018. Parkinson's Disease: Etiology, Neuropathology, and Pathogenesis. In *Parkinson's Disease: Pathogenesis and Clinical Aspects*, 3–26. Codon Publications. doi.org/10.15586/codonpublications.parkinsonsdisease.2018.ch1

Martínez-Martín, P., Rodríguez-Blázquez, C., Mario Alvarez, Arakaki, T., Arillo, V. C., Chaná, P., Fernández, W., Garretto, N., Martínez-Castrillo, J. C., Rodríguez-Violante, M., Serrano-Dueñas, M., Ballesteros, D., Rojo-Abuin, J. M., Chaudhuri, K. R., & Merello, M. 2015. Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale. *Parkinsonism & Related Disorders, 21*(1), 50–54. doi.org/10.1016/j.parkreldis.2014.10.026

Salat, D., & Tolosa, E. 2013. Levodopa in the Treatment of Parkinson's Disease: Current Status and New Developments. *Journal of Parkinson's Disease, 3*(3), 255–269. doi.org/10.3233/JPD-130186

Nguyen, H. D., Tran, K. P., Thomassey, S., & Hamad, M. 2021 Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management *International Journal of Information Management,* 57 . doi.org/10.1016/j.ijinfomgt.2020.102282

Pan, S., Hu, R., Long, G., Jiang, J., Yao, L., & Zhang, C. 2018 Adversarially Regularized Graph Autoencoder for Graph Embedding. doi.org/10.48550/arXiv.1802.04407

Skorvanek, M., Martinez-Martin, P., Kovacs, N., Rodriguez-Violante, M., Corvol, J., Taba, P., Seppi, K., Levin, O., Schrag, A., Foltynie, T., Alvarez-Sanchez, M., Arakaki, T., Aschermann, Z., Aviles-Olmos, I., Benchetrit, E., Benoit, C., Bergareche-Yarza, A., Cervantes-Arriaga, A., Chade, A., … Stebbins, G. T. 2017. Differences in MDS-UPDRS Scores Based on Hoehn and Yahr Stage and Disease Duration. *Movement Disorders Clinical Practice, 4*(4), 536–544. doi.org/10.1002/mdc3.12476

Stahl, D. 2024. New horizons in prediction modelling using machine learning in older people's healthcare research. *Age and Ageing, 53*(9). doi.org/10.1093/ageing/afae201

Kirsch, L., Dane, S., Adam, S., & Dardov, V. 2023. AMP®-Parkinson's Disease Progression Prediction. https://kaggle.com/competitions/amp-parkinsons-disease-progression-prediction. Kaggle.