UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2025-020
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 11 March 2025 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN SCIENCES EXACTES ET NATURELLES

by

Rebecca LOO TING JIIN
Born on 1 July 1989 in Selangor, (Malaysia)

CROSS-COHORT STATISTICAL AND MACHINE LEARNING ANALYSIS OF COMPLICATIONS IN PARKINSON'S DISEASE

Dissertation defence committee

Dr. Enrico Glaab, dissertation supervisor
Assistant Professor, Université du Luxembourg

Dr. Jorge Goncalves, Chairman
Professor, Université du Luxembourg

Dr. Guy Fagherazzi, Vice Chairman
Director of the Department of Precision Health, Luxembourg Institute of Health

Dr. Ramón Díaz-Uriarte
Professor, Universidad Autónoma de Madrid

Dr. Jean-Christophe Corvol
Professor, Paris Brain Institute

UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2025-020

Faculty of Science, Technology and Medicine

DISSERTATION

Presented on 11 March 2025 in Luxembourg
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG
EN SCIENCES EXACTES ET NATURELLES

by

Rebecca LOO TING JIIN
Born on 1 July 1989 in Selangor (Malaysia)

Cross-cohort Statistical and Machine Learning Analysis of Complications in Parkinson's Disease

Affidavit

I hereby confirm that the PhD thesis entitled “Cross-cohort Statistical and Machine Learning Analysis of Complications in Parkinson’s Disease” has been written independently and without any other sources than cited.

Rebecca Loo Ting Jiin
Luxembourg
March 2025

Contents

Abstract

Parkinson's disease (PD) is characterized by a wide range of motor and non-motor symptoms, many of which manifest as complications. These include levodopa-induced dyskinesia (LID), motor fluctuations (MF), mild-cognitive impairment in PD (PD-MCI), and patient-reported cognitive decline (PRCD). These complications significantly impact quality of life, contributing to physical disability, reduced independence, increased healthcare costs and caregiver burden. The progression and occurrence of these complications vary considerably among patients. While some PD patients experience these complications early in their disease course, others may not develop them even after prolonged disease duration or exposure to standard therapies like levodopa.

This clinical heterogeneity may result from genetic mutations, demographic characteristics, and individual disease phenotypes, but the underlying mechanisms remain to be explored. Identifying key predictors of these complications is important to understand the factors driving their variability and to develop precision medicine approaches in PD management. Predictive insights can inform therapeutic decision-making, such as adjusting levodopa dosages, optimizing adjunctive therapies, and developing targeted non-pharmacological interventions to mitigate the risk or severity of these complications.

Despite the clinical significance of complications such as LID, MF, PD-MCI, and PRCD, existing predictive models often lack generalizability and robustness, primarily due to biases inherent in single-cohort analyses. These models may not fully capture the complex relationships between clinical variables and PD complications across diverse populations. A multi-cohort analysis addresses these limitations by integrating data from multiple independent studies. This approach increases statistical power with larger sample sizes, reduces cohort-specific biases, and improves model reliability and generalizability. The consistent identification of predictors across diverse populations is a key strength of this analysis, providing more reliable and clinically applicable insights.

This study used machine learning (ML) frameworks to identify potential LID, MF, PD-MCI, and PRCD predictors using data from three independent longitudinal PD cohorts (LuxPARK, PPMI, and ICEBERG). Cross-study normalization was integrated into the ML workflow to enhance the predictive capability and ensure study consistency. This approach mitigates inter-cohort variability, enabling the detection of reliable predictors across diverse cohorts and minimizing the influence of cohort-specific biases.

Incorporating cross-study normalization, interpretable ML, and leave-one-cohort-out validation has enabled the identification of robust and generalizable predictors of these complications. The findings of this study contribute to the understanding of PD complications while providing insights for early detection, risk stratification, and personalized interventions for patients at risk of or experiencing LID, MF, PD-MCI, and PRCD.

Key predictors for PD complications were identified, highlighting distinct and overlapping factors. LID was positively associated with axial symptoms, freezing of gait, and rigidity, while negatively associated predictors included later disease onset, higher body weight, and better visuospatial ability. Predictors of MF included freezing of gait, axial symptoms, and pathogenic GBA and LRRK2 variants, with tremors and later disease onset inversely associated with its development. For the analysis of PD-MCI and PRCD, older age at PD diagnosis, visuospatial deficits, and non-motor symptoms like autonomic dysfunction emerged as significant predictors. Additionally, sex differences were observed in cognitive outcomes, with women displaying better global cognition and less cognitive interference.

Overall, this study offers interpretable ML models for early risk stratification and personalized interventions targeting motor and cognitive complications in PD. Through robust multi-cohort analyses, complications such as LID, MF, PD-MCI, and PRCD can be predicted earlier in the disease course. These findings support the implementation of personalized approaches, including adjusting levodopa dosage according to individual characteristics, optimizing adjuvant therapies, and targeted cognitive interventions for individuals at higher risk. Consequently, these approaches can contribute to enhanced patient outcomes and improved quality of life.

The analysis enables precision medicine in PD management by identifying associations between predictors and these complications. It enables clinicians to stratify patients by risk, design individualized treatment plans, and potentially delay or prevent the onset of complications, thereby preserving quality of life and reducing the burden of advanced disease. Furthermore, integrating predictive models into digital health tools and electronic medical records is a potential benefit, enhancing clinical workflows and decision-making efficiency.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Enrico Glaab, for his guidance, support, and encouragement throughout my PhD journey. His expertise, constructive feedback, and continuous motivation were invaluable to the success of this research. I am also grateful to my thesis CET committee members, Dr. Guy Fagherazzi and Dr. Roland Krause, for their thoughtful comments and suggestions, which significantly improved this work.

I would also like to thank all my colleagues in the Biomedical Data Science group for their support and insightful discussion. Working within such a collaborative and intellectually stimulating environment has been both productive and rewarding.

I would also like to express my gratitude to all the co-authors of the manuscripts. Their clinical perspective, feedback, and contributions were invaluable. Their collaboration has enriched this research and helped refine the manuscripts. I also appreciate the team involved in the pre-publication check (PPC) for their review and helpful suggestions that have strengthened the overall quality of the manuscripts.

On a personal note, I would like to thank my family and friends for their endless support and encouragement throughout this journey. Their continuous support and confidence in me have been important in maintaining my motivation and strength.

Finally, I would like to express my gratitude to all the study participants, without whom this research would not have been possible. Their contributions are greatly valued to the advancement of understanding in the field of neurodegenerative disorders.

List of abbreviations

AdaBoost	Adaptive Boosting
ADL	Activities of daily living
AUC	Area under the curve
AUNBC	Area under the net benefit curve
C-index	Concordance-index
CART	Classification and Regression Trees
CatBoost	Category Boosting
CI	Confidence intervals
CV	Cross-validation
CW-GBoost	Component-wise Gradient Boosting
DAs	Dopamine agonists
DBS	Deep Brain Stimulation
DCA	Decision curve analysis
FIGS	Fast Interpretable Greedy-Tree Sums
FoG	Freezing of gait
GBM	Gradient Boosting Machines
GBoost	Gradient Boosting
GCP	Good clinical practice
GI	Gastrointestinal
GOSDT	Generalized and Scalable Optimal Sparse Decosopm Trees
GOSDT-GUESSES	Fast-Sparse Decision Tree
H&Y	Hoehn & Yahr
HR	Hazard ratio
HS	Hierarchical Shrinkage
HSD	Honestly Significant Difference
JLO	Judgment of Line Orientation
KM	Kaplan-Meier
LEDD	Levodopa Equivalent Daily Dose
LID	Levodopa-induced dyskinesia
LSVM	Survival Linear Support Vector Machine
LuXPARK	Luxembourg Parkinson's Study
MCMC	Markov chain Monte Carlo

MDS-UPDRS	Movement Disorders Society-Unified Parkinson's Disease Rating Scale
MICE	Multivariate Imputation with Chain-Equation
MF	Motor fluctuations
ML	Machine learning
MoCA	Montreal Cognitive Assessment
MRI	Magnetic resonance imaging
MSE	Mean square error
NLSVM	Naive Survival Linear Support Vector Machine
OSDT	Optimal Sparse Decision Trees
PD	Parkinson's disease
PD-MCI	Mild cognitive impairment in Parkinson'd disease
PET	Positron emission tomography
PPMI	Parkinson's Progression Markers Initiative
PRCD	Patient-reported cognitive decline
RBD	Rapid eye movement sleep behaviour disorder
RF	Random Forest
RFE	Recursive feature elimination
SCOPA-AUT	Scales for Outcomes in Parkinson's Disease - Autonomic Dysfunction
SD	Standard deviation
SHAP	SHapley Additive exPlanations
UKPDSBB	UK Parkinson's Disease Society Brain Bank
XGBoost	Extreme Gradient Boosting

List of Figures

List of Tables

Chapter 1

Introduction

This introductory chapter provides a brief overview of the research topic and its significance, a summary of the research background and motivations, and a brief literature review of the methods used in the thesis.

1.1 Background

The study of Parkinson's disease (PD) complications is important for comprehending and managing this complex neurodegenerative disease. These complications cover a wide range of motor and non-motor symptoms [1], including but not limited to motor impairment, cognitive decline, and psychiatric disorders.

Linear regression has traditionally been used to analyze PD complications [2, 3, 4, 5]. However, these methods may only partially capture the disease's complex and dynamic nature. The linear assumptions inherent in classical models are inadequate in addressing the non-linear relationships and high-dimensional data characteristics of PD outcomes [6, 7].

Machine learning (ML) is a promising alternative for studying PD complications. It can uncover complex patterns and relationships within multidimensional data sets and inherent heterogeneity [8, 9]. ML algorithms have demonstrated a remarkable ability to identify subtle patterns and correlations that traditional methods may not detect [10, 11, 12]. ML facilitates a more comprehensive exploration of the risk factors and mechanisms underlying PD complications by addressing the limitations of traditional statistical approaches [13, 14].

Recent studies have shown that ML is effective in analyzing complications related to PD. ML algorithms have been used to identify new biomarkers associated with specific PD symptoms, which can improve early diagnosis and enable personalized treatment strategies [15, 16]. ML techniques demonstrate potential in predicting disease progression [17] and treatment response [16, 18] by integrating longitudinal patient data [14] or integrating with devices [19] to improve the quality of life of PD patients.

Despite ML's evident advantages, challenges remain in applying it to PD research. The ongoing research concerns integrating heterogeneous data sources and accounting for variability among patients with PD [20]. Despite these challenges, methodological advances in ML continue to improve its utility in analyzing the disease's complex and multifaceted nature [21].

Understanding PD complications through the framework of ML not only advances our knowledge of the disease but also has significant clinical implications. ML-driven insights can improve the development of more effective prognostic tools and personalized treatment interventions, thereby improving the quality of life for individuals with PD [22]. Furthermore, ML in the healthcare industry can save costs by optimizing resource allocation and streamlining patient care processes [23].

Integrating ML into the study of PD complications represents a significant shift in biomedical research. Researchers can use the power of ML to analyze complex data sets, making significant progress in discovering the mechanisms underlying PD complications and developing innovative therapeutic strategies.

1.2 State-of-the-art in Parkinson's research

PD is a complex condition that goes beyond its characteristic motor symptoms. It includes a range of complications that significantly affect patients' well-being [24, 25]. Understanding the diverse clinical profile of PD is important for improving treatment strategies and enhancing the quality of life for those affected. Recent research has revealed the various complications associated with PD, highlighting the need for innovative approaches to manage and mitigate these challenges. Various methodologies have been used to investigate the complex nature of PD complications, including clinical scales and ML. Therefore, it is important to have a comprehensive understanding and management of PD-related complications.

Levodopa is the primary medication for PD due to its effectiveness in alleviating motor symptoms [22, 23, 26]. However, its long-term use may lead to motor complications, such as dyskinesias [1, 27, 28, 29]. Despite being the gold standard treatment [30], further exacerbating the complexity of treatment [31], this potential adverse effect should be noted.

Furthermore, the effectiveness of levodopa may be affected by several factors, such as its absorption in the gastrointestinal tract [30], delayed gastric emptying [32], Helicobacter pylori infection [1, 33], and protein intake [34, 35, 36]. Keun et al. (2021) [34] investigated the influence of dietary factors on levodopa absorption, emphasizing the significance of considering dietary habits in managing PD [37, 38, 39]. Understanding the factors influencing levodopa absorption is important for optimizing treatment outcomes and minimizing adverse effects [40, 41].

In addition, adjunctive therapies, such as dopamine agonists [33] and MAO-B inhibitors [1], are commonly prescribed to complement levodopa therapy and extend its efficacy. Deep brain stimulation (DBS) surgery is a therapeutic option for advanced PD cases [27]. It provides symptomatic relief and improves motor function in selected patients.

Although motor symptoms are important for diagnosis and treatment, they only represent a portion of the challenges experienced by individuals with PD, some of whom may have additional complications. Non-motor symptoms, such as cognitive impairment [42, 43, 44], emotional disturbances [33, 45, 46], autonomic dysfunction [26, 47], and sleep disturbances [31, 45], significantly contribute to the disease burden and greatly impact patients' daily lives.

To comprehensively understand PD and its complications, it is important to consider both motor and non-motor symptoms [30, 48]. Given the heterogeneity of PD manifestations, individual variations in symptomatology and treatment response necessitate personalized therapeutic approaches [15, 22].

The research on PD has evolved and shifted towards advanced methodologies, such as longitudinal

studies, cohort analyses, and ML. These approaches offer insights into the dynamics of disease progression and have the potential to identify predictive biomarkers. Several studies have used ML to accurately categorize stages of PD [17, 49, 50], detect early symptoms through wearable technology [19, 51, 52], and assess levodopa responsiveness [16, 53, 54, 55]. These studies highlight the potential of ML techniques in enhancing the clinical management of PD.

PD progression varies significantly among patients, with some experiencing rapid progression while others remain mildly symptomatic for long periods [5, 56]. Comparative assessments and longitudinal analyses have identified subtypes of PD with varying clinical features and disease courses [3, 57], which can inform treatment strategies adapted to the individual patient [18]. Genetic studies have revealed underlying disease mechanisms [14, 48, 58, 59], which can pave the way for personalized interventions.

Cognitive impairment is a common complication of PD, affecting approximately 80% of patients [60]. ML and statistical models integrating diverse datasets have shown potential in investigating cognitive decline [2, 3, 5, 14, 61, 62], enabling early intervention strategies. Prior studies have emphasized the importance of early detection and targeted therapies to minimize the impact of cognitive decline [43, 57, 63].

Examining the complex connection between emotional distress [64, 65] and sleep disturbances [66, 67, 68] is also important for individuals with PD. These issues are frequently experienced by PD patients and are linked to various complications that significantly affect their quality of life [69, 70]. Prior research has consistently shown links between emotional distress and various complications related to PD, such as dyskinesia, cognitive decline, and gastrointestinal issues [42, 44, 45, 47, 48]. Understanding the relationship between emotional well-being and sleep quality in patients with PD is important for developing targeted interventions and improving overall management strategies for this complex neurological disorder.

The field of PD research is undergoing constant development, incorporating multifaceted approaches to understand the complexities of the disease and develop personalized treatment strategies. By implementing ML algorithms, genetic analyses, and longitudinal investigations, researchers have obtained novel insights into the multifaceted complications of PD, thereby establishing the foundational framework necessary for the development of improved clinical care and patient outcomes. It is imperative to recognize the significance of interdisciplinary collaboration and patient-centered care in addressing the diverse challenges associated with PD. This reaffirms our commitment to advancing knowledge and improving patient well-being.

1.3 Data-driven approaches for the study of clinical data

Data-driven methodologies, including ML algorithms, have emerged as important tools for investigating the complexities of PD. This section discusses the role of data-driven approaches, focusing on ML and cross-cohort analyses.

1.3.1 Classification methods

ML classification methods have emerged as powerful tools in PD research, enabling the identification of patterns and predictors from complex clinical datasets to improve disease understanding and prognosis. Classification and regression trees (CART), also known as decision trees, have been recognized as one

of the leading interpretable ML methods for about 40 years since their introduction in 1984 [71]. The tree-based algorithm is straightforward to comprehend and allows for the discovery of novel data [72]. The CART algorithm is considered 'greedy' because it searches for the best outcome without considering past splits, using either a Gini index or information gain as splitting criteria. Freund, Schapire, and Abe (1999) [73] developed Adaptive Boosting (AdaBoost), which uses a boosting technique to transform weak learners into more robust classifiers to improve prediction accuracy and reduce overfitting. AdaBoost can detect outliers based on the weight of a small to moderate number of outliers and is known for its resistance to overfitting.

Several years later, Friedman (2001) [74] developed a more robust algorithm called Gradient Boosting (GBoost) using a general gradient descent "boosting" approach to minimize the loss function across sequential models. Chen and Guestrin (2016) [75] introduced a variant of the GBoost framework that incorporates regularization to penalize the model's complexity, enables parallelization, and combines weak learners that outperform random guessing to form a stronger learner. The algorithm is named Extreme Gradient Boosting (XGBoost), and it is labeled for its robustness and speed, running over ten times faster than GBoost in scikit-learn. CatBoost is a GBoost framework extension developed by Prokhorenkova et al. (2018) [76]. As its name suggests, it is a new algorithm designed to handle categorical features directly. It implements ordered boosting, a modification of the GBoost algorithm, which outperforms XGBoost.

Hu, Rudin, and Seltzer (2019) [77] proposed Optimal Sparse Decision Trees (OSDT) as a viable option for binary predictors, promoting computational efficiency and effective search space pruning. Lin et al. (2020) [78] improved the algorithm with dynamic programming, resulting in Generalized and Scalable Optimal Sparse Decision Trees (GOSDT). Although GOSDT has shown instability in optimizing continuous variables and lacks robustness for imbalanced data, McTavish et al. (2022) [79] improved its performance by incorporating "guess" information derived from black-box models. This modification improved prediction performance and speed by limiting the search space with bounds. This extended version of GOSDT is called Fast Sparse Decision Tree (GOSDT-GUESSES). It can achieve accuracy levels comparable to, or even better than, black box tree ensembles.

In 2022, Tan et al. (2022) [72] proposed Fast Interpretable Greedy-Tree Sums (FIGS), an advancement in tree-based algorithms. FIGS allows additive structure adaptation to limit the number of trees by using a predefined threshold. This algorithm reduces the complexity and computational costs of the tree while still preserving prediction performance. Agarwal et al. (2022) [80] introduced the post-hoc algorithm, Hierarchical Shrinkage (HS), which regularizes trees by shrinking predictions over each node towards the sample means of its ancestor without modifying the tree structure. HS can be applied to any tree-based algorithm to simplify and stabilize the tree, leading to improved prediction performance with a less complex model. Additionally, HS has been shown to improve the interpretability of the model.

1.3.2 Time-to-event analysis

Time-to-event analysis is a widely used method in biomedical research to investigate the effects of clinical factors on the occurrence of events, such as complications of PD. The Cox proportional hazards regression model [81] is commonly used to examine the relationship between covariates and time-to-event data with censoring. However, this method may have limitations when scaling with high-dimensional data [8] and when the proportional hazards assumption is violated. To overcome these limitations, ML techniques have been extended to adapt to high-dimensional censored data and achieve more accurate

predictions of the time from the baseline visit until a patient with PD develops complications based on associated clinical risk factors [7].

Dealing with censored data is one of the challenges in time-to-event analysis. Censored data arises when individuals do not experience the event of interest before the end of the study, withdraw early, or are lost to follow-up, resulting in incomplete observation. This incomplete data, known as right censoring, means that the actual time until the event occurs is still being determined and is not earlier than the observed time for these PD patients. To avoid biased inferences, it is important to consider censored subjects. Therefore, time-to-event methods were designed to handle censored data.

Penalized Cox regression, also known as Cox model regularization, is a technique that applies regularization to improve the accuracy of the Cox model and control model complexity [82]. The two most common regularization techniques are $L1$ regularization (Lasso) and $L2$ regularization (Ridge regression). Lasso tends to keep the coefficient of one predictor among highly correlated predictors while shrinking other coefficients toward zero [83]. Ridge regression assigns similar weights to highly correlated predictors, thereby regularizing the model in cohorts where $p > n$, with p being the number of predictors and n being the sample size of the cohort. Simon et al. (2011) [84] introduced the elastic net penalty, a mixture of $L1$ and $L2$ penalties that combines both strengths to generalize the model.

Tree-based ML approaches have been extended by reformulating CART [85] as a survival tree to analyze censored data [86] and relax the assumptions of classical time-to-event methods. The tree is grown by recursively splitting homogeneous subgroups to maximize the differences in the survival distribution of the subgroups [7]. Randomized Survival Trees, also known as Extremely Randomized Trees or Extra Survival Trees, were proposed by explicitly using entirely random cut-points to grow the trees and combining ensemble averaging to reduce variance [87]. Ishwaran et al. (2008) [85] transformed Random Forest (RF), which combines the predictions from each tree [88] to create a powerful predictive technique for adapting censored data. The robustness of this technique has gained popularity in the field of time-to-event analysis [89].

Gradient boosting machines (GBM) is an ensemble ML method that uses boosting techniques to solve complex relationships between covariates [90]. GBM optimizes the model by building sequential prediction models to achieve better predictive performance while considering censoring in clinical research [91]. Component-wise Gradient Boosting (CW-GBoost) is implemented within GBM to optimize the predictive model [92].

The Survival Linear Support Vector Machine (LSVM) is another popular ML algorithm for time-to-event analysis. The algorithm captures complex and non-linear relationships between predictors and outcomes by maximizing concordance pairs [7, 82]. In this thesis, a simplified version of this method referred to as the naive LSVM (NLSVM), serves as a fundamental approach to LSVM. LSVM emphasizes the computational efficiency aspect of the Survival LSVM, an extension of the NLSVM algorithm.

1.3.3 Cross-cohort analysis

Nowadays, access to large-scale cohorts has become more flexible, encouraging data-driven research from multiple sources. Cross-cohort studies are observational research designs that combine distinct cohorts from different study populations [20]. This research can be conducted using existing cost-effective cohorts, leading to higher statistical power [93, 94], more robust inference, and increased prediction accuracy [95]. Additionally, increasing the sample size can significantly improve the reliability of the conclusions, resulting in more stable results [20]. In the context of complications prognosis in PD,

a cross-cohort study allows for exploring potential risk factors by examining differences and similarities between cohorts. This can ultimately deepen our understanding of PD complications.

Differences in characteristics or exposures across various cohorts from different sites may result in systematic biases, which offset the benefits of enlarging sample size in cross-cohort studies [94]. This bias, known as the "batch effect", leads to inhomogeneities within datasets and may cause features that are not directly comparable [93]. Samples collected at different times or sites may show systematic dissimilarities unrelated to biological differences, such as uncontrollable variations in qualitative measurement across conditions. Therefore, it is important to consider these factors when analyzing the data to ensure accurate results.

In computational biology, integrating multiple cohorts can be challenging due to "batch effects" [95]. It is recommended that data be combined with appropriate methods to adjust for batch effects [96]. Without such adjustments, the results may be misleading, and the data's conclusions may be biased [97]. Additionally, the predictive performance may be unsatisfactory [93]. Therefore, when dealing with such variation, it is recommended that the "batch effect" be adjusted for in cross-cohort studies before proceeding with further analysis [98].

Cross-study normalization was introduced as a data preprocessing technique to mitigate the "batch effects" in a cross-cohort study [93]. The main goal is to compare cohort data [99] by aligning data distribution across studies [20] and improving prediction performance [100]. While the implementation of normalization techniques does improve precision measurement and prediction performance in cross-cohort analyses [101], this efficacy may not be consistently replicable in real-world settings [94].

1.4 Motivation

A comprehensive understanding of the risk factors associated with PD complications is necessary to improve patient care and outcomes. However, identifying and comprehending these risk factors is challenging due to the complexity and diversity of PD-related problems and the heterogeneity of individual characteristics.

Cross-cohort studies, which combine data from multiple cohorts, play a role in addressing these challenges. This methodological approach allows for integrating more extensive and diverse datasets, thereby facilitating a comprehensive investigation of PD-related problems in diverse patient populations.

Nevertheless, disparities in the duration of follow-up and observation periods across different groups can result in a premature withdrawal from the study, thereby introducing bias into the analysis. The variability in outcomes among different groups can complicate the analysis and interpretation of data. Cross-cohort analysis, in which non-shared characteristics are excluded, can result in a limited understanding of PD associated problems.

Despite the challenges associated with analyzing data from multiple sources and applying ML algorithms, there is considerable potential for improving our understanding of PD-related issues. ML algorithms have the capability to analyze extensive clinical data sets, revealing complex patterns and connections that may not be identifiable through traditional statistical methods. By integrating the strengths of these approaches, we can identify risk factors associated with complications in PD, develop models to predict risk for specific outcomes and implement personalized treatment plans.

Chapter 2

Aims and scope of the thesis

2.1 Novel aspects of this work

This thesis explores the challenges related to PD-associated complications and develops predictive models for personalized risk assessment. To achieve this goal, four main research paths will be explored:

1. Integrate cross-cohort analysis and ML techniques
Develop a comprehensive framework integrating cross-cohort analysis and ML techniques to analyze diverse clinical data from multiple PD cohorts. The process involves standardizing and correcting biases specific to each group of individuals, thereby enabling reliable comparisons between different groups and promoting a comprehensive analysis and understanding of the results.

2. Identify key risk factors associated with complications of PD
The analysis aims to identify and rank the primary risk factors associated with complications of PD in individuals from multiple cohorts. The study examines the influence of demographic characteristics, disease severity, evaluations of motor and non-motor functions, and other clinical variables shared among different cohorts. Its goal is to identify persistent patterns and relationships to understand the clinical predictors associated with complications in PD.

3. Develop predictive models
ML techniques are used to create prediction models for evaluating the likelihood of PD complications in particular patients. These models can incorporate various clinical and demographic factors to offer personalized risk evaluations, allowing for early intervention and customized treatment approaches. The study aims to improve the accuracy and generalizability of the models by using ML techniques.

4. Validate findings across cohorts
We assess the relevance and robustness of significant findings and predictive models across multiple cohorts. This involves verifying results and models in diverse participant groups to ensure their robustness and generalizability, which can be achieved through cross-validation techniques. Our goal is to confirm the reliability and applicability of our findings by combining data from diverse cohorts. This will ensure that our conclusions are not specific to any single cohort.

2.2 Thesis structure

The structure of this thesis is outlined as follows:

Chapter 1: Introduction
The introductory chapter presents an overview of the role of ML in analyzing complications related to PD, the state of the art in PD research, and the rationale for incorporating cross-cohort analysis and ML approaches into the study of PD complications.

Chapter 2: Aims and scope of the thesis
This chapter outlines the research aims and objectives, exploring PD-related complications. The chapter describes the scope of the thesis and highlights potential contributions through publication in peer-reviewed journals and poster presentations.

Chapter 3: Cohorts and methods
This chapter outlines the cohorts used in the study, including clinical records, rating scales, and the inclusion criteria applied to select participants for the study of PD-related complications. The strategies used for inclusion criteria, data preparation, and analysis are well illustrated, covering statistical methods and ML algorithms.

Chapter 4: Levodopa-induced dyskinesia in Parkinson's disease: Insights from cross-cohort prognostic analysis using machine learning
This chapter comprehensively analyzes levodopa-induced dyskinesia, covering its underlying mechanisms, observable symptoms, and associated predictors. The study's results and its correlation with demographic characteristics, disease severity, and clinical symptoms are discussed. The implications of these findings for managing this issue in PD patients are also considered.

Chapter 5: Interpretable machine learning for cross-cohort prediction of motor fluctuations in Parkinson's disease
This chapter examines the occurrence of motor fluctuations in PD, including the effects of levodopa treatment and dosages. The chapter presents research findings on motor fluctuations, including predictive models and identified predictors. It also analyzes the implications of these findings for levodopa treatment and proposes strategies for managing them.

Chapter 6: Multi-cohort machine learning identifies predictors of cognitive impairments in Parkinson's disease
This chapter provides an overview of the cognitive decline experienced by individuals diagnosed with PD. It presents research findings on objective and subjective cognitive outcomes and discusses the implications of cognitive impairment for the diagnosis and management of this complication.

Chapter 7: Conclusions and perspectives
The final chapter briefly summarizes the main findings of the thesis, which focused on the analyses of levodopa-induced dyskinesia, motor fluctuations, and cognitive decline in PD. The analysis

assesses the practical importance of these findings and their impact on patient treatment and monitoring. Additionally, it identifies areas for further research and potential strategies for analyzing PD complications.

2.3 Contributions

This thesis presents a collection of original research papers, some of which have been published while others have been submitted. The following section outlines the specific contributions to each of these works, arranged according to the following structure:

Manuscripts included as part of the PhD thesis:

- **Loo, R.T.J.**, Tsurkalenko, O., Klucken, J., Mangone, G., Khoury, F., Vidailhet, M., Corvol, J., Krüger, R. & Glaab, E. (2024). Levodopa-induced dyskinesia in Parkinson's disease: Insights from cross-cohort prognostic analysis using machine learning. *Parkinsonism & Related Disorders*, 126, p. 107054. DOI: 10.1016/j.parkreldis.2024.107054

- **Loo, R.T.J.**, Pavelka, L., Mangone, G., Khoury, F., Vidailhet, M., Corvol, J., Krüger, R. & Glaab, E. Interpretable machine learning for cross-cohort prediction of motor fluctuations in Parkinson's disease. Submitted.

- **Loo, R.T.J.**, Pavelka, L., Mangone, G., Khoury, F., Vidailhet, M., Corvol, J., & Glaab, E. Multi-cohort machine learning identifies predictors of cognitive impairment in Parkinson's disease. Submitted.

Manuscripts not included in the PhD thesis:

- de Lope, E.G., **Loo, R.T.J.**, Rauschenberger, A., Ali, M., Pavelka, L., Marques, M., Gomes, C.P.C., Krüger, R., Glaab, E. (2024). Comprehensive blood metabolomics profiling of Parkinson's disease reveals coordinated alterations in xanthine metabolism. *npj Parkinsons Disease*, 10 (1), p. 68. DOI: 10.1038/s41531-024-00671-9

- **Loo, R.T.J.**, Soudy, M., Nasta, F., Macchi, M., Glaab, E. (2024). Bioinformatics approaches to study molecular sex differences in complex diseases. *Briefings in Bioinformatics*, 25(6), bbae499. DOI: 10.1093/bib/bbae499

Poster presentation:

- **Loo, R.T.J.**, Glaab, E. "Cross-cohort prognosis of levodopa-induced dyskinesia in Parkinson's disease". Basel Computational Biology Conference, Basel (Switzerland). September 13, 2023.

- **Loo, R.T.J.**, Tsurkalenko, O., Klucken, J., Glaab, E. "Predictive machine learning models of levodopa-induced dyskinesia prognosis in Parkinson's disease across cohorts". Life Sciences PhD Day 2023, University of Luxembourg.

Chapter 3

Cohorts and methods

This chapter presents the workflow and approaches used to study the complications of PD. It describes the methodology for preprocessing and analyzing the associated predictors of the outcomes, combining statistical techniques and machine learning (ML) algorithms.

3.1 Cohort and inclusion criteria

The study analyzed data from three longitudinal cohorts: the Luxembourg Parkinson's Study (LuxPARK [102], NCT05266872), the Parkinson's Progression Markers Initiative (PPMI [103, 104], NCT04477785), and the French ICEBERG cohort study (ICEBERG [105], NCT02305147). These cohort datasets include comprehensive clinical assessments of individuals with Parkinson's disease (PD) and its complications, incorporating widely used PD assessment scales and detailed demographic information. LuxPARK has been enrolling participants in Luxembourg and the surrounding areas, which include the borders of Germany, France, and Belgium, since 2015. PPMI has enrolled individuals at approximately 50 sites worldwide since 2010 (`https://www.ppmi-info.org/`). ICEBERG is a longitudinal study that enrolls patients with PD in the early stages of the disease (average 3 years of disease duration or less at baseline). The study involves annual visits for up to four years at the Paris Brain Institute (Institut du Cerveau - ICM).

The research conducted in the LuxPARK, PPMI, and ICEBERG cohorts adheres to strict ethical standards and has been approved by local ethics committees. All individuals in each cohort provided written informed consent at enrollment. The LuxPARK study was approved by the National Research Ethics Committee (CNER ref: 201407/13 and 202304/03). The ICEBERG study, funded by the French National Institute of Health and Medical Research (INSERM), received approval from the local Ethics Committee (RCB: 2014-A00725-42), and all participants provided informed consent before enrollment. The PPMI study was conducted according to the Declaration of Helsinki and Good Clinical Practice (GCP) guidelines, and it received approval from local ethics committees at all participating sites. A comprehensive list of these sites can be found at `https://www.ppmi-info.org/about-ppmi/ppmi-clinical-sites`.

The baseline clinical characteristics of patients from all three cohorts were used to evaluate the potential predictors associated with the development of PD complications in patients with PD, along with their records of complications over four years. The data included outcomes measured from the baseline to the fourth year of follow-up. The most recent follow-up data were used for the time-to-event

analysis to determine the time-to-censoring for individuals without the event. In contrast, events were identified based on data collected throughout the follow-up period.

