Explainable machine learning for the prediction of motor fluctuations and Levodopa-induced dyskinesias in Parkinson's disease

Walter Endrizzi[1,2], Nicole Campese[3], Flavio Ragni[1], Monica Moroni[1], Stefano Bovo[1], Chiara Longo[3], Lorenzo Gios[1], Maria Chiara Malaguti[3], Antonio Uccelli[4], Bruno Giometto[3,5], Giuseppe Jurman[1,6], Venet Osmani[7], the NeuroArtP3 Network†

[1] Data Science for Health, Fondazione Bruno Kessler, Trento, Italy
[2] Department of Cellular, Computational and Integrative Biology, University of Trento, Trento, Italy
[3] Neurology Unit, Azienda sanitaria universitaria integrata del Trentino (ASUIT), Trento, Italy
[4] IRCCS Ospedale Policlinico San Martino, Genoa, Italy
[5] Centro Interdipartimentale di Scienze Mediche (CISMed), Facoltà di Medicina e Chirurgia, Università di Trento, Trento, Italy
[6] Department of Biomedical Sciences, Humanitas University, Milan, Italy
[7] Digital Environment Research Institute, Queen Mary University of London, London, E1 1HH, United Kingdom
† NeuroArtP3 Network details are presented in the Appendix.

Abstract

Background: Motor complications, such as motor fluctuations and Levodopa-induced dyskinesias (LID), significantly impair quality of life in persons with Parkinson's disease (PD) on long-term Levodopa treatment. Predicting their onset is crucial for tailored patient care.

Objectives: To develop and evaluate machine learning (ML) models to forecast the onset of new motor fluctuations and LID in PD patients within three years from baseline assessment.

Methods: A comprehensive ML workflow with repeated Nested Grid Search Cross-Validation was applied to real-world clinical data from a multicentric cohort of 247 PD patients. ML models were rigorously evaluated on the clinically relevant subgroup free of motor complications at baseline. SHAP analysis provided model explainability.

Results: Models achieved moderate predictive power for both LID (SVC: MCC 0.28 ± 0.14) and motor fluctuations (Voting MCC = 0.32 ± 0.18). For LID prediction, the strongest predictors were the Levodopa Equivalent Daily Dose (LEDD), baseline motor fluctuations, and duration of

Levodopa therapy, with risk increasing significantly above a LEDD threshold of 300-400 mg. A critical ablation study revealed that excluding patients with pre-existing complications caused a collapse in model sensitivity, highlighting their essential role in defining the upper bound of predicted risk.

Conclusions: The model-based risk assessment is consistent with established clinical factors. Inclusion of the full spectrum of disease severity, including patients with pre-existing motor complications, in the training set is essential for achieving a robust probabilistic risk scale and reliable model calibration for new-onset prediction.

Introduction

Motor complications are common, troublesome features associated with long-term Levodopa treatment in Parkinson's disease (PD) (1). According to a longitudinal population-based study, up to 50% of persons living with PD experience motor fluctuations and 15% of PD patients develop Levodopa induced dyskinesias (LID) after 5 years of Levodopa treatment, with an incidence increasing to 100% and 55% respectively after 10 years (2). On the other hand, some people with PD do not experience any motor fluctuations, even after many years of Levodopa treatment, which defines the so called long-duration response to Levodopa (3). Several risk factors for motor complications have been identified in PD, with longer disease duration and higher total Levodopa Equivalent Daily Dose (LEDD) emerging as the strongest predictors of both motor fluctuations and LID (1,4). Additionally, female gender and a younger age at disease onset were consistently reported across studies as risk factors for both motor fluctuations and LID (5,6).

Predicting the onset of motor complications in patients with PD is crucial for tailoring patients' care, improving their quality of life, preventing disability and identifying patients who may be eligible for device-assisted therapies (7,8).

To date, evidence on risk factors for motor fluctuations and LID relies primarily on single-cohort studies, limiting the generalizability of findings, and no validated algorithms are currently available to accurately predict their onset. To address these prognostic challenges, machine learning (ML) models, coupled with Explainable Artificial Intelligence (XAI) to foster clinical trust, are being explored with encouraging results (9–11)**.** However, translating these promising tools into practice requires evaluations that demonstrate genuine clinical utility. Rather than relying solely on broad dataset metrics, rigorous assessments should directly test the true clinical challenge: predicting the future onset of motor complications in patients currently free of them.

This study utilizes real-world clinical data from two distinct cohorts to develop and evaluate ML models for predicting the onset of motor fluctuations and LID in persons with PD within three years from baseline assessment.

Materials and methods

Ethics statement and protocol

The study was approved by the Locals Ethics Committees (Ethics Committee of Azienda Provinciale per i Servizi Sanitari of Trento and Ethics Committee of Liguria Region—Protocol number NET-2018-12366666-3-PD). The protocol for this study was published early in 2024 and refers to the NeuroArtP3 (NET-2018-12366666) project (12). Written informed consent was obtained from every participant, and all assessments were performed according to national and international guidelines.

Cohort characteristics

A total of 256 patients with PD diagnosed according to current criteria (13) were recruited from the NeuroArtP3 project at two Northern Italian sites: Azienda Sanitaria Universitaria Integrata del Trentino (formerly APSS - Provincial Health Services, Trento, n=126) and IRCCS San Martino Hospital (HSM, Genoa, n=130). The final analytic cohort consisted of 247 subjects after exclusion of seven patients treated with Deep Brain Stimulation (DBS) or other advanced therapies and two patients with incomplete clinical information.

For each participant, data was collected longitudinally during routine clinical practice for three consecutive years. For the predictive modeling, only variables collected at the first visit (i.e. baseline visit) were used as input features. These baseline features encompassed clinical-demographic features, including clinical rating scales (e.g., the Movement Disorders Society Unified Parkinson's Disease Rating Scale part III subscores - MDS-UPDRS III, Hoehn & Yahr), PD core motor and non-motor features (e.g., presence of impulse control disorder - ICD, sleep disorders, cognitive status), and medication history. A complete list of these variables and their characteristics is provided in Table 1.

The binary outcome labels, indicating whether a patient developed LID or motor fluctuations within three years from the baseline visit were created using the follow-up data. LID were defined as the sudden peak-of-dose emergence of involuntary movement disorders and motor fluctuations as the sudden transition between functional "on" and "off" states (including wearing-off, delayed-on,

no-on, unpredictable off phenomena) (8,14), as assessed by interview during the follow-up visits and reported in the medical charts. Descriptive statistics of these two outcomes are reported in Table 1. A crucial stratification was necessary for this analysis: at baseline, a number of patients already presented with LID (n=34) or motor fluctuations (n=59). The true clinical challenge lies in identifying which patients will develop these symptoms in the future, not in confirming a pre-existing condition. Therefore, to build a model with real-world prognostic utility, the main evaluation of our predictive analyses focused on the subset of patients who did not present motor complications at baseline.

Variable	**Level**	**Count (%)**
Gender (Male = 0; Female=1)	Female	96 (38.9%)
Disease duration	0-5 years	145 (58.7%)
	5-10 years	63 (25.5%)
	> 10 years	39 (15.8%)
Family history (for PD)	Yes	80 (32.4%)
H&Y stage	1	77 (31.2%)
	2	128 (51.8%)
	>2	42 (17.0%)
Motor phenotype	Tremor-dominant	145 (58.7%)
	Akinetic-rigid	102 (41.3%)
Falls	Yes	31 (12.6%)
Freezing of gait	Yes	31 (12.6%)
LID	Yes	34 (13.8%)
Motor fluctuations	Yes	59 (23.9%)
Presence of dysautonomia	Orthostatic hypotension	20 (8.1%)
	Other domains	68 (27.5%)
	None	159 (64.4%)
Cognitive status (n=242)	Normal	217 (87.9%)
	MCI	19 (7.7%)
	Dementia	6 (2.4%)
Anxiety	Yes	61 (24.7%)
Depression	Yes	70 (28.3%)
ICD	Yes	20 (8.1%)

Hyposmia (n=211)	Yes	74 (30.0%)
Sleep/Wake disorders	Yes	101 (41.3%)
Use of Levodopa	Yes	172 (69.6%)
Use of DA	Yes	173 (70.0%)
Use of MAOIs	Yes	147 (59.5%)
LID at follow-up	Yes	96 (38.9%)
Motor fluctuations at follow-up	Yes	113 (45.7%)
Attribute (n)	**Median (Q1, Q3)**	
Age at baseline (n=247)	69 (63, 74)	
Education (n=244)	11 (8, 13)	
Time between symptoms onset and diagnosis (years) (n=238)	1 (0, 2)	
MDS-UPDRS III (ON) (n=237)	17 (11, 27)	
LEDD (n=246)	300 (0, 450)	
Levodopa use, duration (years, n=240)	2 (0, 5)	
DA use, duration (years, n=244)	1 (0, 4)	
MAOIs use, duration (years, n=245)	0 (0, 1)	

Table 1. Characteristics of the study cohort (N=247). Categorical variables are presented as count (percentage). Continuous variables are presented as median (Q1, Q3). In case of missing data, the sample size for the variable of interest is reported in brackets. Legend: DA: dopamine agonist; H&Y: Hoehn and Yahr stage; ICD: impulse control disorder; LEDD: levodopa equivalent daily dose; LID: Levodopa-induced dyskinesia; MAOIs: monoamine oxidase inhibitors.

