








Eye-Tracking and Machine Learning Significance in Parkinson's Disease Symptoms Prediction

Artur Chudzik¹  , Artur Szymański¹ , Jerzy Paweł Nowacki¹ ,
and Andrzej W. Przybyszewski^{1,2} 

¹ Polish-Japanese Academy of Information Technology, Koszykowa 86 Street,
02-008 Warsaw, Poland

{artur.chudzik, artur.szymanski, nowacki, przy}@pjwstk.edu.pl

² Department of Neurology, University of Massachusetts Medical School, 65 Lake Avenue,
Worcester, MA 01655, USA

andrzej.prybyszewski@umassmed.edu

Abstract. Parkinson's disease (PD) is a progressive, neurodegenerative disorder characterized by resting tremor, rigidity, bradykinesia, and postural instability. The standard measure of the PD progression is Unified Parkinson's Disease Rating (UPDRS). Our goal was to predict patients' UPDRS development based on the various groups of patients in the different stages of the disease. We used standard neurological and neuropsychological tests, aligned with eye movements on a dedicated computer system. For predictions, we have applied various machine learning models with different parameters embedded in our dedicated data science framework written in Python and based on the Scikit Learn and Pandas libraries. Models proposed by us reached 75% and 70% of accuracy while predicting subclasses of UPDRS for patients in advanced stages of the disease who respond to treatment, with a global 57% accuracy score for all classes. We have demonstrated that it is possible to use eye movements as a biomarker for the assessment of symptom progression in PD.

Keywords: Eye-tracking · Saccades · Parkinson's Disease · Machine learning

1 Introduction

In Parkinson's disease, we can distinguish multiple therapies, which could be combined. The gold-standard treatment for PD is a pharmacological treatment with Levo-dopa (L-dopa) [1, 2]. Nevertheless, L-dopa associates with long-term disturbances. We can distinguish hypo-hyperkinetic phenomena and psychosis as examples of motor and mood side effects. [3]. The practical and safe procedure, which is lacking these effects and remains the preferred surgical treatment for advanced Parkinson's disease, is Deep Brain Stimulation of the subthalamic nucleus (STN) [4]. However, there is still a necessity of the support for the neurologists in the field of optimum treatment parameters, because due to the huge diversity of cases, even the most experienced doctors could not be sure how the therapy would influent on the patient.

2 Methods

2.1 The Subject of the Study

The research group was composed of 62 patients who have Parkinson's disease and who are under the supervision of the Warsaw Medical University (Warsaw, Poland) neurologists. We differentiate the patients into three groups. The first one, we named the Best Medical Treatment (BMT). In this group, we placed patients who were treated only by the medication. The second one, the Deep Brain Stimulation (DBS), was a group where patients, who had implanted electrodes in the STN during our study. The last group, which was named Post-Operative Patients (POP) aggregates individuals who had had surgery earlier (before the beginning of our research). Every PD patient had three visits, which were underdone approximately every six months. Every patient from the DBS group had his/her visit before surgery.

In order to obtain the countable value of the disease, it is crucial to provide a precise neurological tool which could objectively measure all of the symptoms and determine a score. For the metric of Parkinson's disease advancement, there are two common neurological standards: the Hoehn and Yahr scale and the Unified Parkinson's Disease Rating (UPDRS). The UPD rating scale is the most commonly used in the clinical study of Parkinson's disease [5], and we also decided to adopt it in this study. Altogether with UPDRS, every patients' disease metric was combined with the disease duration, the result of Parkinson's Disease Questionnaire PDQ39 (which is a disease-specific health-related quality-of-life outline), the result of Epworth Sleepiness Scale (which is intended to measure daytime sleepiness), and the parameters of saccadic eye movements, described further.

The mean age of patients was 51.1 ± 10.2 (standard deviation) years. The mean duration of the disease was 11.6 ± 4.3 years. The mean of UPDRS score (for all symptoms) was 33.8 ± 19.4 . The mean of PDQ39 score was 50.5 ± 26.0 , and the mean of Epworth score was 8.7 ± 4.6 .

