Machine-learning models for MDS-UPDRS III Prediction: A comparative study of features, models, and data sources

Vitor Lobo[1], Diogo Branco[1], Tiago Guerreiro[1], Raquel Bouça-Machado[2,3], Joaquim Ferreira[2,3,4] and CNS Physiotherapy Study Group[2]
[1]LASIGE, Faculdade de Ciências, Universidade de Lisboa
[2]CNS—Campus Neurológico, [3]Instituto de Medicina Molecular João Lobo Antunes,
[4]Faculdade de Medicina, Universidade de Lisboa
vitormarqueslobo@gmail.com;djbranco@fc.ul.pt;tjvg@di.fc.ul.pt;raquelbouca@gmail.com;jferreira@medicina.ulisboa.pt

ABSTRACT

Parkinson's disease is the second most common neurodegenerative disease worldwide. Symptoms tend to fluctuate during the day and through disease progression. Clinical evaluations tend to occur spaced in time. Further, the assessments used are mostly subjective. The gold standard for evaluating disease severity is MDS-UPDRS. The increase in sensor usage enabled objective evaluation and continuous monitoring of the disease fluctuations. One of the symptoms that most affect mobility are gait disorders. The use of gait characteristics started to become popular to monitor the disease. However, the approaches used lack in-depth knowledge of machine learning models for disease staging. In our work, we try to estimate the MDS-UPDRS part III score from accelerometer data. We collected data from 74 patients using the Axitvity AX3 device both on the wrist and lower back. We did experiments with different models, features, and windows size. We achieved a 4.26 Mean Absolute Error on the on left out 10% data using both devices with a 2.5-second sliding window and a random forest model for prediction. We contribute with a comparison of the performed experiments and provide, according to our experiments, the optimal models for MDS-UPDRS part III estimation using only accelerometer data.

KEYWORDS

gait, accelerometer, mds-updrs, Parkinson's disease, features, machine learning, models

1 INTRODUCTION

Parkinson's Disease (PD) is a neurodegenerative disease that affects around 1% of the world's population. This disease is characterized by motor and non-motor symptoms [15]. Motor symptoms include bradykinesia, tremor, rigidity, and gait impairment. These are present in the early stages of the disease and worsen as the disease progresses.

Although there is no cure, the available pharmacological and non-pharmacological therapeutic interventions effectively control symptoms. However, as the disease progresses their efficacy tends to reduce and motor complications, such as motor fluctuations and dyskinesia, appear [11]. These have been labeled as 'ON' and 'OFF' stages [4]. To minimize the impact of these fluctuations and inform better the clinicians there is the need to periodically assess the symptoms. Generally, these evaluations require a visit to a clinic or hospital. Clinicians use validated assessments for PD to characterize a patient's current disease stage [9]. These assessments occur spaced in time and can be hard to capture all the fluctuations that may have happened between appointments. Further, instruments used in clinical practice focus on subjective evaluations. Namely, visual assessments during clinical visits that are supported by clinical scales.

The gold standard for evaluating disease severity in PD is the Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS). This is a comprehensive rating scale that assesses both motor and non-motor symptoms associated with Parkinson's [7]. To optimize disease management, close monitoring of symptom fluctuations is crucial. However, today this monitoring is usually performed through medical appointments, every six months, with a mean duration of 30 minutes. Additionally, what published evidence suggests is that patients perform differently during these moments, providing only information about their best capacity, rather than their usual performance in their daily lives.

The democratization of sensors' usage, namely the body-worn devices, that measure acceleration, and angular velocity allowed the increase of objective evaluations [10]. These devices passively monitor patients during clinical evaluation and in free-living environments. Furthermore, allow movement metrics and feature extraction that can be related to motor symptoms or clinical scales used for disease assessments [6]. Gait disorders are one of the symptoms that most affect mobility. Inertial measuring units can help to identify fluctuations. There have been studies that leverage the identification of walking bouts to extract gait metrics like step length or step variability [1, 4].