Preprocessing and statistical analyses

As initial preprocessing step, variables with a missing data proportion exceeding 50% were discarded. The remaining features then underwent a baseline statistical comparison between the clinical outcome groups (LID and motor fluctuations). This comparison, presented in Table S1, was performed using a complete-case analysis, applying the Mann-Whitney U test for continuous data and Pearson's chi-square test for categorical data. Before each training session, the remaining missing values were imputed using the median for numerical features and the mode (most frequent category) for categorical features; then, all numerical features were scaled to a mean of 0 and a standard deviation of 1.

Training

Binary classification of patients was performed to predict the onset of the target outcome (LID or motor fluctuations) in PD patients within three years from baseline evaluations. Five ML models were tested: Random Forest (RF), ExtraTrees Classifier (ETC), Extreme Gradient Boosting (XGB), Logistic Regression (LR), Support Vector Classifier (SVC). In addition, two ensemble techniques were employed: a Voting classifier that aggregated predictions from RF, SVC, XGB, and LR, and a Stacking classifier that used the predictions from XGB, RF, and SVC as input features for a final LR meta-model.

To assess model performance, a Randomized Nested Grid Search Cross Validation (CV) strategy was implemented; the workflow is summarized in Figure 1. The full procedure iterated 30 times to assess stability, with the average and standard deviations across these iterations reported as results. In each of the 30 iterations, the entire dataset was first randomly split into 80% for training and 20% for hold-out testing. To ensure representative data splits, this partitioning was stratified by both the target outcome and the baseline symptom status (i.e., ensuring an equal proportion of patients with baseline motor complications in both the training and test sets).

Within each of the 30 training loops, a repeated (5 repetitions) 3-fold CV approach was employed to find the best model configuration over a predefined parameter space (randomized grid search over 20 parameters combinations; see Table S2). The model configuration achieving the best averaged Matthews Correlation Coefficient (MCC) (15) across these internal validation folds was selected as the optimal architecture for that iteration and refitted on the entire training set.

To ensure the clinical reliability of risk predictions, the optimal model underwent probability calibration prior to final evaluation. Raw classifier outputs were mapped to true posterior probabilities using an internal CV strategy. Specifically, Platt Scaling (16) was applied via a 5-fold CV scheme on the training data. Calibration quality was assessed according to the calibration hierarchy proposed by Van Calster et al. (17): *weak calibration* was quantified via the calibration intercept and slope, while *moderate calibration* was visually inspected using flexible calibration curves to ensure alignment between predicted probabilities and observed frequencies (see Supplementary Materials, Figure S1 and S3).

All model training and optimization procedures were implemented using scikit-learn. All computational steps, including preprocessing and classification, were integrated into a single pipeline to create a clean and reproducible training flow and to prevent data leakage (i.e., ensure

that information from the test set did not influence the training process).

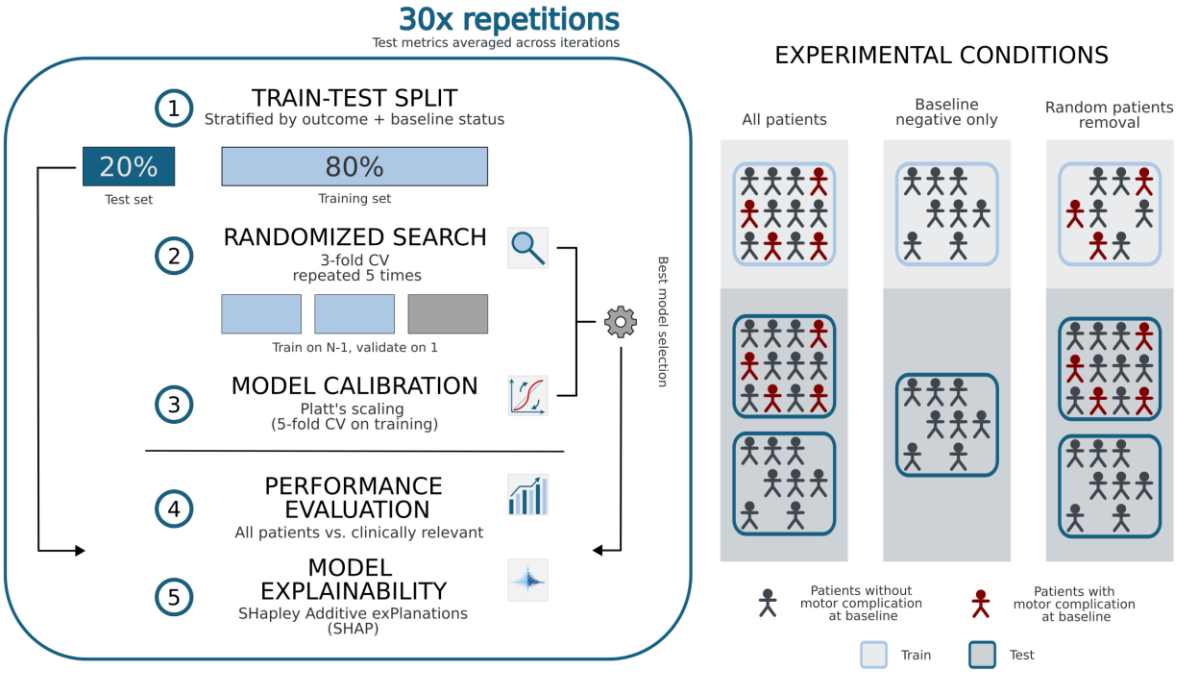

Figure 1. Schematic representation of the repeated nested cross-validation pipeline and experimental conditions. The entire workflow was repeated 30 times, reporting mean ± SD across iterations. Left panel. (1) Train–test split: in each outer iteration, data were split into 80% training and 20% hold-out test, stratified by outcome and baseline motor-complication status. (2) Randomized search: within the training set, hyperparameters were optimized via a randomized search (20 configurations; Table S2) using 3-fold CV (5 repeats), selecting the configuration with the best mean MCC across internal validation folds. (3) Model calibration: the selected model was probability-calibrated with Platt scaling using 5-fold CV on the training data (see Supplementary Materials, Figures S1, S3). (4) Performance evaluation: metrics (MCC, AUC-ROC, F1-score, sensitivity, specificity) were computed on the hold-out test set both globally (all patients) and in the clinically relevant subgroup excluding patients with baseline LID/motor fluctuations. (5) Model explainability: calibrated risk predictions were interpreted with SHAP, aggregating explanations across outer iterations to assess stability of feature contributions. Right panel. Experimental conditions used to probe the role of baseline motor-complication status in learning: training on all patients, training on baseline-negative only, and a random patient-removal control (matched in sample size and preserving the stratification of both outcome and baseline motor complications). CV: cross validation; MCC: Matthew's Correlation Coefficient; LID: Levodopa induced dyskinesia.

Evaluation

Model performance was assessed with a strong emphasis on clinical utility. As described in the training settings, ML classifiers were developed on the entire training population to maximize available information utilization. Performance metrics (MCC, AUC-ROC, F1-Score, Sensitivity and Specificity) were computed on the hold-out test set in two distinct ways:

- Global Evaluation: metrics were calculated on all patients in the test set. These results were inherently biased due to the inclusion of patients who already had the target motor complication at baseline.
- Clinically Relevant Evaluation: all patients presenting LID or motor fluctuations at baseline were excluded from the computation of performance metrics. This analysis specifically assesses the model's ability to forecast the onset of motor complication in patients that did not present LID or motor fluctuations at baseline - the clinically relevant scenario.

Two key methodological decisions underpinned this evaluation framework. First, to ensure the models were evaluated on their ability to predict the onset of LID or motor fluctuations, the baseline value of the target outcome was deliberately excluded from the feature set. This prevented the models from learning a trivial classification shortcut based on the presence or absence of the target outcome at baseline.

Second, to investigate the specific contribution of dyskinetic or fluctuating patients at baseline to the model’s learning process, a controlled ablation study was implemented. Adhering strictly to the Nested Grid Search CV pipeline described above, we first retrained the models solely on the subset of patients who were non-dyskinetic or non-fluctuating at baseline. This aimed to test whether the model could learn onset trajectories exclusively from the feature signals of patients free from LID or motor fluctuations. Because excluding these patients also reduces the numerosity of the training set, we then disentangled the effect of sample-size reduction from the effect of losing information from patients with motor complications at baseline. To do so, in each outer iteration we ran a control experiment in which we removed a random subset of patients equal in size to the excluded LID or motor fluctuation group, preserving the stratification of both outcome and baseline motor complications. Comparing performance across these two scenarios allowed us to determine whether any changes were primarily driven by fewer training examples (quantity of data) or by the loss of informative signals from patients with motor complications at baseline (quality of signal).

Model Explainability

To ensure clinical transparency, model explainability was assessed using SHAP (SHapley Additive exPlanations) (18) to interpret the calibrated risk predictions. SHAP values were computed directly on the final probability outputs using a model-agnostic Kernel Explainer ensuring that feature contributions reflected the actual clinical risk scores. This analysis was performed iteratively across each outer iteration, generating local explanations for the held-out test set. These values were subsequently concatenated to produce a single global summary plot, visualizing the stability and magnitude of feature contributions across the entire study cohort.

Beyond the global feature rankings, we specifically investigated the contribution of LEDD to the prediction of future LID. The aim was to determine how variations in feature value influenced this prediction, and to identify a critical dosage threshold. To explore this non-linear relationship, we generated SHAP dependence plots which illustrate the marginal contribution (SHAP value) of the LEDD feature as its actual value changes across the patient cohort.

