Right on track: Towards improving DBS patient selection and care

Geraedts V.J.

CHAPTER 8

Machine Learning for automated EEG-based classification of cognition during the DBS screening in Parkinson's Disease patients

Geraedts VJ, Koch M, Contarino MF, Middelkoop HAM, Wang H, van Hilten JJ, Bäck THW, Tannemaat MR

Under review

Abstract

Background

A downside of Deep Brain Stimulation (DBS) for Parkinson's Disease (PD) is that cognitive function may deteriorate postoperatively. Accurate cognitive assessment is crucial in determining DBS eligibility, but interpretability of this assessment is limited due to external influences.

Objective

To explore EEG as complementary biomarker for cognition using a Machine Learning (ML) pipeline to classify DBS candidates.

Methods

A fully automated ML pipeline was applied to 112 PD patients, taking EEG time-series as input and predicted class-labels as output. No arbitrary choices were made during the entire process. The most extreme cognitive performance scores were selected for class differentiation, i.e. best cognitive performance (high-COG, n=20) vs. worst cognitive function (low-COG, n=20). 16674 features were extracted per patient; feature-selection was performed using a Boruta algorithm. A random forest classifier was modelled and 10-fold cross-validation with implemented Bayesian optimization procedure was performed to ensure generalizability. The predicted class-probabilities of the entire cohort were compared to actual cognitive performance.

Results

The final model differentiated both groups with a mean (SD) accuracy of 0.92 (0.02), whereas a model using only occipital peak frequency achieved an accuracy of 0.67 (0.06). The class-probabilities and actual cognitive performance were negatively linearly correlated (β = -0.23 (95%CI (-0.29, -0.18))).

Conclusion

These findings indicate particularly high accuracies when using a compound of automatically extracted EEG biomarkers to classify PD patients according to cognition and is superior to a single spectral EEG feature. Automated EEG assessment may have utility for cognitive profiling of PD patients during the DBS screening.

Introduction

Parkinson's Disease (PD) is the fastest growing neurological disorder worldwide,[1] with both characteristic motor and non-motor symptoms. Patients who develop motor complications may be eligible for These patients may be eligible for Deep Brain Stimulation (DBS), an invasive surgical intervention which is highly effective in relieving motor complications and improves quality of life.[2, 3] Despite good effects on motor functioning and substantial relief of motor complications refractory to oral medication,[3, 4] DBS does not improve cognitive symptoms and some deterioration can be observed in cognitive domains [5, 6] and neuropsychiatric functioning after surgery.[7, 8] The screening process for DBS therefore entails an extensive evaluation of cognitive and neuropsychiatric functioning to rule out severe impairment prior to surgery, in order to determine DBS eligibility.[9, 10] However, accurate evaluations of cognition are limited by factors such as intellectual status,[11] while performance tasks may be subject to misinterpretation due to e.g. motor impairment, fatigue, mood disorder, stress, and personal motivation, which may render results less valid.[12, 13] In addition, neuropsychological screening is time-consuming and stressful for patients. Consequently, there is a need for new biomarkers to complement current neuropsychological assessments of cognition.

A candidate instrument for such complementary assessments is quantitative Electroencephalography (qEEG), which can measure brain activity directly and non-invasively. The utility of qEEG to aid during assessment of cognitive impairment, and even predict cognitive deterioration has been previously established in the general PD population.[14] Particularly spectral features reflecting EEG slowing are related to cognitive deterioration, although recent advances in EEG processing have demonstrated an association of cognitive impairment with connectivity and network dysfunction in cross-sectional studies as well.[15-17] However, these latter metrics have been sparsely studied in comparison to spectral analyses.[14] An extensive evaluation across the numerous possibilities of EEG metrics beyond spectral powers, to determine which metrics have the highest potential for reflecting PD symptoms, is lacking.

A limitation of qEEG analyses is the laborious amount of pre-processing, and particularly, the arbitrary selection of features to include during the final modelling. Traditionally, features from time series are manually selected and computed, which is time-consuming and requires expert knowledge and is therefore difficult to translate to clinical practice. A machine learning (ML) approach may overcome these limitations by providing output, such as a classification of cognitive status, without predefined data-extraction or modelling.[18] Preliminary ML results on determining levels of cognitive severity demonstrated high performance scores, although

limited to predetermined (spectral) features only. These models still require a large degree of pre-processing and manual feature-extraction.[19] Ideally, the ML approach is extended to a fully automated ML pipeline, deemed a 'sequence of data processing components'.[20] Within a ML pipeline, the EEG time series are delivered as input, after which an automated algorithm extracts a large number of features, selects those features which are needed to create a representative EEG profile, and learns and optimizes a ML model, without any intervention in between. Such a pipeline limits the necessity of making arbitrary choices, makes the entire process more efficient, and increases the likelihood of identifying novel biomarkers.

Given the need for complementary objective screening instruments to evaluate cognition during the DBS screening, the aim of our study was to evaluate the utility of a qEEG ML pipeline for determining cognitive status in these patients. To this end, the most 'extreme' DBS candidates were selected to build a supervised learning model, i.e. best vs. worst cognitive scores after a comprehensive neuropsychological test battery. The model could then be applied to evaluate the remaining DBS candidates, during which the association between ML-predictions and the actual levels of cognitive function could be studied.