Research using these gait characteristics as a marker for PD has demonstrated the potential for monitoring the disease in several ways [2]. While the use of these gait characteristics has become a popular approach for monitoring PD, novel research has started to analyze signal processing metrics that could also be of use for this purpose. In a 2019 study, the contributions of signal-based features and gait characteristics for the classification of PD were analyzed [13]. Another emerging method to stage PD is the use of total scores of the entire MDS-UPDRS or subparts of the scale. Specifically, MDS-UPDRS III scores have been empirically demonstrated as a good metric for monitoring the progression of PD [12]. As such, several studies have focused on the prediction of this score to monitor disease progression. A recent example of this approach for the monitoring of PD progression is the 2021 study that leveraged a convolutional neural network (CNN) model trained using inertial data collected from the lower back during gait to estimate MDS-UPDRS III scores [14]. While these results are promising, the authors suggest that a comparison with traditional feature-engineered machine learning models could be an avenue for future work, towards

Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia

the deployment of such technologies for continuous monitoring of PD. Other studies have revealed that it is possible to estimate PD progression using gait data collected with accelerometers [8]. However, the relative efficacy and effect of different approaches to data collection and processing, and machine learning pipeline design still lack consensus and clear comparisons that could help inform future research in this field.

In our work, we try to estimate the MDS-UPDRS part III from accelerometer data. We collected the data using the Axitvity AX3 device both on the wrist and lower back [3]. Our dataset contains data collected from 74 patients (HY between 2 and 4) at Campus Neurológico (CNS), a tertiary specialized movement disorders center in Portugal. The final subset of data contained 267 instances of gait from 104 evaluation sessions. We did different experiments with 4 models (Random Forest, XGBoost, SVM, Linear Regression), and 59 features from the statistical, spectral, and temporal domains. Furthermore, we used non-overlapping window sizes of 2.5 and 5 seconds. To validate the trained models we used Leave One Subject Out (LOSO) cross-validation.

Our results showed that the best configuration, with the lowest prediction error on the left out of 10% data, achieved a 4.26 MAE, with the Random Forest model, and a 2.5-second sliding window using combined data from the wrist and lower back. For all of the selected models, the configurations that achieved the best results using either of the validation schemes used data collected from the lower back or both sensors. Most models performed better using a 5-second window length, with the exception of the xgboost model. The best-performing linear regression and SVM-based models used the SURF and relieF feature selection methods.

Therefore, we contribute with the comparison of different models, features, sensor placement, and window sizes. We provide, according to our experiments, the optimal models for MDS-UPDRS part III estimation using only accelerometer data.

2 METHODS

The MDS-UPDRS III estimation was performed using different approaches to data collection, signal processing, and using different machine learning pipelines. In this section, we describe the steps taken together with the variables for each step, in order to enable a comparison between different design decisions and their effect on the estimation of the disease stage.

2.1 Data Collection

We collected data from 74 patients with PD at CNS from periodic evaluations conducted by trained physiotherapists. Each participant wore an Axivity AX3 on the wrist and lower back during a set of clinical assessments. Accelerometer data was set to record at 100 Hz. Our dataset includes 267 instances of gait from 104 evaluation sessions of the 10-meter walk. MDS-UPDRS were also applied for each patient in each session. Among these patients, 49 were male and 23 were female, while the gender of the remaining 2 patients was not reported. The average patient age was 70.4 years (SD=13.12). The average weight was 71.76 kg (SD=13.89) and the average height was 166.49 cm (SD=9.26). Finally, the average MDS-UPDRS III score was 40.92 (SD=14.31) and 2.57 (SD=0.97) for the H&Y scale.