Results

LID Prediction

Model evaluation for future LID prediction was primarily based on the MCC, a metric chosen for its robustness in imbalanced classification (prevalence: 39% total, 31% evaluation subset). While the highest MCC scores were generally achieved when training and testing on the full cohort (SVC MCC = 0.46 ± 0.11; XGB MCC = 0.42 ± 0.11), these performances were considered inflated by the inclusion in the evaluation set of patients with LID at baseline. Therefore, to maximize clinical relevance, final model selection was based on performance specifically within the subset of patients who were free from LID at baseline. The best-performing models in this clinically relevant analysis were the SVC and XGB, which were selected for detailed evaluation (see Supplementary Table S3 for all performance metrics).

Focusing on the task of predicting only the new onset of LID, the two models achieved moderate predictive power (SVC MCC 0.28 ± 0.14; XGB MCC 0.26 ± 0.16). However, when the training strategy was altered to entirely exclude patients free from LID at baseline, a severe performance drop was observed, particularly for the tree-based model (XGB MCC 0.02 ± 0.11).

To understand the key drivers behind the predictions of the SVC model, we conducted a SHAP analysis. The global feature importance is visualized in the summary plot (Figure 2A), which reveals that LEDD, motor fluctuations at baseline, and duration of Levodopa therapy were the three most influential features for SVC in predicting dyskinesia. For all three features, larger values (indicated by red dots) corresponded to positive SHAP values, indicating a strong contribution toward predicting the presence of future LID. Conversely, low values for these features (blue dots) were associated with negative SHAP values, driving the model to predict absence of LID within three years from the baseline.

Given its high importance and clinical relevance, the LEDD feature was analyzed in greater detail to investigate whether a specific dosage threshold is associated with high risk of future LID. The SHAP dependence plot (Figure 2B) revealed that doses below approximately 300-400 mg cluster around or below a SHAP value of zero, suggesting a minimal impact on the model's prediction. In contrast, at doses above this 300-400 mg threshold, the SHAP values become almost exclusively positive, indicating that higher Levodopa doses significantly increase the model's predicted risk of developing LID.

This interpretability analysis was also performed on the XGB model. The results were highly consistent with SVC: the SHAP analysis identified the same key drivers of risk, and the LEDD dependence analysis revealed an identical 300-400 mg risk threshold. The full SHAP summary and dependence plots for the XGB model are provided in the Appendix (Figure S2).

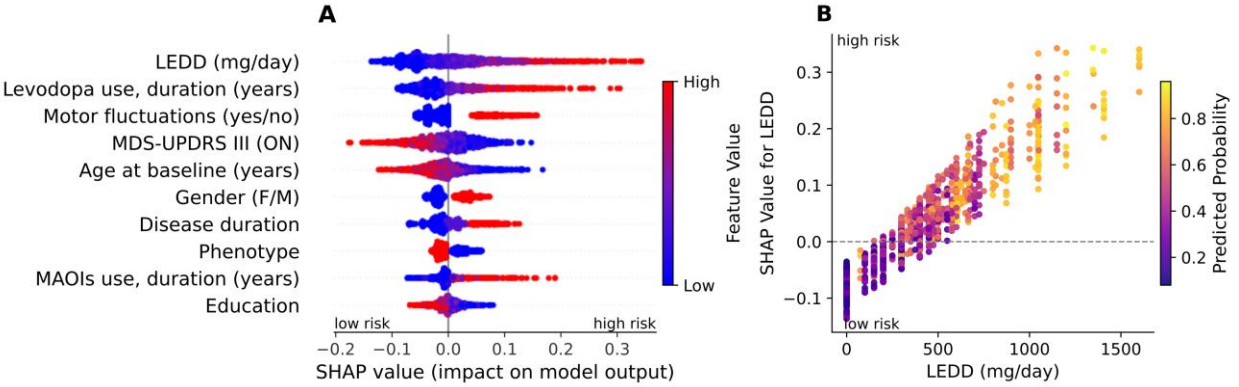

Figure 2. SHAP analysis of SVC for the prediction of future LID. A. Global feature importance and impact for the SVC model revealed by SHAP. Features are ranked in descending order of importance. Each dot represents a patient from the test set. The horizontal axis shows the SHAP value, indicating the feature's contribution to predicting LID (positive values) or absence of LID (negative values) within the third year. Color indicates the feature's original value (blue=low, red=high), revealing the direction and magnitude of its effect. **B.** Detailed analysis of the LEDD feature using a SHAP dependence plot for the SVC model. The x-axis plots the patient's actual LEDD on its original scale, while the y-axis plots the SHAP value, indicating the feature's contribution to the prediction. A potential dosage threshold is visible around 300-400 mg, where the feature's impact (SHAP value) transitions from negative (low risk) to positive (high risk). LEDD: Levodopa equivalent daily dose; LID: Levodopa-induced dyskinesia; MAOIs: monoamine oxidase inhibitors.

Motor fluctuations prediction

For the prediction of motor fluctuations, across all tested classifiers (Supplementary Table S4), the Voting ensemble and RF were the best performing models and were selected for further analysis. Consistent with the LID task, performance metrics were significantly inflated when evaluated on the full cohort (Table 6, MCC Voting = 0.51 ± 0.12, MCC RF = 0.53 ± 0.11). Focusing on the clinically relevant prediction of new onset in patients free of motor complications at baseline, a substantial performance drop was observed for both models (Voting MCC = 0.32 ± 0.18, RF MCC = 0.32 ± 0.18).

When models were trained directly on patients without baseline motor fluctuations, a further performance decline was observed (Voting MCC = 0.08 ± 0.15, RF MCC = 0.10 ± 0.15), mirroring the trend observed in the LID analysis.

To identify the key drivers for the motor fluctuations prediction, we performed a SHAP analysis on the best performing models. The resulting global feature importance for the Voting ensemble model is shown in the summary plot (Figure 3), which reveals a SHAP pattern consistent with that of the RF model (Figure S4).

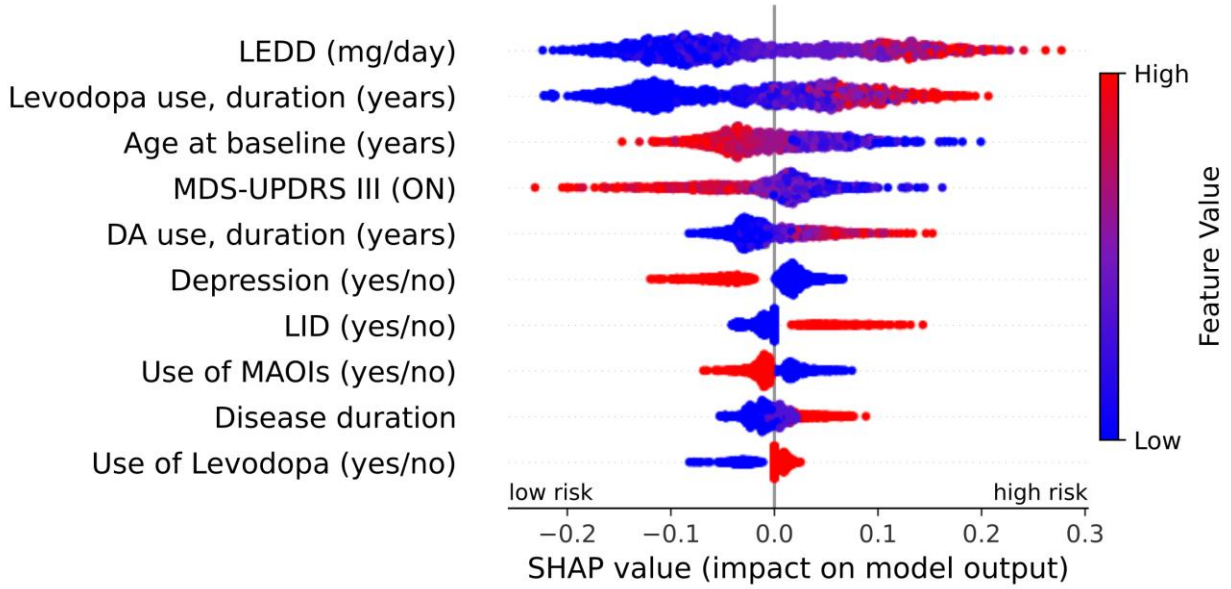

Figure 3. SHAP analysis of Voting ensemble for prediction of future motor fluctuations. Global feature importance and impact for the Voting ensemble revealed by SHAP. Features are ranked in descending order of importance. Each dot represents a patient from the test set. The horizontal axis shows the SHAP value, indicating the feature's contribution to predicting the presence (positive values) or absence of motor fluctuations (negative values) within the third year. Colors indicate the feature's original value (blue=low, red=high), revealing the direction and magnitude of its effect. DA: dopamine agonist; LEDD: levodopa equivalent daily dose; LID: Levodopa-induced dyskinesia; MAOIs: monoamine oxidase inhibitors.

Impact of Training Cohort Composition

A controlled ablation study was performed for both clinical tasks to investigate the drivers of the performance collapse observed when training exclusively on patients without motor complications at baseline.

First, the impact of sample size was assessed. When a random subset of patients - matched in exact size to the group of individuals with complications at baseline - was removed from training (Supplementary Table S3, S4), predictive performance remained robust and comparable to the full-cohort training for both outcomes. This suggests that the deterioration observed after excluding patients with baseline motor complications was not attributable to the reduced cohort size, but rather to the loss of specific clinical information.