Methods

Study participants

All consecutive patients who underwent preoperative screenings for DBS at the Leiden University Medical Center (LUMC) between September 2015 and June 2019 were included in the study. All patients fulfilled the criteria for clinically established PD.[21] The study was approved by the local medical ethics committee and all patients gave written informed consent.

EEG acquisition, pre-processing and analysis

EEG acquisition and pre-processing has been described elsewhere.[17] Recordings were made with 21 Ag/AgCl EEG electrodes according to standard 10-20 positions. Patients used their medication according to their individual schedules. Data were re-referenced towards a source derivation approaching the surface Laplacian derivation [22] to amplify spatial resolution.[23] After visual confirmation of artefact-free signals, five consecutive non-overlapping 4096-point epochs were selected for offline analysis in American Standard Code for Information Interchange (ASCII) format. Recordings with less than five epochs were excluded from analyses. Brainwave software was used for computation of clinically used peak frequencies ((BrainWave version 0.9.152.12.26, C.J. Stam; available at http://home.kpn.nl/stam7883/brainwave.html).

Group composition

From the comprehensive neuropsychological evaluations, six neuropsychological domains were identified according to the Diagnostic and Statistical Manual of mental disorders (5th edition, DSM-V).[24] According to DSM-V consensus guidelines, the following cognitive tests were selected for each domain: (1) 'Learning and Memory': Cambridge Cognitive Examination (CAMCOG) memory section,[25] Rey Auditory Verbal Learning Test (RAVLT),[26] and Wechsler Memory Scale (WMS);[27] (2) 'Executive Functioning': CAMCOG abstract reasoning, Digit Cancellation Test (DCT),[28] digit span,[29] Word-colour Stroop Test (Stroop) 3,[30] Trail Making Test (TMT) B;[31] (3) 'Psychomotor speed': Stroop 1 and 2, and TMT A; (4) 'Language': CAMCOG language section and verbal fluency; (5) 'Perceptive-motoric functioning': CAMCOG perception and CAMCOG praxis, and (6) 'Neuropsychiatric status': Becks Depression Inventory (BDI)[32] and Hospital Anxiety and Depression Scale (HADS) A-D.[33] All individual test-scores were standardised (Z-transformed) and averaged per domain for direct comparability. In case of missing data, an average of the remaining test-scores within the pertaining domain was used rather than imputing data, as long as ≥ 2 test-scores remained per domain (except for the domain 'Language' which contains only two tests and for which no data was imputed). A composite Z-score was derived from averaging all domains, if data from ≥ 4 domains were available. Higher Z-scores indicate better cognitive functioning. From the entire dataset, the most extreme patients in terms of cognitive performance were selected: either the highest cognitive composite scores (high-COG, n=20) or the lowest scores (low-COG, n=20). All other patients were classified as 'intermediate cognitive performance (int-COG). Given the nature of the cohort (i.e. DBS candidates who had already underwent a clinical pre-screening),[10] it was deemed unlikely that a sufficient number of patients would fulfil the criteria for either PD Dementia (PDD) or Mild Cognitive Impairment (MCI) and these classes were therefore deemed unsuitable to use for classification purposes.

Secondary outcomes included: motor function (Movement Disorders Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) part III (range 0-132)),[34] and non-dopaminergic functioning (SEverity of Non-dopaminergic Symptoms in Parkinson's Disease (SENS-PD) scale (range 0-54)),[35] and level II criteria for PD-MCI.[36]

ML Pipeline

A previously reported ML pipeline approach was used for time series classification purposes.[37, 38] Originally developed and applied in the automotive industry to classify time series originating from vehicle-data (i.e. predicting damaged parts after a low-speed crash[37, 38]), the approach was further applied to time series originating from EEGs, particularly to evaluate different ML approaches for classification of PD patients according to their cognitive performance.[39] The resulting ML pipeline consists of four phases: (1) feature-extraction,

(2) feature-selection, (3) training of a classifier, and (4) hyperparameter optimization. All four steps are completely automated, with the EEG time series as input and the class-labels (i.e. high-COG or low-COG) as output. The library 'Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests' (tsfresh) was used to extract features from the time series,[40, 41] resulting in 16674 features per EEG (794 comprehensive features for each of the 21 time series).[42] Feature selection was performed using the Boruta algorithm, by testing the variable importance (VIMP) of each feature against that of 'shadow features', which are created by random shuffling of the real features. The VIMP of shadow and real features are obtained from a random forest model trained thereon. A real feature would be selected if its VIMP frequently dominates the maximal VIMP of shadow features, in multiple independent trials.[42] After feature-selection, this feature set is used to train a Random Forest Classifier (RFC). A RFC is an ensemble of decision trees; the resulting decision is the majority vote from all decision trees.[43] The hyperparameters of the RFC, such as the number of decision trees and their individual tree depths, are optimized with a variant of Bayesian Optimization technique called Mixed Integer Parallel Efficient Global Optimization (MIP-EGO)[44, 45] for mixed-integer categorical search spaces.[46] To ensure generalizability of the RFC, a cross-validation procedure was adopted: the data is randomly split into 10 folds, after which training was performed on 9 folds and tested on the remaining fold. This process was repeated until each fold has served as test set; the average of all test scores of the computations represents the final score. A secondary assessment of interval validity was based on a combination of cross-validation and split-sample validation: cross-validated model-training based on 50% of the data and validated on the remaining sample. This approach was repeated for 60-90% of the data used for model-building with the remaining sample used for internal validation purposes, although it should be noted that cross-validation is superior to split-sample validation to assess internal validity especially for small sample sizes.[47] A detailed description of the applied ML Pipeline is published elsewhere.[39] Since all four steps are fully automated, no arbitrary choices on feature-extraction or feature-selection were made during the model-building-process.