We conducted a comprehensive analysis to evaluate the advantages and disadvantages of integrating data from multiple cohorts. This process involved assessing each individual cohort (LuxPARK, PPMI, and ICEBERG) and evaluating multiple cohort combinations. The combinations included cross-cohort analyses using a combination of training, test, and hold-out test sets from all three cohorts. We implemented a leave-one-cohort-out validation approach, in which models were trained and tested using data from two of the three cohorts, with the remaining cohort being used as an external hold-out test set. This strategy enabled the assessment of the generalizability of the models across independent datasets.

3.2 Data preprocessing

Data preprocessing was conducted before analyzing PD complications to ensure the dataset's reliability and uniformity. This included variable aggregation, missing value imputation, categorical encoding, cross-study normalization approaches, feature selection, and undersampling.

3.2.1 Variable aggregation

Variable aggregation is a data preprocessing technique that combines related variables into a single aggregate score, typically through summation. The purpose of variable aggregation is to simplify data representation, reduce dimensionality, improve the comprehensiveness of understanding PD complications, and collect more robust predictive features for ML. Several items from the MDS-UPDRS Parts II and III (ON) clinical assessment were aggregated into combined features. Table 3.1 provides detailed information on how the variable aggregation was performed, which included summing the scores of certain functionally related MDS-UPDRS items. Additionally, to prevent replication in the analysis, the individual MDS-UPDRS variables used to generate the aggregated features were removed from the ML studies unless specified.

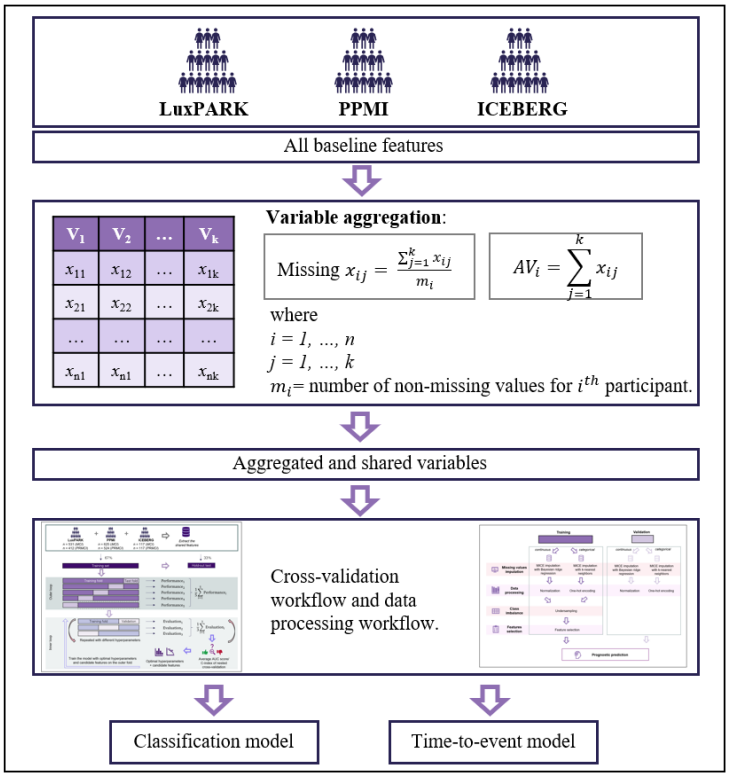
However, the absence of values for specific items can potentially skew the aggregated variables for that individual. To address this issue, we calculate the mean of each individual's non-missing values of the associated variables (see Figure 3.1). This approach ensures that the aggregation is conducted separately for each person, without being influenced by data from others. Addressing missing values at the individual level before variable aggregation, this approach minimizes potential bias and preserves the reliability and representativeness of the aggregate data.

The average for the missing item for i^{th} participant can be expressed as:

$$\text{missing } x_{ij} = \frac{\sum_{j=1}^{k} x_{ij}}{m_j}$$

where
$i = 1, ..., n$
$j = 1, ..., k$
m_i = number of non-missing values for i^{th} participant.
The k is the number of variables used for variable aggregation, and n is the total number of samples.

Figure 3.1 Variable aggregation in data processing and model development.



Workflow for data processing and model development, including variable aggregation and the nested cross-validation workflow for model training and evaluation.

Subsequently, the aggregated variable was calculated as the summation of the items:

$$\text{aggregated variable, } AV_i = \sum_{j=1}^{k} x_{ij}$$

where x_{ij} represents the individual items being aggregated for the i^{th} participant.

The selection of specific items from the MDS-UPDRS to be averaged and retained for each participant was based on clinical evidence. For instance, the average score of items 3.17a to 3.17e assessed different aspects of tremor as clinically relevant. Although facial expression is used for variable aggregation for bradykinesia, we included it in the analysis as the research suggests that the efficacy of levodopa in treating bradykinesia may not significantly affect facial expressivity [106]. In summary, this approach enables a focused analysis of the distinct contributions of bradykinesia and facial expressions to developing PD complications.

Table 3.1 Data preprocessing of the aggregation of MDS-UPDRS variables. Individual MDS-UPDRS items retained for the analysis are marked with an asterisk '*'.

Derived variable	MDS-UPDRS items included		Imputation
Rigidity upper extremities	3.3b	Rigidity - RUE	Average of items 3.3a to 3.3e
	3.3c	Rigidity - LUE	
Rigidity lower extremities	3.3d	Rigidity - RLE	Average of items 3.3a to 3.3e
	3.3e	Rigidity - LLE	
Total rigidity	3.3a	Rigidity - Neck	Average of items 3.3a to 3.3e
	3.3b	Rigidity - RUE	
	3.3c	Rigidity - LUE	
	3.3d	Rigidity - RLE	
	3.3e	Rigidity - LLE	
Bradykinesia score	3.2*	Facial expression	Average of items 3.2, 3.4 to 3.8, and 3.14
	3.4a	Finger tapping right hand	
	3.4b	Finger tapping left hand	
	3.5a	Hand movements - right hand	
	3.5b	Hand movements - left hand	
	3.6a	Pronation-supination - right hand	
	3.6b	Pronation-supination - left hand	
	3.7a	Toe tapping - right foot	
	3.7b	Toe tapping - left foot	
	3.8a	Leg agility - right leg	
	3.8b	Leg agility - left leg	
	3.14*	Global spontaneity of movement	
Axial symptoms score	2.12*	Walking and balance	Average of items 2.12, 2.13, 3.1, 3.9, 3.11 to 3.13
	2.13*	Freezing	
	3.1	Speech	
	3.9*	Arising from chair	
	3.11	Freezing of gait	
	3.12	Postural stability	
	3.13*	Posture	
Selective axial symptoms score	2.12*	Walking and balance	Average of items 2.12, 3.9, 3.12, and 3.13
	3.9*	Arising from chair	
	3.12	Postural stability	
	3.13*	Posture	
Freezing of gait	2.13*	Freezing	Average of items 2.13, 3.11
	3.11	Freezing of gait	
Rest tremor score	2.10	Tremor	Average of items 3.17
	3.17a	Rest tremor amplitude - RUE	
	3.17b	Rest tremor amplitude - LUE	
	3.17c	Rest tremor amplitude - RLE	

Continuation of **Table 3.1** Data pre-processing of the aggregation of MDS-UPDRS variables. Individual MDS-UPDRS items retained for the analysis are marked with an asterisk '*'.

Derived variable	MDS-UPDRS items included		Missing data strategy
	3.17d	Rest tremor amplitude - LLE	
	3.17e	Rest tremor amplitude - lip/jaw	
	3.18	Constancy of rest tremor	
Tremor score	2.10	Tremor	Average of items 3.17
	3.15a	Postural tremor - right hand	
	3.15b	Postural tremor - left hand	
	3.16a	Kinetic tremor - right hand	
	3.16b	Kinetic tremor - left hand	
	3.17a	Rest tremor amplitude - RUE	
	3.17b	Rest tremor amplitude - LUE	
	3.17c	Rest tremor amplitude - RLE	
	3.17d	Rest tremor amplitude - LLE	
	3.17e	Rest tremor amplitude - lip/jaw	
	3.18	Constancy of rest tremor	
Rest tremor amplitude score	2.10	Tremor	Average of items 3.17
	3.17a	Rest tremor amplitude - RUE	
	3.17b	Rest tremor amplitude - LUE	
	3.17c	Rest tremor amplitude - RLE	
	3.17d	Rest tremor amplitude - LLE	
	3.17e	Rest tremor amplitude - lip/jaw	

3.2.2 Missing value imputation

Missing values represent a common challenge in real-world datasets, often reducing the sample size for analysis. This reduction has the potential to reduce statistical power and introduce bias [71]. To address this issue, missing value imputation techniques are commonly used to preserve the sample size, reduce measurement error, and improve the reliability of analyses. Ultimately, these techniques support more informed and accurate decision-making.

Multiple imputation is a method used to address missing data across multiple variables. It models the missingness mechanism based on observed data and is known for its efficient convergence compared to many other Markov chain Monte Carlo (MCMC) methods [107]. Multiple Imputation by Chained Equations (MICE) [107, 108] uses multiple regression to predict missing values for each feature iteratively, incorporating information from other variables until convergence or the maximum iteration limit is reached. MICE can manage various variables, including continuous, binary, and nominal data types [108] while preserving the correlation structure among features. To mitigate the risk of data leakage [109], it is necessary to implement missing value imputation within each cross-validation (CV) fold.

Our study's clinical data from three cohorts had missing values for specific characteristics. Our study excluded variables with missing values exceeding 50% in each training or testing CV partition set from the analysis. Imputation was conducted separately for each CV partition to avoid data leakage. The study used Bayesian ridge regression within the MICE imputation framework to handle missing

values in continuous variables. This approach considers the uncertainty inherent in the imputation process and effectively captures complex interactions within the data. For categorical variables, MICE imputation with the k-nearest neighbors approach was used to preserve the structural properties of categorical data.

3.2.3 Categorical encoding

Data often includes categorical variables, such as gender and symptom severity level (normal, slight, mild, moderate, or severe). Appropriate encoding techniques must be used to represent categorical variables in a format the algorithm can understand and handle in ML algorithms [110].

Encoding techniques should be applied before implementing ML, such as ordinal encoding, one-hot encoding, dummy encoding, effect encoding, binary encoding, and target encoding. One-hot encoding is the most popular technique for categorical features [76] and is used in the analysis of the thesis. It involves creating a new variable for each level of the categorical feature, resulting in higher dimensionality. The variable contains either 0 or 1, representing absence or presence.

The CatBoost algorithm is designed explicitly for categorical variables. Therefore, it is recommended to avoid using any categorical encoding schemes as they can negatively impact both the training speed and the resulting quality of the model.

3.2.4 Cross-study normalization

The presence of variability among cohorts, resulting from discrepancies in data collection protocols and sample populations, can lead to the presence of a "batch effect", which can complicate data analysis. Cross-study normalization methods have been proposed to mitigate "batch effects" and standardize data from different cohorts [93]. This preprocessing step ensures comparability across cohorts [101] by adjusting the data distribution across studies [20]. Although normalization techniques can improve precision measurement [98] and prediction performance [100] in cross-cohort scenarios, their benefits may only sometimes be presented in real-world applications [94].

Hornung et al. (2017) [20] proposed using normalization techniques to mitigate "batch effects" in training and testing datasets to improve prediction performance [111]. It is important to note that impaired prediction performance may occur in scenarios involving imbalanced data or high levels of heterogeneity within the training set. Furthermore, addressing such issues improves the robustness and generalizability of predictive models across diverse datasets and ensures reliable outcomes in real-world applications.

The study in this thesis compared various cross-study normalization methods for continuous variables within each cohort. The methods used were mean-centering [98], standardization, Quantile normalization [111, 112], ComBat [93], Ratio-A [95], and M-ComBat [97]. The aim was to ensure data comparability and reliability across multiple cohorts. To maintain the integrity of the validation process and prevent data leakage, we performed cross-study normalization within each CV fold and on the hold-out test set. This approach ensured normalization procedures were applied independently to each fold, avoiding biases introduced by sharing information between training and test sets.

Mean-centering

Mean-centering is a process that centers the data distribution around zero by subtracting the mean value of each variable across cohorts from its corresponding data points in each cohort's batch [98]. This involves defining the number of cohorts as B. The resulting normalized data reflects the deviations from the mean and follows a location-and-scale model [96].

$$x_{ib} - \bar{x}_b$$

where
$i = 1, ..., n$,
$x_{ib} =$ data to be normalized, and
$\bar{x}_b =$ average of feature in b^{th} cohort with $b = 1, ..., B$.

Standardization

Standardization, known as z-score normalization, scales each variable to have a mean of zero and a standard deviation (SD) of one [98]. The data is normalized by subtracting each feature from the average of the corresponding cohort and then scaling it by the SD of that cohort.

$$\frac{x - \bar{x}_b}{\sigma_b}$$

where
$i = 1, ..., n$,
$x_{ib} =$ data to be normalized,
$\bar{x}_b =$ average of feature in b^{th} cohort with $b = 1, ..., B$, and
$\sigma_b =$ standard deviation of feature in b^{th} cohort with $b = 1, ..., B$.

Quantile normalization

Quantile normalization is a method that aligns the quantiles of variables across cohorts to equalize their empirical distribution. This is achieved through a "documentation by value" approach that adjusts the data to have the same cumulative distribution function across all samples [111, 112]. The process involves replacing the original values with artificial values using the following steps:

1. Sort each column independently: For each feature, the values are sorted in ascending order. Keep the original ranks to restore the data later.

2. Compute the average rank values: For each row index (across all columns), the mean of the sorted values is computed. This process establishes a reference distribution, where each rank position is associated with an averaged value.

3. Replace values with corresponding mean rank values: Assign each value in the original data set the mean rank value from the reference distribution based on its original rank.

ComBat

ComBat is an empirical Bayes method commonly used to eliminate "batch effects" by adjusting data at the cohort level based on an empirical Bayes framework [93]. The method estimates prior and posterior distributions to adjust features across cohorts ($b = 1, ..., B$). The estimation procedure involves the following steps:

1. ComBat fits a linear model for each feature, with cohort variables as covariates.
2. The mean and variance estimation is predicted based on the fitted linear model for each feature across all cohorts. This is the first stage of empirical Bayes estimation based on the prior distribution. The data is then shrunk to the mean and covariance to obtain the normalized data.
3. The empirical Bayes estimator uses prior distribution estimation to estimate the posterior cohort-specific parameters: mean and variance from a fitted linear model (second stage of empirical Bayes). Afterward, posterior estimators normalized the data by shrinking it towards a common mean and covariance.

Ratio-A

Ratio-A normalization requires dividing each observation by the mean of the corresponding cohort [95]:

$$\frac{x_{ib}}{\bar{x}_b}$$

where
$i = 1, ..., n$,
$x_{ib} =$ data to be normalized, and
$\bar{x}_b =$ average of feature in b^{th} cohort,
with $b = 1, ..., B$.

M-ComBat

M-ComBat is an extension of the ComBat method that applies the same empirical Bayes framework and introduces the concept of a "gold standard" [97]. The first two steps of M-ComBat are the same as those of ComBat normalization. However, in step 3, the data points are shifted to the mean and SD of the "gold standard" instead of the overall mean and SD. This approach ensures that the data is adjusted to the parameters of the most representative cohort.

3.2.5 Class imbalance

In real-world data, the distribution of the outcome of interest is usually highly skewed, with one class significantly underrepresented in another. For example, in the PD study, the group of patients with a complication may be much smaller than those without the complication of interest.

Class imbalance is a significant problem in conventional ML, causing bias and reducing the model's generalizability [113]. Algorithms are typically optimized by minimizing error without considering the distribution of the outcome variable. As a result, they may overpredict the majority class while

incurring high misclassification costs for the minority class [114]. Undersampling techniques can be initiated to address this issue by balancing the target outcome's class distribution, improving model performance, and ensuring higher accuracy in predicting both classes.

Undersampling is a technique that reduces the number of subjects in the majority class to match the number of subjects in the minority class, which results in a balanced class distribution for the outcome. Unfortunately, it results in the loss of subjects from the majority class. The most straightforward approach for undersampling is randomly selecting a subset of the majority class and then combining it with another class as the training set in each CV fold.

3.2.6 Feature selection

Feature selection aims to identify the most relevant and informative variables from a data set while excluding irrelevant or redundant variables. This process reduces dimensionality and improves model performance, interpretability, and generalization [75]. Additionally, high-dimensional data can result in high computation costs, leading to an overfitting model and a complex model that increases model interpretability challenges [115].

In this thesis, two feature reduction methods have been used, namely Recursive Feature Elimination (RFE) [109] and a wrapper method, which is a stepwise sequential feature selection [8]. RFE is used in ML algorithms, such as AdaBoost, CART, CatBoost, GBoost, and XGBoost, which return the model's coefficients to measure the features' importance score. On the other hand, stepwise sequential feature selection combines forward and backward selection techniques to achieve an optimal area under the curve (AUC) or concordance index (C-index) while ensuring a less complex model.

To prevent overfitting and ensure the robustness of the feature selection process, we implemented CV and performed feature selection within each nested CV fold. This approach allowed us to evaluate the generalizability of the selected features across different data subsets and obtain more reliable results. The most predictive features were identified using a permutation importance score [116]. This procedure helped to remove uninformative or redundant variables from the models while maintaining the predictive ability of the models in terms of the cross-validated AUC/C index. The final selection of the most predictive features was then used in the outer loop to train the model.

3.3 Machine learning analysis for PD complication classification

A range of machine learning (ML) techniques, focusing on interpretable, tree rule-based methods, were used to predict the occurrence of PD complications in PD patients over up to four years of follow-up based on clinical data from three PD cohorts. Nine tree-based ML classification algorithms were evaluated, including Adaptive Boosting (AdaBoost) [73, 117], Classification and Regression Trees (CART) [71], and Category Boosting (CatBoost) [76], C4.5 [118], Fast Interpretable Greedy-Tree Sums (FIGS) [72], Fast-Sparse Decision Tree (GOSDT-GUESSES) [79], Gradient Boosting (GBoost) [74], Hierarchical Shrinkage (HS) [80], and Extreme Gradient Boosting (XGBoost) [75]. These classification algorithms were used to create PD complication prognosis models and to evaluate the most predictive features of these models as potential risk and protective factors for PD complications using baseline clinical data from PD patients.

For the classification analysis of all cohorts, we analyzed data over four years to ensure consistency

and comparability, as the ICEBERG cohort offers up to four years of follow-up. This time frame is consistent with clinical practice, where significant changes in treatment response and complications often occur within the first few years of therapy. However, focusing solely on this period may limit the study by not capturing long-term trends and outcomes beyond four years. To mitigate this limitation, we conducted a time-to-event analysis, as described in the next section, evaluating events and outcomes over different periods. This approach provides a more comprehensive understanding of the data despite the consistent follow-up period used for classification analysis.

3.4 Time-to-event machine learning analysis

This study used time-to-event analysis as an additional approach to identify predictive clinical features associated with the likelihood of developing complications in PD. This approach is commonly used in biomedical research to examine the impact of medical factors on events observed and followed over a period of time. The data are subject to censoring, which means that some patients may have yet to experience the event of interest by the end of the study, resulting in incomplete observations [81]. The following time-to-event methods were used to analyze time-to-event data: Component Wise Gradient Boosting (CW-GBoost) [92], Survival Trees [85] and Extra Survival Trees [87], Survival Gradient Boosting (Survival GBoost) [90], Survival Support Vector Machine (LSVM) [82], Naive LSVM (NLSVM) [7], Penalized Cox Regression [83, 84], and Random Survival Forests (Survival RF) [86, 88, 89].

3.5 Cross-validation

The cross-validation (CV) workflow shown in Figure 3.2, along with the learning algorithms described in sections 3.3 and 3.4, was used to quantitatively evaluate different ML approaches. The goal was to investigate how clinical predictors relate to the risk and time from baseline to PD complication after the first clinical visit. We focused on clinical features that could serve as common predictors across the LuxPARK, PPMI, and ICEBERG datasets for integrative analyses of different cohorts.

The prognostic model was developed using all baseline clinical variables shared by all cohorts without prior feature selection. The robustness and reliability of the model were improved by ensuring that all common features were included. During the training phase, a CV framework was used to comprehensively validate and evaluate the model's performance across various patient populations. The study aims to provide a reliable framework for predicting outcomes in patients with PD based on all common baseline clinical parameters.

A CV workflow was used to optimize the model parameters without overfitting and to predict the occurrence of PD complications over the next four years for classification and until the end of follow-up for time-to-event analysis. Stratified random sampling was used to divide the data samples into training and test sets. The training set comprised 67% of the samples, stratified by outcome proportion (also stratified by cohort for cross-cohort and leave-one-cohort-out analyses), with the remaining 33% allocated to a hold-out test group (leaving out one cohort as the hold-out test set for the leave-one-cohort-out analysis). A grid search was conducted within the inner loop of the nested CV framework, using a 3-fold CV strategy to optimize the hyperparameters of the prediction models. The list of hyperparameters that were tuned for each model is detailed in Table 3.2. Additionally, two feature selection methods, bidirectional stepwise feature selection [8] and recursive feature elimination (RFE) [109], were applied

within the nested CV. A permutation importance score was used to determine the most predictive features [116]. This process preserved the model's predictive power, as measured by the cross-validated AUC/C index, by removing redundant or uninformative variables. The model was trained using the final selection of the best predictive features and optimized hyperparameter by the outer loop 5-fold CV.

Table 3.2 List of hyperparameters for classification and time-to-event algorithms.

Algorithms	Description	Hyperparameters
Classification:		
AdaBoost	Maximum number of estimators at which boosting is terminated	n_estimators: {1, 2, 3, 4, 5, 10}
CART	Maximum depth or depth limit of the trees	max_depth: {1, 2, 3, 4, 5, 10}
CatBoost	Maximum depth or depth limit of the trees	max_depth: {1, 2, 3}
C4.5	Maximum total number of rules across all trees	max_rules: {1, 2, 3, 4, 5, 10}
FIGS	Maximum total number of rules across all trees	max_rules: {1, 2, 3, 4, 5, 10}
GOSDT-GUESSES	Trade-off between tree complexity and model accuracy	regularization: {0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3}
GBoost	The number of boosting stages to perform	n_estimators: {1, 2, 3, 4, 5, 10}
HS	Maximum total number of rules across all trees	max_rules: {1, 2, 3, 4, 5, 10}
XGBoost	Maximum depth or depth limit of the trees	max_depth: {1, 2, 3}
Time-to-event:		
CW-GBoost	Learning rate shrinks the contribution of each base learner. Values in range [0.0, inf]: trade-off between learning_rate and n_estimators	learning_rate: {0.0001, 0.001, 0.01, 0.1, 0.2, 0.4, 0.6, 0.8, 1}
Extra Survival	Maximum depth or depth limit of the trees	max_depth: {1, 2, 3, 4, 5, 10}
Survival Gboost	Maximum depth or depth limit of the trees	max_depth: {1, 2, 3, 4, 5, 10}
LSVM	Weight of penalizing the squared hinge loss in the objective function	alpha: {1, 2, 3, 4}
NLSVM	Weight of penalizing the squared hinge loss in the objective function	alpha: {1, 2, 3, 4}
Penalized Cox	ElasticNet mixing parameter	l1_ratio: {0.0001, 0.001, 0.01, 0.1, 0.2, 0.4, 0.6, 0.8, 1}
Survival RF	Maximum depth or depth limit of the trees	max_depth: {1, 2, 3, 4, 5, 10}
Survival Trees	Maximum depth or depth limit of the trees	max_depth: {1, 2, 3, 4, 5, 10}

Model performance was improved by identifying the hyperparameters and predictive features that produced the highest average AUC score for PD complication risk categorization and the highest average concordance index (C-index) for time-to-complication analysis within the nested loop of the CV. Hyperparameter tuning and feature selection within a CV framework can be used to optimize model performance. These techniques involve iteratively adjusting hyperparameters and selecting relevant features based on the model's performance on different subsets of the data. By integrating hyperparameter tuning and feature selection into CV, we aim to improve model generalization and reduce the risk of overfitting, thereby improving the robustness and accuracy of our predictive models. The data preparation procedures, including missing value imputation, cross-study normalization, categorical encoding [110], undersampling [113, 114], and feature selection, were performed during each CV iteration's training and validation sets [109] to prevent data leakage. The external hold-out test set was also subjected to independent application of the data preparation methods. Figure 3.3 visually represents the entire data processing and analysis cycle.

Furthermore, patients with baseline PD complication outcome were excluded from the test set

evaluation in the CV workflow to ensure that the ML models were evaluated for their ability to predict the onset of PD complication in individuals without prior onset of the outcome, thereby minimizing potential bias that could arise from pre-existing conditions. This step ensures that the models focus on identifying predictive features and patterns associated with the initial development of PD complications, thereby enhancing their applicability to patient population at risk for PD complications rather than those already affected. It also avoided optimistic predictive performance estimates by ensuring that the test set truly represented novel cases, free from information indirectly available to the model during training.

The *optimized/best* model in this thesis is defined as achieving the highest average cross-validated AUCs for classification tasks and C-indices for time-to-event analyses. These metrics ensure robust evaluation of the model's predictive performance, with the AUC assessing classification predictive capability and the C-index evaluating the concordance between risk scores and observed event times.

Figure 3.2 Machine learning and cross-validation workflow.

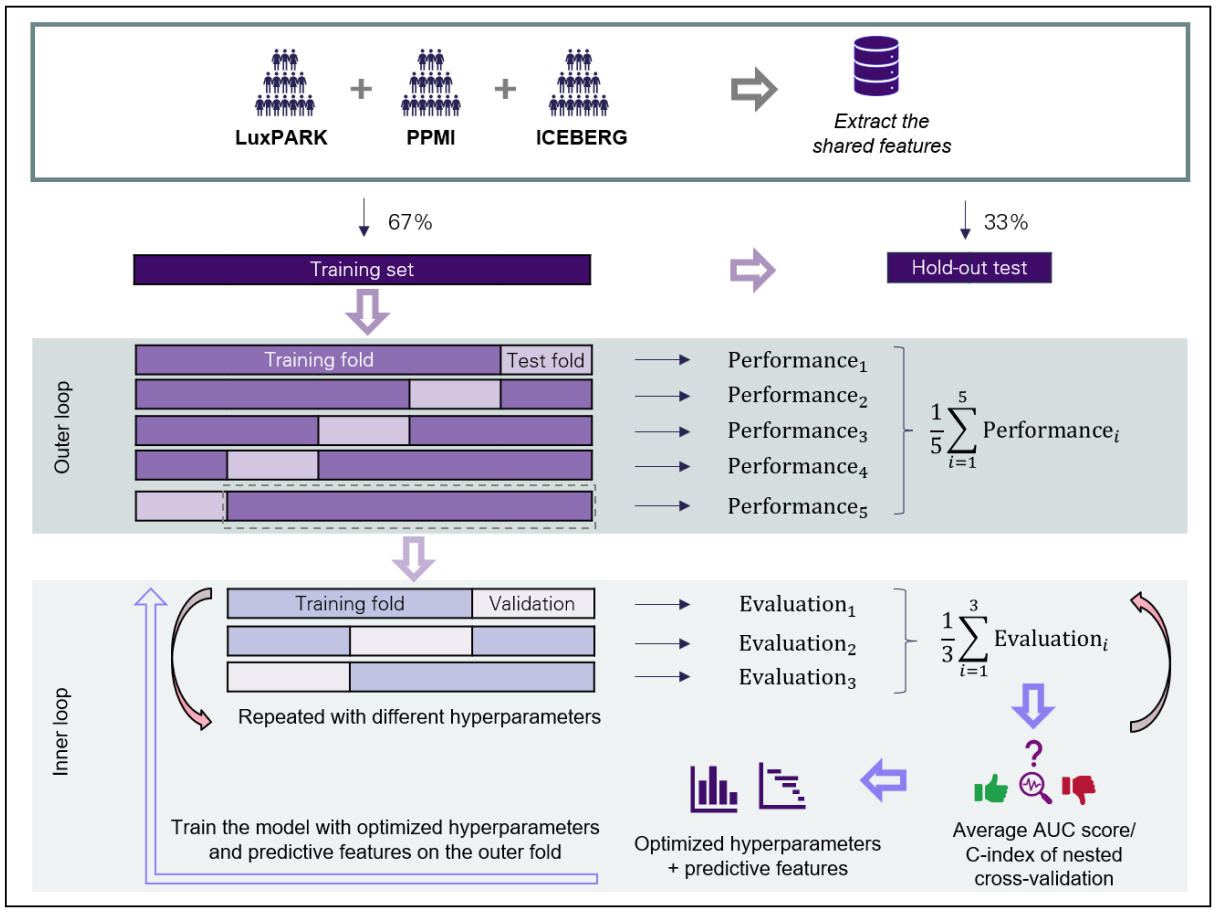
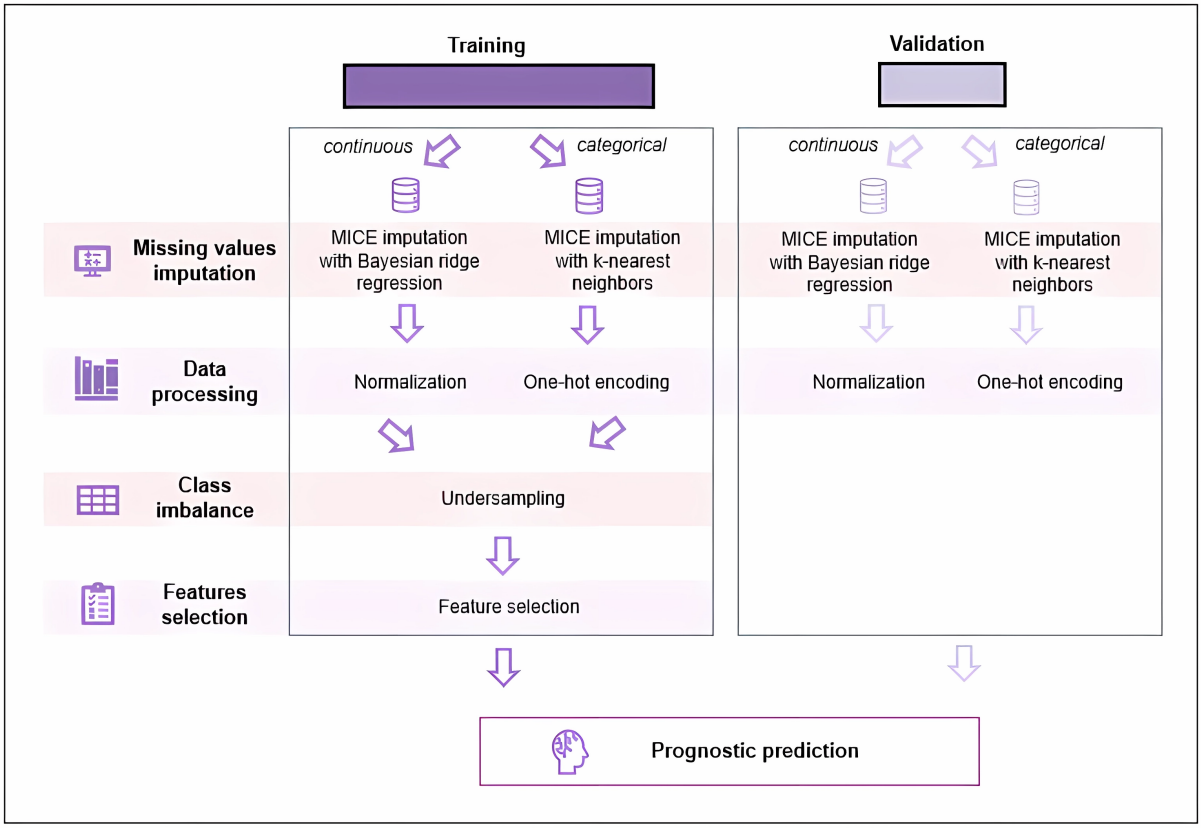


Illustration of the machine learning and cross-validation workflow, which involves training and evaluating a PD complication prognostic model using 5-fold cross-validation on the training set to assess average performance. A 3-fold nested cross-validation within the training set was used to optimize hyperparameters and select predictive features to optimize the model's predictive ability.

Figure 3.3 Data processing and cross-validation workflow.



Data processing and analysis workflow in each cross-validation cycle to optimize and evaluate the PD complication prognostic model. The workflow covers missing values imputation, cross-study normalization, one-hot encoding, undersampling, and feature selection.