2.2 Data Pre-Processing

In order to isolate gait instances, the selected data files were segmented using the annotated timestamps for the 3 trials of the 10-meter walk test. Visualization of each of the segmented gait instances was then created in order to exclude session data that contained sensor failures and misalignment, or mismatched timestamps. During this step, the vector magnitude of the accelerometry signal was computed and appended to each segment using the traditional euclidean vector norm formula $\sqrt{x^2 + y^2 + z^2}$. To avoid the possible temporal drift associated with the process, a resampling step was performed after segmentation to ensure even sampling, as required for the extraction of some of the used Time and Frequency domain features. Finally, all segments were filtered using a fourth-order, digital low pass Butterworth filter with a cut-off frequency of 20 Hz in order to remove possible "machine noise" [5].

2.3 Evaluated Models and Features

We used 16 statistical, 26 temporal, and 17 spectral domain features, with a total of 59. They were computed from all accelerometry axes and vector magnitude. A sliding window technique was used to segment the signal into non-overlapping windows from which the features were extracted. Different feature data frames were then created using 2.5 and 5-second windows, both of which were previously used in the literature [14], in order to assess the effect of window size on the estimation task. During this feature extraction process, MDS-UPDRS III scores were also computed and appended to the corresponding windows for both data frames. The first step toward feature selection was to use a variance filter to exclude features with low (<0.025%) or zero variance which lowered the feature space from 2081 to 266 in the 2.5-second window and 3081 to 452 in the 5-second window. While this reduction may seem drastic, it is to be expected because of the way Time Series Feature Extraction Library works, computing the same feature several times for different frequencies for example which results in a large number of feature columns with hardly any variability, and thus, descriptive power. A further feature selection step was performed using four different feature selection methods that implement different strategies for feature ranking. Each of these feature selection algorithms was used to rank and select the top 10/25/50 features to be used for the regression task using the linear regression algorithm, and with the support vector-based model. The complete feature subset was also used for these models, in order to establish a baseline comparison with the remaining tree-based models that are less affected by the number of features due to their capability to perform intrinsic feature selection.

For each model, a set of parameters were selected and used in a grid search procedure to test all possible combinations. This procedure was then carried out for each sensor placement and the combined sensors, and for the different sliding window lengths used during feature extraction, in order to compare the effect of these variables for the estimation task. Leave One Subject Out (LOSO) cross-validation was used during the grid search procedures in order to avoid overfitting and optimize the models for generalizability. Finally, the optimal models for each combination of these variables were saved and used for the ensuing validation tasks. To validate the trained models, the original dataset was split into training and testing subsets. The training subset comprised 90% of the data and was used during the grid-search procedure for training the models using LOSO cross-validation. The remaining 10% of the data was then used as a validation set to test the model's performance on unseen data from patients whose data the model had already seen, providing information on

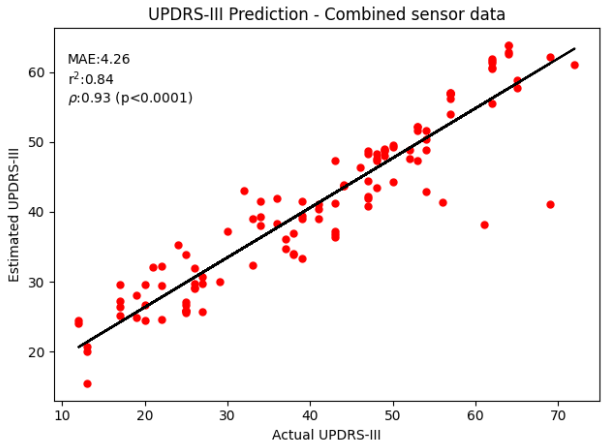


Figure 1: Overall optimal predictions on the 10% of left out data using a Random Forest model on data collected from both sensors and a 2.5s sliding window. Each point represents a window.

the model's ability to estimate MDS-UPDRS III scores for patients that were already known to these models. These steps yield two different scores for each of the optimal models using the same Mean Absolute Error (MAE) evaluation metric: the average MAE for all LOSO splits during training and the MAE for the held-out validation set. For the purpose of this study, this metric is defined as the mean absolute difference between real (x) and estimated (y) MDS-UPDRS III scores over the number of samples used for estimation.