2.2 Eye Movements

The often diagnosed impairment of automatic behavioral responses accompanies the slowness of initiation of voluntary movements in individuals with PD [6]. An individual set of behavioral tasks may provide insight into the neural control of response suppression with the usage of motor impairments analysis based on saccadic eye movements [7, 8]. Saccades are a quick, simultaneous movement of both eyes between two or more phases of fixation in the same direction, and can be measured quickly and precisely.

We choose this marker because of the considerable understanding of the neural circuitry controlling the planning and execution of saccadic eye movements [9].

During this study, we have used a head-mounted saccadometer JAZZ-pursuit (oberconsulting.com), which was able to measure the reflexive saccades (RS) in the high frequency (1000 Hz). We have chosen this device because it is optimized for easy set-up and provides minimal intrusiveness while can keep stable 1 kHz frequency of measurement. During the experiment, we created a task of the horizontal reflexive saccades

analysis. Used hardware allowed us to obtain high accuracy and precision in eye tracking and the compensation of possible subjects' head movements relative to the monitor. Thus subjects did not need to be positioned in an unnatural chinrest, which has a positive influence on the ergonomics of the experiment. However, we asked patients to use a headrest in order to minimize the head motion because they could have a significant influence on the accuracy of the high-frequency measurements. Each patient was seated in front of the monitor at a distance of 60–70 cm.

The patients' task was to fix their eyes on the spot placed in the middle of the screen (0°). Then, the spot changed color into one of the possible variations and shift horizontally to one of the possible directions: 10° to the left, or 10° to the right, after arbitrary time ranging between 0.5–1.5 s. During that task, we measured the fast eye movements of the patient, according to the spot color transition.

When the transition was from white to green, it was a signal for the execution of RS. We also prepared an additional protocol for antisaccades (AS) measurement. In this task, the individual was asked to make a saccade in the direction away from the stimulus. A signal for that was when the spot changed color from white to red. After that, the central marker was hidden, and one of the two peripheral targets was shown. The selection was made randomly, with the same probability.

According to the task, each patient looked at the spots and followed them as they moved (in the RS task) or made opposite direction eye movement (in the AS task). After that, the target remained still for 0.1 s before the next experiment initialization.

In each test, the subject had to perform twenty saccades and antisaccades in a row twice. The recording of the first session (marked as S1) was with the patient who has temporarily disabled treatment (without medicine/with the disabled neurostimulator). In the next session (marked as S3), the patient took medication and had a break for one half to one hour, and then the same experiments were performed (with L-dopa/with enabled neurostimulator).

In this experiment, we have investigated only RS data using the following population parameters averaged for both eyes: mean latency ($RSLat \pm SD$), mean amplitude ($RSAmplitude \pm SD$), mean of the maximum velocity ($RSPVel \pm SD$), and mean duration of the saccades ($RSDur \pm SD$).

2.3 Dataset

We implemented a dedicated database for measurements storage that was designed and maintained by Polish-Japanese Academy of Information Technology (Warsaw, Poland). For operations described further, we flattened the data, placing every experiment in each row, with the results and metadata in the separated columns. Therefore it could be represented by a single table represented by comma-separated-values, which is a universal format for computational engines. The basic structure of the dataset contains 374 observations, and each has 13 variables, which are: *Duration* - the duration of the disease; *UPDRS*, *PDQ39*, *Epworth* - the score for each test; *RSLat*, *RSDur*, *RSAmplitude*, *RSPVel* - the parameters of recorded saccades; *Session*, *Visit* - the indexes of the experiment; and *BMT*, *DBS*, *POP* - boolean variables which describe the kind of patients' therapy.

2.4 Computational Learning Theory

Our data contains a set of N training samples of the form $(x_1, y_1), \dots, (x_n, y_n)$ such that x_i is the feature vector of the i -th sample with the class denoted by y_i . Thus, it is possible to use a supervised learning algorithm which seeks for a function $g : X \rightarrow Y$, where the X is the input space, and the Y is the output space. The g function is an element of some space of possible functions G , known as the hypothesis space.