3.6 Model interpretability

To improve the interpretability of our PD complication prediction model, we used SHAP (SHapley Additive exPlanations) value analysis, a widely used method for model interpretation introduced by Lundberg and Lee (2017) [119]. Unlike model-specific approaches, SHAP analysis provides a model-agnostic framework for *post hoc* interpretation [80, 115], making it adaptable to different ML frameworks. Using SHAP analysis, we quantified the predictive contribution of each feature to the outcome regardless of the underlying model and evaluated their influence on specific outcome predictions. SHAP values provide insight into the contribution and impact of each feature on model predictions. These values are computed by evaluating the average change in prediction when a given predictor is added to all possible model subsets. This method provides a more in-depth understanding of PD complication prediction models' underlying mechanisms, enhancing their interpretability and facilitating clinical decision-making.

Interpreting ML models is important for understanding the models' decisions. This is because interpreting the models involves identifying and explaining the key predictors that influence the

models' predictive performance. This contributes to enhancing confidence in model predictions for decision-making processes and reduces the gap between data scientists and clinicians, who require a comprehensive understanding of model output.

In the study, we also used SHAP values to calculate the hazard ratio (HR) in the time-to-event models proposed by Sundrani and Lu (2021) [120] on the hold-out validation set. This hazard ratio (HR) quantifies the relative risk of developing PD complications over time based on various predictive factors. It provides insight into significant associations and potential risk factors contributing to these complications. The interpretation of HR thus provides insights into the influence of predictors on PD complication development and how changes in predictors influence the hazard of experiencing these complications.

To further improve the hazard ratio interpretation, we calculated confidence intervals (CI) using the percentile method at a 5% significance level. This involved applying bootstrapping techniques to resample data with replacement, generating a hazard ratios distribution. This distribution derived CI to estimate the range within which the true HR is likely to fall, thereby providing a measure of uncertainty associated with HR estimates.

Additionally, the log-rank test was used to identify an optimal threshold for continuous variables, thereby allowing the categorization of patients into two distinct groups for HR analysis. The test helped identify the value along the continuous variable that yielded the most prominent log-rank test statistic, indicating a notable difference in conversion times between resulting groups. This optimal threshold represents a meaningful division point where patients showed distinct outcomes concerning PD complications. Stratifying patients based on this threshold enabled us to analyze the HR between the groups, offering insights into the relative risk of PD complication onset associated with varying levels of the continuous variable. A HR greater than 1 indicates an increased risk associated with a particular predictor, while a ratio less than 1 suggests a reduced risk. The statistical significance of HR was determined using CI.

3.7 Model performance evaluation

AUC is a widely used performance metric for evaluating predictive ability in ML classification [121]. The AUC score is a comprehensive measure of the actual positive rate (sensitivity) versus the false positive rate (1-specificity), providing an intuitive measure to evaluate the model's predictive power. The AUC score ranges from 0 to 1, with a higher value indicating better predictive performance.

In time-to-event analysis, two outcomes are considered, which need to be considered in the model performance evaluation. The C-index is widely used to measure the proportion of concordance pairs among comparable pairs [7]. In this context, considering a pair of patients (i, j), the individuals with higher risk on event $(\eta_i > \eta_j)$ are expected to experience the event of interest sooner, $(T_i < T_j)$ and vice versa. Concordance pairs are defined as the individuals with $(\eta_i > \eta_j : T_i < T_j)$ or $(\eta_i < \eta_j : T_i > T_j)$, and discordant otherwise. The C-index is calculated as

$$C = \frac{\sum_{i,j} I(T_i > T_j) \cdot I(\eta_i < \eta_j) \cdot \delta_j}{\sum_{i,j} I(T_i > T_j) \cdot \delta_j}$$

where $I(\cdot)$ is an indicator if its argument is true and δ_j is the censoring status of subject-j. The C-index ranges from 0 to 1, with a higher C-index indicating better predictive power.

Higher AUC or C-index values indicate better predictive performance and range from 0 to 1. AUC/C-index values of 0.9 or higher are considered outstanding, values between 0.8 and 0.9 are considered excellent, and values between 0.7 and 0.8 are considered acceptable. Conversely, an AUC or C-index of 0.5 represents a random prediction model with no predictive power.

To statistically evaluate and compare the performance of the optimized models in the same cohort analyses, we applied DeLong's test [122], along with its extension by Kang et al. (2015) [123], known as the one-shot nonparametric approach, to the hold-out test set. DeLong's test is commonly used to compare the performance of classifiers, especially concerning the AUC value. It determines the statistical significance of the difference between two AUC scores obtained from the same cohort. However, this approach is less practical for larger sample sizes due to its high computational cost. Therefore, to address this limitation, a superior algorithm has been introduced that reduces the complexity from quadratic to linearithmic order [122], and this revised version was implemented in this study. It is important to note that DeLong's test was initially designed for dichotomous outcomes, not continuous outcomes, as in the case of time-to-complication analyses. Therefore, Kang et al. (2015) [123] extended DeLong's test to fit the C-index, known as the one-shot nonparametric approach. This estimator is computationally efficient because it does not require resampling. Both DeLong's test and the one-shot nonparametric approach are asymptotically standard and normally distributed under the null hypothesis of

$$\hat{\theta}^{(1)} = \hat{\theta}^{(2)}$$

where $\hat{\theta}^{(1)}$ represents the performance metric of model_1 and $\hat{\theta}^{(2)}$ represents the performance metric of model_2, to determine if the difference deviated significantly from zero.

Performance metrics for all cross-study normalization methods were compared on optimized non-normalized and normalized models. *P*-values obtained from within-cohort comparisons were adjusted using the Benjamini-Hochberg method to account for multiple hypothesis testing.

In assessing predictive model performance through CV, we applied the Bayesian signed-rank test [124], a valuable tool for comparing performance metrics across different predictive models. This statistical test is particularly advantageous as it evaluates the null hypothesis that there is no difference in performance between the two models. By estimating probabilities of hypotheses, this test provides a more intuitive interpretation of results, allowing us to gauge the likelihood of one model being superior to another.

Specifically, our analysis computed performance metrics such as the AUC or the C-index for each optimized model during cross-validation. These metrics served as the basis for calculating performance differences between pairs of models across 5-fold CV iterations [125]. The Bayesian signed-rank test identified the probability that the model consistently showed superior predictive performance across the validation sets. This approach is robust as it considers performance variations across multiple folds, comprehensively assessing model efficacy and generalizability.

Ensuring the stability of a prediction model is important for its generalizability to new data or future PD patients. In this context, stability refers to the robustness of the model against perturbations in the training set, which improves its reliability. We calculate the SD of performance metrics obtained from the nested cross-validation within each k-fold cross-validation iteration to quantify stability. This calculation uses the corresponding optimal hyperparameters identified during the nested cross-validation process. This approach enables the assessment of the consistency of the model's performance across different subsets of the data, thereby providing insights into its reliability under varying conditions [8].

3.8 Comparison of selected features across different cohorts

To compare feature selection results in single-cohort analyses, we computed feature selection statistics to identify features consistently selected as predictors of PD complications across cohorts and methods. Before feature selection, categorical variables were subjected to one-hot encoding, generating multiple binary features representing different categories. To avoid counting duplicates, if multiple selected features were derived from the same categorical variable within each CV fold, they were treated as a single feature.

We calculated the percentage of times a predictive feature was selected in each of the 5-fold CVs for each cohort. We then calculated the average percentage for each cohort, providing a consolidated measure of the predictive utility of each feature. The average percentages were then compared across the three cohorts for the optimized PD complication classification models and the time to complication analysis. The goal was to identify consistent predictors of PD complications across methods and cohorts that could serve as biomarkers for cross-cohort PD complication prediction.

To comprehensively compare variables across our study's three cohorts, we used parametric and nonparametric statistical tests: Analysis of Variance (ANOVA) and the Kruskal-Wallis test. ANOVA was used when the assumption was met, including the normality of residuals (assessed using the Shapiro-Wilk test), the homogeneity of variances (evaluated using Levene's test), and the independence of observations. When these assumptions were met, ANOVA allowed for the simultaneous comparison of means across multiple groups, providing insight into potential differences in variable distributions between cohorts. Conversely, in cases where the assumption was violated, the Kruskal-Wallis test, a nonparametric alternative to ANOVA, was applied. The Kruskal-Wallis test evaluates whether the distribution of variables significantly differs across cohorts without assuming normality, making it robust for non-normally distributed data.

To gain further insights into specific differences between pairs of cohorts, pairwise comparisons were conducted using the Dunn test, a *post hoc* test applied after the Kruskal-Wallis test with Bonferroni correction. This correction helps mitigate the issue of multiple comparisons, providing a more stringent criterion for statistical significance. Similarly, when ANOVA indicated significant differences, Tukey's Honestly Significant Difference (HSD) test was used as a *post hoc* test. Tukey's HSD is well-suited for ANOVA results, mainly when dealing with normally distributed data, as it offers a reliable means of identifying which specific cohorts show statistically significant differences in variable distributions. These combined approaches allowed for a comprehensive and robust assessment of variable distributions across cohorts, considering normality assumptions and multiple comparison adjustments.

3.9 Univariate analysis

Univariate analysis provides valuable insights into the distribution and significance of variables across different cohorts, aiding in the identification of potential predictors and predictors associated with PD complications.

Univariate hypothesis testing was used to investigate potential statistical associations between baseline clinical parameters and the occurrence of PD complications during the 4-year visit. When the underlying assumption of normality was not met, we used the Mann-Whitney U test to assess the statistical significance of differences in continuous variables between independent groups. For normally

distributed variables, we used the two-sample *t*-test. Categorical variables were assessed using Fisher's exact test.

To examine the duration until the onset of PD complications, we initially examined the median conversion time, representing the point at which 50% of PD patients experience the event of interest. We used the log-rank test to compare Kaplan-Meier (KM) curves among different groups or subgroups of PD patients. This statistical tool assesses whether significant differences exist in time-to-event outcomes between groups and considers censored data. It enables event time to be compared between heterogeneous patient populations, considering the frequency of problems and the time interval before they occur.

Correlation analysis is fundamental in examining the relationships between variables relevant to PD complications. Spearman correlation analysis assesses the correlation between two continuous variables, providing insight into potential linear or monotonic relationships. In addition, point biserial correlation analysis is used to examine the correlation between binary and continuous or ordinal variables. Matthew's Correlation Coefficient (MCC) is applied to evaluate the correlation between two binary variables, providing a measure of the strength and direction of the association. Additionally, Kendall's tau correlation coefficient is used to analyze the correlation between two ordinal variables, assessing non-linear relationships and providing valuable insight into the ordinal variables. This study aims to investigate the multifaceted relationship between various clinical parameters and PD complications using these correlation analysis techniques, thus contributing to a comprehensive understanding of the disease. Statistical significance was defined as a *p*-value of less than 0.05.

3.10 Validation of clinical utility measures

Evaluation metrics such as the AUC or C-index provide valuable information about a model's discrimination performance. However, they do not offer direct insights into the clinical consequences of decisions. Decision curve analysis (DCA) addresses this gap by evaluating the clinical utility of a predictive model. It compares the model's net benefit, considering the threshold probability for making clinical decisions. DCA is a required tool in ML analysis, as it assists in the informed decision-making process regarding the practicality of predictive models for clinical utility [126].

To achieve this, we used the hold-out test set to apply DCA. This method identifies the range of threshold probabilities at which the model provides a higher net benefit than alternative approaches, such as treating all patients, treating none, or using a different model. Clinical utility is the net benefit, which is desirable when positive [127]. This analysis assists clinicians in comprehending the practical implications of using a predictive model in a real-world setting. It ensures the model performs well statistically and improves clinical outcomes and decision-making.

In addition to DCA, the area under the net benefit curve (AUNBC) was also considered in the analysis. This metric provides a summary measure of the overall benefit of the model across a range of threshold probabilities, as introduced by Zhang et al. (2018) [128]. A larger AUNBC indicates more excellent clinical utility and substantial benefit in clinical decision-making. The AUNBC was estimated using the trapezoidal rule for numerical integration:

$$\text{AUNBC} \approx \sum_{i=1}^{k-1} \frac{1}{2} \times \Delta x \times [NB(x_i) + NB(x_{i+1})]$$

where
k is the maximum number of step size between threshold values,
$NB(x_i)$ is the net benefit at threshold x_i, and
Δx is the step size between threshold values.

To establish the differences between the two models in AUNBC, we used bootstrapping hypothesis testing for *p*-value estimation with 1,000 replicates. This process involves resampling the data to obtain an empirical approximation of the differences in AUNBC sampling distribution. This method allows for the estimation of *p*-values without relying on strict distributional assumptions. The following steps were used in conducting the bootstrapping hypothesis test:

1. For each replicate, the dataset was randomly sampled with replacement to create a new dataset of the same sample size.

2. The AUNBC was computed for each resampled dataset for both models.

3. The difference in AUNBC between the two models was determined for each replicate.

4. The *p*-value was estimated by calculating the proportion of replicates where the difference in AUNBC was greater than the observed difference in the data without resampling.

The equation for estimating the *p*-value can be written as:

$$p\text{-value} = \frac{1}{1,000} I(|t_i| > |t_0|)$$

where
t_i is the difference in AUNBC between the two models,
t_0 is the observed difference in AUNBC between the two models,
I is an indicator function that is 1 if the condition is true, and 0 otherwise.

This method ensures a robust estimation of the *p*-value, thereby providing a reliable measure of whether the difference in the performance of the two models is statistically significant. The threshold for statistical significance was set at 0.05 in the analysis.

In addition to DCA, a calibration analysis was performed to assess the agreement between predicted probabilities and observed outcomes using the hold-out test set. Calibration analysis is a process for assessing the reliability of predictive models [127]. It evaluates how well the predicted probabilities from a model align with the observed outcomes, providing valuable insights into the model's prediction capability. For time-to-event outcomes, the observed conversion probabilities of the event occurring at four years were calculated by fitting a KM function [129].

Moreover, we conducted a comprehensive analysis of the predicted probabilities and observed outcomes for classification and the observed conversion probabilities for time-to-event analysis, examining the slope of the predicted probabilities versus the observed outcomes and the mean squared error (MSE). The calibration slope was obtained by fitting a linear regression model in which the predicted probabilities were regressed against the observed outcomes. A well-calibrated model shows a close match between the predicted probabilities and the actual values in the analysis, as indicated by a slope closer to 1 and a lower MSE.

Chapter 4

Levodopa-induced dyskinesia in Parkinson's disease: Insights from cross-cohort prognostic analysis using machine learning

Levodopa is considered the standard treatment for Parkinson's disease (PD) [30]. It is a precursor to dopamine [1] and is important in managing motor symptoms. Levodopa crosses the blood-brain barrier and is converted into dopamine within the brain [30], which restores dopamine levels and alleviates motor dysfunction.

The long-term use of levodopa is effective in managing the motor symptoms of PD. However, it is associated with various adverse effects that significantly impact patients' quality of life. These effects include nausea, hallucinations, cognitive impairment, and motor complications. Nausea, a common gastrointestinal (GI) side effect, is frequently reported by patients initiating levodopa therapy [32]. Hallucinations, particularly visual hallucinations, may also occur, leading to perceptual disturbances and delusions [31]. Additionally, cognitive impairment is a significant concern, as some patients may show a poorer response to levodopa, experiencing memory difficulties, confusion, and impaired executive function [42, 130].

Levodopa-induced dyskinesia (LID) is a common complication of long-term levodopa treatment in PD [1, 26]. LID tends to become more frequent and severe over time [131]. Levodopa significantly relieves motor impairments such as bradykinesia and rigidity [19, 31]. However, prolonged use can develop involuntary and hyperkinetic movements, including dystonia, myoclonus, chorea, or ballism [28, 29, 132]. The management of LID is a substantial challenge in the treatment of PD, particularly when it manifests subsequent to levodopa administration. It has been demonstrated that in addition to levodopa, dopamine agonist monotherapy can also contribute to the development of dyskinesia [133]. Individuals without PD and those with other neurological diseases typically do not experience dyskinesia following prolonged treatment with levodopa [23, 134].

It is important to note that levodopa administration is temporally associated with LID. In the OFF state, characterized by low dopamine levels, individuals may experience dystonia, which involves

involuntary muscle contractions leading to abnormal postures or movements [135]. Dyskinesia often occurs early in the morning [58] and can be particularly distressing and disabling for patients, often affecting the lower limbs [134]. These movements, which are involuntary and uncontrollable, can disrupt daily activities and impair functional independence. LID can be refractory to treatment [23]. In advanced stages of PD, patients may be unable to care for themselves without appropriate LID management [136]. Managing LID can be a clinical challenge, often requiring dose adjustments, medication combinations, or advanced therapies such as Deep Brain Stimulation (DBS). It is important to note that reducing levodopa treatment may worsen motor side effects [29].

LID presents a significant challenge in the long-term management of PD. Its prevalence is closely linked to disease duration and daily levodopa dosage, which show stronger correlations than the cumulative amount of levodopa received [137]. Several studies have shown that approximately 30% to 50% of PD patients develop dyskinesia within the first five years of levodopa treatment [29, 131, 138], with a median onset time of 5.9 years [139], and this percentage increases to over 90% after ten years of treatment [1, 58]. In advanced stages of PD, 80% of patients treated with levodopa develop LID, with 30% experiencing LID within three years of starting levodopa therapy [22]. Research indicates that within an average treatment duration of 20.5 months, 20-30% of patients develop LID [135]. Notably, younger patients aged 40-49 years face a higher risk of LID, with approximately 50% to 70% experiencing dyskinesia after five years of levodopa treatment, compared to 42% in older patients aged 50-59 years [22]. Additionally, individuals with early-onset PD show a higher incidence of LID compared to those with late-onset PD [15]. Mild-to-moderate dyskinesia is observed in over 60% of cases, while severe or incapacitating dyskinesia is experienced by around 16% of individuals [140].

Previous studies have used various methodologies to investigate LID in PD, providing insights into its pathophysiology and risk factors and developing predictive models using clinical data. The emergence of Machine Learning (ML) techniques has significantly improved LID prognosis by integrating various factors, such as clinical, demographic, and genetic variables, to customize treatment strategies. Advanced analytical methods, such as logistic regression, Cox proportional hazards regression modeling [132], and multivariate analysis [15], have been used to investigate the complex relationship between clinical and genetic variables and their association with LID. For example, one study [141] used logistic regression to analyze motor complications in 91 patients based on the duration of levodopa therapy. Another study [140] used logistic regression focused on clinical profiles and disability to identify risk factors for LID in 110 PD patients treated with levodopa. This methodology was also used to achieve an AUC of 0.817 based on data from 430 PD patients receiving levodopa [138]. Cox proportional hazards regression models were used to investigate the impact of white matter connectivity networks among 30 LID+ patients, 47 LID- patients, and 28 controls [142]. Additionally, the relationship between cognitive dysfunction and LID development over time was assessed using data from 121 PD subjects undergoing comprehensive neuropsychological assessments in the ON state [62]. Robust statistical techniques are important for understanding the complex dynamics of LID progression and associated factors in PD populations.

LID pathophysiology is complex and influenced by various factors, reflecting the interplay between disease progression, treatment modalities, and individual patient characteristics. The stage and severity of PD are important in determining the risk of developing LID and guiding therapeutic decisions [17]. Disease progression in PD increases the risk of levodopa-related complications, which can significantly affect patients' functional abilities [3]. Additionally, there is evidence that early-onset PD [131] and specific motor symptoms such as rigidity and bradykinesia [19] may increase susceptibility to LID.

Resting tremor, a hallmark motor symptom of PD characterized by involuntary limb shaking even at rest, has been investigated for its potential role in modifying the risk of LID. Some studies suggest that resting tremor might have a protective effect against the development of LID, suggesting a negative correlation between resting tremor and dyskinetic symptoms in levodopa-treated PD patients [58, 134].

Moreover, non-motor symptoms, including cognitive impairment, depression, and anxiety, have been demonstrated to be associated with motor symptoms and have been shown to increase the risk of developing LID in patients with PD [15, 62, 132, 143]. Weight loss, frequently observed in PD patients [22], may be influenced by levodopa usage and dyskinesia [25]. Levodopa administration may also affect nutritional status dose-dependently [33]. Furthermore, it has been found that genetic factors, such as *GBA* mutations, are associated with an increased risk of motor complications in patients with PD [47, 137]. It is important to comprehend these complex factors and their interactions to develop precise interventions and personalized management strategies for LID in PD.

Pharmacological interventions are important for managing LID by modifying dopaminergic transmission to minimize peak-dose dyskinesias. These interventions include adjusting levodopa dosages and using adjunctive medications such as dopamine receptor agonists. Although MAO inhibitors are effective as an adjunct therapy in both early and advanced PD cases, their use is limited due to side effects [1]. Another pharmacological option for LID management is dopamine agonists (DAs), which delay the onset of motor complications when prescribed early in PD. However, DAs may be less well-tolerated than levodopa, particularly among older patients, due to the increased prevalence of adverse effects such as nausea and hypotension, as well as central adverse effects such as hallucinations [1].

For patients with severe and persistent LID, DBS surgery presents an alternative treatment option. DBS effectively alleviates dyskinetic symptoms and improves motor function [1]. Early DBS intervention has been associated with favorable outcomes and improved quality of life [27]. Early initiation of treatment has also been shown to increase life expectancy among PD patients, highlighting the importance of timely intervention in disease management [141].

Reducing levodopa dosage during dopaminergic therapy is a common initial step in managing dyskinesia [29]. However, its impact on parkinsonism requires careful consideration. Evidence regarding whether delaying levodopa treatment can postpone dyskinesia onset remains inconclusive [28]. These findings highlight the continued need for innovative and personalized therapeutic approaches to manage LID and improve overall PD patient outcomes.

The management of LID in PD patients requires the consideration of GI tract issues. Individuals receiving levodopa therapy must be aware of the potential impact of these issues on medication absorption and efficacy [32]. Studies have reported a high prevalence of GI symptoms in PD patients. Specifically, 88.9% of patients experience these symptoms before PD onset, detectable up to 20 years before diagnosis [25, 144]. Delayed gastric emptying is a common GI symptom in PD and has been identified as a significant factor contributing to levodopa malabsorption and subsequent motor complications. Delayed gastric emptying has been shown to negatively affect levodopa transport and absorption, potentially leading to medication overload and worsening motor symptoms [30]. The extended transit time of levodopa through the GI tract can potentially hinder its delivery to absorption sites, thereby reducing its therapeutic effects [145]. Effective management strategies for GI dysfunction often involve dietary modifications to improve gastric motility [146], optimizing medication administration to improve absorption rates, and additional treatments targeting underlying GI dysfunction [147].

Effective management of LID in PD requires careful consideration of nutritional factors due to their

significant impact on medication absorption and effectiveness [148]. Dietary factors are important in levodopa [149], the cornerstone of PD treatment. Various dietary components, such as protein and vitamins, play a role in influencing the absorption and utilization of levodopa, thus affecting treatment outcomes [30, 35]. Protein intake is a significant factor for PD patients because dietary protein can compete with levodopa for absorption in the small intestine [150]. High-protein meals have been shown to reduce levodopa absorption, leading to fluctuations in medication levels and potentially compromising symptom control [30].

Specific vitamins such as folic acid, B-6, and B-12 can also interfere with levodopa metabolism, reducing effectiveness [34, 151, 152]. Incorporating vitamin C into levodopa plus carbidopa solutions can prevent degradation and maintain medication stability [34]. Furthermore, dietary fiber and caffeine have been identified as potential stimulants of levodopa absorption [39, 153], improving the clinical response to levodopa and reducing the risk of LID in patients with PD. Patients should maintain a balanced diet of fruits, vegetables, and whole grains to ensure adequate nutrient intake without negatively affecting levodopa absorption [34].

4.1 Rationale for the study

A comprehensive understanding of LID plays a role in optimizing the management of PD. We can gain valuable insights to anticipate and address this challenging aspect of PD treatment by exploring the underlying mechanisms, identifying risk factors, and predicting LID occurrences. Studying LID can lead to the development of personalized treatment strategies based on individual patient characteristics and disease trajectories. We can identify potential biomarkers, genetic factors, and clinical predictors associated with LID susceptibility and stratify patients based on their risk profile. This enables the implementation of preventive measures or early interventions, effectively mitigating the development or progression of dyskinetic symptoms, leading to a better quality of life for individuals living with PD.

Previous studies have primarily focused on single-cohort analyses, which may introduce potential cohort-specific biases. This study aims to address this research gap by implementing a cross-cohort analysis. This approach enables the identification of more robust and generalizable predictors associated with LID. By integrating cross-study normalization, this approach accounts for variability across different patient populations, thereby enhancing the reliability of the findings. Implementing of cross-cohort analysis offers significant advantages in refining treatment regimens and improving patient care. This approach provides a more comprehensive understanding of predictors across diverse groups, enhancing the reliability of the findings.

4.2 Objective of the current study

This chapter aims to analyze LID in PD comprehensively. This will be achieved through a multifaceted approach incorporating cross-cohort analysis, ML techniques, and identifying key predictors associated with LID complications. The specific objectives of this chapter are outlined as follows:

1. Integrate cross-cohort analysis and cross-study normalization:
Examine the effectiveness of cross-cohort analysis, incorporating cross-study normalization techniques, to combine data from multiple independent cohorts. This approach identifies predictive features

associated with LID across diverse patient populations. Utilizing ML techniques to analyze these complex datasets reveals novel associations between clinical, demographic, and genetic factors and the onset of LID.

2. Develop predictive models using ML techniques:
Fine-tune model parameters and features to optimize predictive accuracy and ensure generalizability across various patient cohorts. This enables effective stratification of individuals based on LID risk.

3. Identify key predictors associated with PD Complications, Specifically LID:
To integrate heterogeneous data sources using advanced ML algorithms to identify predictive features for LID and facilitate the development of highly accurate prognostic models. This chapter aims to investigate potential interactions among different predictors and highlight their overall impact on the development of LID, providing insight into the multifactorial nature of this complication in PD.

It provides insights into the multifactorial nature of this complication, improves our understanding of LID development, predictive capabilities, and ultimately contributes to developing targeted interventions and personalized treatment strategies for PD patients at risk of developing LID.

4.3 Research methodology

The methodology used to study LID in PD is comprehensive, integrating clinical data, ML techniques, and cross-cohort analysis. Section 3.1 outlines the criteria for cohort selection. The eligible cohorts included demographic information, disease severity assessments, clinical examinations, and gene mutation profiles. This structured approach ensures uniformity in selecting participants across cohorts and facilitates collecting relevant data for LID research.

4.3.1 Inclusion criteria

The analysis of LID in PD involved assessments using standardized tools and clinical observations. Patients diagnosed with PD underwent evaluation using the Movement Disorders Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) during clinical visits. Patients were categorized as having PD with LID if they met specific criteria during assessments. To be classified as having PD with LID, the criteria included scoring ≥ 1 on either item 4.1 (amount of time spent with dyskinesias) or item 4.2 (functional impact of dyskinesias) of the MDS-UPDRS Part IV scale. The presence of dyskinesia observed during the clinical motor examination was also considered.

The total MDS-UPDRS Part IV score was excluded from further analysis to prevent duplication of dyskinesia evaluation. It is important to note that the LuxPARK cohort was assessed exclusively during the ON state. Therefore, the analysis is limited to data collected during this phase. Assessments conducted during the OFF state were excluded from consideration due to this cohort's specific protocol. The study's inclusion criteria for participants were:
Inclusion criteria (1): The diagnosis of PD must be confirmed according to the UK Parkinson's Disease Society Brain Bank Diagnostic Criteria (UKPDSBB) criteria [154]. Alternatively, subjects must have at least two of the following: rest tremor, bradykinesia, or rigidity with either rest tremor or bradykinesia,

or a single asymmetric rest tremor or asymmetric bradykinesia [103].
Inclusion criteria (2): Patients must be categorized into two groups based on LID symptoms: those who present with LID symptoms within four years of the baseline clinical visit (LID+) and those who do not show LID symptoms during these four years (LID-). These criteria ensure a focused and relevant study population for investigating LID in PD.

Table 4.1 shows the number of PD patients who met the study's inclusion criteria. The table also presents the percentage of events (LID+) used for classification and time-to-event analysis. The timeframe for LID classification was set at a four-year follow-up period from the baseline clinical visit. For the time-to-LID analysis, the outcome was defined as the time until the LID occurrence or until the last follow-up if the event was censored. This approach allowed us to account for the varying lengths of follow-up and provided a comprehensive understanding of the time-to-LID development across different patient cohorts.

Table 4.1 Number of patients meeting inclusion criteria for LID analysis.

Cohort	Inclusion criteria (1)	Inclusion criteria (2)	Events (LID Classification)	Events (Time-to-LID)
LuxPARK	706	356	210 (59.0%)	222 (62.4%)
PPMI	796	484	173 (35.7%)	348 (71.9%)
ICEBERG	162	113	36 (31.9%)	36 (31.9%)
Total	1664	953	419 (44.0%)	606 (63.6%)

Number of patients who met the inclusion criteria and the distribution of the events. The "Events" columns show the total number and percentage of subjects who developed LID during the specified period. LID classification was performed up to 4 years of follow-up, and time-to-LID analysis was performed up to the last available follow-up visit for each patient.

4.3.2 Machine learning framework

To initiate the LID analysis in PD, structured data preprocessing (Section 3.2) played a role in ensuring data quality and enabling effective modeling. An element of this step was variable aggregation (see Section 3.2.1), which involved consolidating relevant variables to reduce data dimensionality while preserving important information related to LID and PD. The dataset was subsequently simplified for subsequent analyses, enhancing computational efficiency and facilitating the interpretation of results.

After variable aggregation, a robust cross-validation (CV) framework was implemented to validate ML models effectively. Several techniques were applied within each fold of the CV process to address data complexities. Initially, incomplete data points were handled using missing value imputation techniques (Section 3.2.2) to ensure a comprehensive and reliable dataset for analysis. Subsequently, categorical variables were converted into numerical representations using categorical encoding techniques (Section 3.2.3) to enable accurate processing by ML algorithms.

Additionally, cross-study normalization techniques (Section 3.2.4) were applied to continuous variables to mitigate biases and ensure consistency in the analysis. This technique standardizes data across different cohorts, which is particularly important when dealing with datasets from diverse sources or cohorts with distinct data distributions. These data preprocessing steps established a basis for the ML analyses that focused on LID classification and time-to-LID analyses in patients with PD.

We implemented undersampling techniques (Section 3.2.5) specifically on the training set to address the class imbalance challenge in clinical datasets. This approach helped to balance the distribution of

outcomes within the training data, preventing ML models from showing bias towards the majority class and improving their ability to generalize to unseen data.

The procedure for selecting features (Section 3.2.6) was integrated into a nested CV framework. This framework combined feature selection and hyperparameter tuning within the nested CV process to discover predictive features while optimizing model generalization and robustness. Within each fold of the nested CV, feature selection techniques were used to identify the most informative and relevant features associated with LID. The techniques were developed to decrease dimensionality, focusing on the most relevant features associated with LID while mitigating the risk of overfitting.

ML classification techniques were used to classify patients into two groups based on their 4-year follow-up visits: LID+ (those showing LID) and LID- (those without LID). The objective was to apply different ML algorithms, as described in Section 3.3, within a CV framework (Section 3.5) to ensure the robustness and generalizability of the models across different patient cohorts.

In addition to LID classification, the study also included time-to-LID analysis, considering the duration until the onset of LID symptoms for each patient up to their last follow-up visit. This phase used ML models, including time-to-event analysis techniques outlined in Section 3.4, to estimate the risk score of LID development over time based on individual patient characteristics.

The goal of integrating ML analyses for both LID classification and time-to-LID analyses was to provide a comprehensive understanding of the factors contributing to the development of LID in PD. These analyses facilitated accurate patient classification into LID+ and LID- groups and enabled the prediction of key milestones related to LID onset. These insights are of particular value to clinicians, as they may facilitate the personalized development of treatment strategies and the early interventions customized to the specific needs of individual patients.

This study used model interpretability techniques, such as SHAP values (Section 3.6), to understand the ML models used for LID classification and time-to-LID analyses. The SHAP values helped to identify the features that significantly influence the decision-making process of the models, enhancing transparency and aiding in understanding how the models derive their predictions. By using SHAP values, we gained insights into how the ML models work and improved their predictions' trustworthiness.

The predictors identified through the time-to-LID model are also important in understanding PD-associated complications. We used hazard ratios (HR) derived from integrating SHAP values analysis to deepen our understanding of these predictors (see Section 3.6). HR provides a quantitative measure of the relative risk associated with each predictive feature, offering valuable insights into the likelihood of developing LID over time based on various contributing factors.

This study developed two distinct prediction models to analyze the onset of LID in PD patients. The first model, known as *comprehensive* model, included all baseline clinical features shared across the cohorts without prior feature selection. This model aimed to use as much information as possible from the available data to capture a range of factors contributing to LID prediction.

In contrast, the second model, the *refined* model, focused on a subset of clinical features that excluded baseline dyskinesia and levodopa medication. The *refined* model aimed to identify additional risk factors or biomarkers that might independently contribute to LID development by excluding baseline LID symptoms or levodopa treatment.

The rationale for developing both models was to comprehensively explore different aspects of LID prediction. The *comprehensive* model considered all available features, including baseline dyskinesia and levodopa medication, providing a comprehensive view. However, the *refined* model allowed for a

more subtle analysis by isolating other potential predictors. Comparing these models' performance and feature importance provided insights into the relative contributions of different clinical features and their predictive power for LID onset in PD patients.

