Parkinson's Disease Progression Prediction

1st Elias Villalvazo-Avila
Tecnológico de Monterrey
M.Sc. in Computer Science
Estado de Mexico, Mexico
A01798097@tec.mx

***Abstract*—Parkinson's Disease (PD) is a disabling disorder that affects movements, cognition, sleep, and other normal functions. Currently, there is no cure and the disease worsens over time. However, research indicates that protein abnormalities play a key role in the onset and worsening of this disease. The main objective of this study is to predict the course of Parkinson's disease using protein abundance data. This work centers on the creation of a solution for this challenge. The solution is based on the CRISP-DM methodology with the goal of generating a predictive model that can be applied in the dataset provided for this challenge. Two methods were proposed to tackle the problematic: a Random Forest (RF) regressor, and a Shallow Neural Network with 4 dense layers. The final scores achieved by each methods in terms of mean square error and mean absolute error are 56.30 and 5.28 for the RF regressor, and 55.51 and 5.11 for the Shallow Neural Network, respectively, being the neural network the model that achieved the best score.**

***Index Terms*—Parkinson's disease, time series, tabular data, data science**

I. Introduction

Parkinson's disease (PD) is a medical condition that is characterized by a combination of symptoms including involuntary shaking (tremor) of the limbs at rest, stiffness of the muscles (rigidity), slow or reduced ability to move the limbs and facial muscles (bradykinesia), and general muscular weakness [1]. This disease usually occurs after the age of sixty but it can happen earlier; when the disease presents in people who are fifty or younger, it is known as young onsert Parkinson's disease (YOPD), which corresponds to 10 to 20% of peoples with PD. PD is 50% more common in men than in women. The reasons behind this are still not understood.

The severity of PD is often measured with the Melvin Yahr and Margaret Hoehn scale, which characterises the disease in five stages: stage 1 is associated with mild tremor or rigidity on one side of the body. In stage 2, tremor, rigidity, and bradykinesia occur on both sides of the body, without any loss of balance. Stage 3 includes all the symptoms from the previous stage, plus balance difficulty, loss of posture control, and hunching over. In stage 4, functional disability increases, but some independent functions are still possible. Some severe symptoms occur in stage 5 that patients require a wheelchair or are bedridden without assistance.

There is another scale to measure the progression of the diseases, the MDS-UPDRS (Unified Parkinson's Disease Rating Scale), a scale introduced in 1987, is a 4-part questionnaire that rates the signs signs and symptoms of Parkinson's Disease. It consists of four different sections, where each question goes from a score of 0 to 4, being 0 a normal state, and 4 the most severe impairment [2]. Nowadays, this scale is widely used for the analysis of the clinical situation of a patient with PD.

It is estimated that by 2037, 1.6 million people in the U.S. will have PD, yet, it is not clear how this disease is acquired by the individuals. It is estimated to have a 30 to 40 percent rate of heritability. Multiple genes may be involved in idiopathic late-onset Parkinson's disease. However, many researches increasingly believe that the root of PD is related to environmental factors and their combination with genetic factors [3]. In addition, some scientists believe that the genes involved in Parkinson's may be activated by exposure to an environmental agent, such as solvents, pesticides, or viruses, and that such agents may prompt the onset of the disease even in those individuals who carry none of the suspected genes. There is no current cure—and the disease worsens over time. The symptoms begin slowly, most often presented as stage 1 tremor, then it progresses to stage 3 in a period of five to ten years. This progression seemingly appears to be caused by the deterioration of several of the four brain structures called basal ganglia, related to the depletion of dopamine. Research indicates that protein or peptide abnormalities play a key role in the onset and worsening of this disease. Gaining a better understanding of this—with the help of data science—could provide important clues for the development of new pharmacotherapies to slow the progression or cure Parkinson's disease. Many of the currently-available treatments affect dopamine levels in the brain [1].