The task itself could be a multiclass classification. Our goal is to predict a level of the disease measured as UPDRS value binned into intervals, for different groups of patients. For the predictions, we have created a dedicated machine learning framework, written in Python and based on two libraries: Scikit Learn [11] and Pandas, which are high-quality, well-documented collection of canonical tools for data processing and machine learning. We have chosen well-known models that implement different multiclass strategies, such as K Neighbors Classifier, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier. The framework allowed us to find an optimal solution by the examination of multiple algorithms with different parameters.

2.5 Hypothesis

Our goal was to predict Parkinson's disease progression in the advanced stage, based on the data obtained from the patients in the different treatment and stage of this disease. This task is non-trivial because there are significant differences between symptom developments and the effects of different treatments in individual PD patients.

As a training dataset, we used patients from the BMT group (3rd visit), DBS (3rd visit), and POP (1st visit). The independent test set consisted of the POP group from the second visit.

3 Results

Our framework was responsible for every step of data processing in order to evaluate the best model based on given data. Therefore, we implemented procedures which were responsible for the creation of the Profiling Report, Correlation Matrices, Missing Data Imputation, Data Discretization, One-Hot Encoding or Data Normalisation of selected variables and Machine Learning Algorithms Evaluation for various parameters.

3.1 Profiling

Our dataset consists of 374 observations, where each has 13 variables. First, we generated a Pearson correlation coefficient matrix, where each cell in the table shows the correlation between two variables. It is a useful tool that proves if a correlation between the parameters from measurements and the symptoms suggest that some close relations exist (Fig. 1).

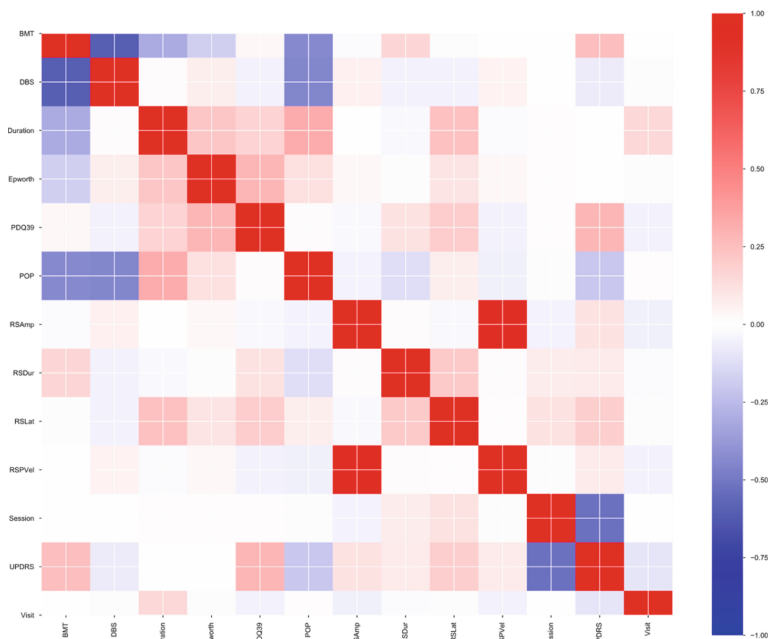


Fig. 1. Correlation matrix. The similarity between all paired parameters is correlated with the color of the cell in the matrix (red shade represents high similarity; blue stands contrariwise). (Color figure online)

3.2 Pre-processing

In the report, we noticed that some of the values are empty. *Epworth* column had 5 (1.3%) zeros; *RSamp*, *RSDur*, and *RSLat* columns had 6 (1.6%) zeros. Because of the Epworth scale ranges between 0–24, we decided not to apply data imputation on that column. When we analyzed records related to a single patient, we noticed, that on an early stage of the disease (duration about 7 ± 0.3 years) two of them indeed reported no problems with the daytime sleepiness, yet visits in the following years revealed a linear increase in the results. For missing parameters of the saccades, we applied the imputation transformer for completing missing values which replaced missing values using the mean along each column.