Second, evaluation of the resulting models highlighted a dissociation between discrimination and threshold-based decision performance. While rank-based metrics were relatively preserved (e.g., SVC AUC-ROC for LID remained stable at 0.67 ± 0.08, compared to 0.70 ± 0.09 in the random patient removal), decision-based metrics collapsed (MCC dropping to ~ 0.10 and Sensitivity decreasing by >50%). This contrast indicates that while the models could still rank patients by risk, they failed to assign calibrated probabilities high enough to cross the decision threshold.

Notably, the magnitude of this dissociation varied across model architectures. While boosting and randomized ensembles (XGB, ETC) exhibited the most severe collapse in overall decision-making capability (e.g., XGB MCC for the LID task dropping to 0.02 ± 0.11), RF and linear models (LR, SVC) retained comparatively higher stability (0.17 ± 0.15 and 0.14 ± 0.17 respectively on the LID task). However, the substantial overlap in performance variance across all algorithms indicates that while linear and averaging strategies appeared slightly more resilient, the generalized loss of sensitivity (dropping below 0.25 across all architectures) confirms that no algorithm was immune to the exclusion of patients with pre-existing motor complications.

Discussion

Levodopa-induced dyskinesias (LID)

Consistently with previous studies (4), a longer cumulative exposure to Levodopa, a higher LEDD and the presence of motor fluctuations at baseline were the strongest determinants of LID at follow-up in our cohort. Moreover, a longer disease duration at baseline was associated with a higher subsequent risk of LID. These results are consistent with observations by Cilia and colleagues, showing that the disease duration and the cumulative duration of the exposure to Levodopa, rather than the disease duration at the time of Levodopa initiation, are the stronger predictors of both motor fluctuations and LID in PD (4). A higher LEDD has been consistently reported as a risk factor for the development of LID, with critical LEDD thresholds lying between 365 and 400 mg daily, according to different studies (4,19). This association was confirmed by our model, which showed a substantial increase of the risk of LID for LEDD values greater than 300-400 mg daily. Paralleling previous observations (19–21), we showed that female PD patients are at higher risk of LID. Both lower body weight and gender-related differences in the pharmacokinetics of Levodopa may underlie this finding (22). Similarly, we confirm previous observations on the association between a younger age at disease onset and an increased risk of LID (4,19,23–25). In young-onset PD, pre- and post-synaptic neuronal plasticity mechanisms are supposed to effectively compensate the dopaminergic neuronal loss until the very late stages of the disease, in which a more severe depletion of dopaminergic cells occurs (4,26,27). Young-onset PD patients are therefore supposed to present with a more severe dopaminergic loss compared to older-onset persons with PD with similar severity of motor features, thus accounting for an increased risk of motor fluctuations (4,26).

In our cohort, higher MDS-UPDRS III scores at baseline were associated with a lower subsequent risk of motor fluctuations. This is not consistent with what reported by a previous study by Eusebi and colleagues (20). It is worth noting that the tremor items have a greater relative weight in determining the total score MDS-UPDRS III compared to akinesia and rigidity items. Therefore, persons presenting a tremor-dominant phenotype tend to score higher at the MDS-UPRDS III compared to rigid-akinetic phenotypes. In our cohort, PD patients with a tremor-dominant phenotype show a lower risk of future LID. The association observed between MDS-UPDRS III scores, and the risk of LID may be therefore driven by differences in clinical phenotypes rather than in disease severity.

Motor fluctuations

In our model, motor fluctuations at follow-up were associated with a higher LEDD, a longer cumulative exposure to Levodopa, and a younger age at baseline. Similarly to LID, both the higher LEDD, a longer cumulative exposure to Levodopa and a longer disease duration, are well established risk factors for motor fluctuations in PD (4,21). In our cohort, motor fluctuations were associated with lower MDS-UPDRS III scores. Prior studies report conflicting findings, showing either a positive (2,5,28) or no association between MDS-UPDRS III scores (21). As previously speculated for LID, the higher MDS-UPDRS III motor scores at baseline may reflect differences in the clinical phenotype rather than in disease severity alone and therefore justify our observation. Depression emerged as a protective factor for motor fluctuations in our models. Non-dopaminergic pathways, including serotoninergic and noradrenergic pathways, seem to be involved in the pathophysiology of motor fluctuations in PD. Persons with depression may be treated with drugs acting on these neurotransmitters, which may therefore exert an indirect effect on motor fluctuations (29). In our cohort the presence of LID at baseline was associated with an increased risk of subsequent motor fluctuations. Motor fluctuations and LID are strongly associated phenomena that frequently co-occur in advanced stage PD (30). From a pathophysiological perspective, shared pathological changes affecting the basal ganglia dopaminergic networks may favor both motor fluctuations and LID. In fact, as the nigrostriatal denervation becomes more severe with disease progression, changes in the dopaminergic receptor sensitivity to Levodopa occur, making PD patients more dependent on the pulsatile oral administration of the drug. This may lead to motor fluctuations with sudden transitions from “on” and “off” states, which are typically counteracted by an increase of the oral Levodopa intake that may in turn favor the development of LID (4,26).

The role of patients with baseline motor complications in model training

A critical finding of our ablation study was that the inclusion of patients with pre-existing motor complications in the training set is essential for model sensitivity, even when the clinical aim is to predict new onset in patients currently free of these complications.

The dissociation observed between preserved discrimination (AUC-ROC, which reflects the model's intrinsic ranking ability independent of threshold) and collapsed decision-making (MCC) provides insight into the underlying learning mechanism. The cohort without baseline motor complications provided sufficient information for the models to learn the relative risk ranking but lacked the examples necessary to scale these predictions into high probabilities.

In the absence of patients with evident motor complications acting as high-confidence reference cases, the models learned a compressed probability distribution. This uncertainty was partly driven by an artificially lowered baseline prevalence; excluding prevalent cases skewed the study cohort and reduced the overall probability of a positive outcome. To account for this lower prevalence and the subtler clinical presentation of newly incident cases, models adjusted their overall risk estimates downwards. Consequently, even the highest-risk patients received conservative scores that rarely crossed the clinical threshold for detection.

This suggests that patients with pre-existing complications serve as anchor examples, establishing the upper bound of the risk probability scale. Without these extremes the model lacks a reference for high-confidence prediction. This dependency was observed consistently across model architectures: no single algorithm demonstrated immunity to the loss of these reference cases. These results support the intuition that to confidently detect the subtle signs of onset in patients free from motor complications, model training benefits from including the full spectrum of disease severity, even advanced cases, to establish a robust probabilistic frame of reference.

Strengths

Identifying individuals at risk of developing motor complications is crucial to improve quality of life, prevent disability, and tailor both pharmacological and non-pharmacological management of motor and non-motor symptoms in PD (31). Although several risk factors for LID and motor fluctuations are well-documented, a comprehensive tool for systematic risk stratification is still lacking. To date, ML efforts to model these predictors in PD have remained largely exploratory (9).

In this multicentric longitudinal cohort study, we address this gap by introducing a robust ML workflow designed to develop and stress-test predictive classifiers for early onset of LID and motor fluctuations in PD. Our findings align with existing evidence on risk factors for motor fluctuations, while providing a more comprehensive, multidimensional assessment of the contribution of individual clinical and demographic variables. Beyond standard performance metrics, the proposed methodology prioritizes clinical utility by stress-testing the workflow for model instability and miscalibration—two key barriers to clinical translation. By integrating SHAP-based explainability and evaluating performance on a clinically relevant test population, the workflow provides a transparent and unbiased assessment of future risk. This design ensures that the model's discriminative power reflects true predictive utility for new-onset complications, supporting clinical decision-making without overinflating performance.

Limitations

Our study has some limitations. First, although we leveraged two independently curated, prospectively collected datasets to ensure high-quality data, the rigorous cleaning process required to maintain methodological homogeneity over the three-year follow-up reduced the effective sample size. Second, while our multicenter approach is a strength, we acknowledge that international heterogeneity, including variations in drug availability, treatment habits, and socio-economic factors, may influence the development of motor complications; consequently, validation in larger and more geographically diverse cohorts is needed to confirm the generalizability of our findings. Third, due to regulatory and ethical restriction, genetic data were available for only a minority of patients and were excluded to preserve statistical power. Fourth, missing baseline body weight for a substantial proportion of the cohort precluded the calculation of weight-adjusted LEDDs. Finally, information on gastrointestinal dysfunction was modeled as part of a pooled variable with other autonomic symptoms. Although this was necessary to preserve sample size, gastrointestinal-specific disturbances (e.g. delayed or erratic gastric emptying, constipation and other symptoms) may selectively impact Levodopa pharmacokinetics and therefore motor fluctuations and LID. While the autonomic dysfunction proxy captures general dysautonomia, future research should disentangle these features to isolate the specific predictive contribution of gastrointestinal symptoms on motor outcomes.

References

1. Freitas ME, Hess CW, Fox SH. Motor complications of dopaminergic medications in Parkinson's disease. Semin Neurol. 2017 Apr;37(2):147–57.

2. Kim HJ, Mason S, Foltynie T, Winder-Rhodes S, Barker RA, Williams-Gray CH. Motor complications in Parkinson's disease: 13-year follow-up of the CamPaIGN cohort. Mov Disord. 2020 Jan;35(1):185–90.

3. Anderson E, Nutt J. The long-duration response to levodopa: phenomenology, potential mechanisms and clinical implications. Parkinsonism Relat Disord. 2011 Sept 1;17(8):587–92.

4. Cilia R, Akpalu A, Sarfo FS, Cham M, Amboni M, Cereda E, et al. The modern pre-levodopa era of Parkinson's disease: insights into motor complications from sub-Saharan Africa. Brain. 2014 Oct;137(Pt 10):2731–42.