The Accelerating Medicines Partnership® Parkinson's Disease (AMP®PD) is a public-private partnership between government, industry, and nonprofits that is managed thorugh the Foundation of the National Institutes of Health (FNIH) [4]. The Partnership created the AMP PD Knowledge Platform, which includes a deep molecular characterization and longitudinal clinical profiling of Parkinson's disease patients, with the goal of identifying and validating diagnostic, prognostic, and/or disease progression biomarkers for Parkinson's disease.

The main objective of this study is to predict the course of Parkinson's disease using protein abundance data. The complete set of proteins involved in PD remains an open research question and any proteins that have predictive value are likely worth investigating further. The solution is based on the CRISP-DM methodology with the goal of generating a predictive model that can be applied in the dataset provided for this challenge.

The research questions that we expect to respond at the end of this contribution are:

- Which variables contain and provide the greatest amount of information to the predictive model?
- Is there any indicator that could predict the course of PD?
- Which model generates the best results? Why?
- Is there any correlation between variables?

II. Related Work

PD has been subject of study of several works, and this problematic has been assessed from different perspectives. For instance, the work by Islam et *al.* [5] presents a model used to predict disease's severity remotely by reviewing the motor performance of individuals with PD. In this study, participants were asked to perform hand movements, and features such as tapping frequency, period, amplitude, acceleration, among others, were used to feed an XGBoost regressor, obtaining promising results. The work from Want et *al.* [6] proposed the use of a deep-learning architecture for the analysis of drawings and writings acquired through a digital pen. This contribution proposes a deep-learning solution to the complex selection of appropriate characteristics acquired through complex processing which has been used by previous works. The proposed network combines a LSTM with Convolutional Neural Network (CNN) and utilises the tremor signals to learn the temporal and spatial characteristics of each tremor. In result, the proposed model generates state-of-the-art scores, using a reduced list of less-complex attributes. In terms of image processing, the work from Tran et *al.* [7] proposes the analysis of the retinal fundus as a diagnostic screening for PD using different machine learning and deep learning techniques. The work concludes that deep neural networks outperforms conventional machine learning models in the prediction of PD in retinal fundus images, and that this disease can be predicted even before the onset symptoms, which is valuable to initiate early intervention.

As observed in this section, there are multiple ways and strategies proposed throughout the years for the detection of PD, making it a very versatile problematic that can be assessed from different perspectives.

III. Data and Methods

A. Data Validation

The dataset used in this contribuition was obtained from the Kaggle challenge named: "AMP®-Parkinson's Disease Progression Prediction", publicly available in *https://www.kaggle.com/*. The core of the dataset consists of protein abundance values derived from mass spectrometry readings of cerebrospinal fluid (CSF) samples gathered from several hundred patients. Each patient contributed several samples over the course of multiple years while they also took assessments of PD severity. Three main files are used in this challenge:

train_peptides.csv, which contains mass spectrometry data at the peptide level. The attributes for this dataset are listed below:

- visit_id: ID code for the visit.
- visit_month: The month of the visit, relative to the first visit by the patient.
- patient_id: An ID code for the patient.
- UniProt: The UniProt ID code for the associated protein. There are often several peptides per protein.
- Peptide: The sequence of amino acids included in the peptide.
- PeptideAbundance: The frequency of the amino acid in the sample.

train_proteins.csv contains protein expression frequencies aggregated from the peptide level data.

- visit_id: ID code for the visit.
- visit_month: The month of the visit, relative to the first visit by the patient.
- patient_id: An ID code for the patient.
- UniProt: The UniProt ID code for the associated protein. There are often several peptides per protein.
- NPX: Normalized protein expression. The frequency of the protein's occurrence in the sample. May not have a 1:1 relationship with the component peptides as some proteins contain repeated copies of a given peptide.

train_clinical_data.csv

- visit_id: ID code for the visit.
- visit_month: The month of the visit, relative to the first visit by the patient.
- patient_id: An ID code for the patient.
- updrs_[1-4]: The patient's score for part N of the Unified Parkinson's Disease Rating Scale. Higher numbers indicate more severe symptoms. Each sub-section covers a distinct category of symptoms, such as mood and behavior for Part 1 and motor functions for Part 3.
- upd23b_clinical_state_on_medication: Whether or not the patient was taking medication during the UPDRS assessment. Expected to mainly affect the scores for Part 3. It's common for patients to take the motor function exam twice in a single month, both with and without medication.

