*Challenges of Trustable AI and Added-Value on Health*
*B. Séroussi et al. (Eds.)*

*doi:10.3233/SHTI220405*

# Utilizing a Non-Motor Symptoms Questionnaire and Machine Learning to Differentiate Movement Disorders

Alexander BRENNER[a], Lucas PLAGWITZ[a], Michael FUJARSKI[a], Tobias WARNECKE[b] and Julian VARGHESE[a]
[a] *Institute of Medical Informatics, University of Münster, Münster, Germany*
[b] *Department of Neurology and Neurorehabilitation, Klinikum Osnabrück – Academic teaching hospital of the University of Münster, Osnabrück, Germany*

**Abstract.** Parkinson's disease (PD) is a common neurodegenerative disorder that severely impacts quality of life as the condition progresses. Early diagnosis and treatment is important to reduce burden and costs. Here, we evaluate the diagnostic potential of the Non-Motor symptoms (NMS) questionnaire by the International Parkinson and Movement Disorder Society based on patient-completed answers from a large single-center prospective study. In this study data from 489 study participants consisting of a PD group, a healthy control (HC) group and patients with differential diagnosis (DD) have been recorded with a smartphone-based system. Evaluation of the study data has shown a significant difference in NMS between the representative groups. Cross-validation of Machine Learning based classification achieves balanced accuracy scores of 88.7% in PD vs. HC, 72.1% in PD vs. DD and 82.6% when discriminating between all movement disorders (PD + DD) and the HC group. The results indicate potentially high feature importance of a simple self-administered questionnaire that could support early diagnosis.

**Keywords.** Parkinson's Disease, movement disorders, artificial intelligence

## 1. Introduction

Worldwide prevalence of Parkinson's disease (PD) has more than doubled over the last two decades [1]. It affects movement with symptoms such as tremor, slowness and rigidity. However, non-motor symptoms are often reported as early indicators of the disease that can severely impact patient's quality of life [2]. Currently, PD is diagnosed primarily on the basis of clinical examination and nuclear imaging, along with other movement disorders. Approaching the need for comprehensive evaluation of mHealth-based biomarkers, the Smart Device System (SDS) has been implemented and applied in a three-year prospective study [3]. From 2018 to 2021, a total of 489 participants, including a broad range of different PD progress states according to Hoehn and Yahr [4] have been recorded. Each session comprises a series of sensor measures of movements, speech recordings, tablet-based tasks and smartphone-based questionnaires. With the recent conclusion of the aforementioned study, here, we focus on diagnostic potential of the Non-Motor symptoms (NMS) questionnaire by the International Parkinson and Movement Disorder Society as preliminary findings of the study.

While the NMS are routinely used to document early symptoms in PD, we elaborate on the diagnostic potential by including PD patients, healthy controls and other movement disorders as well. Our research question is whether we can reliably detect and discriminate PD from healthy controls and other movement disorders using the smartphone-based questionnaire data from our study dataset. We therefore evaluated Machine Learning models for three different classification tasks and pointed out possible important feature groups in the patient data. With the analysis of the questionnaire, which is a simple patient reported outcome, we have performed an important step in the evaluation of our study with the objective of finding digital biomarkers to aid PD diagnosis or screening in the future.

## 2. Methods

The study has been registered (ClinicalTrials.gov ID: NCT03638479) and approved by the ethical board of the University of Münster and the physician's chamber of Westphalia-Lippe (Reference number: 2018-328-f-S). Our study data consist of PD patients, a HC group and patients with DD. All diagnoses were confirmed by neurologists and reviewed by one senior movement disorder expert. The participant sample is summarized in Table 1. The first step of examination provides short questionnaires about medical history and non-motor symptoms (NMS) via a smartphone app. The medical history questionnaire holds information about age, height, weight, kinship with PD and effect of alcohol on tremor (further details provided in Varghese et al. [3]). The self-completed symptoms questionnaire is based on Chaudhuri et al. [5] and composed of 30 yes/no items. It represents different areas of possible PD symptoms.

First, we analyzed the NMS score, i.e., the sum of yes responses per participant, to investigate simple group separation. For comparability, we calculated statistical significance between the representative groups. In order to investigate the supportive potential in a clinical examination scenario, we trained and evaluated different Machine Learning algorithms via cross-validation. Given the questionnaire data, three relevant classification tasks were approached: (1) PD vs. HC, (2) Movement disorders (PD + DD) vs. HC and (3) PD vs. DD. To obtain stable results, a 5-fold cross-validation was randomized in 5 replicates, so that an average score over 25 test sets was produced. Three Machine Learning estimators were evaluated: (1) a multi-layer perceptron neural network (MLP), (2) a support-vector machine (SVM) and (3) CatBoost [6], which is based on gradient-boosted decision trees. All classifiers were trained with balanced weights, as class distributions are unequal in our dataset. Similarly, comparison of classification performance is based on balanced accuracy, precision and recall. Further details on hyper-parameter settings can be found in the supplement [7]. Based on the overall best performing classifier, feature importance was analyzed via permutation importance analysis using Scikit-Learn (0.24.2) [8]. Since NMS questions are grouped by relevant domains and thus certain questions are linked together, we modified the algorithm to consider and permute features group-wise. The selected feature-groups are based on the originally intended domains [5], details are provided in the supplements [7].