Application of the pipeline to EEG data

Both occipital and global peak frequencies, routinely used for clinical purposes, were used as standard-features. All five epochs were averaged per patient, in order to obtain more robust time series and to limit intra-individual variability.[39] The features from each individual computation-run were selected and combined. The resulting model with the combined features was evaluated for model performance. A comparison was drawn between a model using only the occipital peak frequency as a single classifying feature and the ML Pipeline using a combination of the routinely-used peak frequency and the automatically extracted features from the EEG time series.

The final selected model with the best-classifying performance was then applied to the unclassified patients (i.e. those with 'intermediate' cognitive performance scores) and the predicted probabilities of being classified as low-COG were calculated for all patients. A linear regression model was fitted with these predicted probabilities as an outcome, and the composite global cognitive score subdivided into three splines in accordance with the original cognitive classification as independent variables.

Statistical analysis

Demographic, clinical, and neuropsychological variables, as well as electrophysiological spectral features, were compared between the high-COG and low-COG groups using Student T-tests if normally distributed, and Mann-Whitney U tests if not-normally distributed in case of continuous variables, and Pearson's χ^2 Tests in case of categorical data. The ML Pipeline, as well as a model using only occipital peak frequency as classifying feature, was evaluated using accuracy, sensitivity, and specificity metrics.

Missing values, other than cognitive performance scores, were imputed using multiple imputation with five iterations in case of ≤15% missing data.
All analyses were performed using IBM Statistical Package for the Social Sciences (SPSS) 25 Software (SPSS inc., Chicago, Illinois, USA).

Data availability

Anonymized data may be shared upon request.

Results

Patient characteristics

A total of 112 patients were included. Patients classified as high-COG were younger, and with a younger age-at-onset than low-COG patients. Non-dopaminergic disease severity, as well as motor functioning during 'ON' was better in high-COG patients, whereas motor functioning during 'OFF' did not differ (see table 8.1). Composite cognitive Z scores were inherently different between the high-COG and low-COG groups with approximately 1.5 standard deviations (SD) difference (mean (SD) 0.78 (0.57) vs. -0.78 (0.54), respectively). High-COG patients had similarly better scores for the domains 'Learning and Memory', 'Perceptive-motoric functioning', 'Executive functioning', and 'Language'. Strikingly, scores for the domains 'Neuropsychiatric functioning' and 'Psychomotoric speed' were lower for the high-COG patients than for the low-COG patients.

Table 8.1. Demographic and clinical characteristics

	High-COG	Low-COG	P *	Int-COG
N	20	20		72
Age [a]	59.5 (54.6 – 66.4)	67.8 (60.1 – 72.1)	0.004	63.5 (57.7 – 68.0)
Age at onset [b]	48.2 (9.3)	55.4 (9.6)	0.023	51.1 (10.7)
% Female (n) [c]	45 (9)	10 (2)	0.031	37.5 (27)
MDS-UPDRS III 'ON' [a]	18.5 (11 – 22.5)	23 (19 – 36)	0.012	20.5 (13.3 – 30)
MDS-UPDRS III 'OFF' [a]	46.5 (39.3 – 55.5)	48.5 (41 – 57)	0.718	44 (36 – 55)
SENS-PD [b]	9.2 (4.0)	15.3 (4.8)	<0.001	12.4 (4.8)
Z Psychomotoric speed [a]	-0.71 (-0.97 - -0.38)	0.55 (-0.27 - 1.30)	<0.001	-0.23 (-0.60 – 0.18)
Z Language [a]	0.88 (0.50 - 1.24)	-0.93 (-2.11 - - 0.45)	<0.001	0.04 (-0.35 – 0.53)
Z Neuropsychiatric functioning [a]	-0.40 (-0.78 – 0.28)	0.16 (-0.39 – 0.41)	0.108	-0.12 (-0.42 – 0.37)
Z Executive functioning [a]	0.59 (0.28 – 0.74)	-0.71 (-1.64 - -0.35)	<0.001	0.08 (-0.23 – 0.40)
Z Perceptive-motoric functioning [a]	0.40 (0.40 – 0.76)	-1.35 (-1.61 - -0.63)	<0.001	0.40 (-0.06 – 0.76)
Z Learning and Memory [a]	0.92 (0.34 – 1.07)	-0.79 (-1.83 - -0.32)	<0.001	0.06 (-0.28 – 0.50)
Z Global Cognition [b]	0.78 (0.57)	-0.78 (0.54)	<0.001	0.09 (0.22)
% PD-MCI (≥2 domains ≤ -1.5 SD) (n)	0	30 (6)		0
% PD-MCI ((≥2 domains (-1, -1.5) SD (n)	0	15 (3)		3 (2)