To assess the predictive performance of the models, comparisons were made between the *comprehensive* and *refined* models and between the non-normalized and normalized models within the same cohort analysis. As discussed in Section 3.7, we used hold-out AUC or C-index metrics to evaluate the classification and time-to-event analysis.

Comparing the *comprehensive* and *refined* models was intended to identify specific features significantly influencing LID prediction. Additionally, evaluating the performance of non-normalized versus normalized models within the same cohort provided insight into how cross-study normalization procedures affected model predictive ability. These comparisons were important for comprehending model behavior across various feature sets and ensuring reliable predictive performance in predicting LID onset for PD patients.

In addition to evaluating the predictive performance of different models, we conducted cross-validated performance metric comparisons using the Bayesian signed-rank test discussed in Section 3.7. This method provides probabilities indicating the superiority of one model over another based on their performance metrics. By leveraging Bayesian techniques, we obtain more interpretable results regarding the relative strengths of predictive models in capturing the complexities of LID onset prediction in PD patients. This approach goes beyond simplistic comparisons to offer probabilistic insights into the efficacy and reliability of models across multiple validation sets, thereby enhancing the robustness of our analyses.

The stability of the *optimized* model for each cohort analysis was thoroughly evaluated to ensure reliable and consistent predictive performance across various iterations and data subsets. A stable model shows robustness in its ability to generalize well to unseen or future data samples, providing confidence that the model's performance is not overly sensitive to specific subsets of the data and can consistently perform well on new, unseen data. Conversely, models lacking stability may show higher variance or overfitting issues. Stability analysis identifies such issues, ensuring the model's performance remains consistent across validation sets and increasing its reliability in real-world settings.

we conducted statistical analyses to assess their predictive capabilities and compare selected features across the 5-fold CV in a single cohort analysis. The analysis involved calculating percentages to compare the selected features' frequency across each CV fold. This approach provided valuable insights into the consistency and variability of feature selection across multiple model training and evaluation iterations. The study highlighted key features that significantly contribute to predictive performance by identifying stable predictors across various folds, as discussed in Section 3.8.

4.3.3 Statistical analysis

The study included univariate analysis to assess the relationships between each predictor and the outcomes within four years (LID+ and LID-). The significance levels of these tests were used to identify predictors strongly associated with LID development within the specified timeframe. Additionally, the study investigated whether these associations varied significantly across different cohort groups.

Furthermore, a correlation analysis was conducted to examine the relationships between different predictors among features. Correlation coefficients were calculated to measure the strength and direction of linear or monotonic relationships between pairs of predictors. This analysis aimed to identify highly

correlated features that could potentially impact model interpretability. The detailed methods for the univariate and correlation analysis can be found in Section 3.9 of the study.

4.3.4 Clinical utility analysis

To assess the clinical utility and reliability of the prediction models for LID, we conducted decision curve analysis (DCA) and calibration analysis, with detailed methodologies provided in Section 3.10. The DCA approach ensures that our models show satisfactory predictive performance and offer meaningful benefits for clinical decision-making and PD patient outcomes.

The study assessed the clinical utility of the predictive models for LID by measuring the area under the net benefit curve (AUNBC). This metric quantifies the effectiveness of the models regarding clinical decision-making, whereby the net benefits of different intervention strategies are compared. The AUNBC was calculated for the *optimized* models and then compared to the area under the curve for the "treat all" strategy. This comparison is intended to illustrate the additional benefit of utilizing predictive models instead of a baseline strategy wherein all patients receive intervention irrespective of their predicted risk. A higher AUNBC for the *optimized* models indicates they are more clinically helpful in managing LID in PD.

Furthermore, the study applied bootstrapping hypothesis testing to calculate p-values and assess the significance of the observed differences. The null hypothesis proposed that there would be no significant difference in the AUNBC between the two *optimized* models. The p-values were adjusted using the Benjamini-Hochberg procedure to accommodate the multiple comparisons. This adjustment ensures that the reported p-values accurately reflect the probability of the differences, thereby validating the *optimized* models' improved performance in predicting LID in PD.

Similarly, a calibration analysis was conducted to assess the degree of correspondence between the predicted probabilities and the actual outcomes for LID classification, as well as the predicted conversion probabilities and the observed conversion probabilities at year-4 for the time-to-LID model. The evaluation measures the slope and the MSE of the calibration curve, thereby providing insights into the models' accuracy and reliability in predicting LID. A slope of approximately 1 and a lower MSE indicate superior calibration, suggesting a robust alignment between predicted and observed outcomes. This calibration analysis reinforces the robustness and clinical utility of the predictive models.

4.3.5 Code availability

R (v4.2.1) was used for data processing, normalization, and statistical analyses, while Python-3.8.6-GCCcore-10.2.0 was used for ML predictions. The open-source code is available in the GitLab repository under the MIT license at
https://gitlab.com/uniluxembourg/lcsb/biomedical-data-science/bds/ml_dyskinesia.

4.4 Results

4.4.1 Individual cohort analyses

The *optimized* models for each cohort were selected based on the highest average AUC scores for the LID classification model and the highest average C-index values for the time-to-LID model during 5-fold

CV, as highlighted in Tables 4.2, 4.3, A.1, and A.2, respectively. The hold-out performance metrics for the *optimized comprehensive* and *refined* LID classification models, including precision, recall, F-score, accuracy, and balanced accuracy, closely followed the hold-out AUC trends (these additional metrics are not shown in the thesis). These metrics offer insights into the robustness and generalizability of predictive models for LID across diverse patient cohorts. Table 4.4 presents the results of DeLong's and one-shot nonparametric tests, which compare the hold-out predictive performance (AUC/C-index) of the *optimized* models in each cohort.

The *comprehensive* LID classification model's predictive performance was evaluated using AUC scores across the LuxPARK, PPMI, and ICEBERG cohorts. In the LuxPARK cohort, the *optimized comprehensive* model achieved an average cross-validated AUC of 0.735 (SD 0.091) and a corresponding hold-out AUC of 0.678, utilizing 37 features. The hold-out prediction performance of the model showed consistency, with balanced accuracy matching the AUC and a precision of 0.686, highlighting its effectiveness in making accurate predictions while managing class imbalances. However, the *optimized refined* LID classification model in LuxPARK also showed satisfactory performance, with an average cross-validated AUC of 0.706 (SD 0.051) and a hold-out AUC of 0.655, utilizing a reduced feature set 29. The DeLong test for LuxPARK yielded a non-significant p-value of 0.925, indicating no significant difference between the hold-out AUC of the *optimized comprehensive* and *refined* models.

In the PPMI cohort, the *optimized comprehensive* LID classification model showed an average cross-validated AUC of 0.629 (SD 0.071) and a hold-out AUC up to 0.656. In contrast, the *optimized refined* model in PPMI showed an average cross-validated AUC of 0.639 (SD 0.091) and a hold-out AUC of 0.659, with 9 features. DeLong's test for PPMI yielded a significant p-value of 0.036, indicating that the *optimized refined* model outperforms the *optimized comprehensive* LID classification model in terms of hold-out AUC. The *refined* LID classification model showed improved hold-out predictive performance compared to the *comprehensive* model, particularly in precision, recall, and F-score, showing a better balance classification. While both models have similar balanced accuracy, the *refined* model has higher recall and F-score, highlighting its improved ability to identify LID outcomes while maintaining precision correctly.

Similarly, the *optimized comprehensive* LID classification model in the ICEBERG cohort showed an average cross-validated AUC up to 0.595 (SD 0.144) and a hold-out AUC up to 0.652. The *optimized refined* model achieved an average cross-validated AUC of 0.558 (SD 0.122) and a hold-out AUC of 0.515, utilizing only two features. The GOSDT-GUESSES-*optimized* model achieved a hold-out AUC of 0.718 despite a lower average cross-validated AUC. The DeLong test for ICEBERG yielded a non-significant p-value of 0.776, indicating no significant difference between the hold-out AUC of the *optimized comprehensive* and *refined* models.

The *optimized comprehensive* model showed varied performance across cohorts regarding the time-to-LID model. In LuxPARK, the average cross-validated C-index was 0.714 (SD 0.027), while the hold-out C-index was 0.577, utilizing 11 features. By comparison, another model using 13 predictors achieved an average cross-validated C-index of 0.667 (SD 0.055) and a hold-out C-index of 0.651. The *optimized refined* time-to-LID model in LuxPARK showed a slightly lower average cross-validated C-index of 0.701 (SD 0.051) with a comparable hold-out C-index of 0.648. The one-shot nonparametric test for LuxPARK indicated that the *optimized refined* model outperformed the *comprehensive* model regarding the hold-out C-index with p-values of 0.020.

In the PPMI cohort, the *optimized comprehensive* time-to-LID model displayed satisfactory perfor-

mance metrics, boasting an average cross-validated C-index of 0.697 (SD 0.041) and a hold-out C-index of 0.663, all achieved using a feature set 28. Conversely, the *optimized refined* time-to-LID model in PPMI showed comparable results, with cross-validated and hold-out C-index values of 0.688 (SD 0.048) and 0.662, respectively, utilizing a slightly more extensive feature set of 32. Interestingly, the one-shot nonparametric test for PPMI indicated no significant difference (p-value 0.908) between the hold-out C-index values of the *optimized comprehensive* and *refined* models, highlighting their similar predictive capabilities in this specific cohort.

However, the *optimized comprehensive* time-to-LID model showed moderate performance in the ICEBERG cohort, with an average cross-validated C-index of 0.576 (SD 0.115) and a hold-out C-index up to 0.585, using four features. The *optimized refined* time-to-LID model had a similar hold-out C-index (p-values 0.367), reaching 0.572. This was accompanied by a slightly lower cross-validated C-index of 0.586 (SD 0.108), utilizing nine features.

The Bayesian signed-rank test probabilities provide insights into models' relative cross-validated predictive performance across cohorts. About *optimized comprehensive* LID classification, the probabilities indicate a high likelihood (1.00) of the LuxPARK model's predictive superiority over PPMI, a similarly high likelihood (0.97) over ICEBERG, and a moderate likelihood (0.68) of PPMI's superiority over ICEBERG, as shown in Figure 4.1. For the *optimized refined* LID classification models (see Figure A.1), the probabilities indicate a high likelihood (0.93) of LuxPARK's model being superior to PPMI, an even higher likelihood (0.99) over ICEBERG, and again, a high likelihood (0.99) of PPMI's model being superior to ICEBERG.

In the *optimized comprehensive* time-to-LID models, the probabilities strongly suggest that LuxPARK's model is highly likely (0.90) to outperform PPMI in cross-validated C-index, even more (0.99) than ICEBERG. Similarly, there is a high probability (0.99) that PPMI's model performs better than ICEBERG's model in this context (see Figure 4.2). Finally, for the *optimized refined* time-to-LID models, displayed in Figure A.2, the probabilities indicate a moderate likelihood (0.67) of LuxPARK's model being superior to PPMI, a high likelihood (0.97) over ICEBERG and a similar high likelihood (0.97) of PPMI's model being superior to ICEBERG.

Assessing the stability of predictive models is used to understand their reliability across diverse cohorts or variations within the same dataset. In this study, we measured the stability of the *comprehensive* and *refined* LID classification models (Figures A.3 and A.4) and the *comprehensive* and *refined* time-to-LID models (Figures A.5 and A.6) across each cohort analysis using the SD of their performance metrics from the cross-validation framework. Notably, the ICEBERG cohort analysis displayed lower stability, indicated by higher SD, suggesting more variability in performance. Conversely, consistent stability trends among algorithms were observed in LuxPARK and PPMI cohorts, underscoring their stable predictive performance across different analyses.

Table 4.2 Predictive performance metrics for *comprehensive* LID classification in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**
AdaBoost	0.647 (0.079)	0.532	4 (6)	0.623 (0.059)	0.582	2 (6)	0.481 (0.122)	0.512	1 (2)
CART	0.610 (0.059)	0.561	2 (3)	0.607 (0.043)	0.555	9 (10)	0.509 (0.069)	0.512	1 (2)
CatBoost	0.664 (0.068)	0.577	5 (6)	***0.626 (0.060)***	***0.656***	***14 (24)***	***0.546 (0.078)***	***0.652***	***8 (16)***
C4.5	0.607 (0.069)	0.559	2 (7)	**0.629 (0.071)**	**0.570**	**5 (11)**	0.540 (0.097)	0.390	1 (6)
FIGS	0.584 (0.121)	0.539	7 (11)	0.602 (0.070)	0.557	3 (9)	0.482 (0.092)	0.512	1 (2)
GOSDT-GUESS	0.612 (0.132)	0.562	10 (10)	0.592 (0.057)	0.521	21 (36)	0.521 (0.197)	0.448	8 (10)
GBoost	0.650 (0.089)	0.641	16 (24)	0.628 (0.032)	0.621	17 (24)	0.533 (0.073)	0.607	6 (7)
HS	0.620 (0.061)	0.530	5 (9)	0.597 (0.077)	0.557	3 (9)	0.482 (0.092)	0.512	1 (2)
XGBoost	***0.735 (0.091)***	***0.678***	***37 (59)***	0.602 (0.052)	0.621	9 (9)	**0.595 (0.144)**	**0.533**	**13 (14)**

An overview of the *comprehensive* LID prognostic classification's predictive performance statistics summarizes the *comprehensive* LID prognostic classification's predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 4.3 Predictive performance metrics for *comprehensive* time-to-LID in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features
CW-GBoost	***0.667 (0.055)***	***0.651***	***13 (17)***	0.696 (0.033)	0.640	14 (21)	**0.576 (0.115)**	**0.451**	**4 (11)**
Extra Survival	0.699 (0.101)	0.610	11 (11)	0.694 (0.054)	0.651	70 (97)	0.548 (0.088)	0.542	9 (9)
Survival GBoost	0.648 (0.069)	0.647	119 (120)	0.669 (0.074)	0.643	17 (22)	***0.570 (0.132)***	***0.585***	***49 (58)***
LSVM	0.628 (0.043)	0.614	15 (15)	0.670 (0.040)	0.652	35 (35)	0.531 (0.116)	0.557	103 (104)
NLSVM	0.642 (0.017)	0.618	20 (20)	0.673 (0.045)	0.650	29 (29)	0.561 (0.142)	0.452	10 (10)
Penalized Cox	0.685 (0.046)	0.532	1 (5)	***0.697 (0.041)***	***0.663***	***28 (51)***	0.551 (0.092)	0.517	4 (4)
Survival RF	**0.714 (0.027)**	**0.577**	**11 (11)**	0.672 (0.061)	0.650	46 (72)	0.543 (0.114)	0.549	10 (33)
Survival Trees	0.613 (0.052)	0.582	4 (7)	0.644 (0.098)	0.622	8 (12)	0.564 (0.087)	0.498	1 (3)

An overview of the *comprehensive* time-to-LID predictive performance statistics summarizes the *comprehensive* time-to-LID predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-index in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out C-index is indicated in *italics*. The column labeled 'Number of features' displays the number of features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

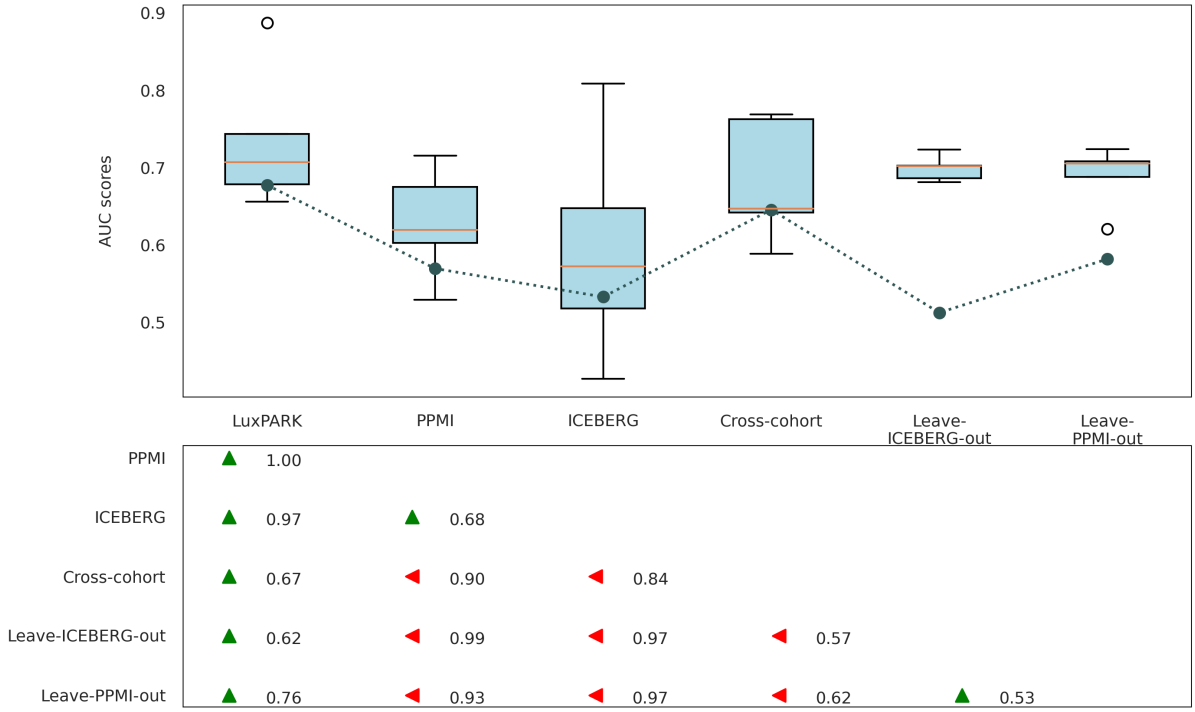
Table 4.4 Significance testing of hold-out predictive metrics between normalized and unnormalized models for LID in multi-cohort analyses.

Cohort	LID classification	Time-to-LID
LuxPARK	0.925	0.020
PPMI	0.036	0.908
ICEBERG	0.776	0.367
Cross-cohort	0.092	0.019
Leave-ICEBERG-out	0.399	1.000
Leave-PPMI-out	0.954	0.235

A comparison of the statistical significance of the differences between the hold-out predictive performance metrics for the *optimized comprehensive* and *refined* models across cohorts. The *p*-values for the significance of the difference were calculated using DeLong's test for LID classification and the one-short nonparametric test for time-to-LID analysis. A p-value < 0.05 indicates a significant difference in hold-out predictive performance between the two models.

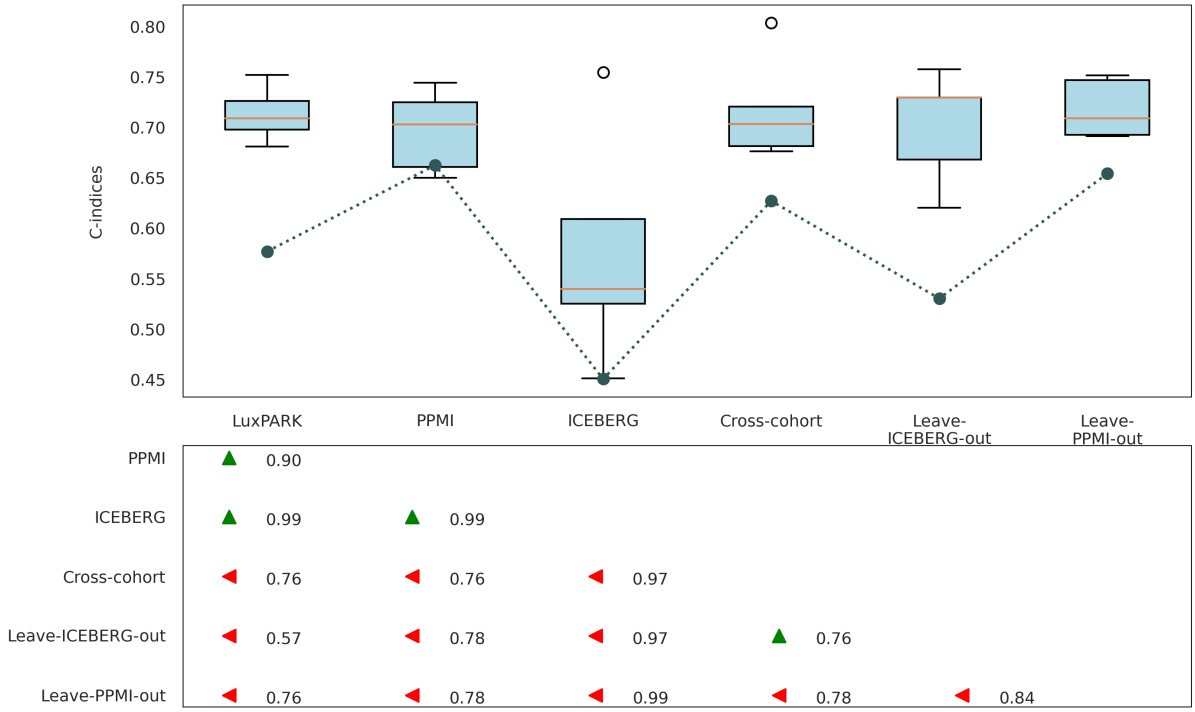
The key predictors derived from the *optimized comprehensive* and *refined* models for LID classification and time-to-LID analyses within each single cohort analysis, LuxPARK, PPMI, and ICEBERG, are detailed in Table 4.5. These predictors, ranked by their average selection percentage during 5-fold cross-validation, emphasize PD complication factors, with PD disease duration and age at PD onset at the top of the list. The importance of these predictors lies in their ability to provide important insights into the predictive

Figure 4.1 Comparison of cross-validated AUC scores for *comprehensive* LID classification models.



A comparison of cross-validated AUC scores and probabilities of better predictive performance for the *optimized comprehensive* LID classification model across cohort analyses. The upper part displays boxplots of the cross-validated AUC scores for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

Figure 4.2 Comparison of cross-validated C-indices for *comprehensive* time-to-LID models.



A comparison of cross-validated C-indices and probabilities of better predictive performance for the *optimized comprehensive* time-to-LID model across cohort analyses. The upper part displays boxplots of the cross-validated C-indices for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

modeling of LID and time-to-LID onset. They serve as fundamental indicators of LID onset in different cohorts, highlighting their consistent relevance and importance across different analyses and cohorts.

4.4.2 Cross-cohort analyses

The predictive performance of the *optimized* LID classification and time-to-LID models was evaluated using 5-fold cross-validation across cross-cohort and leave-one-cohort-out analyses. The results are presented in Tables 4.6 and 4.7 for the *comprehensive* models and Tables A.3 and A.4 for the *refined* models. The hold-out metrics, including precision, recall, F-score, accuracy, and balanced accuracy, showed trends similar to those observed for AUC. Table 4.4 compares hold-out predictive performance (AUC/C-index) between the *optimized comprehensive* and *refined* models in each cohort analysis, including the corresponding *p*-values.

Across the multi-cohort analyses, the cross-cohort analysis of the *optimized comprehensive* LID classification models showed robust predictive performance, with a mean cross-validated AUC of 0.682 (SD 0.080) and a hold-out AUC of 0.646 using 8 features. The CatBoost-*optimized* model performed better than the *optimized* model regarding hold-out AUC and achieving higher precision and F-score values above 0.70 while maintaining robust recall and accuracy. However, the leave-ICEBERG-out analysis showed a slight increase in the average cross-validated area under the curve (AUC) to 0.699 (SD 0.016) and a decline in the hold-out AUC to 0.513, highlighting the challenges in model generalization. The highest hold-out AUC was 0.664, with the average cross-validated AUC of 0.691 (SD 0.052). Similarly, the leave-PPMI-out analysis yielded an average cross-validated AUC of 0.690 (SD 0.040) and a hold-out AUC of 0.582. Another *optimized* model achieved a higher hold-out AUC of 0.627, with an average cross-validated AUC of 0.628 (SD 0.034), suggesting potential areas for improvement in the model when excluding the PPMI cohort from training.

Regarding time-to-LID models, the *optimized comprehensive* analysis achieved a mean cross-validated C-index of 0.718 (SD 0.052) and a hold-out C-index of 0.627 using 36 features in the cross-cohort analysis. In contrast, the leave-ICEBERG-out analysis showed a reduced hold-out C-index of 0.531 with an average cross-validated C-index of 0.701 (SD 0.056) and a more extensive feature set of 57, suggesting the potential for overfitting. The model with a higher hold-out AUC, utilizing fewer predictive features, showed a balance between cross-validated and hold-out performance, achieving an average cross-validated AUC of 0.686 (SD 0.046) and hold-out AUC of 0.684. Conversely, the leave-PPMI-out analysis showed an average cross-validated C-index of 0.719 (SD 0.029) and a competitive hold-out C-index of 0.655, indicating better generalization capabilities.

The *optimized refined* LID classification models were examined using a cross-cohort analysis, which yielded an average cross-validated AUC of 0.688 (SD 0.043) and a hold-out AUC of 0.639 using 5 features. This indicates streamlined predictive power. DeLong's test with a *p*-value of 0.092 showed no significant difference in hold-out predictive performance compared to the *comprehensive* LID classification model. However, the leave-ICEBERG-out approach slightly reduced the hold-out AUC to 0.534 with 6 features and an average cross-validated AUC of 0.692 (SD 0.021), indicating that challenges remain in excluding ICEBERG data from training. The leave-PPMI-out approach yielded a mean cross-validated AUC of 0.679 (SD 0.033) and a hold-out AUC of 0.599 with 9 features. DeLong's test indicated non-significant *p*-values of 0.399 for the leave-ICEBERG-out analysis and 0.954 for the leave-PPMI-out analysis.

Similarly, the *refined* time-to-LID models showed comparable performance in the cross-cohort analysis, with an average cross-validated C-index of 0.715 (SD 0.054) and a hold-out C-index of 0.685. The

Table 4.5 The average percentage of predictors selected in 5-fold cross-validation in LID analysis for classification and time-to-event across LuxPARK, PPMI, and ICEBERG cohorts.

Predictors	*Comprehensive* model			*Refined* model		
	Classification Average in CV (%)	Time-to-LID Average in CV (%)	Overall Average (%)	Classification Average in CV (%)	Time-to-LID Average in CV (%)	Overall Average (%)
Disease duration	86.7	46.7	66.7	80.0	80.0	80.0
Age of onset	73.3	53.3	63.3	80.0	80.0	80.0
MDS-UPDRS I - Urinary problems	53.3	46.7	50.0	46.7	93.3	70.0
MDS-UPDRS I - Sleep problems (night)	66.7	53.3	60.0	46.7	73.3	60.0
BMI (kg/m2)	66.7	33.3	50.0	60.0	66.7	63.3
MDS-UPDRS Part II score	60.0	46.7	53.3	60.0	60.0	60.0
Benton Judgment of Line Orientation	46.7	46.7	46.7	53.3	66.7	60.0
MDS-UPDRS Part I score	53.3	46.7	50.0	40.0	73.3	56.7
SCOPA-AUT Gastrointestinal (GI)	66.7	40.0	53.3	26.7	73.3	50.0
Axial symptoms score	33.3	46.7	40.0	46.7	80.0	63.3
Weight (kg)	46.7	40.0	43.3	33.3	80.0	56.7
Height (cm)	46.7	33.3	40.0	46.7	66.7	56.7
MDS-UPDRS II - Saliva and drooling	40.0	53.3	46.7	33.3	66.7	50.0
SCOPA-AUT Urinary	60.0	26.7	43.3	33.3	73.3	53.3
SCOPA-AUT Thermoregulatory	60.0	26.7	43.3	33.3	73.3	53.3

An overview of the feature selection analysis performed using 5-fold cross-validation compares the average percentage of times clinical features were selected between the *comprehensive* and *refined* models for LID classification and time-to-LID analyses across the LuxPARK, PPMI, and ICEBERG cohorts. The column labeled 'Average in CV (%)' displays the average percentage of times each feature was selected during the 5-fold cross-validation in single cohort analyses within LuxPARK, PPMI, and ICEBERG for both LID and time-to-LID analyses. The column labeled 'Average (%)' represents the mean of the 'Average in CV (%)' values across the cohorts for analyses of LID and time-to-LID. The top 15 predictors are listed in descending order based on their average selection percentages for the *comprehensive* and *refined* models in LID and time-to-LID analyses.

one-shot nonparametric test yielded a significant p-value of 0.019 for the cross-cohort analysis, indicating a significant difference in hold-out predictive performance between the *optimized comprehensive* and *refined* models. However, leave-ICEBERG-out or leave-PPMI-out resulted in lower hold-out C-index values (0.531 and 0.634, respectively), suggesting potential challenges in generalizing time-to-LID models were tested across diverse cohorts with no significant difference to the *comprehensive* model (p-values of 1.000 for leave-ICEBERG-out and 0.235 for leave-PPMI-out analysis).

The Bayesian signed-rank tests offer valuable insights into the cross-validated performance of models across different analyses. Figures 4.1 and A.1 illustrate the comparisons for *comprehensive* and *refined* LID classification. The cross-cohort model consistently performed superior to single cohort analyses, with probabilities exceeding 0.8. However, it showed slightly inferior performance compared to the LuxPARK cohort, with moderate probabilities of 0.67 and 0.62 for the *optimized comprehensive* and *refined* models. This indicates the model's effectiveness in cross-cohort analysis, particularly its superior performance when compared to *optimized* models, specifically from the PPMI and ICEBERG cohorts.

Similarly, the cross-cohort analysis showed a competitive advantage in time-to-LID models illustrated in Figures 4.2 and A.2. It showed probabilities ranging from 0.76 to 1.00 in superiority over single cohort models, indicating reliable and consistent predictive capabilities across varied cohorts. Notably, the *optimized* model from the cross-cohort analysis also outperformed the leave-one-cohort-out analysis with probabilities of 0.76 and above. However, the *optimized* model from the leave-PPMI-out analysis showed superiority over the cross-cohort model with a probability of 0.78 for *comprehensive* time-to-LID models. These findings highlight the robustness of the cross-cohort approach in capturing diverse dataset characteristics and ensuring generalizable predictive performance across different cohorts or datasets.

The stability analysis across cohort scenarios, particularly in cross-cohort comparisons, clarifies the robustness of predictive models. In *comprehensive* LID classification (Figure A.3), numerous algorithms showed notably higher stability in multi-cohort analyses than in single-cohort analyses. This trend extended to *refined* LID classification (Figure A.4) and time-to-LID models (Figures A.5 and A.6), where stable performances were consistently observed across multi-cohort analyses. These findings highlight the reliability of the models, which is important for their practical application across diverse cohorts.

Table 4.6 Predictive performance metrics for *comprehensive* LID classification in multi-cohort analyses.

Algorithm	Cross-cohort			Leave-ICEBERG-out			Leave-PPMI-out		
	Mean (SD)	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**
AdaBoost	**0.682 (0.080)**	**0.646**	**8 (14)**	***0.691 (0.052)***	***0.664***	***4 (7)***	0.667 (0.039)	0.528	4 (5)
CART	0.662 (0.059)	0.647	4 (7)	0.671 (0.016)	0.535	2 (4)	0.648 (0.057)	0.518	15 (21)
CatBoost	***0.675 (0.062)***	***0.702***	***9 (17)***	0.697 (0.045)	0.600	18 (27)	0.675 (0.099)	0.626	8 (18)
C4.5	0.631 (0.111)	0.664	5 (9)	0.669 (0.026)	0.535	2 (4)	***0.628 (0.034)***	***0.627***	***1 (3)***
FIGS	0.659 (0.059)	0.673	3 (5)	0.694 (0.040)	0.606	9 (20)	0.659 (0.051)	0.594	9 (20)
GOSDT-GUESS	0.622 (0.064)	0.613	42 (59)	0.644 (0.018)	0.547	38 (53)	0.623 (0.070)	0.542	27 (43)
GBoost	0.656 (0.054)	0.672	5 (5)	**0.699 (0.016)**	**0.513**	**15 (26)**	0.670 (0.042)	0.619	14 (21)
HS	0.660 (0.033)	0.664	2 (3)	0.692 (0.039)	0.606	9 (20)	0.659 (0.051)	0.594	9 (20)
XGBoost	0.654 (0.032)	0.632	52 (66)	0.690 (0.036)	0.631	35 (64)	**0.690 (0.040)**	**0.582**	**58 (74)**

An overview of the *comprehensive* LID prognostic classification's predictive performance statistics summarizes the *comprehensive* LID prognostic classification's predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each optimized model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 4.7 Predictive performance metrics for *comprehensive* time-to-LID in multi-cohort analyses.