5. Huang Q, Weng H, Weng X, Zhao ZH. Selected risk factors for motor fluctuations in Parkinson's disease: a meta-analysis. Clin Neurol Neurosurg. 2025 Aug;255(108921):108921.

6. Sharma JC, Bachmann CG, Linazasoro G. Classifying risk factors for dyskinesia in Parkinson's disease. Parkinsonism Relat Disord. 2010 Sept;16(8):490–7.

7. Péchevis M, Clarke CE, Vieregge P, Khoshnood B, Deschaseaux-Voinet C, Berdeaux G, et al. Effects of dyskinesias in Parkinson's disease on quality of life and health-related costs: a prospective European study. Eur J Neurol. 2005 Dec 1;12(12):956–63.

8. Aradi SD, Hauser RA. Medical management and prevention of motor complications in Parkinson's disease. Neurotherapeutics. 2020 Oct;17(4):1339–65.

9. Loo RTJ, Tsurkalenko O, Klucken J, Mangone G, Khoury F, Vidailhet M, et al. Levodopa-induced dyskinesia in Parkinson's disease: Insights from cross-cohort prognostic analysis using machine learning. Parkinsonism Relat Disord. 2024 Sept 1;126(107054):107054.

10. Martorell-Marugán J, Carmona-Sáez P, Chierici M, Bandres-Ciga S, Jurman G. Machine Learning applications in the study of Parkinson's disease: A systematic review. Curr Bioinform. 2023 Aug 1;18(7):576–86.

11. Loo RTJ, Pavelka L, Mangone G, Khoury F, Vidailhet M, Corvol JC, et al. Interpretable machine learning for cross-cohort prediction of motor fluctuations in Parkinson's disease. Mov Disord. 2025 Aug;40(8):1604–17.

12. Malaguti MC, Gios L, Giometto B, Longo C, Riello M, Ottaviani D, et al. Artificial intelligence of imaging and clinical neurological data for predictive, preventive and personalized (P3) medicine for Parkinson Disease: The NeuroArtP3 protocol for a multi-center research study. PLoS One. 2024 Mar 14;19(3):e0300127.

13. Postuma RB, Berg D, Stern M, Poewe W, Olanow CW, Oertel W, et al. MDS clinical diagnostic criteria for Parkinson's disease: MDS-PD Clinical Diagnostic Criteria. Mov Disord. 2015 Oct;30(12):1591–601.

14. Jankovic J. Motor fluctuations and dyskinesias in Parkinson's disease: clinical manifestations. Mov Disord. 2005;20 Suppl 11(S11):S11-6.

15. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Min. 2023 Feb 17;16(1):4.

16. Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: *Advances in Large Margin Classifiers*. London, England: MIT Press; 1999.

17. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019 Dec 16;17(1):230.

18. Lundberg S, Lee SI. A unified approach to interpreting model predictions [Internet]. arXiv [cs.AI]. 2017 [cited 2026 Feb 27]. Available from: http://dx.doi.org/10.5555/3295222.3295230

19. Warren Olanow C, Kieburtz K, Rascol O, Poewe W, Schapira AH, Emre M, et al. Factors predictive of the development of Levodopa-induced dyskinesia and wearing-off in Parkinson's disease: Risk Factors forL-Dopa Induced Motor Complications. Mov Disord. 2013 July 1;28(8):1064–71.

20. Eusebi P, Romoli M, Paoletti FP, Tambasco N, Calabresi P, Parnetti L. Risk factors of levodopa-induced dyskinesia in Parkinson's disease: results from the PPMI cohort. NPJ Parkinsons Dis. 2018 Nov 16;4(1):33.

21. Scott NW, Macleod AD, Counsell CE. Motor complications in an incident Parkinson's disease cohort. Eur J Neurol. 2016 Feb 1;23(2):304–12.

22. Conti V, Izzo V, Russillo MC, Picillo M, Amboni M, Scaglione CLM, et al. Gender differences in levodopa pharmacokinetics in levodopa-naïve patients with Parkinson's disease. Front Med (Lausanne). 2022 May 31;9:909936.

23. Schrag A, Quinn N. Dyskinesias and motor fluctuations in Parkinson's disease. A community-based study. Brain. 2000 Nov;123 (Pt 11):2297–305.

24. Kostic V, Przedborski S, Flaster E, Sternic N. Early development of levodopa-induced dyskinesias and response fluctuations in young-onset Parkinson's disease. Neurology. 1991 Feb 1;41(2 (Pt 1)):202–5.

25. Grandas F, Galiano ML, Tabernero C. Risk factors for levodopa-induced dyskinesias in Parkinson's disease. J Neurol. 1999 Dec;246(12):1127–33.

26. de la Fuente-Fernández R, Sossi V, Huang Z, Furtado S, Lu JQ, Calne DB, et al. Levodopa-induced changes in synaptic dopamine levels increase with progression of Parkinson's disease: implications for dyskinesias. Brain. 2004 Dec 25;127(Pt 12):2747–54.

27. Chung SJ, Lee HS, Yoo HS, Lee YH, Lee PH, Sohn YH. Patterns of striatal dopamine depletion in early Parkinson disease: Prognostic relevance: Prognostic relevance. Neurology. 2020 July 21;95(3):e280–90.

28. Bjornestad A, Forsaa EB, Pedersen KF, Tysnes OB, Larsen JP, Alves G. Risk and course of motor complications in a population-based incident Parkinson's disease cohort. Parkinsonism Relat Disord. 2016 Jan 1;22:48–53.

29. Hung AY, Schwarzschild MA. Treatment of Parkinson's disease: what's in the non-dopaminergic pipeline? Neurotherapeutics. 2014 Jan;11(1):34–46.

30. Aslam S, Manfredsson F, Stokes A, Shill H. "Advanced" Parkinson's disease: A review. Parkinsonism Relat Disord. 2024 June;123(106065):106065.

31. de Bie RMA, Katzenschlager R, Swinnen BEKS, Peball M, Lim SY, Mestre TA, et al. Update on treatments for Parkinson's disease motor fluctuations - an International Parkinson and Movement Disorder Society Evidence-Based Medicine review. Mov Disord. 2025 May;40(5):776–94.

Supplementary Materials

	Future LID			Future motor fluctuations		
Attribute	**Negative**	**Positive**	**P-value**	**Negative**	**Positive**	**P-value**
Use of DA (No) N (%)	48 (31.8%)	26 (27.1%)	0.519	41 (30.1%)	33 (29.2%)	0.921
Anxiety (Absent) N (%)	119 (78.8%)	67 (69.8%)	0.147	102 (76.1%)	84 (74.3%)	0.861
Cognitive status (Normal) N (%)	134 (88.7%)	83 (86.6%)	0.635	119 (88.8%)	98 (86.7%)	0.938
Cognitive status (MCI) N (%)	10 (6.6%)	9 (9.4%)		11 (8.2%)	8 (7.1%)	
Cognitive status (Dementia) N (%)	3 (2.0%)	3 (3.1%)		3 (2.2%)	3 (2.7%)	
Depression (Absent) N (%)	107 (70.9%)	70 (72.9%)	0.838	84 (62.7%)	93 (82.3%)	0.001
Presence of dysautonomia (Absent) N (%)	92 (60.9%)	67 (69.8%)	0.166	82 (61.2%)	77 (68.1%)	0.006
Presence of dysautonomia (Other) N (%)	48 (31.8%)	20 (20.8%)		46 (34.3%)	22 (19.5%)	
Presence of dysautonomia (Orthostatic hypotension) N (%)	11 (7.3%)	9 (9.4%)		6 (4.5%)	14 (12.4%)	
Years since diagnosis (0-5) N (%)	106 (70.2%)	39 (40.6%)	< 0.001	96 (71.6%)	49 (43.4%)	< 0.001
Years since diagnosis (5-10) N (%)	34 (22.5%)	29 (30.2%)		29 (21.6%)	34 (30.1%)	
Years since diagnosis (>10) N (%)	11 (7.3%)	28 (29.2%)		9 (6.7%)	30 (26.5%)	
LID (Absent) N (%)	151 (100.0%)	62 (64.6%)	< 0.001	132 (97.8%)	82 (72.6%)	< 0.001
Falls (Absent) N (%)	136 (90.1%)	80 (83.3%)	0.174	120 (89.6%)	96 (85.0%)	0.372
Family history (for PD) (No) N (%)	101 (66.2%)	67 (69.8%)	0.657	83 (61.9%)	84 (74.3%)	0.053
Freezing (Absent) N (%)	137 (90.7%)	79 (82.3%)	0.079	123 (91.8%)	93 (82.3%)	0.040
Gender (Male) N (%)	100 (66.2%)	51 (53.1%)	0.054	86 (64.2%)	65 (57.5%)	0.348
H&Y stage 2 N (%)	76 (50.3%)	52 (54.2%)	0.179	70 (52.2%)	58 (51.3%)	0.098
H&Y stage 1 N (%)	53 (35.1%)	24 (25.0%)		47 (35.1%)	30 (26.5%)	
H&Y stage >2 N (%)	22 (14.6%)	20 (20.8%)		17 (12.7%)	25 (22.1%)	
Hyposmia N (%)	73 (48.3%)	64 (66.7%)	0.005	64 (47.8%)	73 (64.6%)	0.006
ICD N (%)	142 (94.0%)	85 (88.5%)	0.192	125 (93.3%)	102 (90.3%)	0.527
Use of MAOIs (N) (%)	59 (39.1%)	41 (42.7%)	0.664	43 (32.1%)	57 (50.4%)	0.005
Use of Levodopa N (%)	61 (40.4%)	14 (14.6%)	< 0.001	61 (45.5%)	14 (12.4%)	< 0.001
Motor fluctuations (Absent) N (%)	139 (92.1%)	49 (51.0%)	< 0.001	134 (100.0%)	54 (47.8%)	< 0.001
Phenotype (Akinetic) N (%)	56 (37.1%)	46 (47.9%)	0.121	51 (38.1%)	51 (45.1%)	0.320
Sleep/Wake disorders (Absent) N (%)	84 (55.6%)	62 (64.6%)	0.207	73 (54.5%)	73 (64.6%)	0.138