* High-COG (20 patients with highest cognitive scores) vs. Low-COG (20 patients with lowest cognitive scores)
Int-COG = all patients with intermediate cognitive scores
[a] Mann Whitney U tests (median (interquartile range)); [b] Student T tests (mean (standard deviation)); [c] Pearson $\chi 2$ tests
MDS-UPDRS III: Movement Disorders Society – Unified Parkinson's Disease Rating Scale III; SENS-PD: SEverity of Non-dopaminergic Symptoms in Parkinson's Disease

High-COG patients had spectrally faster EEGs than low-COG patients, demonstrated by particularly higher occipital peak frequencies (mean (SD) 9.0 (0.9) vs. 7.8 (1.4) Hz) and lower ratios of slow-over-fast relative powers $((\delta + \theta) / (\alpha 1 + \alpha 2 + \beta))$ (median (interquartile range) 0.69 (0.49 – 0.86) vs. 1.21 (0.57 – 2.20) (table 8.2 and figure 8.1).

Table 8.2 EEG spectral characteristics

	High-COG	Low-COG	P *	Int-COG
Occipital peak frequency [a]	9.0 (0.9)	7.8 (1.4)	0.003	8.4 (1.4)
Total peak frequency [a]	8.8 (0.8)	7.9 (1.4)	0.013	8.2 (1.1)
Relative δ power [b]	0.21 (0.18 – 0.27)	0.24 (0.17 – 0.39)	0.369	0.26 (0.20 – 0.35)
Relative θ power [b]	0.15 (0.11 – 0.20)	0.20 (0.13 – 0.31)	0.068	0.17 (0.12 – 0.26)
Relative α1 power b	0.23 (0.16 – 0.30)	0.16 (0.07 – 0.22)	0.024	0.14 (0.09 – 0.21)
Relative α2 power [b]	0.11 (0.09 – 0.17)	0.07 (0.06 – 0.11)	0.008	0.09 (0.06 – 0.13)
Relative β power [b]	0.19 (0.16 – 0.25)	0.16 (0.12 – 0.23)	0.327	0.19 (0.15 – 0.25)
Slowing ratio $((\delta + \theta) / (\alpha 1 + \alpha 2 + \beta))$ b	0.69 (0.49 – 0.86)	1.21 (0.57 – 2.20)	0.026	1.07 (0.59 – 1.43)

* High-COG vs. Low-COG
[a] Student T-test (mean (standard deviation)); [b] Mann Whitney U test (median (interquartile range))

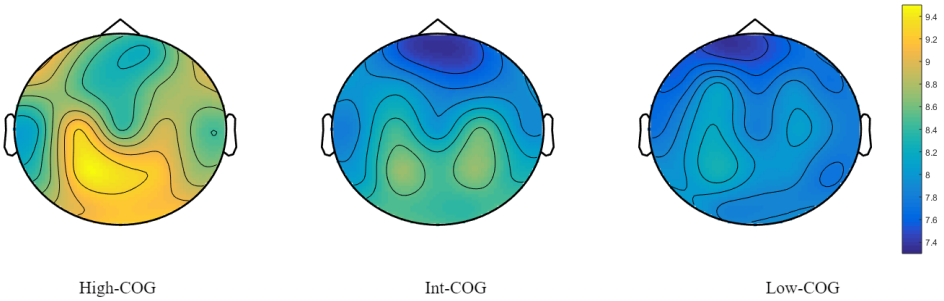


Figure 8.1 Spectral plots (peak-frequency) per cognitive class
Peak frequencies were calculated in Hz. Patients with high cognitive performance scores (high-COG) have spectrally faster EEGs than patients with lower cognitive performance scores (low-COG).

Patients classified as int-COG had clinical, cognitive, and spectral scores situated between low-COG and high-COG scores, respectively.

ML Pipeline performance

The accuracy (mean (SD)) of the average of all individual runs of the pipeline was 0.81 (0.01). After a secondary series of cross-validation runs incorporating all features from the individual runs, the extended model performance increased to 0.92 (0.02). Using only the occipital peak frequency as a classifying feature, the accuracy was lower: 0.67 (0.06) (see table 8.3). The list of features (n=13) selected by the ML pipeline included the clinically used 'occipital peak frequency'. All features were in a VIMP range of 4-15% (see supplementary table 8.1). A combination of cross-validation and split-sample validation demonstrated good internal validity for all splits (see supplementary figure 8.1).

Table 8.3 Machine learning model performances

	Occipital peak frequency only	Mean of all individual cross-validation runs	All features from all cross-validation runs
Accuracy	0.67 (0.06)	0.81 (0.01)	0.92 (0.02)
Sensitivity	0.74 (0.09)	0.82 (0.04)	0.90 (0.04)
Specificity	0.59 (0.04)	0.83 (0.07)	0.94 (0.02)

Data expressed as mean (standard deviation)

Calibration

A scatterplot demonstrating the correlation between actual cognitive functioning and the predicted probability of being classified as low-COG is shown in figure 8.2, demonstrating a negative trend (i.e. a lower probability of being classified as low-COG correlates to better

cognition: β = -0.23 (95%CI -0.29 - -0.18)). Both the high-COG and the low-COG groups contributed to this negative trend (spline-high-COG: β = -0.289 (95% CI -0.37 - -0.20), spline-low-COG: β = -0.26 (95%CI -0.34 - -0.17)), but the int-COG patients, who were not used during model-training, did not (spline-int-COG: β = 0.12 (95%CI -0.05 – 0.30).