For neurological and neuropsychological tests results (Fig. 2), we applied k-bins discretization, which provides a way to partition continuous features into discrete values. Thus those features are a one-hot encoded allowing the model to be more expressive while maintaining interpretability.

For neurological and neuropsychological tests results, we applied k-bins discretization, which provides a way to partition continuous features into discrete values. Thus those features are a one-hot encoded allowing the model to be more expressive while maintaining interpretability.

$$Epworth = (-\infty, 6.00), [6.00, 11.00), [11.00, +\infty)$$

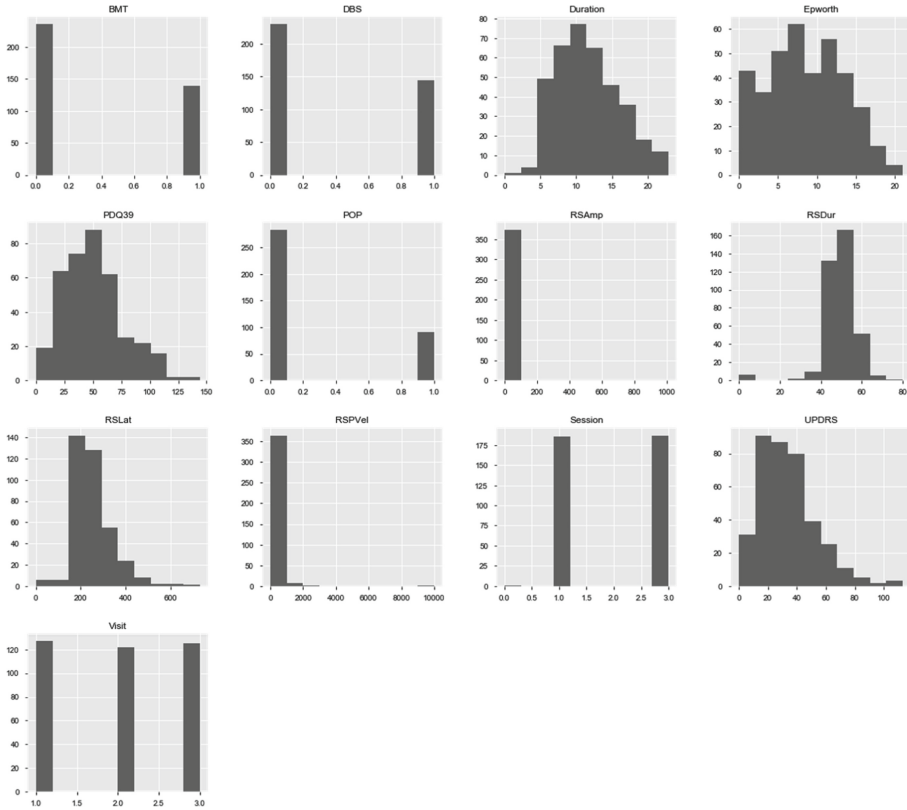


Fig. 2. A histogram, which is a representation of the distribution of data for every measurable property in the data set. Y-Axis represents a number of samples. The X-Axis presents a parameter value.

$$PDQ39 = (-\infty, 25.0), [25.0, 38.0), [38.0, 47.5), [47.5, 58.0), [58.0, 74.0), [74.0, +\infty)$$

Other columns (*Duration*, *RSLat*, *RSDur*, *RSamp*, *RSPVel*) were standardized by removing the mean and scaling to unit variance to keep the subtle representation of the eye movement signal. The calculation of the standard score of X sample is defined by:

$$z = \frac{x - u}{s}$$

where u is the mean of the training samples, and s is the standard deviation of the training samples.

Target value, the *UPDRS* score was optimally divided into four ranges

$$UPDRS = (-\infty, 19.25), [19.25, 30.50), [30.50, 44.00), [44.00, +\infty)$$

This split ensured approximately the same data set size for each class.