Age at baseline Median [Q1, Q3]	70 [63, 75]	68 [61.75, 73]	0.061	70 [64.25, 75]	67 [61, 73]	0.024
DA use, duration Median [Q1, Q3]	0 [0, 3]	3 [0, 7]	< 0.001	0 [0, 3]	3 [0, 7]	< 0.001
Education Median [Q1, Q3]	11.5 [8, 13]	9 [8, 13]	0.407	13 [8, 13]	10 [8, 13]	0.212
IMAOYears Median [Q1, Q3]	0 [0, 1]	0 [0, 2]	0.020	0 [0, 1]	0 [0, 2]	0.174
Levodopa use, duration Median [Q1, Q3]	0 [0, 2]	4 [2, 8]	< 0.001	0 [0, 2]	3.5 [2, 7]	< 0.001
MDS-UPDRS III (ON) Median [Q1, Q3]	19 [11, 29]	16 [12, 21.75]	0.081	19 [12.5, 29]	16 [11, 22]	0.032
YearsSymptomsDiagnosis Median [Q1, Q3]	1 [0, 2]	1 [0, 2]	0.952	1 [0, 2]	1 [0, 2]	0.775
LEDD Median [Q1, Q3]	200 [0, 300]	400 [275, 650]	< 0.001	150 [0, 300]	400 [250, 625]	< 0.001

Table S1: Descriptive statistics of baseline attributes and statistical significance with respect to the outcome. Baseline characteristics stratified by future outcome. Continuous variables are presented as median [Q1, Q3] and were compared using the Mann-Whitney U test. Categorical variables are presented as count (percentage of group total) and were compared using Pearson's chi-square test. For multi-level categorical variables, the p-value tests the overall distribution. P-values <0.001 indicate strong statistical evidence. Statistical tests were performed using complete-case analysis, excluding missing values. DA: dopamine agonist; H&Y: Hoehn and Yahr stage; ICD: impulse control disorder; LEDD: levodopa equivalent daily dose; LID: Levodopa-induced dyskinesia; MAOIs: monoamine oxidase inhibitors.

Model	Hyperparameter	Search Space / Values
Random Forest	max_depth	1 to 110 (step 1)
	min_samples_split	2 to 30 (step 1)
	max_features	log2, sqrt
	n_estimators	100 to 1000 (step 5)
	class_weight	balanced, balanced_subsample
Extra Trees	n_estimators	1 to 800 (step 5)
	max_depth	1 to 40 (step 1)
	max_features	1 to 40 (step 1)
	class_weight	balanced, balanced_subsample
XGBoost	learning_rate	0.01 to 0.1 (Uniform)
	subsample	0.4 to 1.0 (Uniform)
	n_estimators	50 to 250 (Integer)
	max_depth	1 to 20 (step 1)
	gamma	0 to 0.5 (Uniform)
	scale_pos_weight	1, 2, 3, 4, 5, 6, 7, 8, 9
Logistic Regression	C	0.001 to 1000 (Log-Uniform)
	max_iter	500 to 1000 (step 1)
	fit_intercept	True, False
	class_weight	balanced
SVC	C	0.0001 to 10000 (Log-Uniform)
	kernel	linear, poly, rbf, sigmoid
	degree	2, 3, 4
	gamma	0.001 to 1.0 (Uniform)
	coef0	0, 0.5, 1
	shrinking	True, False
	class_weight	balanced

Table S2. Summary of the hyperparameter search spaces and sampling strategies used for model optimization. Ranges for continuous variables are specified with their respective distribution types (Uniform or Log-Uniform), while discrete parameters include the step size or the set of specific categorical values evaluated during the tuning process.

Train cohort composition		All		Non dyskinetic only	All (r. s.)	
Test cohort composition		All	Non dyskinetic only	Non dyskinetic only	All (r. s.)	Non dyskinetic only (r. s.)
Model	Metric					
ETC	AUC	0.77 ± 0.06	0.70 ± 0.08	0.67 ± 0.08	0.78 ± 0.06	0.71 ± 0.08
	F1	0.59 ± 0.07	0.40 ± 0.10	0.11 ± 0.12	0.60 ± 0.08	0.43 ± 0.16
	MCC	0.42 ± 0.11	0.24 ± 0.14	0.03 ± 0.15	0.45 ± 0.11	0.29 ± 0.18
	Sens.	0.51 ± 0.08	0.33 ± 0.10	0.07 ± 0.09	0.50 ± 0.10	0.35 ± 0.15
	Spec.	0.87 ± 0.07	0.87 ± 0.07	0.95 ± 0.05	0.90 ± 0.05	0.90 ± 0.05
LR	AUC	0.78 ± 0.06	0.70 ± 0.07	0.69 ± 0.07	0.78 ± 0.06	0.71 ± 0.09
	F1	0.57 ± 0.08	0.39 ± 0.11	0.21 ± 0.12	0.56 ± 0.11	0.40 ± 0.14
	MCC	0.41 ± 0.11	0.25 ± 0.13	0.17 ± 0.15	0.39 ± 0.12	0.25 ± 0.15
	Sens.	0.48 ± 0.10	0.33 ± 0.12	0.14 ± 0.08	0.47 ± 0.13	0.34 ± 0.14
	Spec.	0.88 ± 0.07	0.88 ± 0.07	0.96 ± 0.05	0.88 ± 0.05	0.88 ± 0.05
RF	AUC	0.80 ± 0.06	0.73 ± 0.08	0.70 ± 0.07	0.80 ± 0.06	0.73 ± 0.09
	F1	0.62 ± 0.07	0.45 ± 0.10	0.28 ± 0.14	0.66 ± 0.06	0.51 ± 0.11
	MCC	0.43 ± 0.11	0.27 ± 0.14	0.14 ± 0.16	0.47 ± 0.09	0.34 ± 0.14
	Sens.	0.59 ± 0.09	0.43 ± 0.11	0.22 ± 0.11	0.61 ± 0.09	0.48 ± 0.14
	Spec.	0.82 ± 0.08	0.82 ± 0.08	0.89 ± 0.07	0.84 ± 0.07	0.84 ± 0.07
STACKING	AUC	0.79 ± 0.06	0.72 ± 0.08	0.69 ± 0.06	0.79 ± 0.06	0.72 ± 0.09
	F1	0.62 ± 0.07	0.45 ± 0.10	0.23 ± 0.14	0.64 ± 0.07	0.49 ± 0.09
	MCC	0.42 ± 0.12	0.27 ± 0.14	0.12 ± 0.14	0.44 ± 0.10	0.31 ± 0.13
	Sens.	0.59 ± 0.09	0.44 ± 0.12	0.17 ± 0.12	0.59 ± 0.09	0.45 ± 0.12
	Spec.	0.82 ± 0.08	0.82 ± 0.08	0.92 ± 0.06	0.83 ± 0.06	0.83 ± 0.06
SVC	AUC	0.79 ± 0.06	0.70 ± 0.08	0.67 ± 0.08	0.78 ± 0.06	0.70 ± 0.09
	F1	0.61 ± 0.07	0.43 ± 0.11	0.19 ± 0.14	0.60 ± 0.08	0.42 ± 0.11
	MCC	**0.44 ± 0.12**	**0.28 ± 0.14**	**0.14 ± 0.17**	**0.43 ± 0.10**	**0.26 ± 0.14**
	Sens.	**0.54 ± 0.09**	**0.37 ± 0.12**	**0.12 ± 0.10**	**0.52 ± 0.09**	**0.35 ± 0.11**

	Spec.	0.87 ± 0.07	0.87 ± 0.07	0.95 ± 0.05	0.88 ± 0.06	0.88 ± 0.06
VOTING	AUC	0.80 ± 0.06	0.72 ± 0.08	0.68 ± 0.08	0.79 ± 0.06	0.72 ± 0.09
	F1	0.63 ± 0.06	0.45 ± 0.10	0.25 ± 0.14	0.64 ± 0.07	0.48 ± 0.12
	MCC	0.43 ± 0.10	0.27 ± 0.13	0.13 ± 0.16	0.45 ± 0.10	0.31 ± 0.15
	Sens.	0.59 ± 0.09	0.43 ± 0.12	0.18 ± 0.11	0.59 ± 0.09	0.45 ± 0.14
	Spec.	0.82 ± 0.08	0.82 ± 0.08	0.91 ± 0.07	0.84 ± 0.06	0.84 ± 0.06
XGB	AUC	0.77 ± 0.06	0.70 ± 0.08	0.66 ± 0.07	0.78 ± 0.06	0.72 ± 0.09
	F1	0.60 ± 0.08	0.45 ± 0.11	0.13 ± 0.11	0.63 ± 0.07	0.49 ± 0.12
	MCC	**0.39 ± 0.13**	**0.26 ± 0.16**	**0.02 ± 0.11**	**0.43 ± 0.11**	**0.31 ± 0.15**
	Sens.	**0.57 ± 0.09**	**0.44 ± 0.11**	**0.09 ± 0.09**	**0.57 ± 0.10**	**0.45 ± 0.16**
	Spec.	0.81 ± 0.08	0.81 ± 0.08	0.92 ± 0.07	0.84 ± 0.07	0.84 ± 0.07

Table S3. Comparative Performance of Machine Learning Models for Predicting LID Onset. The table displays key performance metrics (AUC, F1-score, MCC, Sensitivity, Specificity) under three distinct experimental conditions: models trained on all patients (first two columns), models directly trained only on patients without LID at baseline (third column) and models trained using a random subset of patients matching in size the cohort of non-dyskinetic patients at baseline (columns four and five). Performance is evaluated on both the entire test set (All) and the clinically relevant subset of patients without LID at baseline (Non dyskinetic only). All values are presented as mean ± standard deviation. Sens.: Sensitivity, Spec.: Specificity, r.s.: random subset.