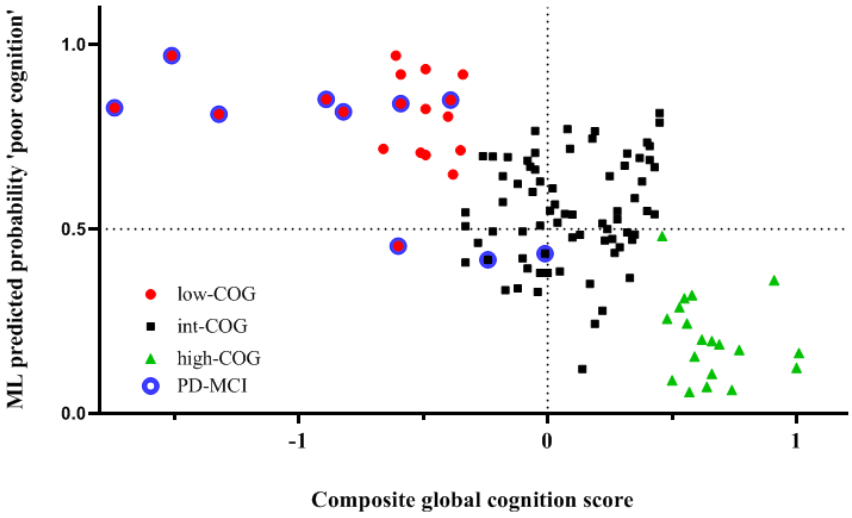


Figure 8.2 Predicted probability of being classified 'low-COG' vs. actual cognitive performance

Discussion

In this study, we show that DBS candidates with PD with either clinically determined 'good' or 'poor' cognition may be classified according to their cognitive function based on a fully automated EEG-assessment.

Contrary to previous studies which highlight singular, or few features to distinguish patients with different levels of cognitive impairment,[15-17, 19, 48] we showed that a compound of multiple EEG-biomarkers provides the highest accuracy in classifying patients.
Our final model performs slightly better than previously reported ML algorithms, which report accuracies between 74% [16] and 88%.[19] Betrouni and colleagues differentiated five groups of PD patients, with different levels of cognitive impairment using support vector machines (accuracy=84%) and k-nearest neighbour models (88%).[19] Although different electrode-densities were used, analyses were limited to spectral features in an effort to prevent

overfitting. As the dataset was subdivided into five different categories based on cognitive clusters, the two groups with worst cognitive function were smallest, containing respectively five and nine patients. In contrast, the results described above demonstrate the advantage of automated feature-extraction and simultaneous analysis to both increase the accuracy and limit the need for laborious pre-processing. Pragmatically, the use of spectral features to reflect EEG slowing is currently still easier to translate to routine clinical practice than applying a ML pipeline to new EEG data, although less accurate. Another study added connectivity metrics, i.e. Phase-Lag-Index (PLI) to spectral features resulting in 396 features (66 spectral- and 330 PLI features).[16] Although the reported accuracies were lower, PLI features discriminated better between PD patients with or without MCI (spectral features: Area-under-the-curve (AUC) = 0.64; PLI features: AUC = 0.74). Our model does not include between-channel connectivity metrics but rather focuses on synchronization patterns within one individual time series. Our accuracy may yet be further increased by including connectivity- or network features. However, the amount of computation runtime increases exponentially when automated models are expanded in such way.[16] Given that the number of features reflecting between-channel connectivity extends several folds beyond the feature-selection delineated here, the computation runtime may become too protracted for practical purposes.[49]

Although the ML pipeline treats all patients within one subgroup equally, despite within-group differences in cognitive functioning, the association between the predicted class-probability and actual cognitive performance follows a linear correlation. This trend is predominantly fuelled by the patients on which the model was trained, i.e. high-COG and low-COG patients. Patients classified as int-COG were poorly predicted and no linear trend could be discerned for this subgroup. The final model including all features from the separate cross-validation was inherently not based on 'unseen data' and therefore runs the risk of overfitting, despite several safeguards such as multiple cross-validation runs and Bayesian hyperparameter optimization. This was unavoidable given the small sample size, and the accuracies from the final model are therefore best interpreted as the best approximated maximum, with accuracies from the averaged cross-validation runs as minimum. The risk of overfitting may also partly explain why the model-performance in the int-COG group was ineffective. Other explanations include the limited variability in the int-COG group (by definition, all patients had cognitive scores within 1.5 SD) and variation in cognitive performance within this limited range is likely to occur regardless of the degree of cortical PD pathology and reflect normal variation also found in the otherwise 'normal' population. Furthermore patients with an 'intermediate' cognition were never included during the initial-model building and therefore constitute a separate class which is unrecognized by the model.

In contrast to previous studies that explored a wide range of cognitive functioning in PD patients, our results focus on PD patients undergoing the screening procedure for DBS. DBS candidates often have a relatively longer disease duration to allow for several treatment options before considering DBS surgery and often have more severe PD symptoms than newly-diagnosed PD patients. Furthermore, severe cognitive impairment is a contraindication for DBS [9, 10] and patients with obvious cognitive deficits will not be referred for screening, indicating that the range of cognitive function is likely much smaller in the DBS population than in the global PD population, emphasizing the sensitivity of this ML pipeline.