**Table 1.** Participant sample. Height in centimeters, weight in kilograms. Values correspond to mean (SD).

| Disease Class | #Samples | Age | Height | Weight | NMS score |
|---|---|---|---|---|---|
| Parkinson's Disease (PD) | 276 | 65.5 (9.6) | 170.9 (15.1) | 78.7 (16.8) | 9.93 (5.25) |
| Differential diagnosis (DD) | 124 | 60.2 (13.4) | 175.6 (9.8) | 83.6 (18.2) | 7.42 (4.92) |
| Healthy control (HC) | 89 | 60.2 (14.8) | 173.5 (8.6) | 79.0 (17.7) | 2.36 (2.69) |

## 3. Results

Table 1 shows the NMS score across the groups. On average NMS scores are lowest in the healthy group (mean 2.36, median 2), followed by the DD group (mean 7.42, median 7) and highest for PD patients (mean 9.93, median 9). In pairwise comparison the groups significantly differ from each other in NMS scores (Mann-Whitney, Kruskal-Wallis and t-test, $P < 0.0001$). Percentages of yes-answers per question are presented in the supplements [7].

To analyze the potential of the questionnaire data in supporting diagnosis, the previously described Machine Learning models were applied to the study data. Classifiers were cross-validated in different settings: (1) using only NMS score, (2) using all NMS items and (3) using the complete questionnaire (except medication). Complete results can be found in the supplements [7]. The overall best performing classifier scores are summarized in Table 2.

**Table 2.** Performance for the tasks (1) PD vs. HC, (2) Movement disorders (PD + DD) vs. HC and (3) PD vs. DD. All results are based on CatBoost classifier. Values correspond to mean (SD). *Balanced Accuracy.

| Task | Input | Accuracy | B. Accuracy* | Precision | Recall |
|---|---|---|---|---|---|
| PD vs. HC | NMS score | 0.817 (0.036) | 0.849 (0.047) | 0.965 (0.032) | 0.787 (0.039) |
| | NMS quest. | 0.886 (0.049) | 0.887 (0.045) | 0.961 (0.025) | 0.885 (0.048) |
| | Full quest. | 0.888 (0.041) | 0.873 (0.047) | 0.947 (0.025) | 0.902 (0.046) |
| PD + DD vs. HC | NMS score | 0.755 (0.040) | 0.816 (0.039) | 0.974 (0.019) | 0.720 (0.050) |
| | NMS quest. | 0.831 (0.036) | 0.810 (0.042) | 0.945 (0.018) | 0.843 (0.044) |
| | Full quest. | 0.860 (0.032) | 0.826 (0.057) | 0.946 (0.024) | 0.879 (0.032) |
| PD vs. DD | NMS score | 0.623 (0.043) | 0.610 (0.037) | 0.772 (0.029) | 0.646 (0.075) |
| | NMS quest. | 0.740 (0.044) | 0.702 (0.054) | 0.820 (0.039) | 0.800 (0.052) |
| | Full quest. | 0.761 (0.040) | 0.721 (0.049) | 0.828 (0.034) | 0.826 (0.045) |

Using the complete questionnaire information, grouped feature importance analysis was performed for task 1, PD vs. HC. The computed information gain per feature-group is depicted in Figure 1.

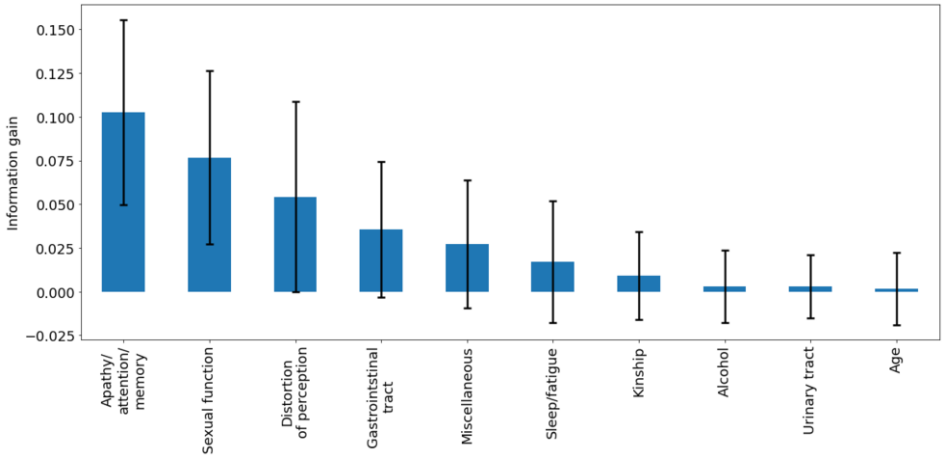


**Figure 1.** Grouped feature importance analysis of the NMS questionnaire. Bar height corresponds to mean information gain (± SD). Top 10 feature groups are displayed.