3.3 Machine Learning

We decided to evaluate multiple algorithms in order to determine the best approach in the meaning of accuracy of the predictions. In the previous research [10], Random Forrest Classifier achieved the high prediction level, being second to Rough Set. This article covers the evaluation of a few machine learning models which were not used in the previous research, such as K Neighbors Classifier, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier. The following sections present each algorithm with evaluated parameters and scores obtained from our machine learning framework.

K Neighbors Classifier

Classifier implementing the k-nearest neighbors' vote with uniform weights achieved maximum accuracy score (0.50) with a parameter defining the number of neighbors on the level of 12 and 13 (Fig. 3).

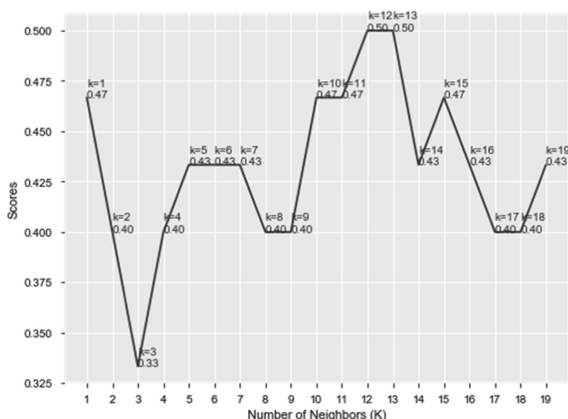


Fig. 3. K Neighbors Classifier accuracy scores for different K values.

Support Vector Classifier

We evaluated the C-Support Vector Classification against different kernels (linear, polynomial, radial-basis, and sigmoid). The “linear” kernel achieved the best accuracy score (0.40) under the penalty parameter $C = 0.3$ (Fig. 4).

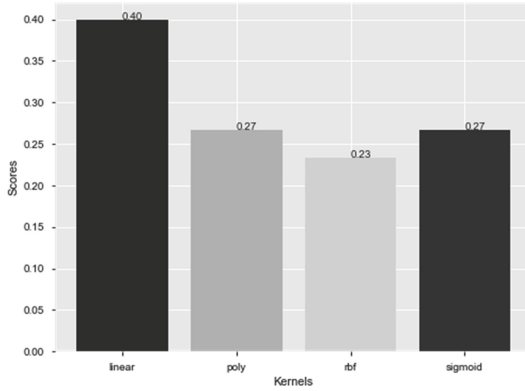


Fig. 4. Support Vector Classifier accuracy scores for different kernels.

Decision Tree Classifier

A decision tree classifier could provide different results when we change the number of features to consider when looking for the best split. Varying them between 1 and the size of all columns of the learning set, the best accuracy (0.50) was for 6, 14, and 15 features (Fig. 5).

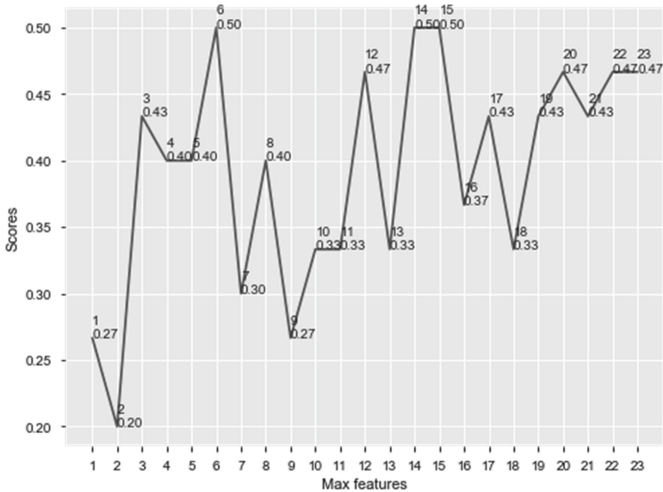


Fig. 5. Decision Tree Classifier accuracy scores for different number of maximum features.

Gradient Boosting Classifier

Gradient Boosting it allows for the optimization of arbitrary differentiable loss functions. We challenged a different number of boosting stages to perform. Gradient boosting is relatively robust to over-fitting, so a large number usually results in better performance.