As with all supervised learning models, the crucial determinant of the models' validity is the correct labelling (either high-COG or low-COG, or another arbitrarily defined label). In this study, an extensive neuropsychological test battery was used to determine cognitive functioning of six consensus-based domains,[24] and a derived composite score reflecting global cognition. However, cognitive (dys)function is not a purely binary classification: performance is rated in a spectrum of possible scores and a derived binary classification may be subject to discussion. In this study, classes of cognitive functioning were determined in a data-driven fashion by taking the twenty best- and worst performing patients from the entire cohort. This was an a priori defined classification, as it was deemed unlikely that there would be sufficient DBS candidates with either MCI or PDD. However, it should be noted that both a classification based on the neuropsychological test battery, and cognitive-screening-tests reported previously,[39] yielded similar model performances suggesting high accuracy regardless of the exact tests used for cognitive profiling.

Our results therefore indicate the utility of using qEEG as complementary biomarker to assess cognitive function, but do not provide an answer towards the pathophysiological mechanism underlying cognitive deficits. We speculate that higher-density source-space setups may provide a better indication of such an underlying mechanism, possibly using Magnetoencephalography (MEG) to better reflect subcortical structures.[18] However, such an approach would have lower practical utility as it would be more difficult to implement high-density EEG or MEG in routine clinical practice. Nevertheless, this study demonstrates the cortical spatial expansion of the mechanism underlying cognitive impairment.

The ultimate ground truth in terms of clinical impact would be a classification based on long-term postoperative cognitive functioning. This data is however not available, whereas patients with poor preoperative functioning, as identified by the neuropsychological test battery, may be rejected for DBS surgery after screening and thus not contribute to follow-up data.

Strengths of our study include the automated ML pipeline which circumvents making arbitrary choices on pre-processing and feature selection, the large number of extracted features, and extensive cognitive profiling on which the initial classification was based. The use of cross-validation warrants the internal validity of our model. To our knowledge, ours is the only cohort of consecutively included DBS candidates with PD with EEG data available in the literature. Given the uniqueness of our cohort, no external validity can therefore be assessed. Despite multiple cross-validation runs, the algorithm was trained on only 20 vs. 20 patients. This constitutes a small sample size to base definitive conclusions on and requires validation in a larger cohort. Nevertheless, our results clearly demonstrate the utility of qEEG during the DBS screening for automated cognitive profiling and the superiority of a compound of EEG features over a single spectral feature.

The classification was based on the most extreme patients with composite scores of six Z-transformed domains. The domains 'Neuropsychiatric functioning' and 'Psychomotoric speed' were paradoxically worse in patients classified as high-COG than low-COG. Also, high-COG patients were younger and had less severe motor- and non-dopaminergic symptoms. However, these factors do not constitute a contra-indication for surgery.

Future studies may confirm the external validity of our model within the population of DBS candidates and evaluate the use of such a ML pipeline on other neurodegenerative diseases with cognitive impairment such as Alzheimer's Disease of Dementia with Lewy Bodies.[50] In such a way, it could be determined whether biomarkers differentiating cognitive subtypes are disease-specific (i.e. different biomarkers for different diseases), or whether there is a neurophysiological compound underlying cognitive impairment across neurodegenerative diseases. Furthermore, the ultimate goal of the ML pipeline would be to determine its utility as a predictor of cognitive deterioration rather than cross-sectional classification of cognitive functioning.

Strikingly, the model proposed here was originally developed for the automotive industry and applied here to a vastly different research field. This suggests that the origin of the time series, i.e. whether a signal originates from an EEG or from a vehicle, is not important during analyses. We speculate that multidisciplinary approaches such as these may advance healthcare-research and valorise these higher-order analysis-techniques through applications in fundamentally different fields.

We emphasize that currently, the EEG analyses described here are not intended to replace the neuropsychological assessments during the DBS screening and should be seen as complementary. However, these results provide strong evidence of the utility of qEEG as a

biomarker for cognitive performance during the DBS screening and may have potential both in clinical practice and for future clinical trials studying disease modifying therapy.

Acknowledgements

The authors would like to thank the members of the DBS team of LUMC/Haga Teaching Hospital (G.E.L. Hendriks, A. Mosch, R. Zutt, C.F. Hoffmann, N.A. van der Gaag) for patient care and the EEG technicians of the LUMC for their help with the data acquisition.