Algorithm	Cross-cohort			Leave-ICEBERG-out			Leave-PPMI-out		
	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features
CW-GBoost	0.712 (0.045)	0.673	14 (25)	0.696 (0.049)	0.639	13 (23)	0.702 (0.022)	0.613	10 (16)
Extra Survival	0.704 (0.045)	0.667	160 (161)	0.698 (0.050)	0.605	160 (161)	0.696 (0.049)	0.651	12 (12)
Survival GBoost	0.704 (0.061)	0.661	24 (49)	***0.686 (0.046)***	***0.684***	***11 (21)***	**0.719 (0.029)**	**0.655**	**13 (22)**
LSVM	**0.718 (0.052)**	**0.627**	**36 (36)**	0.681 (0.048)	0.547	52 (52)	0.701 (0.025)	0.639	147 (147)
NLSVM	0.701 (0.068)	0.652	52 (52)	0.681 (0.052)	0.681	56 (56)	***0.705 (0.016)***	***0.669***	***36 (36)***
Penalized Cox	0.692 (0.055)	0.666	28 (28)	**0.701 (0.056)**	**0.531**	**57 (96)**	0.691 (0.077)	0.551	1 (32)
Survival RF	***0.705 (0.055)***	***0.682***	***134 (139)***	0.701 (0.048)	0.612	104 (136)	0.689 (0.036)	0.663	116 (122)
Survival Trees	0.649 (0.078)	0.646	8 (13)	0.661 (0.042)	0.491	13 (16)	0.654 (0.048)	0.512	11 (17)

An overview of the *comprehensive* time-to-LID predictive performance statistics summarizes the *comprehensive* time-to-LID predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-index in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

4.4.3 Differences between clinical features across cohorts

Significant differences were observed across cohorts in critical aspects of PD progression and symptomatology, as shown in Table 4.8. Firstly, the LuxPARK cohort showed a significantly longer average PD disease duration than the PPMI and ICEBERG cohorts, with differences of 6.7 years and 7.3 years, respectively. About the age of PD onset, LuxPARK showed a significantly younger average age of onset compared to PPMI and ICEBERG, with differences of 3.1 years and 5.2 years, respectively. Furthermore, analyzing motor and non-motor symptom severities using established metrics such as MDS-UPDRS Parts I to III and SCOPA-AUT scores revealed significant differences among the cohorts. The LuxPARK and ICEBERG cohorts showed significantly more severe motor and non-motor symptoms than the PPMI cohort. These distinctions highlight the heterogeneity in disease manifestation and symptomatology across the cohorts.

4.4.4 Comparative evaluation of cross-study integration methods

An assessment of normalized and unnormalized *optimized* models for *comprehensive* and *refined* LID classification and time-to-LID models aims to determine if cross-study normalization can improve hold-out predictive performance. These comparative findings are detailed in Table 4.9. Across both *comprehensive* and *refined* LID classification analyses, DeLong's test p-values indicate that there are no statistically significant differences in hold-out AUC between the *optimized* models with and without

Table 4.8 Comparative analysis of baseline features mean differences across cohorts in LID analysis.

Predictors	LuxPARK vs. PPMI (*p*-values)	LuxPARK vs. ICEBERG (*p*-values)	PPMI vs. ICEBERG (*p*-values)	*p*-values
Age of onset	-3.08 (7.82E-04)	-5.22 (4.06E-05)	-2.14 (0.120)	7.29E-06
Disease duration	6.71 (4.86E-76)	7.33 (3.73E-43)	0.62 (0.168)	3.86E-87
Weight (kg)	-0.01 (1.000)	4.44 (0.078)	4.46 (0.057)	5.26E-02
Height (cm)	-1.66 (0.070)	-1.55 (0.500)	0.10 (1.000)	6.50E-02
BMI (kg/m^2)	0.52 (0.237)	2.03 (2.97E-04)	1.51 (0.013)	4.86E-04
MDS-UPDRS Part I score	8.31 (8.49E-83)	1.88 (1.000)	-6.43 (3.43E-33)	2.02E-92
MDS-UPDRS Part II score	5.89 (5.30E-34)	4.26 (1.79E-04)	-1.63 (1.68E-04)	2.24E-33
MDS-UPDRS Part III (ON) score	16.71 (5.65E-30)	7.21 (0.002)	-9.50 (1.24E-08)	2.09E-29
SCOPA-AUT total score	10.51 (3.91E-93)	3.24 (0.021)	-7.27 (2.98E-27)	1.31E-98

A comparative analysis of the mean differences for baseline features across the LuxPARK, PPMI, and ICEBERG cohorts. The *p*-values indicated statistically significant differences in the average of the predictors between specific cohort pairs, providing insights into cohort-specific variations in predictor distributions in LID analysis.

normalization, regardless of the analysis, including cross-cohort, leave-ICEBERG-out, and leave-PPMI-out.

About time-to-LID models, the results from one-shot nonparametric tests for hold-out C-index mirror the patterns observed in LID classification analyses. The ratio-A normalization technique has significantly improved hold-out performance, particularly in the leave-PPMI-out analysis (see Table 4.10). While not all normalized *optimized* models showed notable superiority over their unnormalized counterparts, there is still improvement in hold-out performance. These findings show the potential of cross-study normalization techniques on model efficacy across multi-cohort analyses.

4.4.5 Associations between clinical features and dyskinesia outcome

The cross-cohort analysis identified the most significant predictors for LID using permutation importance across *optimized comprehensive* and *refined* models for LID classification and time-to-LID analysis. Disease duration emerged as the most significant predictor, consistently ranking highest across models. Other significant predictors included motor fluctuations, levodopa medication intake, and selective axial symptoms, all of which contributed substantially to the prediction models (see Table 4.11).

Across both the *comprehensive* and *refined* LID classification models (Figures 4.3 and A.7), key predictors such as disease duration, motor fluctuation, tremor, age of onset, and body weight consistently emerge as influential features. Additional predictors like levodopa treatment, PD disease severity, rigidity at the lower extremities, thermoregulatory, axial symptoms, and BMI collectively contribute to understanding the complexity of LID development and classification.

Similarly, in the time-to-LID models (Figures 4.4 and A.8), key predictors include motor symptoms (tremor, bradykinesia), rigidity, visuospatial ability (Benton Judgment of Line Orientation (JLO)), age of PD onset, and freezing of gait. These predictors, in conjunction with more subtle predictors such as non-motor symptoms (hygiene, depressed moods, apathy, gastrointestinal), levodopa treatment, *GBA*

mutation, motor fluctuation, PD disease duration, axial symptom, and motor severity (MDS-UPDRS Part III), provide insight into the temporal dynamics of LID onset.

Specific predictors, such as motor fluctuation, PD disease duration, age of PD onset, and specific motor and non-motor symptom scores, consistently feature across multiple models. This highlights their robust influence on both LID classification and time-to-LID analyses. These findings highlight the multifactorial nature of LID and emphasize the importance of considering both motor and non-motor aspects in predictive modeling for LID onset.

Some predictors emerge as particularly noteworthy in the single cohort analyses, particularly in the *optimized comprehensive* LID classification model for the LuxPARK cohort, which achieved a higher hold-out AUC of 0.678 (see Table 4.2), and in the *optimized comprehensive* time-to-LID model for the PPMI cohort, which displayed a higher hold-out C-index of 0.663 (see Table 4.3) across single cohort analyses. Noteworthy motor symptoms, including motor fluctuation, bradykinesia, tremor, rigidity, and gait freezing, significantly influence the *optimized* LID classification model derived from the LuxPARK analysis, as illustrated in Figure A.9 In the *optimized* time-to-LID model derived from the PPMI analysis, levodopa treatment and axial symptoms emerge as the top predictors, as shown in Figure A.10. Moreover, predictors such as the age of PD onset, PD disease duration, gastrointestinal symptoms, and Benton JLO scores are significant in both models, indicating their consistent influence across LID classification and time-to-LID analyses.

A potential association was observed between cardiovascular autonomic dysfunction and the onset of LID. While no statistically significant difference was found based on the log-rank test (p-value = 0.773), PD patients without cardiovascular autonomic dysfunction showed a heightened risk of developing LID starting around year 5, as depicted in Figure A.11. The observed association may be attributed to the overlapping distribution in the Kaplan-Meier (KM) plot.

While a positive relationship between visuospatial ability, as measured by Benton JLO, and LID onset may be evident in the SHAP values plot shown in Figure A.10, further analysis within the PPMI cohort reveals a different pattern of results. This apparent positive relationship may be influenced by the overlapping distributions observed in the KM plot shown in Figure A.12, specifically within the PPMI cohort. This overlap suggests that the predictive power of visuospatial ability, measured by Benton JLO, regarding time-to-LID onset, may be more complex than initially perceived.

Table 4.12 presents the hazard ratios (HR) and their corresponding 95% confidence intervals (CIs) for predictors in the *comprehensive* time-to-LID model within the cross-cohort analysis, as illustrated in Figure 4.4. Among the significant predictors, levodopa treatment stands out with a significant hazard ratio of 1.38 (95% CI 1.14, 1.68), indicating a positive association with the time to LID onset. Notably, there is a significant difference in the time-to-LID distributions for patients with and without baseline levodopa treatment (p-values 0.013). Approximately 50% of levodopa-treated PD patients experienced LID symptoms within 3.37 years from the baseline, highlighting this treatment factor's clinical impact.

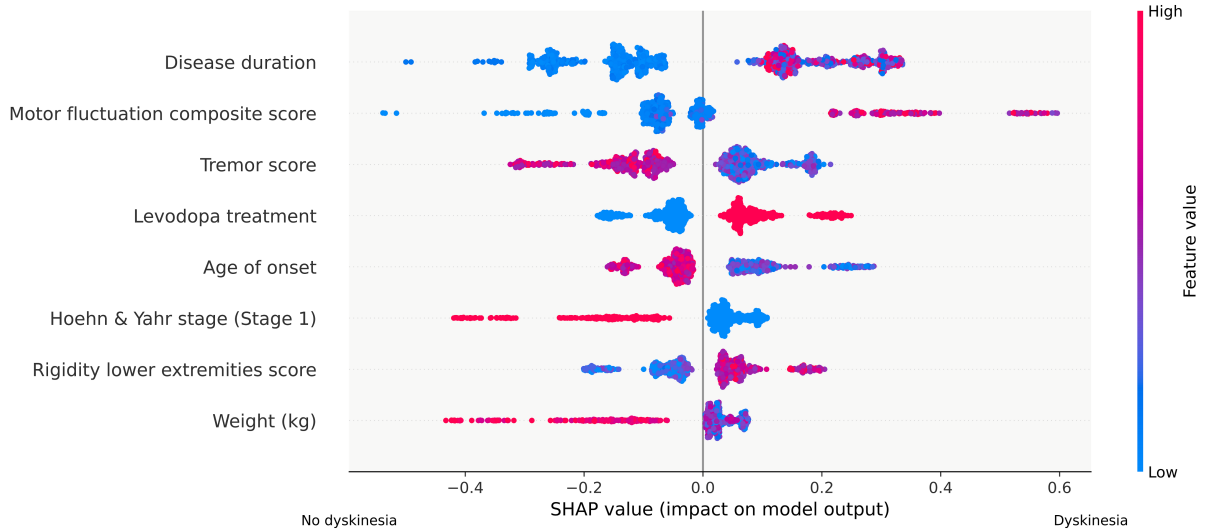
Conversely, the age of PD onset at or above 66 years shows a protective effect against LID development, as evidenced by a HR of 0.84 (95% CI 0.73, 0.97). Furthermore, rigidity at the lower extremities shows a significant HR of 1.27 (95% CI 1.08, 1.52), indicating its role as a predictor for accelerated LID onset. While predictors such as freezing of gait and Benton JLO show HRs close to 1, indicating relatively weaker associations with time-to-LID onset, their significance based on the log-rank test highlights significant differences in LID onset among patient subgroups. These insights contribute to a more nuanced understanding of the multifactorial nature of LID development in PD.

Table 4.9 Significance testing of hold-out predictive metrics between normalized and unnormalized models for LID in multi-cohort analyses.

	***Comprehensive* model**		***Refined* model**	
Cohort	**Normalized vs. Unnormalized**	**Cross-cohort normalization**	**Normalized vs. Unnormalized**	**Cross-cohort normalization**
LID classification:				
Cross-cohort	0.912	Mean-centering	0.683	Quantile
Leave-ICEBERG-out	0.549	M-ComBat	0.981	Mean
Leave-PPMI-out	0.96	Ratio-A	0.719	Ratio-A
Time-to-LID:				
Cross-cohort	0.096	Standardize	0.527	Mean
Leave-ICEBERG-out	0.298	Quantile	0.867	Quantile
Leave-PPMI-out	0.019	Ratio-A	0.955	Mean

A comparison of the statistical significance of the differences between the hold-out predictive performance metrics for the *optimized comprehensive* and *refined* models across cohorts. The *p*-values for the significance of the difference were calculated using DeLong's test for LID classification and the one-short nonparametric test for time-to-LID analysis. A *p*-value < 0.05 indicates a significant difference in hold-out predictive performance between the two models. The normalization method used on the *optimized* model is indicated in the column "Normalization".

Figure 4.3 SHAP values plot for the *optimized comprehensive* LID classification model in cross-cohort analysis.



SHAP value plot displaying the top predictors for the *optimized comprehensive* model in cross-cohort LID prognostic classification. The plot shows the magnitude and direction (positive or negative) of each feature's influence on LID prognosis status as output.

Table 4.10 Predictive performance metrics between normalized and unnormalized models for LID in multi-cohort analyses.

	Cross-cohort			Leave-ICEBERG-out			Leave-PPMI-out		
	Mean (SD)	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**
LID classification: *Comprehensive* model									
Normalized	0.682 (0.083)	0.661	5 (10)	0.699 (0.016)	0.513	15 (26)	0.685 (0.05)	0.595	34 (65)
Unnormalized	0.682 (0.08)	0.646	8 (14)	0.683 (0.038)	0.603	9 (17)	0.69 (0.04)	0.582	58 (74)
LID classification: *Refined* model									
Normalized	0.654 (0.030)	0.612	17 (27)	0.692 (0.021)	0.534	6 (15)	0.679 (0.033)	0.599	9 (20)
Unnormalized	0.688 (0.043)	0.639	5 (9)	0.662 (0.047)	0.548	8 (12)	0.669 (0.045)	0.557	58 (84)
	Mean (SD)	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**
Time-to-LID: *Comprehensive* model									
Normalized	0.718 (0.052)	0.627	36 (36)	0.701 (0.056)	0.531	57 (96)	0.719 (0.029)	0.655	13 (22)
Unnormalized	0.700 (0.042)	0.669	14 (24)	0.687 (0.047)	0.584	154 (156)	0.702 (0.022)	0.613	10 (16)
Time-to-LID: *Refined* model									
Normalized	0.715 (0.054)	0.685	106 (131)	0.695 (0.061)	0.531	50 (80)	0.687 (0.037)	0.634	95 (119)
Unnormalized	0.694 (0.059)	0.677	111 (134)	0.687 (0.044)	0.542	158 (160)	0.680 (0.026)	0.635	110 (115)

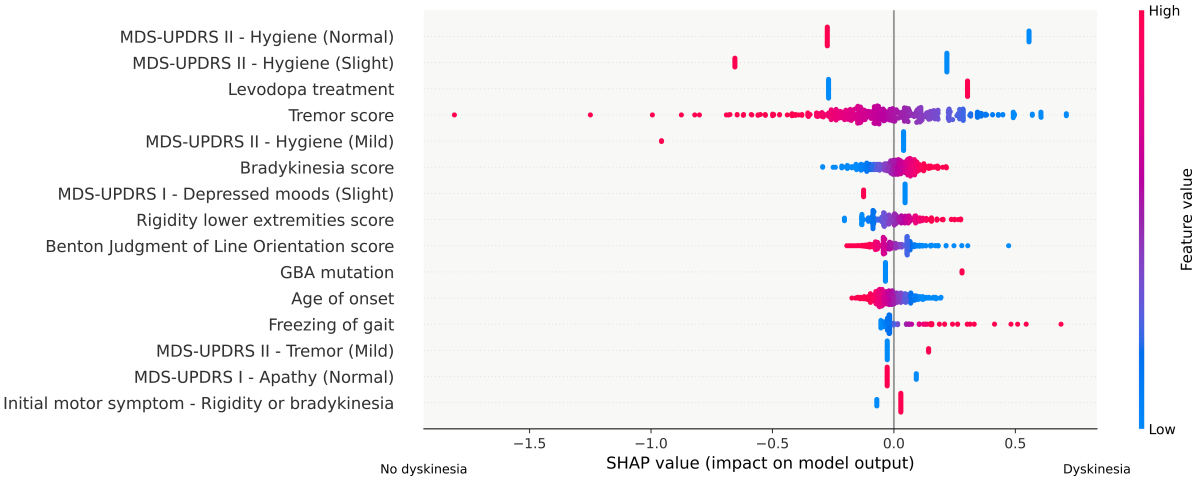
Assessment of the predictive performance of *comprehensive* and *refined* prognostic models for LID, including classification and time-to-LID analyses. The evaluation includes cross-validated and hold-out AUC or C-index calculations for both normalized and unnormalized models, as well as a detailed examination of the number of features used in each model. The column labeled 'Number of features' displays the number of features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 4.11 Top 10 predictors for LID prognosis in cross-cohort analysis.

Predictors	**Average ranks**
Disease duration	1.33
Motor fluctuation composite score	1.67
Levodopa treatment	2.50
Selective axial symptoms	4.00
Tremor	5.33
MDS-UPDRS Part III score (ON)	6.00
Age of onset	6.25
Rigidity lower extremities score	6.67
Weight (kg)	8.67
Benton Judgment of Line Orientation	9.00

A list of predictors for LID prognosis that received the top 10 average ranks in the cross-cohort analysis, using a ranking by permutation importance across the *optimized comprehensive* and *refined* models for LID classification and time-to-LID analysis. The final rank reflects the average from non-missing ranks across the *optimized* models in the cross-cohort analysis.

Figure 4.4 SHAP values plot for the *optimized comprehensive* time-to-LID model in cross-cohort analysis.



SHAP value plot displaying the top 15 predictors for the *optimized comprehensive* model in cross-cohort time-to-LID analysis. The plot shows the magnitude and direction (positive or negative) of each feature's influence on time-to-LID as output.

The HR for bradykinesia in the *comprehensive* time-to-LID model is 0.92 (95% CI 0.78, 1.07), indicating a relatively weak association with the time-to-LID onset. Notably, within this analysis, individuals with bradykinesia scores of less than 20 appear to have a slightly higher risk of LID onset than those with scores of 20 or greater. However, this difference is not statistically significant. This trend is reflected in the KM plot in Figure A.13, where the LID onset curves for these two subgroups show noticeable overlap, particularly at year 1 and year 5. Nevertheless, the HR approaching 1 indicates that the distinction between these bradykinesia subgroups in predicting LID onset may be insignificant.

The results of the correlation analysis, which detail the direction and strength of the relationships between predictors and the outcome of LID, are presented in Table 4.13. Table 4.14 provides further insight into the correlations between predictors, outlining positive or negative associations. Positive correlations suggest that variables tend to increase together, which may be associated with an increased risk of LID onset. Conversely, negative correlations indicate that variables increase in opposite directions, which may be associated with a reduced risk of LID. These findings contribute substantially to our understanding of the complex interplay between LID risk factors.

The analysis of levodopa equivalent daily dose (LEDD) revealed significant differences between PD patients with and without LID, as shown in Table 4.15. The average LEDD for patients without LID (LID-) was 580.7mg (SD 306.71), whereas for patients with LID (LID+), it was significantly higher at 790.8mg (SD 397.99). The p-value from the significance test comparing these groups indicates a statistically significant difference.

Further analysis examined the time-to-LID based on LEDD levels. For PD patients with LEDD $<$400mg, the average time-to-LID was 2.8 years (SD 2.02), while for those with LEDD $\geq$400mg, the average time-to-LID was shorter, at 1.9 years (SD 1.95), indicating a statistically significant difference in the time-to-LID based on LEDD levels. These results showed the impact of higher LEDD on the earlier occurrence of LID in PD patients.

4.4.6 Assessment of clinical utility and calibration

The analysis of the AUNBC for *comprehensive* LID classification, illustrated in the bar plot in Figure 4.5, indicates that the *optimized* model trained using CatBoost achieved a larger area under the net benefit curve than other ML methods. Nevertheless, the observed differences in the AUNBC among these methods were not statistically significant.

For the *refined* LID classification, the bar plot in Figure A.14 shows that the CatBoost-*optimized* model once again achieved the larger AUNBC among the compared methods. In this case, the CatBoost model showed a statistically significant difference in the area under the net benefit curve compared to the *optimized* models by C4.5 and GOSDT-GUESSES. Furthermore, the C4.5 and GOSDT-GUESSES models showed larger areas of net benefit at negative values and smaller areas than the "treat all" model (which assumes that all patients receive the intervention regardless of their predicted risk), further emphasizing the superior performance of the CatBoost-*optimized* model. Moreover, CatBoost shows a calibration slope closer to 1 and a lower MSE for predicted probabilities versus observed LID outcomes, as shown in Table 4.16 (left side).

Figure 4.6 presents a bar plot that illustrates the AUNBC of the *optimized* models for *comprehensive* time-to-LID models. The penalized Cox method-trained model showed the highest AUNBC, with a hold-out C-index of 0.67. The CW-GBoost model follows with a hold-out C-index of 0.67 and shows a statistically significant superiority in clinical utility compared to the *optimized* NLSVM model. Table

Table 4.12 Median conversion times and hazard ratios for *comprehensive* time-to-LID in cross-cohort analysis.

Predictors	**Hazard Ratio (95% CI)**	**Median Conversion (95% CI)**	**Log-rank (*p*-values)**
MDS-UPDRS II - Hygiene (Normal)			
Yes	0.89 (0.05, 6.11)	4.68 (4.19, 5.22)	0.692
No		4.60 (3.05, 5.52)	
MDS-UPDRS II - Hygiene (Slight)			
Yes	0.83 (0.04, 5.84)	4.76 (3.22, 5.77)	0.731
No		4.52 (4.05, 5.04)	
Levodopa treatment			
Yes	1.38 (1.14, 1.68)	3.37 (2.71, 4.99)	0.013
No		5.04 (4.44, 5.76)	
Tremor score			
≥ 3	1.25 (0.90, 1.77)	4.76 (4.19, 5.22)	0.057
< 3		2.34 (1.42, NR)	
MDS-UPDRS II - Hygiene (Mild)			
Yes	1.08 (0.06, 7.33)	2.23 (1.09, 4.61)	0.080
No		4.76 (4.05, 5.09)	
Bradykinesia score			
≥ 20	0.92 (0.78, 1.07)	3.37 (1.38, NR)	0.329
< 20		4.68 (4.19, 5.09)	
MDS-UPDRS I - Depressed moods (Slight)			
Yes	1.05 (0.86, 1.3)	4.61 (3.64, 5.76)	0.595
No		4.60 (3.77, 5.22)	
Rigidity lower extremities score			
≥ 1	1.27 (1.08, 1.52)	4.19 (3.43, 4.76)	0.001
< 1		NR (5.02, NR)	
Benton Judgment of Line Orientation score			
≥ 16	0.91 (0.80, 1.02)	NR (4.99, NR)	3.18E-04
< 16		4.19 (3.52, 4.76)	
GBA mutation			
Yes	1.09 (0.83, 1.45)	5.52 (1.84, 5.92)	0.801
No		4.60 (4.00, 5.02)	
Age of onset			
≥ 66	0.84 (0.73, 0.97)	5.60 (4.28, 5.84)	0.043
< 66		4.19 (3.37, 4.76)	
Freezing of gait			
≥ 1	1.16 (0.96, 1.46)	3.05 (2.00, 4.19)	0.014
< 1		4.86 (4.28, 5.52)	
MDS-UPDRS II - Tremor (Mild)			
Yes	1.29 (0.97, 1.68)	4.28 (3.05, 5.01)	0.179
No		4.76 (3.92, 5.60)	
MDS-UPDRS I - Apathy (Normal)			
Yes	0.95 (0.75, 1.19)	4.86 (4.19, 5.52)	0.255
No		3.85 (2.92, 4.76)	
Initial motor symptom - Rigidity or bradykinesia			
Yes	1.14 (0.92, 1.38)	4.28 (3.68, 4.86)	0.093
No		5.04 (4.05, 5.92)	

Summary of the hazard ratio (HR), median conversion time with 95% confidence interval (CI), and *p*-values from the log-rank test for the top 15 predictors identified in the time-to-LID model in the cross-cohort analysis. The HR provides insights into the risk associated with each predictor, while the median conversion time and log-rank test assess Kaplan-Meier (KM) curve differences between groups. "NR" (not reached) indicates that the LID event did not occur for some participants during the study period.

Table 4.13 Correlation between predictors and LID outcomes in cross-cohort analysis.

Predictors	**Correlation**	***p*-values**
Age of onset	-0.22	4.33E-12
Disease duration	0.34	2.45E-27
Levodopa treatment	0.41	2.57E-38
Weight (kg)	-0.09	4.73E-03
Height (cm)	-0.13	4.84E-05
BMI (kg/m^2)	-0.02	5.49E-01
Benton Judgment of Line Orientation	-0.21	6.87E-10
Hoehn & Yahr stage	0.26	1.15E-16
Axial symptoms	0.34	9.79E-27
Selective axial symptoms	0.26	2.14E-16
Motor fluctuation composite score	0.38	1.97E-23
Freezing of gait	0.28	5.84E-19
Rest tremor	-0.08	6.75E-02
Tremor	-0.10	1.85E-02
Bradykinesia	0.14	9.21E-04
Rigidity lower extremities	0.13	1.35E-03
Rigidity upper extremities	0.03	5.28E-01
Initial motor symptom - Rigidity or bradykinesia	0.02	4.89E-01
Modified Schwab & England ADL	-0.26	1.51E-10
MDS-UPDRS Part I score	0.30	3.05E-21
MDS-UPDRS Part II score	0.36	2.51E-30
MDS-UPDRS Part III score (ON)	0.11	8.52E-03
MDS-UPDRS I - Apathy	0.18	1.46E-08
MDS-UPDRS I - Depressed moods	0.17	1.67E-07
MDS-UPDRS I - Sleep problems (night)	0.23	1.18E-12
MDS-UPDRS I - Urinary problems	0.11	4.49E-04
MDS-UPDRS II - Freezing	0.32	3.23E-23
MDS-UPDRS II - Hygiene	0.20	1.25E-09
MDS-UPDRS II - Saliva and drooling	0.15	3.25E-06
MDS-UPDRS II - Tremor	0.02	4.50E-01
SCOPA-AUT Gastrointestinal (GI)	0.28	5.60E-19
SCOPA-AUT Cardiovascular	0.27	3.85E-17
SCOPA-AUT Urinary	0.17	3.62E-07
GBA mutation	0.06	1.10E-01

The correlation of predictors with LID outcome was measured using the point biserial correlation for continuous or ordinal predictors and the Matthews Correlation Coefficient (MCC) for the binary predictor.

Table 4.14 Correlation analysis results for LID predictors in cross-cohort analysis.

Predictors	Age of onset	Axial symptoms	Benton JLO	Brady-kinesia	Disease duration	Motor fluctua-tion	Weight	Hoehn & Yahr stage	Rest tremor
Axial symptoms	-0.03 (3.43E-01)								
Benton JLO	1.87E-03 (9.57E-01)	-0.21 (1.55E-09)							
Bradykinesia	-0.01 (7.84E-01)	0.55 (3.32E-46)	-0.20 (1.65E-05)						
Disease duration	-0.28 (5.45E-19)	0.51 (5.14E-65)	-0.22 (3.83E-10)	0.31 (1.41E-14)					
Motor fluctuation composite score	-0.24 (8.65E-10)	0.31 (2.87E-16)	-0.02 (6.33E-01)	0.17 (6.76E-05)	0.41 (3.15E-27)				
Weight	-0.07 (2.20E-02)	0.03 (3.95E-01)	0.11 (1.17E-03)	0.06 (1.88E-01)	-0.04 (1.84E-01)	0.03 (4.82E-01)			
Hoehn & Yahr stage	0.06 (5.90E-02)	0.47 (2.12E-53)	-0.13 (2.96E-04)	0.45 (4.06E-30)	0.33 (3.17E-25)	0.19 (1.40E-06)	-0.05 (1.03E-01)		
Rest tremor	0.02 (7.11E-01)	0.01 (8.54E-01)	-0.03 (5.33E-01)	0.18 (1.22E-05)	8.38E-04 (9.84E-01)	-0.03 (4.10E-01)	0.06 (1.56E-01)	0.04 (2.96E-01)	
Rigidity lower ex-tremities	-0.05 (2.31E-01)	0.28 (7.09E-12)	-0.10 (3.81E-02)	0.43 (1.30E-27)	0.15 (1.99E-04)	0.16 (1.96E-04)	0.10 (1.55E-02)	0.22 (9.39E-08)	0.20 (1.99E-06)

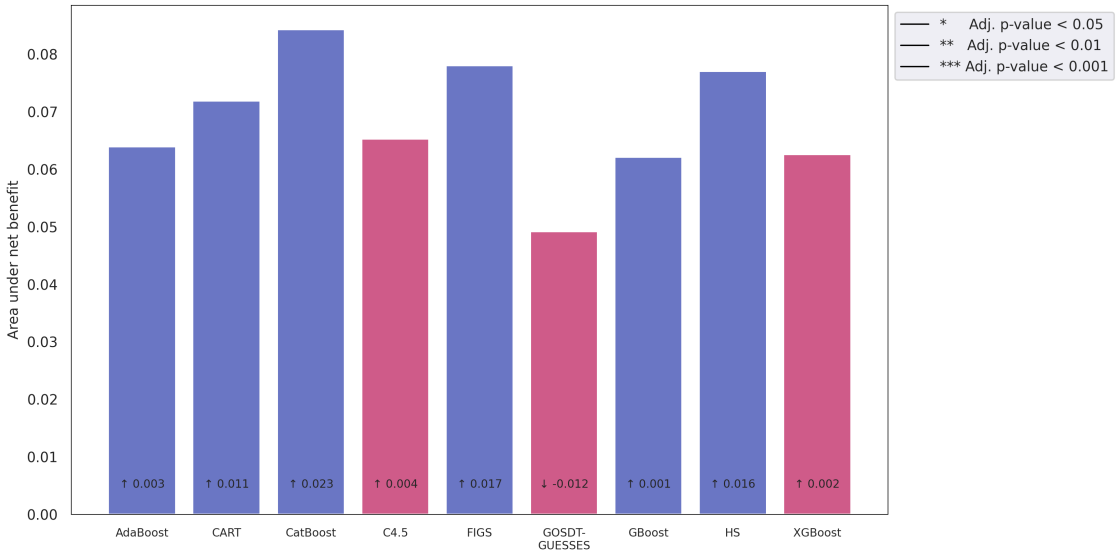
Correlation analysis of predictors was conducted using Spearman correlation for two continuous/ordinal variables, point biserial correlation for continuous/ordinal and binary variables, and Matthews correlation coefficient (MCC) for two binary variables. The correlation coefficients are presented with the *p*-values in brackets.

Table 4.15 Summary statistics for Levodopa Equivalent Daily Dose (LEDD) among PD patients with and without LID.

Statistics	LID-	LID+	Significance test (*p*-values)	LEDD $<400mg$	LEDD $\geq 400mg$	Log-rank test (*p*-values)
n	93	192	5.25E-06	56	229	3.14E-03
Mean (SD)	580.7 (306.71)	790.8 (397.99)		2.8 (2.02)	1.9 (1.95)	

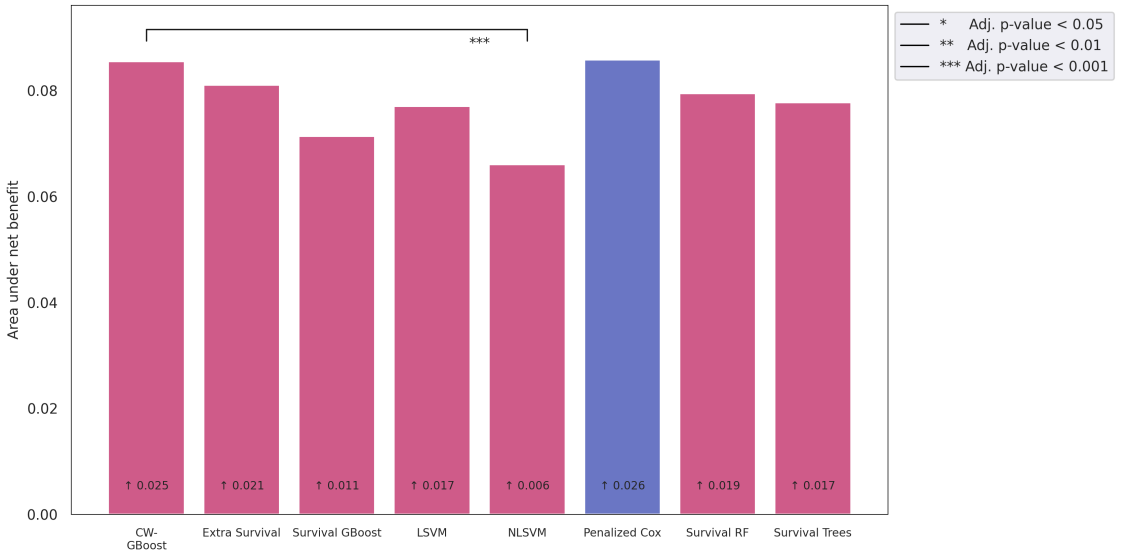
The statistical significance of these differences is indicated by *p*-values from t-tests (for normally distributed data) or Mann-Whitney U-tests (for non-normally distributed data). Additionally, the table presents time-to-LID statistics for PD patients with LEDD $<400mg$ and LEDD $\geq 400mg$, along with *p*-values from log-rank tests comparing these two groups.