3 RESULTS AND DISCUSSION

This section lays out the results from all of the steps taken toward UPDRS III estimation, including data processing, feature extraction and selection, and finally model training and validation results.

3.1 Optimal configurations

The configuration with the lowest prediction error on the left out 10% of data used data from both devices processed using a 2.5-second sliding window and a Random Forest model for prediction, achieving 4.26 MAE and strong correlation ($\rho = 0.93$) as illustrated in Figure 1. The best performing configuration when performing LOSO CV was a Support Vector-based model, using data from both sensors but a 5-second feature extraction window, achieving a MAE of 9.99. While predictions using this model on the validation set were less accurate than some of the other options at 7.94 MAE, it achieved the best balance when considering both of the validation schemes. Table 1 summarizes the optimal results achieved by each model along with the used data sources and sliding window length for the 10% left out and LOSO validation tasks.

3.2 Sensor placement and windows size

Both device placement and window length used during feature extraction significantly impacted the performance of all models. For all of the selected models, the configurations that achieved the best results using either of the validation schemes used data collected from the lower back or both sensors combined. Specifically, all of the non-tree-based models performed better in both

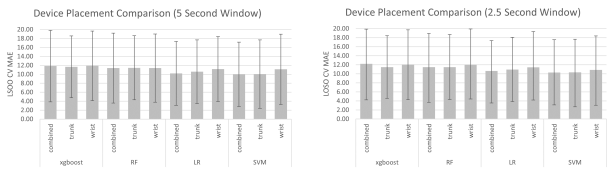
val_m	model	device_placement	win_length	ft_sel	num_fts	loso_mae	val_mae
1	rf	combined	250	-	266	11.50	**4.26**
1	xgboost	trunk	500	-	229	11.67	4.39
1	svm	combined	500	SURF	25	9.99	7.95
1	lin_reg	combined	500	reliefF	25	10.21	8.98
2	rf	combined	500	-	452	11.39	11.39
2	xgboost	trunk	250	-	133	11.49	5.74
2	svm	combined	500	SURF	25	**9.99**	7.95
2	lin_reg	combined	500	reliefF	25	10.21	8.98

Table 1: Optimal configurations used by each model to achieve optimal MAE on the left out 10% of data (val_m => 1) and LOSO (val_m => 2).

validation schemes using data from both sensors, with the exception of the SVM-based model using a 2.5-second window, which compared to the other options using the same window length achieved lower, albeit negligible, validation MAE using data from the wrist. As for the tree-based models, optimal validation MAE was attained by models using both sensors with the 2.5-second sliding windows, and data from the lower back for the same models using the 5-second window. Figures 2a and 2b illustrate the intra and inter-model comparison for both of the validation schemes, using different window lengths. While the fluctuations were relatively low using LOSO CV, most models performed better using a 5-second window length, with the exception of the xgboost model. MAE using the left out 10% of validation data fluctuated more considerably but was also lowest using 5-second windows for all models except RF.

3.3 Optimal parameters

As for model parameters, excluding linear regression, the remaining models had different parameters to achieve the best performance during LOSO CV. For Random Forest (criterion: mae ; max_features: 0.333 ; n_estimators: 250), for xgboost (colsample_bynode: 1; eta: 0.1 ; importance_type: total_gain; max_depth: 3 ; num_parallel_tree: 100 ; tree_method: gpu_hist), and for svm (C: 10 ; epsilon: 0.3 ; gamma: auto ; kernel: rbf). The xgboost was the one that used only the trunk sensor. The others models used both devices. We used a Grid Search procedure that exhaustively tested all parameter combinations for each model, independently of the used device placements and sliding window lengths. The exhaustive nature of the grid search procedure makes this method of parameter optimization computationally expensive. For this reason, and considering that the procedure was used for several models, the used parameter space for each model was not as comprehensive as those used in some other works with a smaller scope and narrower focus. However, the present results should still serve as a good starting point for model tuning in future research.