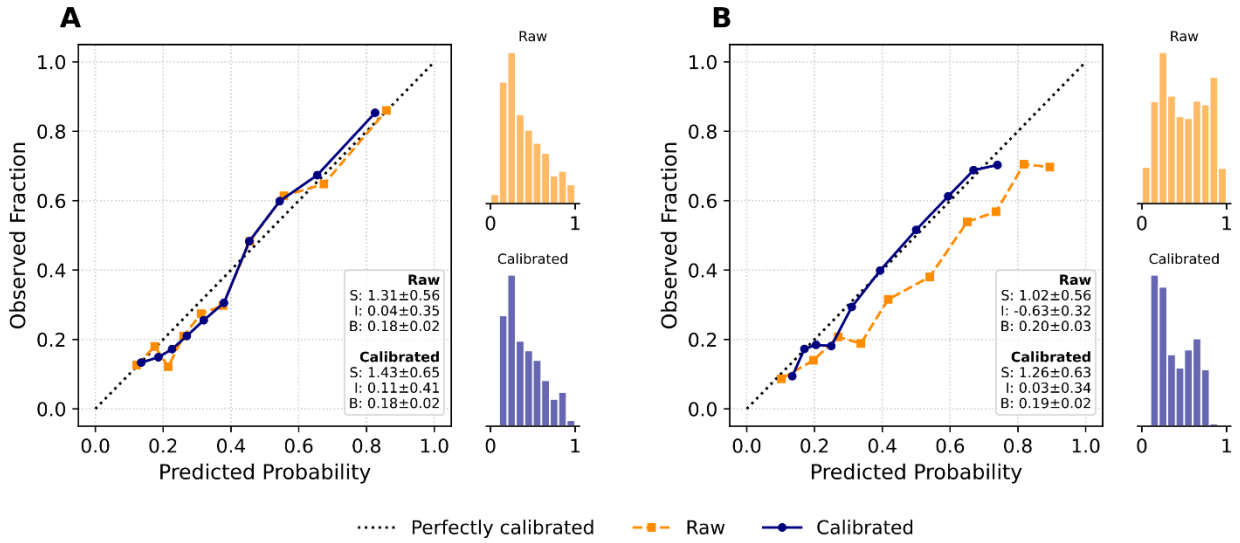

Figure S1. Probability calibration analysis of the optimal predictive models - LID Task. The plots illustrate the alignment between predicted probabilities (x-axis) and observed outcome frequencies (y-axis) for the two best-performing models: (A) Support Vector Classifier (SVC) and (B) Extreme Gradient Boosting (XGB). The orange squares (dashed line) represent the raw, uncalibrated classifier outputs, while the dark blue circles (solid line) depict the posterior probabilities after calibration via Platt Scaling. The diagonal dotted line indicates perfect calibration. Quantile binning (equal sample size per bin) was employed to ensure statistical robustness across the probability range. Inset text reports the quantitative calibration metrics: Slope (S, ideal=1), Intercept (I, ideal=0), and Brier Score (B, lower indicates better accuracy). Marginal histograms display the distribution of predicted probabilities, highlighting the shift from raw scores (orange) to calibrated probabilities (blue). Note that the SVC (A) demonstrates naturally higher resolution (predictions clustered at extremes) compared to the XGB model (B).

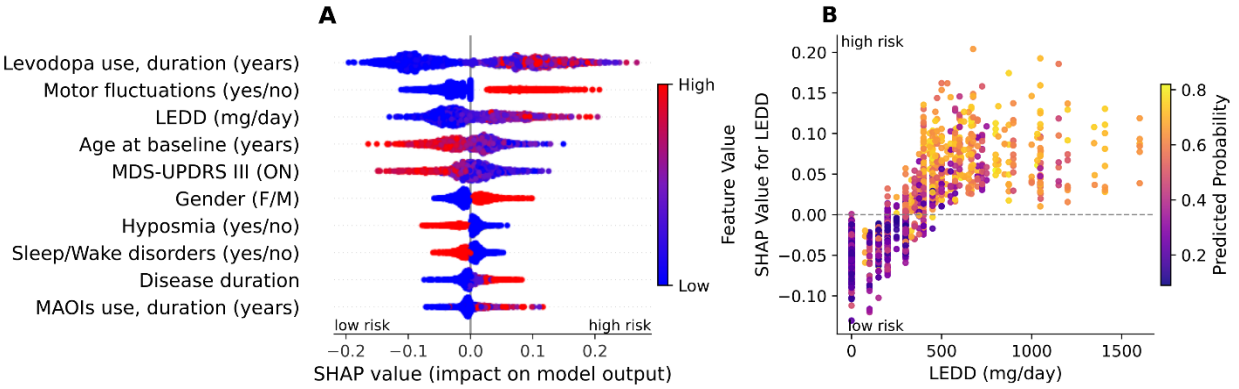

Figure S2. SHAP analysis of XGB for the prediction of future LID. A. Global feature importance and impact for the XGB model revealed by SHAP. Features are ranked in descending order of importance. Each dot represents a patient from the test set. The horizontal axis shows the SHAP value, indicating the feature's contribution to predicting LID (positive values) or absence of LID (negative values) within the third year. Color indicates the feature's original value (blue=low, red=high), revealing the direction and magnitude of its effect. **B.** Detailed analysis of the levodopa equivalent daily dose (LEDD) feature using a SHAP dependence plot for the XGB model. The x-axis plots the patient's actual LEDD on its original scale, while the y-axis plots the SHAP value, indicating the feature's contribution to the prediction. A potential dosage threshold is visible around 300-400 mg, where the feature's impact (SHAP value) transitions from negative (low risk) to positive (high risk). LEDD: levodopa equivalent daily dose; MAOIs: monoamine oxidase inhibitors.

Train cohort composition		All		Non fluctuating only	All (r.s.)	
Test cohort composition		All	Non fluctuating only	Non fluctuating only	All (r.s.)	Non fluctuating only (r.s.)
Model	Metric					
ETC	AUC	0.80 ± 0.06	0.72 ± 0.08	0.68 ± 0.09	0.81 ± 0.06	0.72 ± 0.08
	F1	0.70 ± 0.06	0.49 ± 0.12	0.04 ± 0.11	0.70 ± 0.06	0.49 ± 0.09
	MCC	0.45 ± 0.10	0.28 ± 0.16	-0.01 ± 0.11	0.48 ± 0.11	0.31 ± 0.14
	Sens.	0.69 ± 0.09	0.52 ± 0.15	0.03 ± 0.08	0.70 ± 0.09	0.53 ± 0.16
	Spec.	0.76 ± 0.07	0.76 ± 0.07	0.97 ± 0.06	0.77 ± 0.10	0.77 ± 0.10
LR	AUC	0.82 ± 0.05	0.72 ± 0.09	0.70 ± 0.09	0.83 ± 0.06	0.74 ± 0.09
	F1	0.71 ± 0.05	0.49 ± 0.12	0.24 ± 0.13	0.70 ± 0.06	0.48 ± 0.10
	MCC	0.51 ± 0.08	0.30 ± 0.16	0.16 ± 0.17	0.50 ± 0.10	0.30 ± 0.14
	Sens.	0.67 ± 0.08	0.47 ± 0.14	0.16 ± 0.09	0.66 ± 0.08	0.46 ± 0.12
	Spec.	0.82 ± 0.06	0.82 ± 0.06	0.94 ± 0.06	0.82 ± 0.08	0.82 ± 0.08
RF	AUC	0.81 ± 0.06	0.72 ± 0.09	0.72 ± 0.09	0.83 ± 0.07	0.74 ± 0.10
	F1	0.74 ± 0.06	0.52 ± 0.13	0.18 ± 0.12	0.74 ± 0.06	0.53 ± 0.11
	MCC	0.53 ± 0.11	0.32 ± 0.18	0.10 ± 0.15	0.55 ± 0.12	0.35 ± 0.15
	Sens.	0.73 ± 0.08	0.53 ± 0.15	0.12 ± 0.09	0.74 ± 0.08	0.55 ± 0.15
	Spec.	0.79 ± 0.07	0.79 ± 0.07	0.93 ± 0.06	0.80 ± 0.10	0.80 ± 0.10
STACKING	AUC	0.81 ± 0.05	0.72 ± 0.09	0.70 ± 0.10	0.83 ± 0.06	0.74 ± 0.10
	F1	0.72 ± 0.06	0.50 ± 0.11	0.09 ± 0.12	0.73 ± 0.06	0.51 ± 0.11
	MCC	0.50 ± 0.11	0.29 ± 0.16	0.03 ± 0.16	0.53 ± 0.11	0.34 ± 0.16
	Sens.	0.71 ± 0.07	0.52 ± 0.12	0.06 ± 0.08	0.72 ± 0.07	0.53 ± 0.15
	Spec.	0.78 ± 0.08	0.78 ± 0.08	0.95 ± 0.06	0.80 ± 0.09	0.80 ± 0.09
SVC	AUC	0.81 ± 0.05	0.72 ± 0.08	0.67 ± 0.08	0.83 ± 0.06	0.73 ± 0.09
	F1	0.70 ± 0.06	0.48 ± 0.10	0.14 ± 0.14	0.72 ± 0.09	0.50 ± 0.15
	MCC	0.48 ± 0.11	0.29 ± 0.14	0.10 ± 0.17	0.53 ± 0.14	0.34 ± 0.20
	Sens.	0.67 ± 0.08	0.47 ± 0.11	0.09 ± 0.09	0.68 ± 0.12	0.49 ± 0.18