Figure 4.5 Bar plot of the area under the positive net benefit curve for the *optimized comprehensive* LID classification models in cross-cohort analysis.



The bar plot shows the area under the positive net benefit for different cross-cohort *comprehensive* LID classification models. The lines indicate significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

Figure 4.6 Bar plot of the area under the positive net benefit curve for the *optimized comprehensive* time-to-LID models in cross-cohort analysis.



The bar plot shows the area under the positive net benefit for different cross-cohort *comprehensive* time-to-LID models. The lines indicate significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

Table 4.16 Calibration analysis for LID classification and time-to-LID analyses.

LID classification					**Time-to-LID**				
Algorithm	***Comprehensive***		***Refined***		**Algorithm**	***Comprehensive***		***Refined***	
	Slope	**MSE**	**Slope**	**MSE**		**Slope**	**MSE**	**Slope**	**MSE**
AdaBoost	4.85	0.23	2.01	0.22	CW-GBoost	0.48	0.03	0.39	0.03
CART	0.52	0.23	0.58	0.21	Extra Survival	0.44	0.03	0.48	0.03
CatBoost	0.70	0.20	0.67	0.20	Survival GBoost	0.26	0.05	0.29	0.05
C4.5	0.29	0.33	0.20	0.34	LSVM	0.38	0.08	0.56	0.07
FIGS	0.63	0.20	0.59	0.21	NLSVM	0.59	0.02	0.72	0.02
GOSDT-GUESSES	0.21	0.37	0.10	0.41	Penalized Cox	0.27	0.11	0.36	0.04
GBoost	6.65	0.24	1.22	0.21	Survival RF	0.43	0.03	0.51	0.02
HS	0.63	0.22	0.59	0.21	Survival Trees	0.22	0.06	0.14	0.15
XGBoost	0.37	0.26	0.50	0.22					

Calibration analysis for *comprehensive* and *refined* models in both LID classification and time-to-LID analysis for cross-cohort analysis.

4.16 (right side) shows that although the NLSVM model has a calibration slope closer to 1, it shows significantly lower clinical utility than the CW-GBoost model. However, The CW-GBoost model has the second-closest calibration slope of 1 among the compared models.

The bar plot in Figure A.15 displays the AUNBC of the *optimized* models for *refined* time-to-LID models. The Extra Survival Tree model achieved the highest AUNBC, with a hold-out C-index of 0.68. Subsequently, the Survival RF model achieved a hold-out C-index of 0.685. The NLSVM model shows the highest calibration in the *refined* time-to-LID model and a lower MSE. No statistically significant difference exists in the area under the net benefits among these models.

4.5 Discussion

Predicting the onset of LID in PD patients is challenging due to inter-individual heterogeneity and the many factors influencing disease progression. These challenges are further complicated in cross-cohort studies, where differences in study populations and data collection methods introduce additional variability. Recent advancements in modeling PD progression, such as the statistical progression model in Severson et al. (2021) [155] using contrastive latent variable and personalized input-output hidden Markov models, have highlighted the importance of capturing intra- and inter-individual variability in disease trajectories. While the efficacy of their approach in identifying disease states and assessing their clinical relevance across motor and cognitive outcomes is well-documented, it is less applicable to the cross-cohort prediction of specific complications like LID.

Our study has made significant advances in the field of LID prognosis by addressing the limitations of previous research. Previous studies primarily focused on single-cohort analyses [132, 156, 157, 158], which limit the generalizability of predictive models to specific patient populations. In contrast, our study incorporated cross-cohort analysis to evaluate multivariable ML models under single- and multi-cohort settings. This approach enabled a more comprehensive examination of LID prediction. This comprehensive approach included LID classification within four years and time-to-LID analysis, providing a multifaceted understanding of this PD complication.

The cross-cohort models showed competitive predictive performance, achieving greater robustness and generalizability compared to single-cohort studies. These findings highlight the importance of considering study heterogeneity in developing prediction models that are broadly applicable across diverse patient populations and clinical settings. By mitigating cohort-specific biases and improving predictive reliability, our study emphasizes the value of cross-cohort analysis as a foundation of ML applications in LID prognosis.

Furthermore, our study compared various approaches for model building, feature selection, and cross-study normalization to inform the design of future predictive approaches. Utilizing a two-level CV framework, we ensured that our findings reflect predictive accuracy and methodological robustness. This methodological approach provides practical guidance on selecting the most effective predictive framework for the study, balancing computational integrity with real-world applicability.

4.5.1 Comparative evaluation of predictive models

A comparison of the results across cohorts revealed significant differences in predictive performance. LuxPARK demonstrated consistent superiority in cross-validated performance for LID classification and

time-to-LID analysis. This superiority is probably due to its larger sample size and more heterogeneous patient population, which resulted in a more comprehensive representation of disease variability and improved the model's capability to generalize across diverse subpopulations. In contrast, the smaller sample size and shorter average disease durations in ICEBERG limited the model's capability to capture long-term predictive patterns, resulting in lower performance in these datasets. This observation highlights the role of cohort size and population diversity in model development and prediction accuracy.

To identify additional predictive features associated with LID development, we also considered excluding baseline dyskinesia and levodopa medication use, two well-established predictors of LID. The *refined* models showed hold-out performance that differed from the *comprehensive* models yet achieved significant predictive performance. This finding suggests that the *refined* model may reveal subtle, previously under-explored relationships among predictors, offering more profound insights into disease mechanisms.

Furthermore, cross-cohort integrative analyses were compared with single-cohort analyses, which revealed significant advantages. Cross-cohort models demonstrated superior predictive capabilities for both LID prognosis and time-to-LID analysis, accompanied by greater stability of prediction outcomes across CV cycles. This consistency highlights the ability of cross-cohort approaches to mitigate biases inherent in single-cohort models. Such biases may include overfitting to specific cohort characteristics or failing to account for inter-cohort variability. By pooling data from diverse populations, cross-cohort models increase statistical power and enable the detection of patterns that might be masked in smaller, more homogenous datasets. This integrative approach is particularly valuable for addressing data variability, demographic variation, and differences in clinical protocols across cohorts.

In general, cross-cohort integrative models have proven to be a reliable methodology for predicting LID in PD, offering increased applicability across a range of patient demographics. These findings emphasize the necessity of using multi-cohort approach to improve predictive models' reliability and generalizability. Future studies could expand these approaches by incorporating additional data types, such as genetic or imaging biomarkers [159], to improve the predictive capability of cross-cohort models further and gain deeper mechanistic insights into LID development.

4.5.2 Interpretation of models and predictors

The analyses identified several clinically relevant predictors, enhancing our understanding of PD's LID development. Notably, the administration of levodopa emerged as a key predictor due to its association with LID, a common side effect of levodopa treatment [19]. The hazard ratio of 1.38 for LID onset about levodopa suggests that PD patients receiving levodopa have a 38% higher risk of developing LID compared to those not receiving levodopa at baseline visits. This finding emphasizes the intricate relationship between levodopa treatment and dyskinesia onset, underscoring the need for vigilant monitoring and management in clinical practice.

Previous studies have identified LEDD as a significant predictor of LID [35, 131, 137, 141]. However, the lack of complete and consistent LEDD data for two cohorts limited our ability to include levodopa dosage in our predictive model. To address this, we performed a statistical analysis within the LuxPARK cohort to examine the relationship between LEDD and LID. The analysis showed that patients with LID showed significantly higher levels of LEDD than those without LID within four years. This highlights the association between increased dopaminergic therapy and the development of dyskinesias. Furthermore, the time-to-LID analysis indicated that higher LEDD levels, particularly those above $400\,mg$, increase

the risk of LID and accelerate its onset. These findings emphasize the necessity of carefully managing levodopa dosage to achieve an optimal balance between therapeutic efficacy and the risk of dyskinesia while treating PD symptoms.

Furthermore, our investigation highlighted the significant impact of PD duration on long-term LID outcomes. This aligns with existing hypotheses suggesting that prolonged levodopa therapy [22], alongside the inherent severity of PD itself [35], contributes significantly to exacerbating LID symptoms over time. As PD progresses to advanced stages, there is a rising tendency for the prevalence of LID, indicative of a complex association between LID and disease progression [160]. This association extends beyond mere levodopa dosages, as individuals in later PD stages often experience increased distress, as measured across both motor and non-motor aspects of the disorder.

The MDS-UPDRS parts II and III offer valuable insights into disease progression and symptomatology in PD [143]. Our analysis revealed a significant positive correlation between higher MDS-UPDRS part II scores, which assess activities of daily living (ADL) and self-care tasks, and the risk of developing LID. This connection underscores the impact of functional impairment on LID risk [134], indicating that individuals facing more significant challenges in ADL may be more susceptible to LID.

Similarly, increased MDS-UPDRS part III scores, which evaluate motor aspects, showed a significant positive correlation with LID onset [132]. The presence of more severe motor symptoms corresponds to an increased risk of dyskinesia in PD patients [161]. These findings highlight the relationship between motor dysfunction severity and LID emergence, aligning with previous research on individuals in more advanced disease stages [141].

The study consistently linked increased baseline disease severity [134] and motor dysfunction [138] as predictors of LID in PD. Our analysis further revealed a significant correlation between PD severity and motor symptoms, including axial dysfunction, bradykinesia, motor fluctuation, and rigidity. Each of these symptoms displayed a significant positive association with disease severity, indicating that as PD progresses, the likelihood and severity of these motor impairments, including the risk of developing LID, also increase.

The SHAP value analysis revealed the significance of motor fluctuations, particularly those associated with prolonged levodopa therapy. This finding is consistent with previous studies [28, 29, 35]. These fluctuations typically manifest in patients with PD approximately three to five years after initiating levodopa treatment [22], adding complexity to the development of LID. Characterized by unpredictable changes in motor function throughout the day, these fluctuations signal disease progression and necessitate adjustments in medication management strategies involving levodopa [26]. However, delayed or lower doses of these medications can inadvertently worsen LID symptoms [143], emphasizing the necessity for careful balance in PD pharmacotherapy.

Bradykinesia, a defining symptom of PD characterized by slowness of movement and impaired motor function [22, 106], has emerged as a significant predictor of LID. Several studies have highlighted the significance of bradykinesia in LID, particularly in patients with advanced PD stages [28, 29, 35]. The progressive nature of bradykinesia throughout PD progression reflects underlying neurodegenerative processes and dopaminergic system alterations [134], indicating disease severity. These insights highlight the importance of considering bradykinesia severity and temporal dynamics in predicting LID risk and optimizing PD patient care strategies. Furthermore, decreased handwriting quality, often associated with bradykinesia and identified as a predictor in the single cohort analysis on PPMI, has been linked to increased LID risk.

Tremor and resting tremor are distinct motor phenomena in PD, each with implications for LID development. Our analysis revealed unique associations between tremor types and LID onset. Tremor, which may occur at rest, showed a significant negative correlation with LID onset [132]. This finding can be attributed to patients with tremor-dominant PD who typically require lower levodopa doses [143] and show slower disease progression [162, 163]. Conversely, resting tremor observed during rest periods [136] showed a negative association with LID onset but a significant positive correlation with bradykinesia and rigidity in the lower extremities [19]. This dual correlation highlights the interplay among PD motor symptoms and their roles in LID development. This suggests that patients with predominant tremor symptoms may have a lower risk of LID development than those with other initial motor symptoms, which aligns with previous research [58, 131, 164].

Another significant motor symptom related to LID is rigidity, characterized by muscle stiffness and resistance to movement [165]. Our study showed a positive correlation between rigidity severity and LID risk, particularly as PD progresses. Furthermore, we observed an increased HR of 1.27 for rigidity at lower extremities scores, indicating a 27% higher risk of developing LID associated with rigidity severity. This finding highlights the importance of monitoring and managing rigidity in PD to mitigate the risk of LID and optimize treatment strategies.

Freezing of gait (FoG) presents a significant challenge in managing PD, manifesting as a sudden and temporary inability to walk despite the intention to walk [166]. It has emerged as a potential predictor for the onset of LID, with studies indicating a positive association between FoG severity and LID risk, particularly in advanced stages of PD [167]. The impact of FoG on LID highlights its role as a marker of disease progression and motor complications in PD. Understanding this relationship offers insights into the complex motor impairments experienced by PD patients and aids in developing targeted interventions to manage both symptoms effectively.

In addition to FoG, axial impairments, including postural stability difficulties, gait impairment, and heightened lower extremity rigidity, were identified as predictors for LID development [5]. Our analysis revealed a positive correlation between axial impairment and disease duration, motor fluctuation, disease severity, and rigidity. This indicates that these factors influence the onset and progression of LID. As the disease advances, these symptoms worsen, significantly impacting LID severity. This highlights the interplay between disease progression, motor impairment, and treatment outcomes in PD patients. Addressing axial impairments comprehensively in PD management may help mitigate LID risk and improve overall patient well-being.

The MDS-UPDRS part I scores, which assess non-motor symptoms in PD patients, have revealed specific aspects, such as depressed mood, apathy, and cognition, as potential predictors of LID onset. Higher scores indicating a heavier burden of non-motor symptoms may suggest an increased risk of developing LID [132]. Given their lessened responsiveness to levodopa [143], the MDS-UPDRS part I assessment may offer a more consistent measure of disease severity in treated patients. Individuals with PD who experience more severe non-motor symptoms may also be more likely to be in advanced disease stages [168].

Furthermore, difficulties with hygiene are also associated with LID [169]. Challenges in maintaining personal hygiene may indicate broader issues with motor coordination or cognitive function, both of which can influence the risk of dyskinesia. Furthermore, experiencing light-headedness upon standing, a symptom highlighted in LuxPARK analysis, adds complexity to LID prediction and may be related to dizziness due to levodopa's adverse effects [31], potentially impacting LID onset in PD patients.

Similarly, non-motor symptoms related to pain and sensory changes and participation in hobbies and activities, as observed in the PPMI analysis, offer valuable predictive insights. Alterations in pain perception and activity levels may reflect disease progression stages [170], affecting treatment response and susceptibility to LID. These findings highlight the importance of considering non-motor and motor symptoms in predicting and managing LID risk in PD patients.

The progression of autonomic dysfunction, which is often linked to the severity of PD, can indirectly increase the susceptibility to LID. When evaluating non-motor symptoms as potential LID predictors, early PD features such as gastrointestinal issues, constipation, urinary problems, and sexual dysfunction are associated with both PD severity and increased LID risk [1]. Gastrointestinal symptoms are common in patients with PD, with dietary habits playing a significant role in their development [34]. These symptoms can create barriers to levodopa absorption, such as delayed gastric emptying, which can affect the effectiveness of levodopa therapy [30]. Prolonged gastrointestinal tract impairment further complicates levodopa delivery to its absorption site, potentially contributing to treatment challenges and LID.

Moreover, the severity of urinary dysfunction has been linked to the degeneration of the dopamine-dependent caudate nucleus [168], which may influence motor dysfunction [57]. Saliva and drooling, which PD patients often experience, can lead to difficulties in eating and speaking [32, 33]. While these associations may indirectly link to PD severity [25, 144], specific non-motor symptoms like gastrointestinal issues directly impact levodopa pharmacokinetics and treatment efficacy [145].

Another common aspect of autonomic dysfunction in PD identified in LuxPARK analysis is thermoregulatory autonomic dysfunction. Fluctuations in body temperature regulation, such as excessive sweating or impaired temperature control, may indicate autonomic system dysregulation [25], further exacerbating PD symptoms and affecting LID onset. Additionally, positive correlations have been noted between cardiovascular autonomic dysfunction and LID onset [134]. These findings highlight the multifaceted role of autonomic dysfunction in PD progression and its potential impact on LID susceptibility, emphasizing the need for comprehensive assessments in PD management.

A negative correlation between advanced age at PD onset and the likelihood of experiencing LID symptoms has been confirmed in our analyses, consistent with prior studies [132, 135]. Our analysis revealed a HR of 0.84, indicating a reduced risk of developing LID in patients aged 66 years or older compared to those younger than 66. Early-onset PD patients often show distinct disease mechanisms associated with genetic factors. Consequently, determining the appropriate levodopa dosage is important, with early-onset PD individuals typically requiring a lower initial dose than those with late-onset PD [15]. Customizing treatment protocols based on age at PD onset is therefore recommended.

Our study has clarified the association between *GBA* mutations and LID risk in PD patients. PD patients with *GBA* mutations have been found to show distinct clinical features and disease progression trajectories [5]. Emerging evidence suggests a potential link between *GBA* mutations and an increased risk of developing LID [47], particularly in patients with early-onset PD [48, 143]. Genetic factors, including *GBA* mutations, have been shown to influence PD symptoms and are associated with earlier LID development [137].

Moreover, mutations in the *GBA* gene have been associated with cognitive impairment in PD patients [31, 47, 171], highlighting the significant impact of cognitive factors on PD outcomes. Notably, younger age at PD onset is often associated with a higher degree of cognitive performance [172], including deficits in visuospatial sustained attention [15, 62], highlighting the complex interplay between cognitive

and motor aspects of PD and their role in LID onset.

There is a correlation between cognitive impairment in PD and the onset of LID, which is often attributed to deficiencies in visuospatial function, measured by the Benton JLO test. The negative cognitive impact may be attributable to levodopa therapy [42]. A previous study has indicated that more severe cognitive impairment in PD is associated with a poor response to levodopa treatment [130]. Interestingly, individuals with superior visuospatial abilities are less likely to experience LID-related symptoms, indicating a protective predictor against LID development [24]. This observation is consistent with research indicating a negative correlation between visuospatial abilities and the severity of axial impairment, overall disease severity, and rigidity in the lower extremities [63].

It is of great importance to gain an understanding of the psychological disturbances that can occur in patients with PD, such as emotional dysfunction, fatigue, and sleep disturbances. A previous study has shown that approximately half of PD patients show signs of depression, which can impact cognitive functions [45]. Apathy, which is often linked with clear signs of depression [15], can also correlate to the development of LID. The presence of emotional dysfunction, apathy, and fatigue may serve as potential indicators of LID development, reflecting the multifaceted nature of PD progression. Additionally, excessive daytime sleepiness has been associated with impaired motor function in PD, as evidenced by previous studies [45]. Further research must clarify these complex associations and their implications for managing PD.

Our study has revealed a significant correlation between body weight and the risk of developing LID in patients with PD. Previous studies have consistently shown that lower body weight is associated with a reduced risk of LID onset [133, 134]. This inverse correlation suggests that individuals with lower body weight may be more susceptible to experiencing LID symptoms during levodopa treatment. One potential explanation for this relationship is the impact of body weight on levodopa absorption, given that the condition of the gastrointestinal tract can significantly affect medication absorption rates. Factors such as dietary habits and nutritional status indirectly influence body weight in patients with PD, subsequently affecting their response to levodopa therapy [34]. It is therefore important to recognize the impact of body weight on LID risk to optimize treatment strategies and customize medication doses to individual patient characteristics, including considerations related to body weight and nutritional status.

The findings of this study highlight the significance of integrating clinical data from multiple cohorts to provide valuable insights into LID prognosis across diverse patient populations. The analyses reaffirm established predictors of LID and revealed novel associations, highlighting the utility of interpretable ML approaches in this context. The model robustness observed in the cross-study analyses underlines the efficacy of integrating data from multiple studies, establishing the foundation for more comprehensive and reliable predictive models for LID and PD progression.

4.5.3 Clinical utility and calibration

The findings of this study highlight the significant advantage of integrating clinical data from multiple cohorts in enhancing LID prognosis across diverse patient populations. By integrating diverse data sources, the analyses validated well-established predictors of LID. They revealed novel insights, highlighting the efficacy of interpretable ML approaches in identifying subtle and complex patterns in clinical data. Notably, the improved model robustness observed in the cross-cohort analyses highlights the effectiveness of integrating data from multiple studies to develop predictive tools that are both

reliable and generalizable.

Evaluating the clinical utility and calibration of prediction models revealed that different models offer varying degrees of applicability for clinical decision-making. Among the methods analyzed, CatBoost and Extra Survival Trees were the most promising, showing superior net benefit and calibration performance. These models provide a robust basis for practical implementation in real-world clinical settings, where predictive accuracy, reliability, and ease of integration are all-important. The favorable results of this study suggest that ML-driven tools can offer support in the personalized treatment planning process, such as the early identification of patients at high risk for LID and the ability to make personalized adjustments to therapeutic strategies.

Nevertheless, further validation of these models is necessary. To ensure the generalizability and utility of these tools, independent datasets incorporating a more heterogeneous cohort of participants across various geographic regions and clinical practices must be used. Validation with larger datasets will strengthen confidence in the findings and address limitations associated with small sample sizes, reducing variance in performance estimates and enhancing the robustness of model evaluations. This step is particularly important in establishing the reliability of ML applications for more extensive clinical use.

While this study provided the interpretability of the predictive models by clarifying the relationships between input features and outcomes, future efforts must extend beyond interpretability to address reliability and trustworthiness. In addition, regarding technical and privacy-related challenges, ML implementation in healthcare raises broader ethical questions, such as the risk of biases in model predictions and unequal access to technology. Addressing these issues through strategies such as bias mitigation during model development, comprehensive transparency in deployment processes, and equitable access to predictive tools is recommended. Previous literature has extensively discussed these concerns, emphasizing that AI-based healthcare tools must align with ethical principles and equity to build trust and ensure positive patient outcomes [173, 174]. Ultimately, to achieve successful clinical integration, it is important to address the dual challenges of robust model validation and adherence to ethical standards. By implementing these strategies, such as validation across diverse cohorts and the maintenance of ethical and legal considerations, these tools can ensure that their potential to improve patient care is met with the highest data integrity, fairness, and clinical reliability standards.

4.6 Summary and conclusions

This study introduces three key contributions to the investigation of LID in PD. Firstly, it develops cross-cohort prediction models that enable both prognostic classification and time-to-LID prediction from clinical data. These models provide valuable tools for early intervention and personalized management of PD. Secondly, the models differ from conventional single-variable approaches in that they incorporate multivariable signatures that use complementary clinical descriptors, improving prediction robustness and accuracy. Thirdly, the study uses feature selection and SHAP analysis to improve the interpretability of the models and facilitate more informed clinical decision-making.

The cross-cohort ML models demonstrated considerable predictive capabilities and robustness, consistently showing superior performance in prediction stability and generalizability compared to single-cohort models. The evaluation of these models highlighted the value of integrating nested CV, hyperparameter optimization, and feature selection techniques with cross-study normalization. These

methodologies contributed to the model's ability to address cohort-specific biases while ensuring broader applicability across diverse patient populations.

The interpretability models identified key clinical predictors associated with LID risk, including levodopa medication intake and PD progression markers. SHAP analysis provided insights into the contributions of individual features, thereby confirming the clinical relevance and applicability of the models.

This study highlights the potential of ML in facilitating cross-cohort LID prognosis. The approach paves the way for precision medicine in PD by integrating clinical data from different cohorts, enabling personalized risk prediction and management strategies. These findings can potential to contribute to developing personalized pharmacological interventions, optimized drug dosing regimens, and non-pharmacological strategies, such as targeted physical activity programs. Future research should prioritize validating these models in more extensive and diverse datasets to improve their reliability, applicability, and impact on clinical decision-making.

4.7 Contribution statement

The published manuscript and the supplementary material can be found in *Parkinsonism & Related Disorders*, offering supplementary details to complement the findings presented in this chapter. This study was a collaborative effort, with contributions from all authors in various aspects of the research and manuscript preparation (published in *Parkinsonism & Related Disorders*).

Rebecca Loo Ting Jiin: Conducted the study as the first author, developed the methodology, conducted analysis, created visualizations, validated results, and drafted the original manuscript.

Enrico Glaab: He supervised the study as the corresponding author for the submitted manuscript, contributed to the study's correction, methodology, and investigation, guided the project, acquired funding, and reviewed and edited all sections of the manuscript.

Olena Tsurkalenko: Reviewed and edited the manuscript, providing clinical insights. She also provided the study's variable aggregation list, which involved combining relevant clinical variables into aggregated features to better capture clinical patterns and relationships in the data.

Jochen Klucken, Graziella Mangone, Fouad Khoury, Marie Vidailhet, Jean-Christophe Corvol, and Rejko Krüger: Contributed by reviewing and editing the manuscript, providing insights from clinical perspectives.

Chapter 5

Interpretable machine learning for cross-cohort prediction of motor fluctuations in Parkinson's disease

Motor fluctuations (MF) are a common and significant complication of Parkinson's disease (PD). They are defined as alternating periods of worsening and improving motor function [31]. These may include the reappearance of PD symptoms such as bradykinesia, rigidity, and tremor. MF significantly impacts the quality of life of patients with PD, leading to periods of impaired motor function that interfere with daily activities, increase healthcare costs, and contribute to emotional and psychological distress [23, 175, 24]. One contributing factor to the increased healthcare costs is the higher frequency of freezing of gait (FoG) during OFF periods associated with MF, which can lead to an increased rate of falls. These falls heighten the risk of injury and contribute to higher hospitalization rates and associated costs [175].

The frequency and severity of these fluctuations can vary, significantly impacting the quality of life and complicating disease management [3, 35]. The timing of MF varies considerably among patients [30], with approximately half of PD patients developing MF within five years [139]. In the longer term, more than 90% of PD patients experience motor complications, including MF, after ten years [1]. Current management strategies of MF include deploying complementary medications, implementing dietary modifications [1], and invasive therapies, such as deep brain stimulation (DBS) or pump systems [27, 176].

A recent study used penalized regression in the Cox proportional hazards model to build clinical-pharmacogenetic models to identify clinical and genetic factors that predict the onset of MF. The study achieved an AUC of 0.68 for the clinical model and 0.70 for the clinical-pharmacogenetic model in predicting MF [139]. Significant risk factors for the development of MF include younger age at PD diagnosis [1, 139], and symptoms such as rigidity and bradykinesia [19]. Patients who develop PD at a younger age are more likely to experience MF as their disease progresses. The progression of PD itself is another key factor influencing MF [134].

Genetic factors also play a significant role in the development and severity of MF. Individuals with genetic mutations, such as those in the *GBA1* gene, tend to experience more severe MF than non-carriers [5, 137]. MF and dyskinesias are more prevalent in *GBA*-PD patients than in non-carriers [177], highlighting the influence of genetic predispositions on the progression and treatment response

of PD [47].

Gastrointestinal tract disorders, such as impaired transport and absorption of levodopa, significantly impact the drug's pharmacokinetics and contribute to MF [1, 178, 179]. Dietary factors further influence these fluctuations [37]; a low-protein diet is associated with better motor performance in PD patients experiencing MF [30, 180]. Furthermore, eliminating daytime dietary protein through protein redistribution diets has improved and prolonged motor function in patients with fluctuating levodopa responses [34].

Moreover, complications such as rigidity and bradykinesia are closely correlated with MF [19]. The subtype of PD and specific symptoms can also influence the risk of MF, such as tremor-predominant PD [139], which is associated with a decreased risk of MF. These symptoms reflect the broader impact of disease progression on motor control and highlight the complex interplay of various factors contributing to MF.

5.1 Rationale for the study

A comprehensive understanding of MF is important to optimize the management of PD. Early detection and intervention are important for maintaining stable motor function and reducing the overall burden of PD. Furthermore, the variability in response to PD medications among patients underlines the need for personalized treatment approaches. By identifying the predictors associated with MF, researchers can distinguish patterns and individual patient characteristics that influence the onset and severity of these complications. This study will enable the development of effective management strategies, facilitate early intervention to reduce the severity of these complications, and ultimately improve patient outcomes and overall quality of life.

Previous research has primarily focused on single-cohort analyses, which may have resulted in cohort-specific biases. This study aims to address this limitation by using cross-cohort analysis, which facilitates the identification of MF-associated predictors in a more robust and generalizable manner. Integrating cross-study normalization further addresses variability across different patient populations, enhancing the reliability of the results.

5.2 Objective of the current study

The primary objective of this study is to comprehensively analyze MF in PD through a cross-cohort analysis, leveraging various ML techniques. By integrating data from diverse PD patient populations, this study aims to identify key predictors associated with MF. The specific objectives of this chapter are outlined as follows:

1. Integrate cross-cohort analysis and cross-study normalization:
Aggregate and combine data from multiple cohorts of patients with PD, incorporating cross-study normalization techniques to create a comprehensive and diverse dataset. This will enable the assessment of the consistency of MF patterns across different populations, thereby facilitating the identification of demographic or clinical variations.

2. Develop predictive models using ML techniques:

Utilizing ML algorithms enables the detection of subtle patterns and relationships that traditional statistical methods may overlook. This approach facilitates the development of robust predictive models with favorable predictive capabilities. Based on their demographic and baseline clinical profiles, these models are designed to identify patients at higher risk of developing MF.

3. Identification of key predictors associated with MF:
Identification and ranking of the key predictors associated with MF, along with exploring the interactions and contributions of these predictors to the variability in MF among patients with PD.

The study aims to provide insight into the multifactorial nature of MF in PD by improving predictive capabilities that can be used in clinical settings to predict MF in individual patients. This will facilitate the implementation of early and targeted interventions to manage the disease more effectively. Identifying key predictors and developing predictive models will improve the ability to manage these fluctuations effectively, leading to improved clinical outcomes and quality of life for patients with PD.

5.3 Research methodology

To comprehensively understand MF in PD, this study uses a multi-faceted research methodology that integrates cross-cohort analysis and ML techniques using clinical data from three distinct PD cohorts. The criteria for these cohorts are detailed in Section 3.1. This methodology uses diverse patient data and advanced analytical tools to identify key predictors and develop predictive models. This approach ensures a particular examination of MF patterns, variability among different populations, and the complex relationships between clinical and demographic variables. The study aims to provide insights that can significantly improve the management of MF in PD patients by leveraging comprehensive analysis techniques.

5.3.1 Inclusion criteria

The inclusion criteria for this study on MF in PD are based on the MDS-UPDRS, a comprehensive tool for assessing the severity and progression of PD symptoms. To be included in the study, patients must show evidence of MF as defined by MDS-UPDRS Part IV or undergo relevant assessments to evaluate motor complications. Specifically, patients were classified as having PD with MF if they met the following criteria during the assessments: scoring ≥ 1 on either item 4.3 (time spent in the off state), item 4.4 (functional impact of fluctuations), or item 4.5 (complexity of motor fluctuations) of the MDS-UPDRS Part IV scale. The presence of MF observed during the clinical motor examination was also considered for the inclusion criteria (used for the PPMI and ICEBERG cohorts).

The total MDS-UPDRS Part IV score was excluded from further analysis to prevent the replication of MF evaluation. It is important to note that the LuxPARK cohort was assessed exclusively during the ON state. Therefore, the analysis is limited to data collected during this state, and assessments conducted during the OFF state were excluded. The study's inclusion criteria for participants were as follows:
Inclusion criteria (1): A diagnosis of PD according to the UK Parkinson's Disease Society Brain Bank Diagnostic Criteria (UKPDSBB) criteria [154], or subjects must have at least two of the following: rest tremor, bradykinesia, or rigidity with either rest tremor or bradykinesia, or a single asymmetric rest

tremor or asymmetric bradykinesia [103].
Inclusion criteria (2): Patients with MF are defined as those who present with MF within four years of the baseline clinical visit (MF+) and patients without MF within four years of follow-up (MF-).

Table 5.1 presents the number of PD patients who met the inclusion criteria for studying MF and the percentage of events (MF+) for classification and time-to-event analysis. The classification analysis is conducted on events occurring within four years, consistent with the ICEBERG cohort's follow-up period. This period ensures comparability across all cohorts. This approach is designed to identify those PD patients at an increased risk of developing MF within a shorter timeframe. However, due to the limitations of this timeframe, we also conducted a time-to-event analysis. This analysis offers the advantage of accounting for varying follow-up durations among patients, depending on their total participation time within each cohort study. In the time-to-event analysis, the outcome was defined as the duration until MF was observed or until the last follow-up if the event was censored. This method comprehensively explains the time-to-MF development across different patient cohorts.

Table 5.1 Number of patients meeting inclusion criteria for MF analysis.

Cohort	Inclusion criteria (1)	Inclusion criteria (2)	Events (MF Classification)	Events (Time-to-MF)
LuxPARK	706	395	281 (71.1%)	288 (72.9%)
PPMI	796	485	277 (57.1%)	445 (91.8%)
ICEBERG	162	116	58 (50.0%)	58 (50.0%)
Total	1664	996	616 (61.8%)	791 (79.4%)

Number of patients who met the inclusion criteria and the distribution of the events. The "Events" columns show the total number and percentage of subjects who developed motor fluctuations (MF) during the specified period. MF classification was performed up to 4 years of follow-up, and time-to-MF analysis was performed up to the last available follow-up visit for each patient.

