Title: Discovering Key Proteins and Peptides for Parkinson's Disease Prognosis with Machine Learning and Interpretable Methods

Advisors: Arnout Van Messem, Wesley De Neve

Parkinson's disease is a neurodegenerative disorder that affects millions of people worldwide, leading to the progressive deterioration of motor function, cognitive abilities, and quality of life. Understanding its underlying causes and developing effective treatments are essential for improving patient outcomes. Recently, machine learning and its interpretability have revolutionized industries, including healthcare, offering new approaches to complex medical problems.

This study aims to leverage the capabilities of machine learning and interpretable methods to investigate the underlying causes of Parkinson's disease and identify potential therapeutic target proteins or peptides. We built a machine-learning pipeline to search for key proteins and peptides by combining mutual information with machine-learning models on a Parkinson's disease prognosis dataset of 1,200 protein candidates obtained through cerebrospinal fluid mass spectrometry. To achieve our goal, we applied various preprocessing methods, such as removing skewness, imputing missing values, and standardizing data. We trained five major regression models: Linear Regression, Decision Tree Regression, Elastic Net, Bayesian Ridge, and Gradient Boosting Regression, to predict the Movement Disorder Society - Unified Parkinson's Disease Rating Scale (MDR-UPDRS), using 5-fold cross-validation and grid search hyperparameter optimization. To confirm the performance of our pipeline, we also tested it under five different conditions on the same dataset and compared its effectiveness with the effectiveness of other methods. Our pipeline showed the lowest symmetric mean absolute percentage error, confirming its superior performance. Finally, we used three interpretable methods - Model coefficients or tree importances, Permutation feature importance, and SHapley Additive exPlanations - to filter out important proteins and peptides in predicting MDS-UPDRS, using the optimal models trained through our pipeline. Finally, our pipeline was able to identify 11 key proteins and peptides that have a notable influence on the prognosis of Parkinson's disease.

Our findings represent a meaningful step forward in the understanding of the molecular mechanisms underlying Parkinson's disease and provide a solid foundation for future research. In the next phase of our investigation, we plan to delve deeper into relevant biological databases such as Uniprot and the Protein Data Bank (PDB) to identify the specific receptors affected by the selected proteins and peptides. This information will be invaluable in guiding the development of targeted drug therapies aimed at modulating the activity of these receptors and ultimately improving the prognosis for Parkinson's disease patients.