	Spec.	0.81 ± 0.07	0.81 ± 0.07	0.96 ± 0.06	0.83 ± 0.08	0.83 ± 0.08
VOTING	AUC	0.82 ± 0.05	0.73 ± 0.09	0.70 ± 0.07	0.83 ± 0.06	0.74 ± 0.09
	F1	0.73 ± 0.07	0.52 ± 0.13	0.17 ± 0.14	0.74 ± 0.05	0.51 ± 0.12
	MCC	0.51 ± 0.12	0.32 ± 0.18	0.08 ± 0.15	0.53 ± 0.10	0.33 ± 0.15
	Sens.	0.72 ± 0.09	0.54 ± 0.15	0.12 ± 0.10	0.73 ± 0.07	0.53 ± 0.18
	Spec.	0.79 ± 0.08	0.79 ± 0.08	0.93 ± 0.05	0.80 ± 0.08	0.80 ± 0.08
XGB	AUC	0.79 ± 0.05	0.70 ± 0.08	0.69 ± 0.08	0.80 ± 0.06	0.72 ± 0.10
	F1	0.70 ± 0.06	0.49 ± 0.12	0.06 ± 0.10	0.71 ± 0.06	0.49 ± 0.14
	MCC	0.47 ± 0.10	0.28 ± 0.17	-0.01 ± 0.12	0.49 ± 0.12	0.29 ± 0.18
	Sens.	0.69 ± 0.08	0.51 ± 0.16	0.04 ± 0.08	0.70 ± 0.09	0.52 ± 0.18
	Spec.	0.77 ± 0.08	0.77 ± 0.08	0.95 ± 0.05	0.78 ± 0.10	0.78 ± 0.10

Table S4. Comparative Performance of Machine Learning Models for Predicting Motor Fluctuations Onset. The table displays key performance metrics (AUC, F1-score, MCC, Sensitivity, Specificity) under three distinct experimental conditions: models trained on all patients (first two columns), models directly trained only on patients without fluctuations at baseline (third column) and models trained using a random subset of patients matching in size the cohort of patients without motor fluctuations at baseline (columns four and five). Performance is evaluated on both the entire test set (All) and the clinically relevant subset of patients without motor fluctuations at baseline (Non fluctuating only). All values are presented as mean ± standard deviation. Sens.: Sensitivity, Spec.: Specificity, r.s.: random subset.

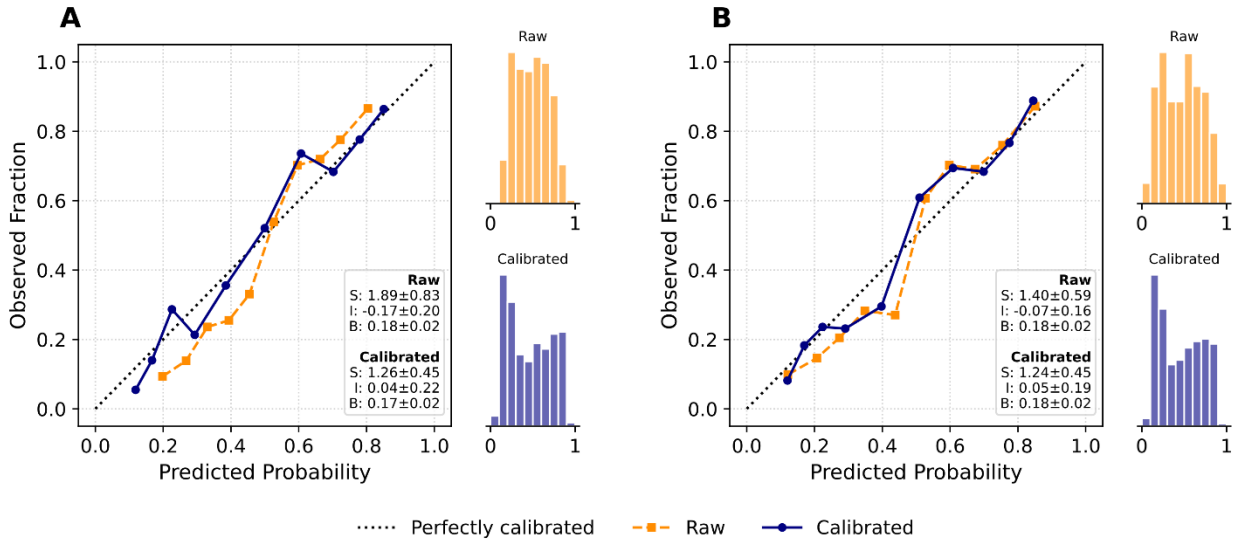

Figure S3. Probability calibration analysis of the optimal predictive models - motor fluctuations task. The plots illustrate the alignment between predicted probabilities (x-axis) and observed outcome frequencies (y-axis) for the two best-performing models: (A) Random Forest and (B) Voting ensemble. The orange squares (dashed line) represent the raw, uncalibrated classifier outputs, while the dark blue circles (solid line) depict the posterior probabilities after calibration via Platt Scaling. The diagonal dotted line indicates perfect calibration. Quantile binning (equal sample size per bin) was employed to ensure statistical robustness across the probability range. Inset text reports the quantitative calibration metrics: Slope (S, ideal=1), Intercept (I, ideal=0), and Brier Score (B, lower indicates better accuracy). Marginal histograms display the distribution of predicted probabilities, highlighting the shift from raw scores (orange) to calibrated probabilities (blue).

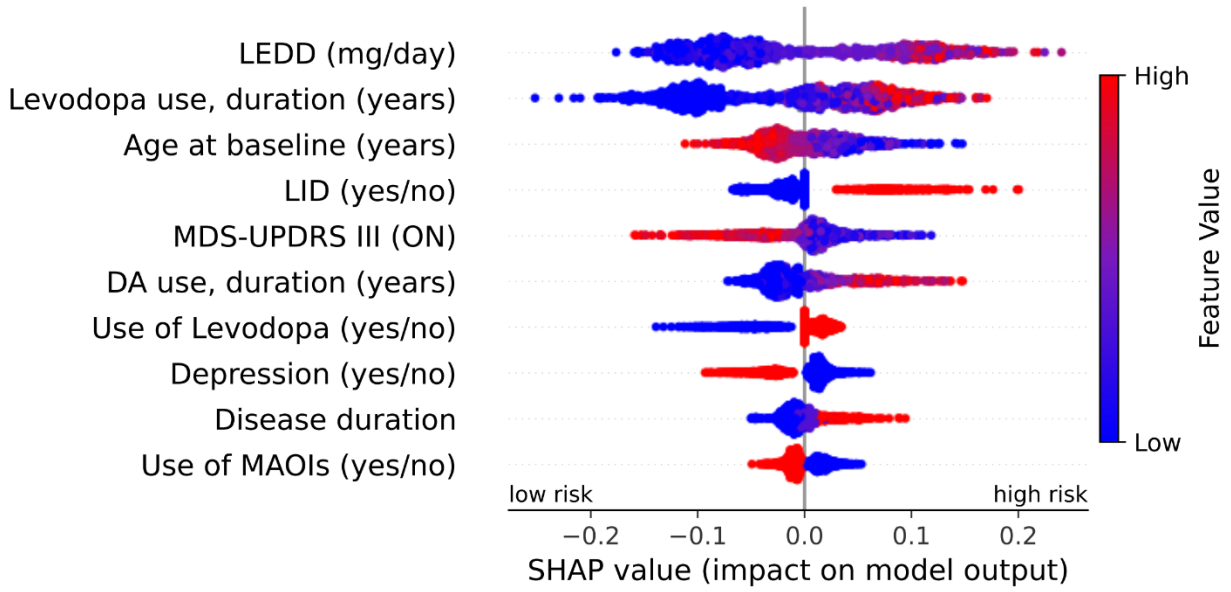

Figure S4. SHAP analysis of Random Forest for the prediction of future motor fluctuations. Global feature importance and impact for the RF model revealed by SHAP. Features are ranked in descending order of importance. Each dot represents a patient from the test set. The horizontal axis shows the SHAP value, indicating the feature's contribution to predicting the presence (positive values) or absence (negative values) of motor fluctuations within the third year. Color indicates the feature's original value (blue=low, red=high), revealing the direction and magnitude of its effect. Legend: DA: dopamine agonist; LEDD: levodopa equivalent daily dose; LID: Levodopa-induced dyskinesia; MAOIs: monoamine oxidase inhibitors.