Parkinson's Disease Progression Prediction Using Protein and Peptide Data Measurements with Deep Learning

Group 6

Daisy Adhikari Lohitha Vanteru Mahe Jabeen Abdul Pranavi Sandrugu

Abstract

Parkinson's disease is a long-term neurological condition that impairs movement, affecting more than 10 million individuals worldwide. People start to have trouble speaking, writing, walking, or performing other basic skills as the dopamine-generating neurons in certain areas of the brain are impaired or perish. As a result, the intensity of the symptoms in the patients increases over time. It is frequently diagnosed using clinical assessments and a progression scale, which typically depend on the skill of the medical professional. Accuracy varies widely across different examiners, and it also takes a long time to diagnose correctly. According to research, this disease's development and progression are significantly influenced by anomalies in proteins or peptides. A thorough evaluation of both motor and non-motor symptoms related to Parkinson's disease is provided by the Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS). Deep Learning can help us understand the intrinsic patterns from protein and peptide data measurements and forecast MDS-UPDR values, which represent the disease's development in Parkinson's patients which could lead to the discovery of novel therapeutic interventions that can either stop the course of Parkinson's disease or treat it. In order to address the multicollinearity issues in the data and to condense the size of the input feature space, principal component analysis (PCA) will be initially used in this study to analyze the featured dataset. The proposed Deep Neural Network model is then fed with the condensed input feature space and an addition of dense connections and batch normalization. The model can be used to identify people who are at high risk of developing Parkinson's disease and will offer insightful information about the underlying causes of the disease. The team is planning to utilize deep learning techniques covered in the course to analyze the complex relationship between protein and peptide levels and disease progression. The results of this project can have significant implications for the early detection and treatment of Parkinson’s disease, which can improve the quality of life of many people affected by this disease.

Dataset Links and Description: This project's dataset has been taken from the Kaggle competition for Parkinson's Disease Prediction is based on the Parkinson's Progression Markers Initiative (PPMI) dataset which is provided by the Michael J. Fox Foundation. The dataset includes protein and peptide measurements along with the clinical, imaging, and biomarker data taken from 2,500 healthy control volunteers and 5,800 Parkinson's disease patients. The dataset includes 20,000 features that indicate the levels of proteins and peptides acquired from mass spectrometry analysis of cerebrospinal fluid (CSF) samples in blood plasma

coupled with details on each patient's age, gender, and disease progression status. Below are the files along with descriptions that have been provided by Kaggle for this competition.

Train_peptides.csv: Mass spectrometry data at the peptide level, with each peptide being a subunit component of proteins. The file includes information on the visit, patient, and peptide abundance.

Train_proteins.csv: Protein expression frequencies aggregated from the peptide level data, including information on the visit, patient, UniProt ID code for the associated protein, and normalized protein expression.

Train_clinical_data.csv: Clinical data for each patient, including information on the visit, patient, and scores for different parts of the Unified Parkinson's Disease Rating Scale, which assesses the severity of PD symptoms.

Supplemental_clinical_data.csv: Clinical records without any associated CSF samples, intended to provide additional context about the typical progression of Parkinson's.

Example_test_files: Data illustrating how the API functions, including the same columns delivered by the API.

Amp_pd_peptide: Files enabling the API, with the expectation that the API will deliver all of the data in under five minutes and reserve less than 0.5 GB of memory.

Public_timeseries_testing_util.py: An optional file that facilitates custom offline API tests.

Technology Proposed for Use: For this project, we suggest using Python as the major programming language along with a number of well-known libraries like TensorFlow, Keras, and Pandas for data preprocessing, analysis, and model training. Python is an established programming language for machine learning and data analysis, and it offers a large library for manipulating and visualizing. A robust framework for creating and implementing deep learning models is offered by Google's open-source machine learning toolkit, TensorFlow. A high-level neural network API called Keras was created in Python and may be used with TensorFlow. Jupyter Notebooks will be used for code development and collaboration as well. To create our predictive models, we want to use deep learning methods like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Hybrid Architectures. To enhance the functionality of our models, we will also employ a variety of data preprocessing techniques, such as feature scaling, normalization, and feature selection.

- Convolutional Neural Networks (CNN) can be used to analyze complex data such as gene expression profiles, imaging data, or proteomic data to predict disease progression or response to treatment. This model can be used to extract features protein data measurements. It can help identify patterns in the data.
- Autoencoders are a type of neural network that can be used to learn a compressed representation of UPDRS data that can be used to predict future UPDRS scores and also can be used to learn a compressed representation of PD progression data.

- Variational Autoencoders are a type of autoencoder that can be used for generative modeling tasks. They can be used to generate synthetic UPDRS data that can be used to augment the limited amount of real-world data available for training predictive models.
- LSTMs are a type of RNN that are particularly good at capturing long-term dependencies in time series data and could be useful for modeling the progression of Parkinson's disease over time.
- RNNs are a type of deep learning model that can be used to predict the future year's MDS-UPDRS score of a patient by taking previous year's RNA-Sequence data as the input. This model can make use of the sequential patterns present in the RNA sequence data from the past to make predictions or gain insights.

Reference Link(s)

[1] "AMP®-Parkinson's Disease Progression Prediction," kaggle.com. https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction

[2] S. Ahmed, M. Komeili, and J. Park, "Predictive modelling of Parkinson's disease progression based on RNA-Sequence with densely connected deep recurrent neural networks," Scientific Reports, vol. 12, no. 1, p. 21469, Dec. 2022, doi: https://doi.org/10.1038/s41598-022-25454-1.

[3] Shahid AH, Singh MP. A deep learning approach for prediction of Parkinson's disease progression. Biomed Eng Lett. 2020 Apr 16;10(2):227-239. doi: 10.1007/s13534-020-00156-7. PMID: 32477610; PMCID: PMC7235154.

[4] Mahmood, A.; Mehroz Khan, M.; Imran, M.; Alhajlah, O.;Dhahri, H.; Karamat, T. End-to-End Deep Learning Method for Detection of Invasive Parkinson's Disease.Diagnostics 2023, 13, 1088. https://doi.org/10.3390/diagnostics13061088