B. Statistical Analysis

The dataset train_clinical_data consists of 2615 entries, each of these entries corresponds to the tracking of different patients through a maximum of 108 months. From this 2615 entries, only 248 different patient id's were found, which is almost equivalent to 10% of entries are unique. In the case of the dataset train_proteins , the total number of entries in the file is 232741. Each of the rows in this file correspond to a certain protein and its quantity during an exam performed at a certain visit by each patient. The number of patients corresponds equally to 248, and the number of unique proteins present in the file are 968. Finally, the file corresponding to the peptides, train_peptides, consists of 981834 different entries. The reason that this file is larger than the other is because a certain proteins can have more than a peptide associated to its type. A summary of the statistical measurements associated to the

TABLE I: Statistical measurements of the numerical variables available in the three main files.

	count	min	max	median	mean	std
visit_month	2615	0	108	24	31.1908	25.1991
updrs_1	2614	0	33	6	7.1106	5.5260
updrs_2	2613	0	40	5	6.7436	6.3232
updrs_3	2590	0	86	19	19.4212	15.0003
updrs_4	1577	0	20	0	1.8618	3.0221
PeptideAbundance	981834	10.9985	178752000	74308	642890	3377989
NPX	232741	84.6082	613851000	113556	2712077	22241550

different numeric types present in these files is presented in Table I

C. Exploratory Data Analysis (EDA)

The first step of the EDA was to check the amount of missing values in each dataset. The datasets of train_proteins and train_peptides did not contain missing information within their rows. However, the dataset train_clinical_data contained several missing values. For instance, the column updrs_1 had 0.038% of missing informaton, column updrs_2, 0.076%, updrs_3 0.95%, updrs_4 39.69%, and upd23b_clinical_state_on_medication 50.74%, being this half of the dataset left empty. However, the rows of the dataset could not be removed since the number of rows (2615) is small.

The second step of the EDA consisted in the analysis of the distribution of the number of visits per time, which can be observed in Figures 1 and 2. It can be observed that the number of visits per month reduces over time. The first 45 months concentrate the majority of the information.

In order to use all the information from the three sources, the information from each file was merged into a single dataframe, which was re-structure to have several entries of month, user id and visit number per patient, based on the number of peptides and proteins present in data. After the transformation, the concentration of all the different peptides and proteins during a certain visit will be considered as an independent variable during the training. The predicted labels were also modified so for each entry in our dataset, we predict all the UPDRS scores in a time frame of 0, 6, 12, and 24 months after the entry. After the merging and the modification of the dataset, we obtained 1068 different entries, each with 1196 features related to the amount of protein, peptides and visit month, which will be used for training and testing.

During this step, a correlation analysis was performed to find any correlation between the UPDRS, the amount of proteins present during exams, and the visit month. However, no correlation was found, as observed in Fig. 3.

Regarding the treatment of the missing values, we explored two options to treat them based on the model we used to create the predictor, which are described in the following section.

D. Proposed Model

In order to find any relationship between the severity of the symptoms and the visit months, a series of histograms that consider the previously-mentioned variables were created, which can be observed in Fig. 1. Unfortunately, the amount of mild symptoms (bottom) remains the most frequent case throughout time.