The scope was between 10 and 1000 estimators, and the highest score (0.43) reveals in the 200 boosting stages (Fig. 6).

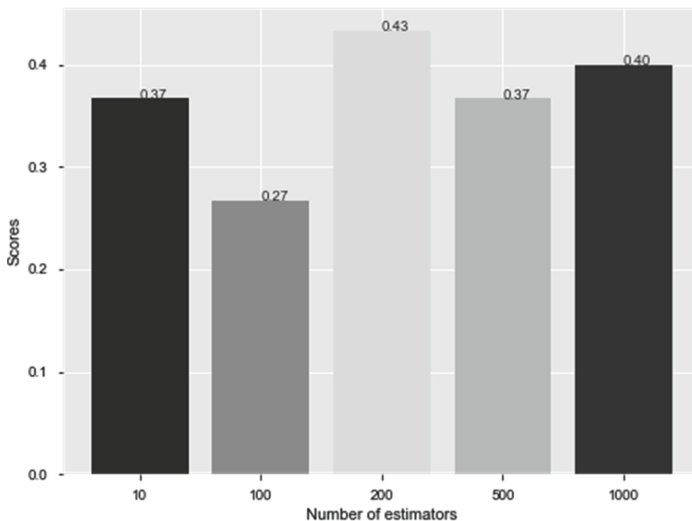


Fig. 6. Gradient Boosting Classifier accuracy scores for different number of estimators.

Random Forest Classifier

In this classifier, we evaluated a parameter which defines the number of trees in the forest. We used Gini impurity as a criterion of the quality of a split. Random Forest Classifier achieved the highest overall accuracy score (0.57) among other machine learning algorithms when the number of estimators exceeded 120 (Table 1, Fig. 7).

Table 1. Confusion matrix based on Random Forest Classifier result.

		Predicted label				Accuracy
		(-Inf, 19.25)	[19.25, 30.5)	[30.5, 44.)	[44., +Inf)	
True label	(-Inf, 19.25)	6	1	1	0	0.75
	[19.25, 30.5)	1	7	2	0	0.70
	[30.5, 44.)	3	2	3	0	0.38
	[44., +Inf)	1	1	1	1	0.25

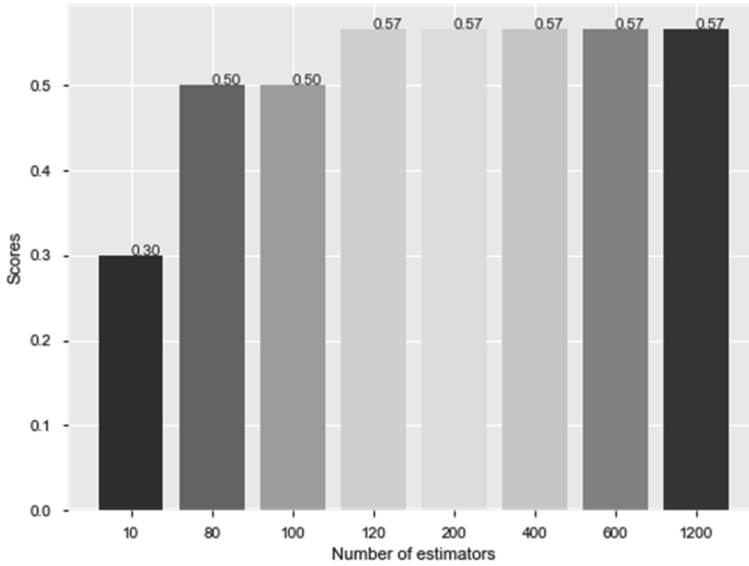


Fig. 7. Random Forest Classifier scores for different number of estimators.

4 Discussion

Examination of the variation of the saccades is a non-invasive method of evaluating the neural networks involved in the control of eye movements. The examples above demonstrated that it is possible to use eye movements as a biomarker for the assessment of symptom progression in PD.