(a) LOSO CV MAE values (Y-axis) for different device placements using 5-second windows.

(b) LOSO CV MAE values (Y-axis) for different device placements using 2.5-second windows.

Figure 2

3.4 Feature importance

For the models that benefited from it, several feature selection methods were tested, along with different numbers of features to select. The best performing linear regression and SVM-based models used the SURF and relieF feature selection methods respectively, both selecting 25 as the optimal number of features. We then selected the top 20 for each model. Among the 8 top performing models across the two tested window lengths, no model used data exclusively from the wrist, and only 3 models used data exclusively from the trunk. As for the remaining models, the majority of top-ranking features were extracted from devices mounted on the lower back. In some cases, no wrist features were ranked among the top 20, which suggests that although these were used for the estimation task, their contribution is minimal, which is in line with the minimal performance gain in these models when compared to their counterparts using data exclusively from the lower back. Features from the anteroposterior plane of movement (z-axis) were the most prevalent among the top 20 extracted from the trunk sensor, consisting of 50 out of the 140 features considered for this analysis. The vertical plane of movement (x-axis) produced the least amount of features among those considered here, with only 22 ranking among the top contributing features. Spectral-domain features were the most prevalent among these, making up almost half of the 140 considered features, with temporal domain features coming in second by a small margin, and temporal features last consisting of a quarter of this total.

3.5 Limitations

The dataset used in this study consisted of data collected from 74 patients. While this number of patients is significant for preliminary results, a larger sample size could improve the estimation task and further validate the present findings. Beyond the volume of data used to train the models, a wider range of MDS-UPDRS III and Hoehn and Yahr scores could also possibly improve the results, by including a wider variety of walking patterns that in smaller sample sizes could be considered outliers and negatively affect performance. Furthermore, the inclusion of a healthy cohort in the dataset could provide a baseline for the models to recognize healthy gait, exacerbating the difference between data from healthy and affected subjects. Therefore, in future work a longitudinal study in free-living environments with a larger sample size to address our limitations and extend our conclusions.

4 CONCLUSIONS

This paper presents a study that compares the different models, features, and window sizes to estimate MDS-UPDRS part III using accerometeler data. One of the most common disorders for people with PD is gait. The increase in sensor usage opened the opportunity for increasing objective evaluations. However, there is a lack of knowledge of the current machine learning approaches. In our work, we compare 4 machine learning models (random forest, xgboost, svm, and linear regression), 59 features (16 statistical domain, 26 spectral domain, and 17 temporal domain), and windows size (2.5 and 5 seconds). To validate our models we used LOSO cross-validation. We showed that the configuration with the lowest prediction error on the left out 10% of data used data from both devices processed using a 2.5-second sliding window and a Random Forest model for prediction, achieving 4.26 MAE. This work opens the opportunity to improve the knowledge of machine learning approaches. However, in future work, there are opportunities for longitudinal studies in free-living environments with larger datasets.

ACKNOWLEDGEMENTS

We would like to thank all the participants that kindly participated in the studies. This project was partially supported by FCT through LASIGE Research Unit funding refs. UIDB/00408/2020 and UIDP/00408/2020 and SFRH/BD/144242/2019 to Diogo Branco., and the WideHealth project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952279.

REFERENCES

[1] Raquel Bouça-Machado, Diogo Branco, Gustavo Fonseca, Raquel Fernandes, Daisy Abreu, Tiago Guerreiro, Joaquim J Ferreira, and CNS Physiotherapy Study group. 2021. Kinematic and clinical outcomes to evaluate the efficacy of a multidisciplinary intervention on functional mobility in Parkinson's disease. *Frontiers in neurology* 12 (2021), 637620.

[2] Raquel Bouça-Machado, Constança Jalles, Daniela Guerreiro, Filipa Pona-Ferreira, Diogo Branco, Tiago Guerreiro, Ricardo Matias, and Joaquim J Ferreira. 2020. Gait kinematic parameters in Parkinson's disease: a systematic review. *Journal of Parkinson's disease* 10, 3 (2020), 843–853.