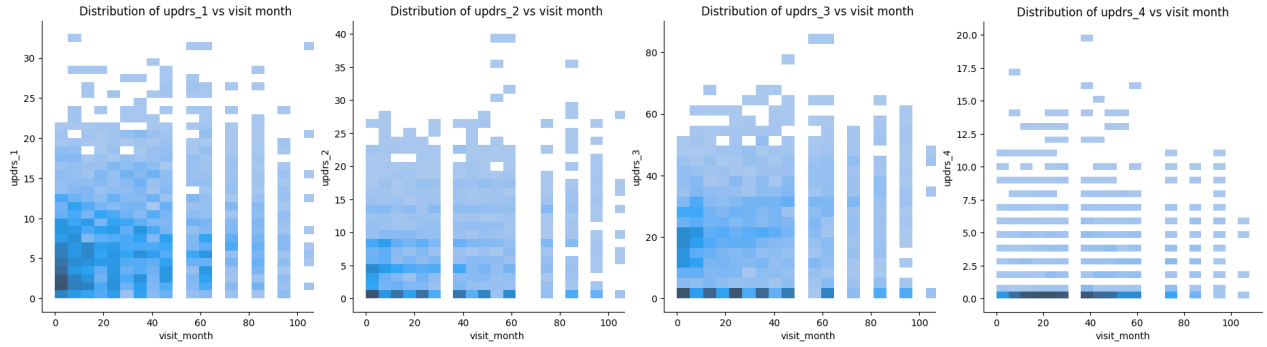
To compare the traditional machine learning models and neural network models, a model of each was proposed to solve this regression problem. To train the models, the data was divided into two partitions for training (80%) and testing (20%), which correspond to 864 and 204 samples, respectively.

For the traditional machine learning model, a Random Forest (RF) Regressor was selected. This model is an estimator that fits several decision trees on various sub-samples of the dataset and uses an averaging or voting mechanism to improve the accuracy of the problem, and to control overfitting. The initial configuration of the model consisted in the creation of 300 trees with a max depth of 16, using all the possible variables for the decision. The selected metric to measure the regression error was the mean square error. A random forest model was created for each UPDRS level, having at the end a total of 4 regressors.

The second model consists of a shallow neural network. One of the benefits from using neural networks is that complex features are generated according to the data that the model was trained with. The model consists of four fully-connected layers with 64, 100, 100 and 16 neurons each. To add non-linearities to our data, ReLU was attached after each layer. One of the limitations of neural networks in comparison with traditional techniques is the large amount of data needed to train the model. Since the dataset used is small, techniques such as the addition of Adam optimizer, dropout layers, and a small training period with a non-deep architecture were used to avoid overfitting of the model.

IV. Results

The RF model was updated from the 300 trees, which was the default parameter, to use only 100 trees. We observed that for all the cases, the mean square error metric achieved a good relationship between number of trees and error around that value, as observed in Fig. 4. The other configuration that was changed was the max depth of the trees; it was limited to 16 nodes. The implementation was done using TensorFlow Decision Forest library. For the treatment of the missing values, we dropped the rows that contained missing information either in the training rows or in the target rows. By doing so, the training and testing data for each regressor (one per updrs label) had a varying training size, as seen in Table II. The mean squared error (MSE) and the mean absolute error (MAE) were the metrics used to measure the performance of this regression problem, these were obtained for each regressor (see columns 4 and 5 in Table II). In general, RF regressor was able to obtain low error rates for the scores 1, 2 and 4. For the score level 3, the error rates are slightly higher. The reason is that, as observed in Table I, the standard deviation of the values related to updrs3 are sparser, compared to the other three levels. This behavior is then reflected on the performance of the model. In average, the MSE and MAE of the model are 56.60 and 5.3, respectively.



(a) updrs1 (b) updrs2 (c) updrs3 (d) updrs4

Fig. 1: Analysis of the updrs score for the four different levels vs the visit month.

For the Neural Network, the model was implemented and the data was rescaled using TensorFlow in Python. The proposed model consisted of four dense layers, with Adam optimization with a learning rate of 0.01 and mean square error loss. Compared to the RF, the missing values in data were not dropped. Instead, these were filled using the median of each column. The reason behind this decision is because neural networks require more data in order no work properly and not overfit. Since the training dataset is small, we must use as much information as possible. Additionally, dropout layers were added in the model to reduce even more the chances of overfitting. The dropout probability used in this model was 0.5, and a total of four layers were added in the model, one after every dense block. The model was trained between 100 and 200 epochs with a training validation split of 80% (945 rows, 1196 columns) and 20%, respectively. The number of epochs that achieved the best score was 150. The MSE and MAE scores achieved for this combination is 55.51 and 5.1 in average for all the updrs predictions. The final results for both models is concentrated in Table III.