We live in the “age of implementation” of the machine learning models. Commercial companies are acquiring data on a large scale in order to present the content which is best-fitted to the end-user, which influences the businesses. However, there is still an emerging necessity of medical data extension because it is a crucial factor in the context of machine learning. Paradoxically, we have access to much fewer data of the medical records (even about ourselves) than most of the managers of the commercial websites who aggregates about our shopping behavior. The modeled data sample, presented in this article, is relatively abundant in the neuroscience scale. Based on the examinations conducted by neurologists, we were able to create predictions based on the different patient population with different treatments for the most advanced stages of the disease. This results supported with further extended and in-depth research could lead to a new approach in the development of a follow-up tool for PD symptoms. As an outcome, this automated mechanism could provide to a doctor an objective opinion about applied therapy symptoms. Hence we can conclude that when the patient is doing significantly worse than others, their treatment is not optimal and should be changed. However, there is still a large field for the model improvements that should lead to more accurate results. The proposed protocol allowed us to evaluate multiple machine learning models in a relatively agile process of data aggregation. Consequently, more complex variations of saccadic tasks can give insight into higher-order eye movement control [12]. Our work is

an evaluation of well-known models that implement different multiclass strategies, such as K Neighbors Classifier, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier in the context of saccades research. In this trial, Random Forest Classifier achieved the highest overall accuracy score, which could lead to a direction in further discoveries in the field of bioinformatics.

5 Conclusions

We believe that the multidisciplinary cooperation between neurologists and information engineering is essential to achieve significant results in the open-science approach. We attempted to predict the longitudinal symptom developments during different treatments based on the neuropsychological data aligned with the parameters of the eye movements. As a training dataset, we used patients from the BMT group (3rd visit), DBS (3rd visit), and POP (1st visit). The independent test set consisted of the POP group from the second visit. For predictions, we used a machine learning framework, written in Python. The best classifier - Random Forest - reached 75% and 70% of accuracy while predicting subclasses of UPDRS for patients in advanced stages of the disease who respond to treatment, with a global 57% accuracy score for all classes. Thanks to collaborative research, we have presented a comparison of different machine learning models that could be useful in the context of bioinformatics. Our direction is to create a new research ecosystem, that would significantly increase (by a factor of 10) the number of attributes and measurements in order to implement deep learning methods.

References

1. Connolly, B.S., Lang, A.E.: Pharmacological treatment of Parkinson disease: a review. *JAMA* **311**(16), 1670–1683 (2014)
2. Goldenberg, M.M.: Medical management of Parkinson's disease. *Pharm. Therapeutics* **33**(10), 590 (2008)
3. Thanvi, B.R., Lo, T.C.N.: Long term motor complications of levodopa: clinical features, mechanisms, and management strategies. *Postgrad. Med. J.* **80**(946), 452–458 (2004)
4. Benabid, A.L., et al.: Deep brain stimulation of the subthalamic nucleus for the treatment of Parkinson's disease. *Lancet Neurol.* **8**(1), 67–81 (2009)
5. Ramaker, C., et al.: Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* **17**(5), 867–876 (2002)
6. Henik, A., et al.: Disinhibition of automatic word reading in Parkinson's disease. *Cortex* **29**(4), 589–599 (1993)
7. Jones, G.M., DeJong, J.D.: Dynamic characteristics of saccadic eye movements in Parkinson's disease. *Exp. Neurol.* **31**(1), 17–31 (1971)
8. White, O.B., et al.: Ocular motor deficits in Parkinson's disease: II. Control of the saccadic and smooth pursuit systems. *Brain* **106**(3), 571–587 (1983)
9. Chan, F., et al.: Deficits in saccadic eye-movement control in Parkinson's disease. *Neuropsychologia* **43**(5), 784–796 (2005)
10. Przybyszewski, A., et al.: Multimodal learning and intelligent prediction of symptom development in individual Parkinson's patients. *Sensors* **16**(9), 1498 (2016)
11. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
12. Nij Bijvank, J.A., et al.: A standardized protocol for quantification of saccadic eye movements: DEMoNS. *PLoS ONE* **13**(7), e0200695 (2018)