[3] Clare L Clarke, Judith Taylor, Linda J Crighton, James A Goodbrand, Marion ET McMurdo, and Miles D Witham. 2017. Validation of the AX3 triaxial accelerometer in older functionally impaired people. *Aging Clinical and Experimental Research* 29, 3 (2017), 451–457.

[4] Silvia Del Din, Alan Godfrey, Brook Galna, Sue Lord, and Lynn Rochester. 2016. Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length. *Journal of neuroengineering and rehabilitation* 13, 1 (2016), 1–12.

[5] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H Granat, Tom White, Vincent T Van Hees, Michael I Trenell, Christoper G Owen, et al. 2017. Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study. *PloS one* 12, 2 (2017), e0169649.

[6] Alberto J Espay, Paolo Bonato, Fatta B Nahab, Walter Maetzler, John M Dean, Jochen Klucken, Bjoern M Eskofier, Aristide Merola, Fay Horak, Anthony E Lang, et al. 2016. Technology in Parkinson's disease: challenges and opportunities. *Movement Disorders* 31, 9 (2016), 1272–1282.

[7] Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B Stern, Richard Dodel, et al. 2008. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society* 23, 15 (2008), 2129–2170.

[8] Murtadha D Hssayeni, Joohi Jimenez-Shahed, Michelle A Burack, and Behnaz Ghoraani. 2021. Ensemble deep model for continuous estimation of Unified Parkinson's Disease Rating Scale III. *Biomedical engineering online* 20, 1 (2021), 1–20.

[9] Anthony E Lang, Shirley Eberly, Christopher G Goetz, Glenn Stebbins, David Oakes, Ken Marek, Bernard Ravina, Caroline M Tanner, Ira Shoulson, and LABS-PD investigators. 2013. Movement disorder society unified Parkinson disease rating scale experiences in daily living: longitudinal changes and correlation with other assessments. *Movement Disorders* 28, 14 (2013), 1980–1986.

[10] Walter Maetzler, Josefa Domingos, Karin Srulijes, Joaquim J Ferreira, and Bastiaan R Bloem. 2013. Quantitative wearable sensors for objective assessment of Parkinson's disease. *Movement Disorders* 28, 12 (2013), 1628–1637.

[11] C Warren Olanow, Yves Agid, Yoshi Mizuno, Alberto Albanese, U Bonucelli, Philip Damier, Justo De Yebenes, Oscar Gershanik, Mark Guttman, F Grandas, et al. 2004. Levodopa in the treatment of Parkinson's disease: current controversies. *Movement disorders* 19, 9 (2004), 997–1005.

[12] Antoine Regnault, Babak Boroojerdi, Juliette Meunier, Massimo Bani, Thomas Morel, and Stefan Cano. 2019. Does the MDS-UPDRS provide the precision to assess progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort. *Journal of neurology* 266, 8 (2019), 1927–1936.

[13] Rana Zia Ur Rehman, Christopher Buckley, Maria Encarna Micó-Amigo, Cameron Kirk, Michael Dunne-Willows, Claudia Mazzà, Jian Qing Shi, Lisa Alcock, Lynn Rochester, and Silvia Del Din. 2020. Accelerometry-based digital gait characteristics for classification of Parkinson's disease: what counts? *IEEE open journal of engineering in medicine and biology* 1 (2020), 65–73.

[14] Rana Zia Ur Rehman, Lynn Rochester, Alison J Yarnall, and Silvia Del Din. 2021. Predicting the Progression of Parkinson's Disease MDS-UPDRS-III Motor Severity Score from Gait Data using Deep Learning. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 249–252.

[15] Ole-Bjørn Tysnes and Anette Storstein. 2017. Epidemiology of Parkinson's disease. *Journal of neural transmission* 124, 8 (2017), 901–905.