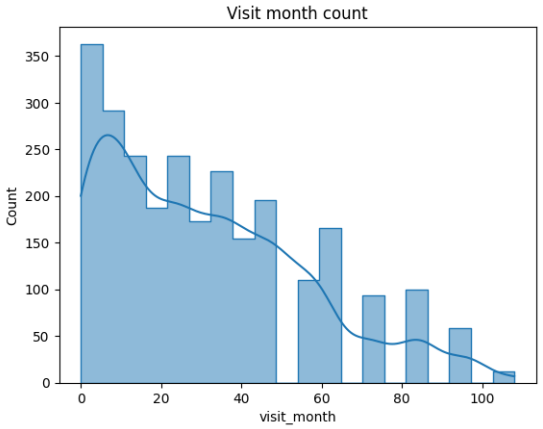


Fig. 2: Count of visits per month

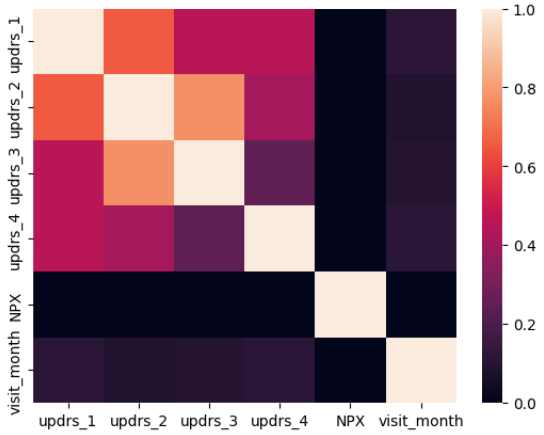


Fig. 3: Correlation matrix between updrs scores from 1 to 4, and the count of proteins.

TABLE II: Number of training and testing samples, and their associated mean square error and mean absolute error for the Random Forest Regressor, per updrs label.

Label	Training size	Testing size	MSE	MAE
updrs_1	828	240	25.80	3.65
updrs_2	851	217	33.67	4.42
updrs_3	823	235	174.95	10.19
updrs_4	450	119	6.61	2.19

V. Discussion

The best score was achieved by the RF forest classifier, achieving a smaller MSE and MAE values. However, the difference between each model is minimum and we can use the two models indistinctly for the updrs measurement

TABLE III: Mean average precision and Mean absolute error for the two models assessed in this contribution.

Model	Mean Square Error	Mean Absolute Error
RF	56.3	5.28
Shallow NN	55.51	5.11

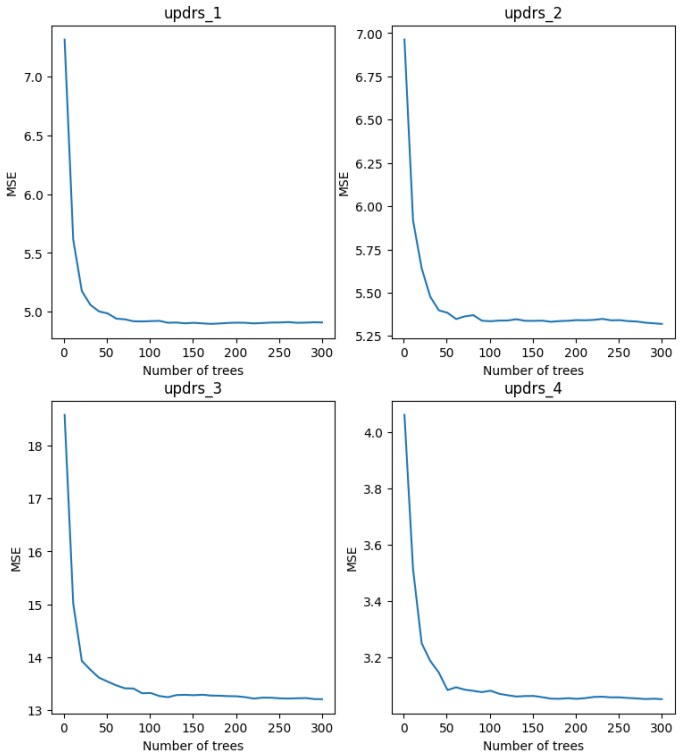


Fig. 4: Number of trees vs. Mean Square Error for the four UPDR scores when training the Random Forest Regressor.