5.3.2 Machine learning framework

We applied the same ML framework previously detailed in Section 4.3.2 to analyze MF in PD further. This framework includes a comprehensive set of processes, such as data preprocessing, imputation of missing values, categorical encoding, cross-cohort normalization, technique for addressing class imbalance, and feature selection with two levels of cross-validation (CV). These steps ensure the consistency and robustness of our analytical approach by leveraging various ML classification and time-to-event algorithms.

ML classification techniques were applied to categorize PD patients into MF+ (those with motor fluctuations during 4-year follow-up visits) and MF- (those without motor fluctuations during the same period). Different ML classification algorithms, as described in Section 3.3, were used within a CV workflow detailed in Section 3.5 to ensure the robustness and generalizability of the models across different PD cohorts.

Furthermore, for the time-to-MF analysis, we focused on predicting the risk of MF onset by considering the duration from the baseline visit until the onset of MF for each patient up to their last follow-up visit. This can be achieved using time-to-event analysis techniques outlined in Section 3.4, which predict the risk score and conversion probability based on patients' characteristics.

Integrating multiple ML techniques into a CV workflow for both MF classification and time-to-MF analysis is intended to provide a comprehensive understanding of the factors contributing to the devel-

opment of MF in PD. These analyses improve predictive capabilities, allowing for the classification of PD patients at higher risk of developing MF during follow-up periods. The findings are valuable for clinicians, enabling personalized treatment strategies and early interventions for high-risk patients. Furthermore, Bayesian signed-rank tests were applied to compare the model's cross-validated performance statistics across cohorts.

We conducted a SHAP values analysis to understand the prediction model's predictors further, as described in Section 3.6. The SHAP values enable the identification of the predictors that significantly influence the decision-making process of the models, thereby enhancing transparency and interpretability. This analysis provides insights into the process of the ML models and allows for the improvement of the model's predictive capability and reliability. By identifying the contribution of each predictor, clinicians can make more informed decisions regarding patient management, which in turn should improve patient outcomes.

Furthermore, we estimated the hazard ratios (HRs) based on SHAP values to quantify the impact of individual predictors on the risk of developing MF. This approach enables the translation of the influence of specific features into a measure of relative risk, thereby providing a better understanding of how various factors contribute to MF. Section 3.6 provides a more detailed description of this measurement. By estimating HRs, we can identify which predictors have stronger associations with increased risk, refining our predictive models and improving their clinical utility. This comprehensive analysis aims to bridge the gap between advanced ML techniques and practical insights for managing MF in patients with PD.

As with the LID analysis, this study also developed two distinct prediction models to analyze the onset of MF in PD patients. The initial model was designated the *comprehensive* model, incorporating all baseline clinical characteristics shared across the cohorts without prior feature selection. This model aimed to use as much information as possible from the available data to capture various factors contributing to MF prediction. The *comprehensive* model was designed to provide a comprehensive overview of the predictors involved in developing MF by incorporating a wide range of clinical features.

The second model, the *refined* model focused on a subset of clinical features that excluded baseline MF and levodopa medication. The *refined* model was designed to identify additional predictors that might independently contribute to developing MF and levodopa treatment during model training. This approach enabled the identification of novel predictors that could offer insights related to MF independent of the effects of levodopa medication.

In developing these two models, the *comprehensive* model aimed to use the full scope of available clinical data to predict MF, while the *refined* model seeks to uncover additional predictors that could provide novel insights into the understanding and management of MF in PD. This dual approach improves the robustness and depth of the analysis, offering an understanding of the predictors of MF.

To assess the predictive performance of the *comprehensive* and *refined* models, we compared the hold-out AUC for MF classification models and the hold-out C-index for time-to-MF models within the same cohort analysis, applying the statistical test detailed in Section 3.7. These metrics provide a robust evaluation of the model's predictive capabilities and discriminatory power in predicting the onset of MF in patients with PD.

In addition, we assessed the performance of non-normalized versus normalized models within the same cohort. Normalization techniques are important in ensuring that the features have comparable scales, as this may significantly impact the performance of ML models. The hold-out AUC and C-

index were compared to *optimized* non-normalized and normalized models with the highest average cross-validated metrics to determine the impact of normalization on predictive performance.

The stability of the *optimized* model for each cohort analysis was assessed to ensure consistent performance. Data perturbations in the cross-validated training set further assess the stability of the models. This stability analysis examined the standard deviation of performance metrics such as the AUC for binary classification and the C-index for time-to-MF models across CV folds corresponding to the optimal hyperparameters. A low standard deviation is indicative of consistency and reliability. This stability analysis ensures that the models are stable, reliable, and generalizable, enhancing their capability to predict MF in PD patients across diverse clinical settings.

To gain further insight into the predictors of MF, we compared the selected features across the 5-fold CV within the single cohort analyses. This process involved calculating the frequency percentages of each selected feature across the folds within the CV process. We identified key predictors that consistently contribute to the model's performance across different cohorts by examining the consistency of feature selection frequency across multiple model training and evaluation iterations. The stable predictors show higher consistency in feature selection and are discussed in detail in Section 3.8. This analysis provides valuable insights into the robustness and reliability of the features that predict MF, enhancing the overall understanding of their impact on PD.

5.3.3 Statistical analysis

In this study, we conducted univariate analysis to investigate the relationships between each predictor and the occurrence of MF within four years (MF+ and MF-). This analysis allowed us to assess whether these associations varied significantly across different cohorts and the outcome, providing insights into potential cohort-specific factors influencing MF. Additionally, we performed correlation analysis to examine the relationships between predictors among features and quantify the correlation coefficients. By exploring these interrelationships, we gained a deeper understanding of how individual predictors may interact and influence the development of MF in PD. These methods, described in Section 3.9, contribute to evaluating the predictors in MF outcomes and improve the overall understanding of the factors associated with MF in PD.

5.3.4 Clinical utility analysis

We applied decision curve analysis (DCA) and calibration analysis, as discussed in Section 3.10, to gain further insight into the performance and reliability of our *optimized* predictive models for MF in PD. Through the DCA, the clinical utility of our models was evaluated by the net benefit across various threshold probabilities. This analysis illustrated the potential advantages of integrating our models into clinical practice, thereby assisting in diagnosis decisions for PD patients with MF.

The study also measured the area under the net benefit curve (AUNBC) to assess further the clinical utility of the predictive models for MF. This approach quantifies the effectiveness of the models in a clinical decision-making context by comparing the net benefit of different intervention strategies. The AUNBC was calculated for the *optimized* models and compared to the AUNBC for the "treat all" strategy. This comparison is intended to show the additional value of the predictive models over a baseline strategy, in which all patients receive intervention regardless of their predicted risk. A higher AUNBC for the *optimized* models indicates more excellent clinical utility regarding MF in PD.

Moreover, bootstrapping hypothesis testing was used to measure *p*-values and evaluate the significance of the observed differences. The null hypothesis was formulated to indicate no significant discrepancy in the AUNBC between the two *optimized* models. The *p*-values were adjusted using the Benjamini-Hochberg procedure to address the multiple comparisons. This adjustment ensures that the reported *p*-values accurately reflect the likelihood of actual differences, thereby confirming the reliability of the *optimized* models' superior performance in predicting MF in PD.

In addition, a calibration analysis was conducted to assess the comparison of the predicted probabilities with the actual outcomes for MF classification and the predicted conversion probabilities with the observed conversion probabilities at year 4 for the time-to-MF model. The calibration was evaluated by measuring the slope and the MSE of the calibration curve. These metrics provided insights into the accuracy and reliability of the models in predicting MF. A slope close to 1 and a lower MSE indicate better calibration, suggesting that the predicted probabilities are close to the observed outcomes. This calibration analysis further verifies the robustness of the predictive models.

5.3.5 Code availability

R (v4.2.1) was used for data processing, normalization, and statistical analyses, while Python-3.8.6-GCCcore-10.2.0 was used for ML predictions. The open-source code is available in the GitLab repository under the MIT license at
https://gitlab.com/uniluxembourg/lcsb/biomedical-data-science/bds/ml-motor-fluctuations.

5.4 Results

5.4.1 Individual cohort analyses

The *optimized* models for predicting MF in PD were selected based on the highest average cross-validated AUC scores for classification models and the highest average cross-validated C-indices for time-to-MF models across different cohorts. In the LuxPARK cohort, the *comprehensive* MF classification model achieved an average cross-validated AUC of 0.638 (SD 0.069) with a hold-out AUC of 0.554. According to permutation importance analysis, this model used 10 features with predictive impact (see Table 5.2). Compared to this, another model with a lower average cross-validated AUC of 0.561 (SD 0.152) achieved a slightly higher hold-out AUC of 0.600. The *refined* classification model showed a similar average cross-validated AUC, reaching 0.625 (SD 0.074), while the hold-out AUC increased to 0.6, utilizing 5 features (see Table B.1). For the analysis of time-to-MF, the *comprehensive* model achieved an average cross-validated C-index of 0.669 (SD 0.134) with a hold-out C-index of 0.606 using 9 predictors (see Table 5.3). The *refined* time-to-MF model showed an average cross-validated C-index of 0.555 (SD 0.114) with a hold-out C-index of 0.665 (see Table B.2). The *p*-values for comparing *comprehensive* and *refined* models were 0.639 for classification and 0.139 for time-to-MF, as shown in Table 5.4, indicating no statistically significant differences at a 5% significance level when excluding baseline MF and levodopa medication from model training.

In the PPMI cohort, the *comprehensive* classification model achieved an average cross-validated AUC of 0.711 (SD 0.044) with a hold-out AUC of 0.710, utilizing 4 features. Utilizing 32 features, the *refined* model showed an average cross-validated AUC of 0.700 (SD 0.048) with a hold-out AUC of 0.631. The *comprehensive* MF classification model achieved higher hold-out predictive performance than the *refined*

model across key metrics, including precision, recall, f-score, accuracy, and balanced accuracy (data are not shown in the thesis). The *refined* model, while showing slightly lower performance, maintained consistent results across all metrics. In the context of time-to-MF analysis, the *comprehensive* model achieved an average cross-validated C-index of 0.711 (SD 0.047) with a hold-out C-index of 0.718. The *refined* time-to-MF model showed an average cross-validated C-index of 0.692 (SD 0.053) with a hold-out C-index of 0.704. The *p*-values from DeLong's and one-shot nonparametric tests indicated no statistically significant differences between the *comprehensive* and *refined* models in PPMI cohort analyses. Notably, the PPMI cohort showed higher average cross-validated AUC and C-index, along with smaller SD, compared to the LuxPARK and ICEBERG cohorts. This result indicates that the predictive models are more stable and reliable in the PPMI cohort.

The *comprehensive* classification model showed an average cross-validated AUC of 0.689 (SD 0.129) for the ICEBERG cohort, with a hold-out AUC of 0.531. This model used 7 features with predictive impact, as determined by permutation importance analysis. The *refined* model showed comparable predictive performance. In the context of time-to-MF analysis, the *comprehensive* model achieved an average cross-validated C-index of 0.628 (SD 0.093) with a hold-out C-index of 0.687, utilizing 12 predictors. The *refined* time-to-MF model showed an average cross-validated C-index of 0.577 (SD 0.057) with a hold-out C-index of 0.517, with 11 predictive features. The *p*-values for comparing the *comprehensive* and *refined* models indicated no statistically significant differences.

Table 5.2 Predictive performance metrics for *comprehensive* MF classification in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features
AdaBoost	0.566 (0.131)	0.530	2 (5)	***0.711 (0.044)***	***0.710***	***4 (9)***	0.530 (0.106)	0.539	3 (3)
CART	0.554 (0.107)	0.565	5 (8)	0.686 (0.064)	0.606	2 (5)	0.582 (0.125)	0.592	4 (9)
CatBoost	0.532 (0.052)	0.570	7 (10)	0.711 (0.054)	0.681	6 (8)	0.608 (0.122)	0.531	12 (19)
C4.5	0.556 (0.096)	0.550	4 (7)	0.673 (0.050)	0.623	3 (5)	**0.689 (0.129)**	**0.531**	**7 (13)**
FIGS	0.556 (0.111)	0.579	2 (7)	0.675 (0.047)	0.606	2 (5)	0.607 (0.080)	0.542	3 (6)
GOSDT-GUESSES	**0.638 (0.069)**	**0.554**	**10 (11)**	0.657 (0.060)	0.571	12 (12)	0.539 (0.202)	0.556	11 (12)
GBoost	0.578 (0.111)	0.544	5 (5)	0.689 (0.039)	0.591	15 (20)	***0.520 (0.213)***	***0.611***	***5 (8)***
HS	0.529 (0.136)	0.487	1 (1)	0.671 (0.049)	0.606	2 (5)	0.607 (0.080)	0.594	2 (6)
XGBoost	***0.561 (0.152)***	***0.600***	***18 (28)***	0.676 (0.069)	0.652	33 (53)	0.501 (0.205)	0.528	24 (36)

An overview of the *comprehensive* MF prognostic classification's predictive performance statistics summarizes the *comprehensive* MF prognostic classification's predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 5.3 Predictive performance metrics for *comprehensive* time-to-MF in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features
CW-GBoost	**0.669 (0.134)**	**0.606**	**9 (9)**	0.708 (0.022)	0.705	11 (23)	0.587 (0.094)	0.594	5 (13)
Extra Survival	***0.587 (0.139)***	***0.629***	***10 (10)***	**0.711 (0.047)**	**0.718**	**108 (114)**	0.575 (0.145)	0.642	10 (10)
Survival GBoost	0.598 (0.081)	0.588	81 (88)	0.681 (0.044)	0.684	12 (17)	***0.628 (0.093)***	***0.687***	***12 (19)***
LSVM	0.605 (0.161)	0.608	8 (8)	0.679 (0.06)	0.726	27 (27)	0.482 (0.068)	0.600	11 (11)
NLSVM	0.532 (0.173)	0.601	16 (16)	***0.695 (0.043)***	***0.736***	***30 (30)***	0.501 (0.105)	0.661	15 (15)
Penalized Cox	0.533 (0.144)	0.565	1 (2)	0.700 (0.036)	0.710	29 (37)	0.518 (0.049)	0.624	14 (14)
Survival RF	0.569 (0.109)	0.626	9 (9)	0.687 (0.013)	0.716	65 (94)	0.596 (0.046)	0.603	11 (11)
Survival Trees	0.555 (0.05)	0.595	14 (20)	0.640 (0.032)	0.637	6 (7)	0.538 (0.054)	0.633	6 (8)

An overview of the *comprehensive* time-to-MF predictive performance statistics summarizes the *comprehensive* time-to-MF predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-index in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out C-index is indicated *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 5.4 Significance testing of hold-out predictive metrics between normalized and unnormalized models for MF in multi-cohort analyses.

Cohort	MF classification	time-to-MF
LuxPARK	0.639	0.139
PPMI	0.217	0.223
ICEBERG	0.883	0.080
Cross-cohort	0.594	0.401
Leave-ICEBERG-out	0.090	0.080
Leave-PPMI-out	0.618	0.004
Leave-LuxPARK-out	0.587	0.950

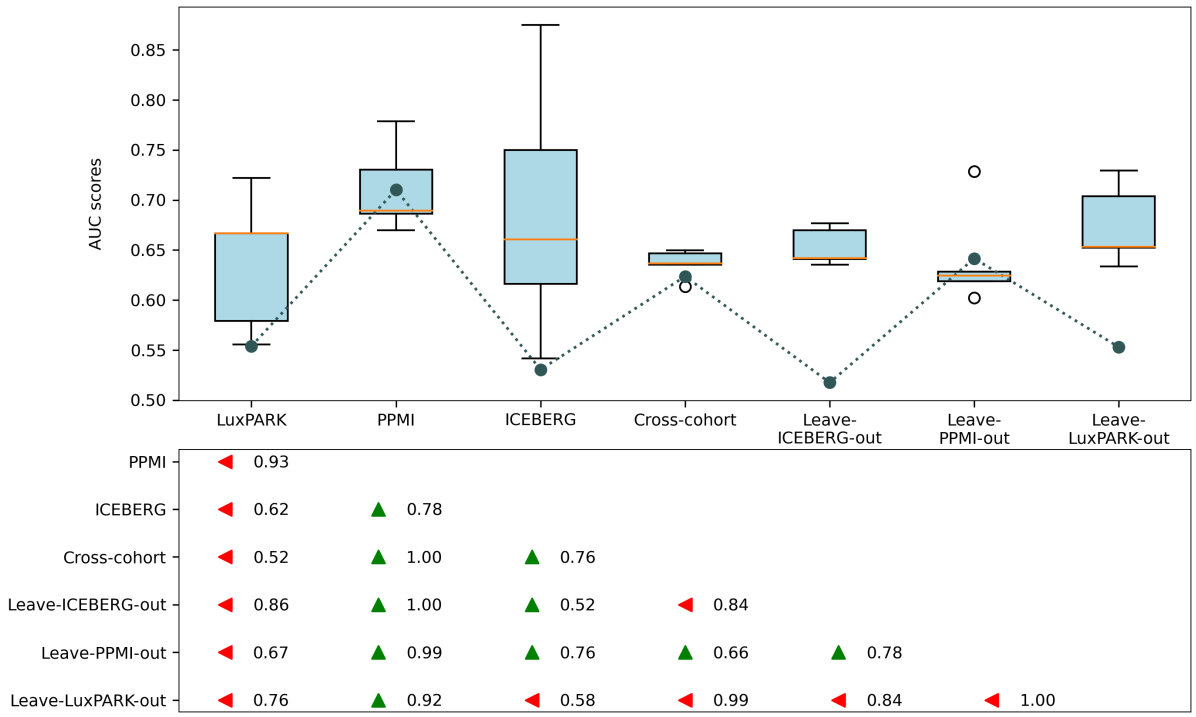
A comparison of the statistical significance of the differences between the hold-out predictive performance metrics for the *optimized comprehensive* and *refined* models across cohorts. The p-values for the significance of the difference were calculated using DeLong's test for MF classification and the one-short nonparametric test for time-to-MF analysis. A p-value < 0.05 indicates a significant difference in hold-out predictive performance between the two models.

The Bayesian signed-rank test was applied to compare the cross-validated predictive performance of the *optimized comprehensive* and *refined* models across the cohorts. In the case of the MF classification models, the PPMI cohort consistently showed superior predictive performance. In the *comprehensive* model (see Figure 5.1), PPMI showed superior performance to LuxPARK with a probability of 0.93 and

to ICEBERG with a probability of 0.78, indicating that PPMI and ICEBERG showed similar performance. The probability of ICEBERG being superior to LuxPARK was 0.62. In the *refined* model (see Figure B.1), PPMI showed superior performance to LuxPARK with a probability of 0.97 and to ICEBERG with a probability of 0.58. The probability of ICEBERG outperforming LuxPARK was 0.90.

In the time-to-MF models, the PPMI cohort showed superior performance. In the *comprehensive* time-to-MF model (see Figure 5.2), PPMI showed superior performance to LuxPARK with a probability of 0.67 and to ICEBERG with a probability of 0.90. The probability of LuxPARK demonstrating superiority over ICEBERG was 0.62. In the *refined* model (see Figure B.2), PPMI showed even greater superiority, with a probability of 1.00 over LuxPARK and ICEBERG. The probability of ICEBERG being superior to LuxPARK was 0.58.

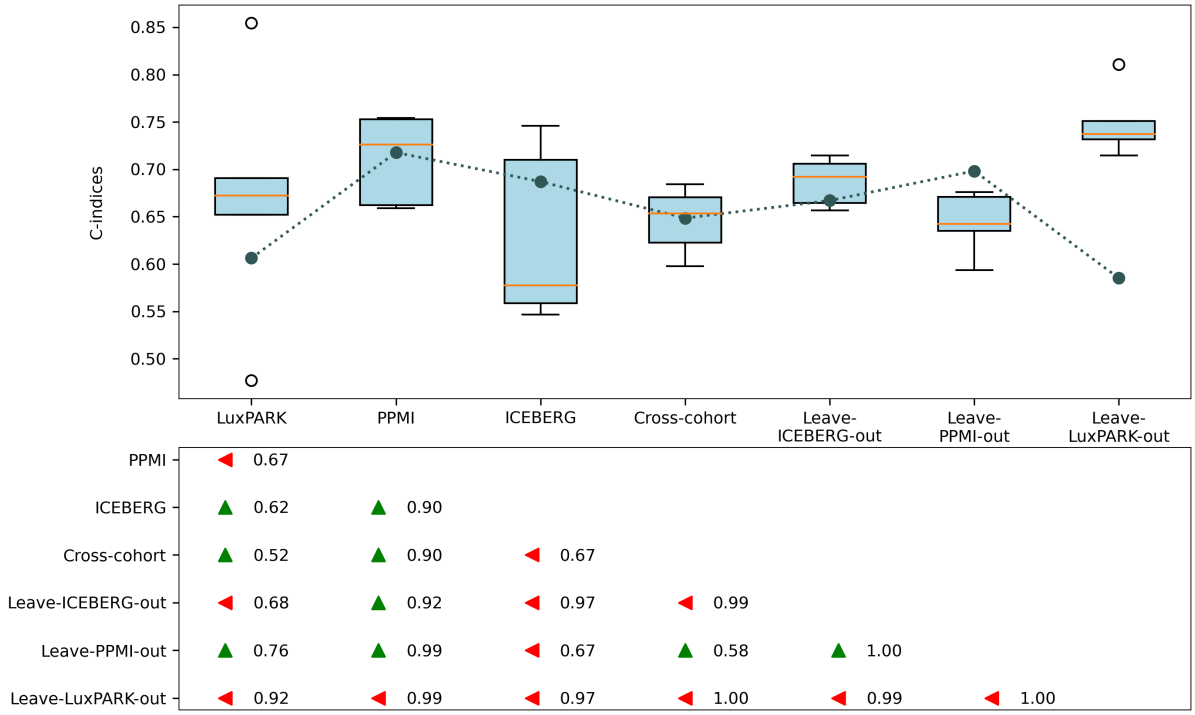
Figure 5.1 Comparison of cross-validated AUC values for the *comprehensive* MF classification models.



A comparison of cross-validated AUC scores and probabilities of better predictive performance for the *optimized comprehensive* MF classification model across cohort analyses. The upper part displays boxplots of the cross-validated AUC scores for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

To determine the stability of our predictive models across different cohorts and data perturbations during 5-fold CV, we examined the standard deviation of the predictive performance metrics corresponding to the optimal hyperparameters. The stability results for the *comprehensive* MF classification (Figure B.3), *refined* MF classification (Figure B.4), *comprehensive* time-to-MF (Figure B.5), and *refined* time-to-MF (Figure B.6) models across each cohort analysis indicate that the PPMI cohort consistently showed lower standard deviations in predictive performance. This indicates that the models trained

Figure 5.2 Comparison of cross-validated C-indices for the *comprehensive* time-to-MF models.



A comparison of cross-validated C-indices and probabilities of better predictive performance for the *optimized comprehensive* time-to-MF model across cohort analyses. The upper part displays boxplots of the cross-validated C-indices for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

on the PPMI cohort were more stable and reliable than those trained on the LuxPARK and ICEBERG cohorts, regardless of whether the *comprehensive* or *refined* models.

Table 5.5 presents the top predictors from the *optimized comprehensive* and *refined* models for MF classification and time-to-MF analyses across each cohort. The predictors, ranked by their average selection percentages during a 5-fold CV, identify the most influential predictors in predicting MF in PD. In all three cohort analyses, disease duration and age of PD onset were consistently identified as the key predictors of MF development. Furthermore, body weight, axial symptoms, MDS-UPDRS Parts I and II, Benton JLO, and autonomic dysfunction, including GI, were identified as significant predictors, reflecting their association with overall health and cognitive function. Also, axial symptoms and various indicators from MDS-UPDRS Part I, Part II, and SCOPA-AUT significantly predicted MF in PD. It is important to note that these predictive features are not independent of each other, even though their joint inclusion in the predictive models suggests an added value in capturing multiple interrelated features. In particular, disease duration is a key determining factor for many of these variables, including axial symptoms, MDS-UPDRS scores, SCOPA-AUT, and generally all features related to disease severity. This interconnectedness reflects the multifaceted progressive nature of PD and highlights the complex relationships among these predictors.

5.4.2 Cross-cohort analyses

In addition to the single-cohort analyses, we conducted multi-cohort analyses to improve the generalization and robustness of the MF prediction models and extend their applicability to different patient populations. Specifically, both CV analyses involving samples from all cohorts (termed cross-cohort analysis), leave-ICEBERG-out, leave-PPMI-out, and leave-LuxPARK-out analyses were performed.

In the cross-cohort analysis, the *comprehensive* MF classification model achieved an average cross-validated AUC of 0.637 (SD 0.014) and a hold-out AUC of 0.624, utilizing 8 predictive features (see Table 5.6). The *refined* classification model, utilizing 7 features, showed an average cross-validated AUC of 0.625 (SD 0.014) and a hold-out AUC of 0.591 (see Table B.3). Both the *comprehensive* and *refined* MF classification models showed consistent hold-out prediction performance across key metrics, including precision, recall, F-score, accuracy, and balanced accuracy. In the analysis of time-to-MF, the *comprehensive* model achieved an average cross-validated C-index of 0.646 (SD 0.035) and a hold-out C-index of 0.649 with 8 features (see Table 5.7). The *refined* model showed an average cross-validated C-index of 0.646 (SD 0.021) and a hold-out C-index of 0.633 with 14 features (see Table B.4). The p-values for comparing the *comprehensive* and *refined* models were 0.594 for classification and 0.401 for time-to-MF (see Table 5.4), indicating no statistically significant difference in the models when excluding the baseline MF and levodopa medication during model training.

In the leave-ICEBERG-out analysis, the *comprehensive* classification model showed an average cross-validated AUC of 0.653 (SD 0.019) and a hold-out AUC of 0.518, using 14 features. The *refined* classification model showed an average cross-validated AUC of 0.650 (SD 0.036) and a hold-out of 0.586, with 17 features. The *comprehensive* model achieved an average cross-validated C-index of 0.687 (SD 0.025) and a hold-out C-index of 0.667 with 16 features in the time-to-MF analysis. The *refined* model showed an average cross-validated C-index of 0.687 (SD 0.031) and a hold-out C-index of 0.572. The p-values for these comparisons indicate no significant differences in the hold-out performance in the models when excluding the levodopa medication.

Table 5.5 The average percentage of predictors selected in 5-fold cross-validation for MF classification and time-to-MF analyses across LuxPARK, PPMI, and ICEBERG cohorts.

Predictors	*Comprehensive* model			*Refined* model		
	Classification Average in CV (%)	Time-to-MF Average in CV (%)	Overall Average (%)	Classification Average in CV (%)	Time-to-MF Average in CV (%)	Overall Average (%)
Disease duration since PD diagnosis (years)	86.67	46.67	66.67	80.00	80.00	80.00
Age at PD diagnosis	73.33	53.33	63.33	80.00	80.00	80.00
MDS-UPDRS I - Urinary problems	53.33	46.67	50.00	46.67	93.33	70.00
MDS-UPDRS I - Sleep problems (night)	66.67	53.33	60.00	46.67	73.33	60.00
BMI (kg/m^2)	66.67	33.33	50.00	60.00	66.67	63.33
MDS-UPDRS Part II score	60.00	46.67	53.33	60.00	60.00	60.00
Benton Judgment of Line Orientation	46.67	46.67	46.67	53.33	66.67	60.00
MDS-UPDRS Part I score	53.33	46.67	50.00	40.00	73.33	56.67
SCOPA-AUT Gastrointestinal (GI)	66.67	40.00	53.33	26.67	73.33	50.00
Axial symptoms	33.33	46.67	40.00	46.67	80.00	63.33
Weight (kg)	46.67	40.00	43.33	33.33	80.00	56.67
Height (cm)	46.67	33.33	40.00	46.67	66.67	56.67
MDS-UPDRS II - Saliva and drooling	40.00	53.33	46.67	33.33	66.67	50.00
SCOPA-AUT Urinary	60.00	26.67	43.33	33.33	73.33	53.33
SCOPA-AUT Thermoregulatory	60.00	26.67	43.33	33.33	73.33	53.33

An overview of the feature selection analysis performed using 5-fold cross-validation compares the average percentage of times clinical features were selected between the *comprehensive* and *refined* models for MF classification and time-to-MF analyses across the LuxPARK, PPMI, and ICEBERG cohorts. The column labeled 'Average in CV (%)' displays the average percentage of times each feature was selected during the 5-fold cross-validation in single cohort analyses within LuxPARK, PPMI, and ICEBERG for both MF and time-to-MF analyses. The column labeled 'Average (%)' represents the mean of the 'Average in CV (%)' values across the cohorts for analyses of MF and time-to-MF. The top 15 predictors are listed in descending order based on their average selection percentages for the *comprehensive* and *refined* models in MF and time-to-MF analyses.

In the leave-PPMI-out analysis, the *comprehensive* classification model achieved an average cross-validated AUC of 0.641 (SD 0.050) and a hold-out AUC of 0.642. The *refined* classification model showed an average cross-validated AUC of 0.635 (SD 0.064) and a hold-out AUC of 0.647, with 19 features. The *comprehensive* model achieved an average cross-validated C-index of 0.644 (SD 0.033) and a hold-out C-index of 0.698 in the time-to-MF analysis, utilizing 21 features. The *refined* model yielded an average cross-validated C-index of 0.623 (SD 0.033) and a hold-out C-index of 0.661. The p-values for these comparisons were 0.618 for classification and 0.004 for time-to-MF, indicating significant differences between the hold-out performance of the *comprehensive* and *refined* time-to-MF models.

In the leave-LuxPARK-out analysis, we observed higher average cross-validated performance metrics than other multi-cohort analyses. However, this observation did not translate into similarly high hold-out predictive performance. In particular, the average cross-validated AUC for the *optimized comprehensive* MF classification model was 0.675 (SD 0.040), whereas the hold-out AUC was lower at 0.553. Similarly, the average cross-validated C-index for the *optimized comprehensive* time-to-MF model was 0.749 (SD 0.037), but the hold-out C-index decreased to 0.585. These findings suggest that while the model performed satisfactorily during CV, it tended to overfit when applied to the hold-out LuxPARK cohort.

The Bayesian signed-rank test was used to assess the cross-validated performance of models across different multi-cohort analyses. The results were then summarized with hold-out AUC/C-index values, providing a comprehensive evaluation. For the MF classification models, the cross-cohort *refined* model showed a probability (0.66) of being superior in cross-validated AUC compared to the leave-PPMI-out analysis, with a hold-out AUC of 0.642.

The leave-ICEBERG-out analysis showed a high probability of superiority in cross-validated C-index for both *comprehensive* and *refined* models in the case of the time-to-MF models. However, this analysis yielded lower hold-out C-index values (0.667 for *comprehensive*, 0.572 for *refined*) than the cross-cohort hold-out C-index (0.649 and 0.633, respectively). Conversely, the leave-PPMI-out *comprehensive* model showed the performance with a high probability of superiority in cross-validated C-index and a higher hold-out C-index of 0.698, indicating robust performance across both metrics.

Overall, while the leave-ICEBERG-out analysis showed superior cross-validated performance, it often showed lower hold-out performance than the cross-cohort analysis. In contrast, the leave-PPMI-out analysis showed performance with superior cross-validated performance and higher hold-out performance in specific models. This emphasizes the necessity of considering cross-validated and hold-out metrics when evaluating model performance across different cohort analyses.

A comparison of the cross-validated performance between the multi-cohort analysis and single cohort analysis revealed that PPMI showed superior performance to the cross-cohort analysis. In contrast, the cross-cohort analysis performed better than LuxPARK in the *comprehensive* MF classification model. Similarly, for the *refined* MF classification model, PPMI showed superiority over the cross-cohort analysis, whereas the cross-cohort analysis showed superior performance over LuxPARK. In the *comprehensive* time-to-MF model, PPMI showed superiority over the cross-cohort analysis, while LuxPARK and ICEBERG showed comparable performance to the cross-cohort analysis. In the *refined* time-to-MF model, PPMI was superior to the cross-cohort analysis, whereas the cross-cohort analysis was superior to ICEBERG.

The stability analysis indicates that the multi-cohort analysis showed consistently lower standard deviations in performance metrics. These results indicate that multi-cohort analyses are more stable than single-cohort analyses. This indicates that integrating data from multiple cohorts may result in

more stable predictive models. In contrast, the single-cohort analyses, particularly for LuxPARK and ICEBERG, showed higher standard deviations, indicating less stable performance. The PPMI cohort showed comparable stability to the multi-cohort analysis, with lower standard deviations in its predictive performance.

Table 5.6 Predictive performance metrics for *comprehensive* MF classification in multi-cohort analyses.