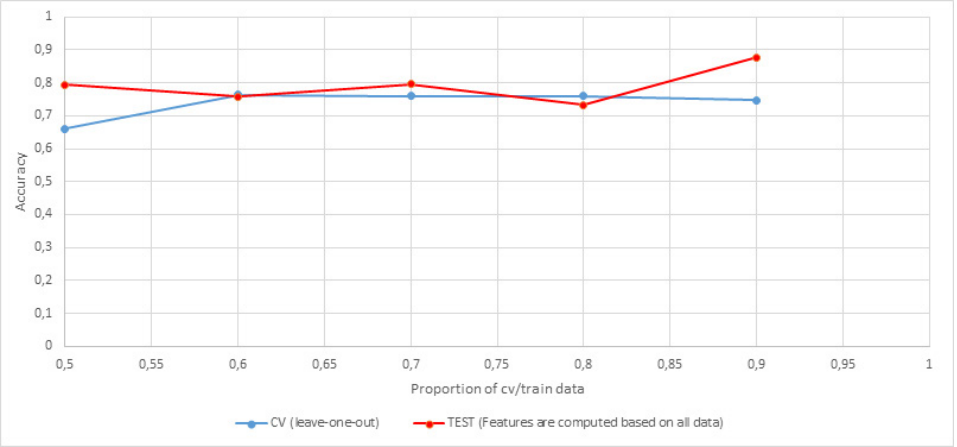
References

1. Collaborators GBDPsD. Global, regional, and national burden of Parkinson's disease, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 2018;17:939-953.
2. Ahlskog JE, Muenter MD. Frequency of levodopa-related dyskinesias and motor fluctuations as estimated from the cumulative literature. Mov Disord 2001;16:448-458.
3. Deuschl G, Agid Y. Subthalamic neurostimulation for Parkinson's disease with early fluctuations: balancing the risks and benefits. Lancet Neurol 2013;12:1025-1034.
4. Okun MS, Gallo BV, Mandybur G, et al. Subthalamic deep brain stimulation with a constant-current device in Parkinson's disease: an open-label randomised controlled trial. Lancet Neurol 2012;11:140-149.
5. Weaver FM, Follett K, Stern M, et al. Bilateral Deep Brain Stimulation vs Best Medical Therapy for Patients With Advanced Parkinson Disease A Randomized Controlled Trial. Jama-J Am Med Assoc 2009;301:63-73.
6. Contarino MF, Daniele A, Sibilia AH, et al. Cognitive outcome 5 years after bilateral chronic stimulation of subthalamic nucleus in patients with Parkinson's disease. J Neurol Neurosurg Psychiatry 2007;78:248-252.
7. Smeding HM, Speelman JD, Huizenga HM, Schuurman PR, Schmand B. Predictors of cognitive and psychosocial outcome after STN DBS in Parkinson's Disease. J Neurol Neurosurg Psychiatry 2011;82:754-760.
8. Drapier D, Drapier S, Sauleau P, et al. Does subthalamic nucleus stimulation induce apathy in Parkinson's disease? J Neurol 2006;253:1083-1091.
9. Lang AE, Houeto JL, Krack P, et al. Deep brain stimulation: preoperative issues. Mov Disord 2006;21 Suppl 14:S171-196.
10. Geraedts VJ, Kuijf ML, van Hilten JJ, Marinus J, Oosterloo M, Contarino MF. Selecting candidates for Deep Brain Stimulation in Parkinson's disease: the role of patients' expectations. Parkinsonism Relat Disord 2019.
11. Duncan JS. Conventional and clinimetric approahces to individualization of antiepileptic drug therapy. In: Meinardi HC, J. A.; Baker, G. A.; da Silva A. M., ed. Quantitative assessment in epilepsy care. Porto, Portugal: Springer Science+Business Media, LLC, 1993.
12. Duckworth AL, Quinn PD, Lynam DR, Loeber R, Stouthamer-Loeber M. Role of test motivation in intelligence testing. Proceedings of the National Academy of Sciences of the United States of America 2011;108:7716-7720.
13. Duckworth AL, Yeager DS. Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. Educ Res 2015;44:237-251.
14. Geraedts VJ, Boon LI, Marinus J, et al. Clinical correlates of quantitative EEG in Parkinson disease: A systematic review. Neurology 2018;91:871-883.
15. Utianski RL, Caviness JN, van Straaten ECW, et al. Graph theory network function in parkinson's disease assessed with electroencephalography. Clinical Neurophysiology 2016;127:2228-2236.
16. Chaturvedi M, Bogaarts JG, Kozak Cozac VV, et al. Phase lag index and spectral power as QEEG features for identification of patients with mild cognitive impairment in Parkinson's disease. Clin Neurophysiol 2019;130:1937-1944.
17. Geraedts VJ, Marinus J, Gouw AA, et al. Quantitative EEG reflects non-dopaminergic disease severity in Parkinson's disease. Clin Neurophysiol 2018;129:1748-1755.