## 4. Discussion

We have evaluated the diagnostic accuracy of self-completed questionnaire data from our recently closed single-center prospective study for the patient groups PD, DD and HC. By testing different classification models and subsets of questions, we explored a robust set of features for disease prediction. For PD vs. HC the highest balanced accuracy score of 88.7% is achieved using NMS questions without additional information. Full questionnaire data resulted in a score of 87.3%. Given that additional data does not improve classification performance, we believe that the NMS questionnaire itself provides highest information gain in identifying PD patients. Our observations are similar in the task PD + DD vs. HC. Here, balanced classification accuracy for the NMS questionnaire, 81.0%, and the full questionnaire, 82.6%, are almost equal. It also shows that discrimination between healthy samples and PD is more accurate than discrimination between the healthy group and all movement disorders. This is to be expected as the questionnaire has been design to cover PD specific NMS. While the most challenging task - PD vs. DD - achieves lowest balanced accuracy overall, the highest score of 72.1% using the full questionnaire still discriminates both groups reasonably well. Here, the score for just the NMS questionnaire is slightly lower with 70.2%. It is noteworthy that even the aggregated NMS score already can reasonably divide the target groups, especially PD vs. HC with 84.9% balanced accuracy. Nonetheless, how PD affects a person can highly vary from patient to patient. Finding specific symptoms and subgroups can influence diagnosis and consequently treatment.

In our feature importance analysis for PD vs. HC, the top three groups in terms of information gain are apathy/attention/memory, sexual function and distortion of perception. This observation partially confirms the findings of previous studies [5,9]. In contrast, comparing to a comprehensive Italian study with a structured interview and an extended NMS questionnaire, fatigue and anxiety have shown to be more prevalent in PD patients [10]. Differences can arise from various causes, i.e. differences in answering (self-completed or structured), differences between study groups, while also some NMS generally become more common with higher age [2].

A limitation of our study is that it is conducted at a single site and thus bias due to specific site population, examination setting or target diagnoses can not be ruled out. Still, we believe the study data provides a suitable representation, as participants have been recruited from a large tertiary care center and include various types and stages of PD and other movement disorders. An important issue in Machine Learning remains the lack of interpretability and transparency when using complex models or feature computations. As a result, there is often mistrust in using models as decision support, even if they are able to make accurate predictions when measured against the gold standard. In our case, we worked with structured tabular data so that input features remained interpretable. Further, the feature importance analysis gave us insights about what questions were relevant for the decision process. We have demonstrated that the NMS yes/no answers were sufficient to classify PD from healthy controls with 88.7% accuracy. Overall it can be concluded, that a simple symptoms questionnaire could be of high importance for early patient screening. With the inclusion of additional sensor-based modalities from the study we will further continue the analysis.

## 5. Conclusion

A simple symptoms questionnaire and supervised Machine Learning provide high classification accuracy in the domain of Movement Disorders, in particular Parkinson's disease. While this questionnaire is known in practice to document non motor symptoms of Parkinson's disease we show its diagnostic potential against healthy participants and even differential diagnoses via Machine Learning. It could pave the way for early patient screening utilizing mHealth technologies. Continuing the work, we will further analyze the study data including further sensor-based modalities by means of dimensionality reduction and clustering techniques in order to find possible subgroups of the disease.

## References

[1] Rocca WA. The burden of Parkinson's disease: a worldwide perspective. The Lancet Neurology. 2018;17(11):928-9.
[2] Pfeiffer RF. Non-motor symptoms in Parkinson's disease. Parkinsonism & related disorders. 2016;22:S119-22.
[3] Varghese J, Niewöhner S, Soto-Rey I, Schipmann-Miletić S, Warneke N, Warnecke T, et al. A smart device system to identify new phenotypical characteristics in movement disorders. Frontiers in neurology. 2019;10:48.
[4] Bhidayasiri R, Tarsy D. Parkinson's disease: Hoehn and Yahr scale. In: Movement Disorders: A Video Atlas. Springer; 2012. p. 4-5.
[5] Chaudhuri KR, Martinez-Martin P, Schapira AH, Stocchi F, Sethi K, Odin P, et al. International multicenter pilot study of the first comprehensive self-completed nonmotor symptoms questionnaire for Parkinson's disease: the NMSQuest study. Movement disorders: official journal of the Movement Disorder Society. 2006;21(7):916-23.
[6] Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:181011363. 2018.
[7] Brenner A, Plagwitz L, Fujarski M, Warnecke T, Varghese J. Online supplement SDS MIE 2022; 2022. [cited 2022-01-23]. Available from: https://uni-muenster.sciebo.de/s/XIGIMyI1gLMsEbY.
[8] Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning; 2013. p. 108-22.
[9] Khoo TK, Yarnall AJ, Duncan GW, Coleman S, O'Brien JT, Brooks DJ, et al. The spectrum of nonmotor symptoms in early Parkinson disease. Neurology. 2013;80(3):276-81.
[10] Barone P, Antonini A, Colosimo C, Marconi R, Morgante L, Avarello TP, et al. The PRIAMO study: a multicenter assessment of nonmotor symptoms and their impact on quality of life in Parkinson's disease. Movement disorders: official journal of the Movement Disorder Society. 2009;24(11):1641-9.