prediction. However, there are some remarks that must be considered: First, the data used to train the Neural Network model contained 53.20% of missing values that are related to some proteins or peptides that were not measured at a certain visit. Therefore, the training information will have a certain bias induced by the way missing values were handled by each model. Despite that, the two models achieved a similar score with two different missing-values-handling strategies, which can confirm that such bias may not be prejudicial. Second, the dataset is small, therefore large Neural Network models will overfit. Strategies such as Adam, dropout, and a small learning rate helped the training of the network presented in this work. Complementary Data and Imputing strategies need to be assessed in future work

VI. Conclusion

In this work, we tested two methods for the prediction of the unified Parkinson's disease rating scale based on information from protein and peptide measurements acquired during the periodic evaluation of 248 patients during a period between 0 and 108 months. The dataset was obtained from the Kaggle challenge named "AMP®-Parkinson's Disease Progression Prediction". The results given in this contribution demonstrate that the updrs values can be predicted from the protein and peptide information by using a random forest regressor, or by the implementation of neural networks. However, there may still exist a bias caused by the limited information in the dataset. Therefore future work is necessary to implement data augmentation strategies and to add more information for a more accurate accuracy.

References

[1] P. Ross, John Alan and P. Singer, Sanford S., "Parkinson's disease," 2022, accession Number: 86194383; Author: Ross, John Alan, PhD; Singer, Sanford S., PhD; Subject Term: Parkinson's disease; Subject Term: Parkinsonian disorders; Subject Term: Movement disorders; Subject Term: Brain diseases; Subject Term: Brain degeneration; Subject Term: Dementia; Number of Pages: 6p.; Document Type: Article; Publication Type: Encyclopedia; Full Text Word Count: 3585.

[2] J. Kulisevsky, M. Luquin, J. Arbelo, J. Burguera, F. Carrillo, A. Castro, J. Chacón, P. García-Ruiz, E. Lezcano, P. Mir, J. Martinez-Castrillo, I. Martínez-Torres, V. Puente, A. Sesar, F. Valldeoriola-Serra, and R. Yañez, "Enfermedad de parkinson avanzada. características clínicas y tratamiento (parte i)," *Neurología*, vol. 28, no. 8, pp. 503–521, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0213485313001175

[3] A. A. Hicks, H. Pétursson, T. Jónsson, H. Stefánsson, H. S. Jóhannsdóttir, J. Sainz, M. L. Frigge, A. Kong, J. R. Gulcher, K. Stefánsson, and S. Sveinbjörnsdóttir, "A susceptibility gene for late-onset idiopathic parkinson's disease," *Annals of Neurology*, vol. 52, no. 5, pp. 549–555, 2002. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.10324

[4] [Online]. Available: https://amp-pd.org/

[5] M. S. Islam, W. Rahman, A. Abdelkader, P. T. Yang, S. Lee, J. L. Adams, R. B. Schneider, E. R. Dorsey, and E. Hoque, "Using ai to measure parkinson's disease severity at home," 2023.

[6] X. Wang, J. Huang, M. Chatzakou, K. Medijainen, P. Taba, A. Toomela, S. Nomm, and M. Ruzhansky, "A light-weight cnn model for efficient parkinson's disease diagnostics," 2023.

[7] C. Tran, K. Shen, K. Liu, and R. Fang, "Deep learning predicts prevalent and incident parkinson's disease from uk biobank fundus imaging," 2023.