18. Bonanni L. The democratic aspect of machine learning: Limitations and opportunities for Parkinson's disease. Mov Disord 2019;34:164-166.
19. Betrouni N, Delval A, Chaton L, et al. Electroencephalography-based machine learning for cognitive profiling in Parkinson's disease: Preliminary results. Mov Disord 2019;34:210-217.
20. Geron A. Hands-on Machine Learning with Scikit-Learn & TensorFlow : Concepts, Tools, and Techniques to build Intelligent Systems2017.
21. Postuma RB, Berg D, Stern M, et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 2015;30:1591-1601.
22. Hjorth B. Source derivation simplifies topographical EEG interpretation. American Journal of EEG Technology 1980;20:121-132.
23. Burle B, Spieser L, Roger C, Casini L, Hasbroucq T, Vidal F. Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view. Int J Psychophysiol 2015;97:210-220.
24. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (5th ed. ed.). Arlington: VA: American Psychiatric Publishing, 2013.
25. Huppert FA, Brayne C, Gill C, Paykel ES, Beardsall L. CAMCOG--a concise neuropsychological test to assist dementia diagnosis: socio-demographic determinants in an elderly population sample. The British journal of clinical psychology 1995;34 (Pt 4):529-541.
26. Vakil E, Blachstein H. Rey Auditory-Verbal Learning Test: structure analysis. Journal of clinical psychology 1993;49:883-890.
27. Wechsler D. Wechsler memory scale. San Antonio, TX, US: Psychological Corporation, 1945.
28. Dekker R, Mulder JL, Dekker PH. De ontwikkeling van vijf nieuwe Nederlandstalige tests. Leiden: PITS, 2007.
29. Richardson JT. Measures of short-term memory: a historical review. Cortex; a journal devoted to the study of the nervous system and behavior 2007;43:635-650.
30. Scarpina F, Tagini S. The Stroop Color and Word Test. Front Psychol 2017;8:557-557.
31. Tombaugh TN. Trail Making Test A and B: Normative data stratified by age and education. Archives of Clinical Neuropsychology 2004;19:203-214.
32. Beck AT, Steer RA, Ball R, Ranieri W. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. Journal of personality assessment 1996;67:588-597.
33. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta psychiatrica Scandinavica 1983;67:361-370.
34. Goetz CG, Tilley BC, Shaftman SR, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008;23:2129-2170.
35. van der Heeden JF, Marinus J, Martinez-Martin P, van Hilten JJ. Evaluation of severity of predominantly non-dopaminergic symptoms in Parkinson's disease: The SENS-PD scale. Parkinsonism Relat Disord 2016;25:39-44.
36. Litvan I, Goldman JG, Troster AI, et al. Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. Mov Disord 2012;27:349-356.
37. Koch M, Bäck T. Machine Learning for Predicting the Impact Point of a Low Speed Vehicle Crash. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); 2018 17-20 Dec. 2018: 1432-1437.

38. Koch M, Wang H, Bäck T. Machine Learning for Predicting the Damaged Parts of a Low Speed Vehicle Crash. 13th International Conference on Digital Information Management 2018: 179-184.
39. Koch M, Geraedts V, Wang H, Tannemaat MR, Bäck T. Automated Machine Learning for EEG-Based Classification of Parkinson's Disease Patients. 2019 IEEE International Conference on Big Data. Los Angeles2019: 4845-4852.
40. Christ M, Kempa-Liehr AW, Feindt M. Distributed and parallel time series feature extraction for industrial big data applications. arXiv e-prints [serial online] 2016. Available at: https://ui.adsabs.harvard.edu/abs/2016arXiv1610077717C. Accessed October 01, 2016.
41. Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing 2018;307:72-77.
42. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. 2010 2010;36:13 %J Journal of Statistical Software.
43. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition: Springer New York, 2009.
44. Wang H, Emmerich M, Bäck T. Cooling Strategies for the Moment-Generating Function in Bayesian Global Optimization. 2018 IEEE Congress on Evolutionary Computation (CEC); 2018 8-13 July 2018: 1-8.
45. Wang H, Stein Bv, Emmerich M, Back T. A new acquisition function for Bayesian optimization based on the moment-generating function. 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2017 5-8 Oct. 2017: 507-512.
46. Yang K, Blom Kvd, Bäck T, Emmerich M. Towards single- and multiobjective Bayesian global optimization for mixed integer problems. 2019;2070:020044.
47. Steyerberg EW. Validation in prediction research: the waste by data splitting. Journal of clinical epidemiology 2018;103:131-133.
48. Klassen BT, Hentz JG, Shill HA, et al. Quantitative EEG as a predictive biomarker for Parkinson disease dementia. Neurology 2011;77:118-124.
49. García-Martín E, Rodrigues CF, Riley G, Grahn H. Estimation of energy consumption in machine learning. Journal of Parallel and Distributed Computing 2019;134:75-88.
50. Dauwan M, van der Zande JJ, van Dellen E, et al. Random forest to differentiate dementia with Lewy bodies from Alzheimer's disease. Alzheimers Dement (Amst) 2016;4:99-106.

Supplementary material



Supplementary Figure 8.1 Split-sample vs cross-validation

Supplementary Table 8.1 Model features

'T6-Cz__fft_coefficient__coeff_77__attr_"imag"'
'Pz-Cz__fft_aggregated__aggtype_"skew"'
'O2-Cz__fft_coefficient__coeff_77__attr_"imag"'
'O2-Cz__energy_ratio_by_chunks__num_segments_10__segment_focus_3'
'Pz-Cz__fft_coefficient__coeff_96__attr_"imag"'
'T3-Cz__fft_coefficient__coeff_98__attr_"abs"'
'Pz-Cz__fft_coefficient__coeff_59__attr_"angle"'
'Occipital peak frequency'
'P3-Cz__fft_coefficient__coeff_89__attr_"real"'
'O1-Cz__ar_coefficient__k_10__coeff_2'
'O1-Cz__fft_coefficient__coeff_55__coeff__attr_"angle"'
'T4-Cz__cwt_coefficients__widths_(2, 5, 10, 20)__coeff_14__w_10'
'Pz-Cz__fft_coefficient__coeff_68__attr_"imag"'

All featured were derived from the library 'Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests' (tsfresh) Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing 2018;307:72-77.