Algorithm	Cross-cohort			Leave-ICEBERG-out		
	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features
AdaBoost	0.628 (0.024)	0.606	4 (10)	0.645 (0.035)	0.589	4 (7)
CART	0.616 (0.023)	0.611	3 (9)	***0.632 (0.064)***	***0.679***	***14 (24)***
CatBoost	**0.637 (0.014)**	**0.624**	**8 (8)**	**0.653 (0.019)**	**0.518**	**14 (22)**
C4.5	0.628 (0.026)	0.624	3 (6)	0.641 (0.031)	0.629	4 (9)
FIGS	0.625 (0.038)	0.636	5 (9)	0.640 (0.036)	0.609	5 (8)
GOSDT-GUESSES	0.612 (0.058)	0.582	32 (44)	0.636 (0.026)	0.456	19 (19)
GBoost	0.629 (0.046)	0.624	17 (31)	0.650 (0.043)	0.608	18 (29)
HS	0.628 (0.055)	0.625	2 (7)	0.640 (0.036)	0.609	5 (8)
XGBoost	***0.623 (0.036)***	***0.647***	***18 (18)***	0.646 (0.059)	0.634	35 (48)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
AdaBoost	0.610 (0.075)	0.564	5 (11)	0.668 (0.035)	0.573	4 (6)
CART	0.604 (0.055)	0.571	18 (37)	0.654 (0.057)	0.569	11 (15)
CatBoost	0.636 (0.057)	0.625	15 (27)	**0.675 (0.040)**	**0.553**	**9 (18)**
C4.5	0.604 (0.037)	0.603	3 (7)	0.647 (0.046)	0.534	8 (11)
FIGS	0.599 (0.047)	0.598	2 (5)	0.667 (0.041)	0.571	7 (19)
GOSDT-GUESSES	0.594 (0.040)	0.574	35 (52)	***0.642 (0.042)***	***0.581***	***19 (19)***
GBoost	***0.595 (0.042)***	***0.660***	***17 (35)***	0.660 (0.050)	0.579	5 (9)
HS	0.599 (0.047)	0.598	2 (5)	0.667 (0.041)	0.571	7 (19)
XGBoost	**0.641 (0.050)**	**0.642**	**41 (63)**	0.661 (0.021)	0.558	56 (70)

An overview of the *comprehensive* MF prognostic classification's predictive performance statistics summarizes the *comprehensive* MF prognostic classification's predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

5.4.3 Differences between clinical features across cohorts

The analysis of baseline features revealed significant differences among the LuxPARK, PPMI, and ICEBERG cohorts. These differences highlight variations in PD patient demographics and clinical characteristics, as shown in Table 5.8. Three cohorts showed no significant difference in age at PD diagnosis. Furthermore, LuxPARK patients showed a significantly longer disease duration than those in the PPMI and ICEBERG cohorts, with averages of 4.2 years and 4.8 years longer, respectively.

Comparative analyses of body weight showed no statistically significant differences between LuxPARK and PPMI patients. However, patients in the ICEBERG cohort showed a significantly lower average body weight than the LuxPARK and PPMI cohorts, with average differences of 5.1kg and 5.4kg, respectively. A similar observation on the BMI, with ICEBERG patients showed a significantly lower average BMI than LuxPARK and PPMI, with differences of an average of 2.4 and 1.7, respectively.

Regarding the MDS-UPDRS scores, patients in the PPMI cohort showed significantly lower scores on average across all parts than the LuxPARK and ICEBERG cohorts. This finding highlights significant

Table 5.7 Predictive performance metrics for *comprehensive* time-to-MF in multi-cohort analyses.

Algorithm	Cross-cohort			Leave-ICEBERG-out		
	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features
CW-GBoost	**0.646 (0.035)**	**0.649**	**8 (9)**	***0.687 (0.025)***	***0.667***	***16 (27)***
Extra Survival	0.64 (0.045)	0.680	158 (160)	0.685 (0.011)	0.588	144 (150)
Survival GBoost	0.64 (0.032)	0.654	18 (29)	0.66 (0.069)	0.656	11 (24)
LSVM	***0.627 (0.038)***	***0.687***	***42 (42)***	0.674 (0.033)	0.589	57 (57)
NLSVM	0.614 (0.033)	0.648	47 (47)	0.672 (0.043)	0.612	55 (55)
Penalized Cox	0.589 (0.082)	0.614	2 (3)	0.671 (0.046)	0.500	1 (3)
Survival RF	0.641 (0.034)	0.664	90 (129)	0.681 (0.037)	0.565	24 (55)
Survival Trees	0.59 (0.058)	0.589	2 (2)	0.629 (0.064)	0.489	4 (6)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
CW-GBoost	0.63 (0.065)	0.698	13 (24)	***0.734 (0.032)***	***0.625***	***15 (24)***
Extra Survival	***0.644 (0.033)***	***0.698***	***21 (21)***	0.729 (0.032)	0.617	123 (123)
Survival GBoost	0.616 (0.043)	0.690	12 (21)	0.712 (0.024)	0.560	29 (57)
LSVM	0.632 (0.036)	0.696	40 (40)	0.725 (0.039)	0.572	53 (53)
NLSVM	0.625 (0.045)	0.661	32 (32)	0.727 (0.039)	0.586	54 (54)
Penalized Cox	0.607 (0.042)	0.590	2 (4)	**0.749 (0.037)**	**0.585**	**83 (118)**
Survival RF	0.619 (0.03)	0.697	101 (115)	0.728 (0.034)	0.623	71 (103)
Survival Trees	0.597 (0.043)	0.642	10 (18)	0.665 (0.036)	0.614	8 (10)

An overview of the *comprehensive* time-to-MF predictive performance statistics summarizes the *comprehensive* time-to-MF predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-index in the cross-validation of the cohort analyses are highlighted in bold. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

variations in baseline clinical characteristics among the cohorts. In particular, the average score for MDS-UPDRS Part I was 8.5 points higher for LuxPARK than for PPMI, while PPMI scored 6.4 points lower than that of ICEBERG. For MDS-UPDRS Part II, LuxPARK scored 6.4 points higher than PPMI and 4.9 points higher than ICEBERG, while PPMI showed scores lower than ICEBERG. A comparable pattern was observed in MDS-UPDRS Part III (ON) and SCOPA-AUT scores; LuxPARK patients consistently scored higher than PPMI and ICEBERG, while PPMI showed lower scores than ICEBERG in these assessments.

Table 5.8 Comparative analysis of baseline features mean differences across cohorts in MF analysis.

Predictors	**LuxPARK vs. PPMI (*p*-values)**	**LuxPARK vs. ICEBERG (*p*-values)**	**PPMI vs. ICEBERG (*p*-values)**	***p*-values**
Age at PD diagnosis	-	-	-	1.52E-01
Disease duration	4.19 (7.75E-41)	4.76 (1.04E-24)	0.57 (0.236)	7.51E-48
Weight (kg)	-0.32 (1.000)	5.11 (0.025)	5.43 (0.018)	1.69E-02
Height (cm)	-2.33 (0.002)	-1.81 (0.276)	0.52 (1.000)	2.90E-03
BMI (kg/m^2)	0.66 (0.059)	2.36 (1.19E-05)	1.70 (0.004)	1.82E-05
MDS-UPDRS Part I score	8.52 (2.29E-89)	2.16 (0.829)	-6.37 (3.95E-33)	1.14E-97
MDS-UPDRS Part II score	6.40 (1.17E-38)	4.90 (5.50E-06)	-1.50 (6.52E-04)	6.39E-38
MDS-UPDRS Part III (ON) score	17.50 (9.57E-31)	8.19 (2.39E-04)	-9.31 (8.14E-08)	1.84E-30
SCOPA-AUT total score	10.64 (5.71E-101)	3.54 (0.004)	-7.10 (1.75E-26)	4.60E-105

A comparative analysis of the mean differences for baseline features across the LuxPARK, PPMI, and ICEBERG cohorts. The *p*-values indicated statistically significant differences in the average of the predictors between specific cohort pairs, providing insights into cohort-specific variations in predictor distributions in motor fluctuations analysis.

5.4.4 Comparative evaluation of cross-study integration

Since data from different studies may be affected by different biases and study-specific data distributions, a comparative evaluation of cross-study normalization methods for MF analysis is conducted to determine the most favorable approaches to model performance. A statistically significant difference in hold-out performance was observed between the normalized models and those without additional normalization (p = 0.011), particularly in the cross-cohort *comprehensive* time-to-MF models (Table 5.9). This resulted in improved predictive performance (see Table 5.10), indicating that while cross-study normalization techniques can improve predictive capability by adjusting for study-specific sources of bias and variability, they do not always result in significant performance improvements.

5.4.5 Associations of clinical features with motor fluctuations outcome

The associations of clinical features with MF were analyzed to identify the most predictive features and to highlight their roles as potential risk or protective factors. This analysis also examined the direction of their statistical associations with MF outcomes. In the cross-cohort analyses, we identified levodopa medication, disease duration, the presence of dyskinesia (a common PD complication involving involuntary movements), tremor, and age at PD diagnosis as the most frequently selected features (Table

Table 5.9 Significance testing of hold-out predictive metrics between normalized and unnormalized for MF models in multi-cohort analyses.

	***Comprehensive* model**		***Refined* model**	
Cohort	**Normalized vs. Unnormalized**	**Cross-cohort normalization**	**Normalized vs. Unnormalized**	**Cross-cohort normalization**
MF classification:				
Cross-cohort	0.998	Mean-centering	0.925	Quantile
Leave-ICEBERG-out	0.158	M-ComBat	0.937	M-ComBat
Leave-PPMI-out	0.656	ComBat	0.949	Ratio-A
Leave-LuxPARK-out	0.838	M-ComBat	0.751	Standardize
Time-to-MF:				
Cross-cohort	0.011	Mean-centering	0.449	Quantile
Leave-ICEBERG-out	0.190	ComBat	0.892	ComBat
Leave-PPMI-out	0.855	M-ComBat	0.891	M-ComBat
Leave-LuxPARK-out	0.553	Standardize	0.467	Standardize

A comparison of the statistical significance of the differences between the hold-out predictive performance metrics for the *optimized comprehensive* and *refined* models across cohorts. The *p*-values for the significance of the difference were calculated using DeLong's test for MF classification and the one-shot nonparametric test for time-to-MF analysis. A *p*-value < 0.05 indicates a significant difference in hold-out predictive performance between the two models. The normalization method used on the *optimized* model is indicated in the column "Normalization".

5.11). Levodopa medication, disease duration, and dyskinesia consistently showed high predictive impact and were associated with a higher likelihood of MF development. In contrast, resting tremor and Benton JLO (higher scores indicate better visuospatial ability) were negatively associated with MF (Figure 5.3, Figure B.7, and Table 5.13).

Similar to the classification models, the *comprehensive* and *refined* time-to-MF models, illustrated in Figures 5.4 and B.8, identify the predictors with higher influence in the predictive model. These include levodopa treatment, disease duration, dyskinesia, and tremor. Furthermore, specific symptoms from the MDS-UPDRS, such as fatigue and sleep disorders, are identified as influential. Furthermore, other predictors, including axial symptoms, also contribute to determining the time-to-MF. These findings highlight the multifactorial nature of MF in predicting their onset.

The single cohort analysis revealed that the PPMI cohort showed higher cross-validated and hold-out AUC/C-index values in both MF classification and time-to-MF analyses. The SHAP values illustrate the predictors of these *optimized* model plots in Figures B.9 and B.10 for the *optimized comprehensive* MF classification and time-to-MF models in PPMI, respectively. As in the cross-cohort analysis, key predictors observed in PPMI include the age of PD diagnosis, levodopa treatment, axial symptoms, disease duration, as well as the pathogenic *GBA* and *LRRK2* variants.

Table 5.12 presents the HRs and their 95% confidence intervals (CIs) for predictors in the *comprehensive* time-to-MF model within the cross-cohort analysis. The corresponding SHAP values for these predictors are illustrated in Figure 5.4. Although a key predictor, levodopa did not display significant differences in the time-to-MF analysis.

Table 5.10 Predictive performance metrics between normalized and unnormalized for MF models in multi-cohort analyses.

MF classification						
	Cross-cohort			**Leave-ICEBERG-out**		
	Mean (SD)	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**
***Comprehensive* model**						
Normalized	0.637 (0.014)	0.624	8 (8)	0.653 (0.019)	0.518	14 (22)
Unnormalized	0.628 (0.049)	0.663	8 (10)	0.639 (0.064)	0.632	43 (67)
***Refined* model**						
Normalized	0.625 (0.014)	0.591	7 (17)	0.650 (0.036)	0.586	17 (28)
Unnormalized	0.619 (0.030)	0.640	2 (4)	0.622 (0.052)	0.623	10 (17)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
***Comprehensive* model**						
Normalized	0.641 (0.05)	0.642	41 (63)	0.675 (0.04)	0.553	9 (18)
Unnormalized	0.636 (0.057)	0.625	15 (27)	0.674 (0.051)	0.591	9 (17)
***Refined* model**						
Normalized	0.635 (0.064)	0.647	19 (33)	0.651 (0.022)	0.581	5 (7)
Unnormalized	0.633 (0.036)	0.622	47 (53)	0.642 (0.079)	0.560	15 (26)
Time-to-MF						
	Cross-cohort			**Leave-ICEBERG-out**		
	Mean (SD)	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**
***Comprehensive* model**						
Normalized	0.643 (0.055)	0.694	13 (23)	0.687 (0.025)	0.667	16 (27)
Unnormalized	0.646 (0.035)	0.649	8 (9)	0.672 (0.043)	0.612	55 (55)
***Refined* model**						
Normalized	0.646 (0.021)	0.633	14 (26)	0.687 (0.031)	0.572	50 (91)
Unnormalized	0.636 (0.046)	0.648	6 (25)	0.668 (0.031)	0.580	15 (29)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
***Comprehensive* model**						
Normalized	0.644 (0.033)	0.698	21 (21)	0.749 (0.037)	0.585	83 (118)
Unnormalized	0.632 (0.036)	0.696	40 (40)	0.727 (0.023)	0.600	80 (108)
***Refined* model**						
Normalized	0.623 (0.033)	0.661	45 (45)	0.738 (0.043)	0.585	86 (119)
Unnormalized	0.613 (0.055)	0.660	44 (44)	0.721 (0.039)	0.598	46 (46)

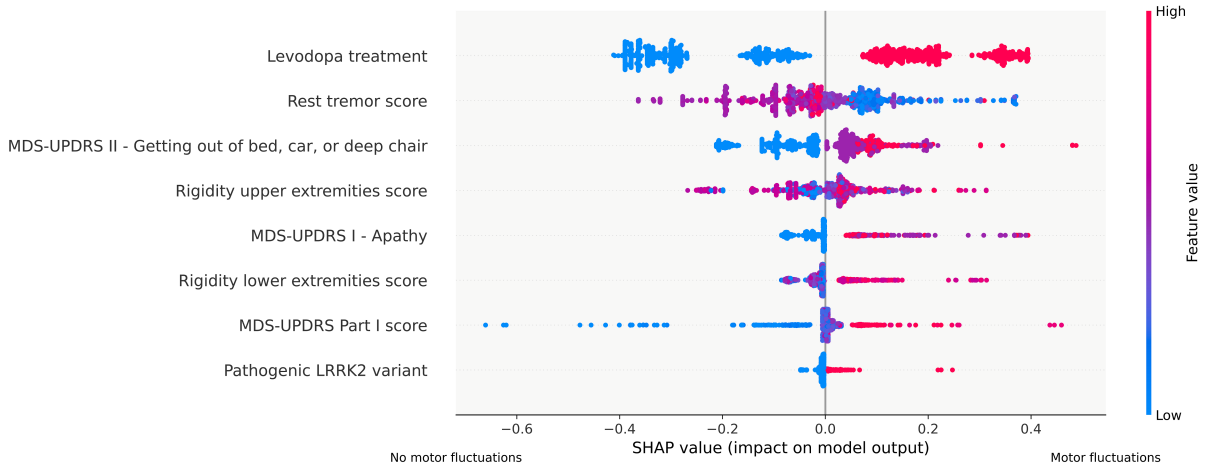
Assessment of the predictive performance of *comprehensive* and *refined* prognostic models for MF, including classification and time-to-MF analyses. The evaluation includes cross-validated and hold-out AUC or C-index calculations for both normalized and non-normalized models and a detailed examination of the number of features used in each model. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 5.11 Top 10 predictors for MF prognosis analysis.

Predictors	**Average ranks**
Dyskinesia	1.0
Levodopa treatment	1.5
Rest tremor	2.0
MDS-UPDRS Part I score	3.0
Tremor	3.5
MDS-UPDRS IV - Painful OFF-state dystonia	3.7
Disease duration	3.7
Rigidity upper extremities	4.0
MDS-UPDRS Part II score	4.0
MDS-UPDRS II - Freezing	4.0

List of predictors for motor fluctuations prognosis which received the top 10 average ranks in the cross-cohort analysis, using a ranking by permutation importance across the *optimized comprehensive* and *refined* models for motor fluctuations classification and time-to-MF analysis. The final rank reflects the average from non-missing ranks across the *optimized* models in the cross-cohort analysis.

Figure 5.3 SHAP values plot for the *optimized comprehensive* MF classification model in cross-cohort analysis.



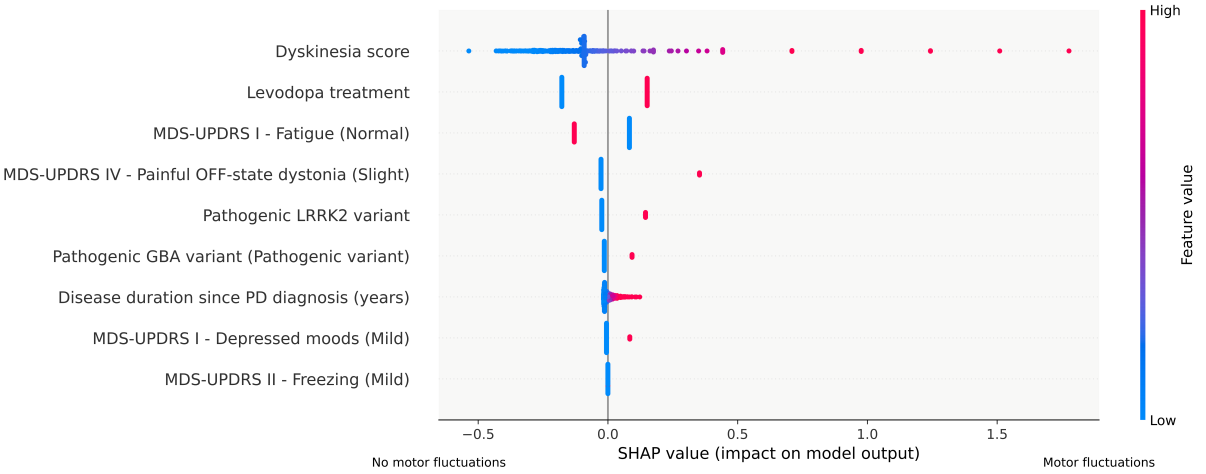
SHAP value plot displaying the top 15 predictors for the *optimized comprehensive* model in cross-cohort motor fluctuations prognostic classification. The plot shows the magnitude and direction (positive or negative) of each feature's influence on motor fluctuations prognosis status as output.

Table 5.12 Median conversion times and hazard ratios of *optimized comprehensive* time-to-MF model in cross-cohort analysis.

Predictors	**Hazard Ratio (95% CI)**	**Median Conversion (95% CI)**	**Log-rank (p-values)**
Dyskinesia			
≥1	1 (0.8, 2.37)	0.95 (0, 2.44)	1.44E-06
<1		4.07 (3.77, 4.36)	
Levodopa treatment			
Yes	1.14 (1, 1.56)	3.01 (2.27, 4.07)	0.446
No		4.19 (3.77, 4.51)	
MDS-UPDRS I - Fatigue (Normal)			
Yes	0.7 (0.51, 0.94)	4.44 (4.03, 4.76)	0.116
No		3.03 (2.34, 4)	
MDS-UPDRS IV - Painful OFF-state dystonia (Slight)			
Yes	5.06 (1.72, 16.23)	0 (0, 0)	9.87E-08
No		4.01 (3.44, 4.28)	
Pathogenic *LRRK2* variant			
Yes	2.02 (1.41, 2.92)	0.99 (0.11, 2.03)	2.15E-06
No		4.11 (3.84, 4.44)	
Pathogenic *GBA* variant (Pathogenic variant)			
Yes	3.78 (2.13, 6.26)	1.02 (0.28, 2.26)	5.45E-07
No		4.11 (3.79, 4.44)	
Disease duration			
≥2	1.1 (0.98, 1.49)	2.76 (2.03, 4.01)	0.011
<2		4.28 (4, 4.68)	
MDS-UPDRS I - Depressed moods (Mild)			
Yes	1.45 (1, 2.57)	3.01 (0, 5.02)	0.167
No		4.03 (3.31, 4.28)	

Summary of the hazard ratio (HR), median conversion time with 95% confidence interval (CI), and p-values from the log-rank test for the top 15 predictors identified in the time-to-MF model in the cross-cohort analysis. The HR provides insights into the risk associated with each predictor, while the median conversion time and log-rank test assess Kaplan-Meier (KM) curve differences between groups. "NR" (not reached) indicates that the MF event did not occur for some participants during the study period.

Figure 5.4 SHAP values plot for the *optimized comprehensive* time-to-MF model in cross-cohort analysis.



SHAP value plot displaying the top 15 predictors for the *optimized comprehensive* model in cross-cohort time-to-MF analysis. The plot shows the magnitude and direction (positive or negative) of each feature's influence on time-to-MF as output.

For the duration of the PD, the hazard ratio is 1.10 (95% CI: 0.98, 1.49), with a p-value of 0.011. Figure B.12 illustrates the KM plot, demonstrating that PD patients with a disease duration of 2 years or more are at a higher risk of MF. However, the overlapping of the KM curves in more extended follow-up visits resulted in no statistically significant difference in the HRs for these two groups of patients, despite a low p-value of the log-rank test and longer median conversion time for the patients with shorter disease duration during the baseline visits. Moreover, the disease duration significantly correlates with other features, including age at PD diagnosis, axial symptoms, bradykinesia, dyskinesia, the Hoehn & Yahr (H&Y) stage, and rigidity (Table 5.14), highlighting the complexity of PD progression.

The HR is significant for MDS-UPDRS IV - painful off-state dystonia (HR of 5.1), with a p-value from the log-rank test of 9.87E-08. The KM plot in Figure B.13 establishes a significant separation between the two curves, suggesting that patients with dystonia are at a heightened risk of experiencing MF compared to those without the condition. Genetic factors were also significant predictors, including pathogenic *LRRK2* variants (HR of 2.0, 95% CI: (1.4, 2.9)), and pathogenic *GBA* variants (HR of 3.8, 95% CI: (2.1, 6.3)).

The correlation analysis of predictors to MF outcome, as shown in Table 5.13, reveals several key insights consistent with previous findings from the SHAP value analysis and HR results. The correlation of the predictors, detailed in Table 5.14, highlights interactions of the predictors. The disease duration is significantly associated with levodopa medication, H&Y stage, axial symptoms, bradykinesia, dyskinesia, and rigidity in the lower extremities.

Furthermore, higher H&Y stages are associated with poorer visuospatial ability, increased severity of motor symptoms, and rigidity in patients. This dual association highlights the progressive nature of PD, whereby higher disease stages impact motor function and visuospatial abilities.

Levodopa equivalent daily dose (LEDD) analysis revealed significant differences between PD patients with and without MF over the 4-year study period, as shown in Table 5.15. PD patients without MF (MF-) had an average baseline LEDD of 587.2mg (SD 293.1), whereas patients with MF (MF+) had a

Table 5.13 Correlation between predictors and MF outcomes in cross-cohort analysis.

Predictors	**Correlation**	***p*-values**
Age at PD diagnosis	-0.18	1.54E-08
Disease duration since PD diagnosis (years)	0.30	6.69E-22
Gender	-0.03	5.98E-02
Levodopa treatment	0.35	1.09E-28
Weight (kg)	0.01	8.41E-01
Height (cm)	-0.06	4.81E-02
BMI (kg/m2)	0.05	1.25E-01
Hoehn & Yahr stage	0.24	4.55E-14
MOCA (adjusted for education)	-0.06	1.66E-01
Benton Judgment of Line Orientation	-0.12	4.28E-04
REM sleep behavior disorder (RBD)	0.12	1.54E-04
Resting tremor	-0.15	2.72E-06
Initial motor symptom - Rigidity or bradykinesia	0.04	1.68E-01
Axial symptoms	0.28	6.02E-19
Selective axial symptoms	0.21	8.44E-12
Freezing of gait	0.25	8.16E-16
Tremor	-0.11	5.29E-03
Rest tremor	-0.10	1.52E-02
Rest tremor amplitude	-0.06	1.65E-01
Rigidity upper extremities	0.06	1.40E-01
Rigidity lower extremities	0.11	6.41E-03
Total rigidity	0.09	1.88E-02
Bradykinesia	0.14	4.74E-04
Dyskinesia	0.22	3.38E-09
MDS-UPDRS Part I score	0.26	4.16E-16
MDS-UPDRS Part II score	0.29	1.10E-20
MDS-UPDRS Part III score (ON)	0.09	1.86E-02
SCOPA-AUT Gastrointestinal (GI)	0.18	1.71E-08
SCOPA-AUT Urinary	0.07	3.48E-02
SCOPA-AUT Cardiovascular	0.14	1.10E-05
SCOPA-AUT Thermoregulatory	0.20	1.60E-10
SCOPA-AUT Sexual dysfunction	0.04	3.16E-01
SCOPA-AUT Total score	0.17	1.12E-07
Family history of PD	-0.03	3.35E-01
Pathogenic *LRRK2* variant	0.09	9.11E-03
Pathogenic *GBA* variant	0.11	8.47E-04

The correlation of predictors with motor fluctuations outcome was measured using the point biserial correlation for continuous or ordinal predictors and the Matthews correlation coefficient (MCC) for the binary predictor.

Table 5.14 Correlation analysis results for MF predictors in cross-cohort analysis.

Predictors	Levodopa treatment	Age at diagnosis	Disease duration	Weight (kg)	Hoehn & Yahr stage	Axial symptoms	Bradykinesia	Dyskinesia	Tremor	Rigidity lower extremities
Age at diagnosis	0 (9.2E-01)									
Disease duration	0.47 (5.4E-55)	-0.22 (3.4E-12)								
Weight (kg)	-0.03 (4.3E-01)	-0.07 (1.9E-02)	-0.05 (1.3E-01)							
Hoehn & Yahr stage	0.34 (1.5E-27)	0.14 (2.0E-05)	0.35 (3.6E-29)	-0.06 (5.5E-02)						
Axial symptoms	0.42 (1.7E-44)	0.03 (2.8E-01)	0.48 (3.9E-57)	0.02 (5.9E-01)	0.5 (2.3E-64)					
Bradykinesia	0.18 (7.9E-06)	0.05 (2.0E-01)	0.26 (6.9E-11)	0.04 (3.2E-01)	0.47 (5.7E-36)	0.56 (2.5E-53)				
Dyskinesia	0.2 (4.4E-08)	-0.21 (1.1E-08)	0.4 (3.0E-28)	-0.04 (2.6E-01)	0.13 (6.9E-04)	0.25 (1.3E-11)	0.09 (3.1E-02)			
Tremor	0.02 (6.6E-01)	0.05 (1.7E-01)	-0.01 (7.7E-01)	0.07 (7.9E-02)	0.09 (2.6E-02)	0.06 (1.5E-01)	0.26 (2.5E-11)	-0.06 (1.3E-01)		
Rigidity lower extremities	0.11 (8.1E-03)	-0.04 (2.6E-01)	0.16 (5.7E-05)	0.12 (3.1E-03)	0.26 (2.5E-11)	0.31 (1.1E-15)	0.45 (1.8E-32)	0.07 (8.8E-02)	0.21 (1.4E-07)	
Benton JLO	-0.15 (5.5E-06)	-0.01 (7.3E-01)	-0.24 (6.5E-13)	0.09 (8.1E-03)	-0.15 (9.3E-06)	-0.24 (1.5E-12)	-0.22 (4.7E-07)	-0.05 (2.6E-01)	-0.11 (1.3E-02)	-0.12 (5.6E-03)

Correlation analysis of predictors was conducted using Spearman correlation for two continuous/ordinal variables, point biserial correlation for continuous/ordinal and binary variables, and Matthews correlation coefficient (MCC) for two binary variables. The correlation coefficients are presented with the p-values in brackets.

Table 5.15 Summary statistics for Levodopa Equivalent Daily Dose (LEDD) among patients with PD with and without motor fluctuations within 4-year follow-up.

Statistics	MF-	MF+	Significance test (p-values)	LEDD <400mg	LEDD ≥400mg	Log-rank test (p-values)
n	69	246	4.54E-04	58	257	3.24E-03
Mean (SD)	587.2 (293.1)	780.7 (411.89)		2.3 (1.96)	1.3 (1.85)	

The statistical significance of these differences is indicated by p-values from t-tests (for normally distributed data) or Mann-Whitney U-tests (for non-normally distributed data). Additionally, the table presents time-to-MF statistics for PD patients with LEDD <400mg and LEDD ≥400mg, along with *p*-values from log-rank tests comparing these two groups.

significantly higher mean baseline LEDD of $780.7 mg$ (SD 411.89). A statistical comparison between these groups yielded a *p*-value indicating a statistically significant difference.

For time-to-MF based on baseline LEDD levels, PD patients with LEDD $<400 mg$ had a mean time-to-MF of 2.3 years (SD 1.96). In contrast, patients with LEDD $\geq 400 mg$ had a shorter average time-to-MF of 1.3 years (SD 1.85). The log-rank test for these groups yielded a *p*-value of 3.24E-03, indicating a significant difference in time-to-event outcomes based on LEDD levels.

5.4.6 Assessment of clinical utility and calibration

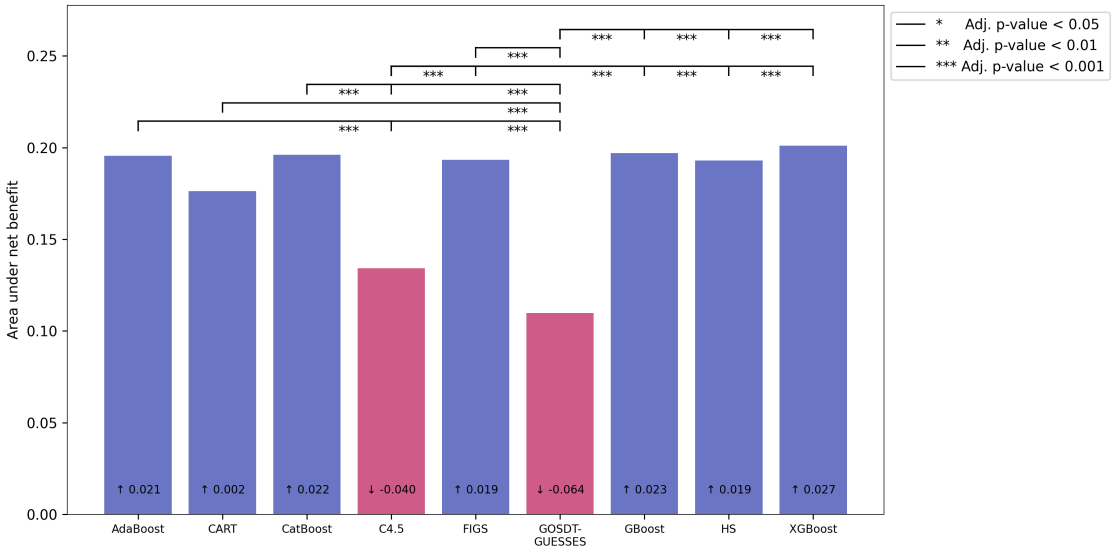
Table 5.16 Calibration analysis for MF classification and time-to-MF analyses.

MF classification					Time-to-MF				
Algorithm	*Comprehensive*		*Refined*		Algorithm	*Comprehensive*		*Refined*	
	Slope	MSE	Slope	MSE		Slope	MSE	Slope	MSE
AdaBoost	1.97	0.23	1.13	0.23	CW-GBoost	0.67	0.04	0.66	0.07
CART	0.40	0.26	0.53	0.25	Extra Survival	0.62	0.04	0.63	0.04
CatBoost	0.70	0.23	0.76	0.24	Survival GBoost	0.40	0.05	0.38	0.06
C4.5	0.25	0.38	0.17	0.41	LSVM	0.82	0.04	0.64	0.05
FIGS	0.57	0.24	0.62	0.24	NLSVM	0.64	0.05	0.68	0.05
GOSDT-GUESSES	0.17	0.42	0.04	0.48	Penalized Cox	5979.06	0.06	5979.06	0.06
GBoost	1.09	0.23	2.96	0.24	Survival RF	0.77	0.04	0.61	0.08
HS	0.63	0.24	0.62	0.24	Survival Trees	0.20	0.14	0.19	0.11
XGBoost	0.70	0.23	0.65	0.24					

Calibration analysis for *comprehensive* and *refined* models in both motor fluctuations classification and time-to-MF analysis for cross-cohort analysis with the calibration slope and mean square error (MSE).

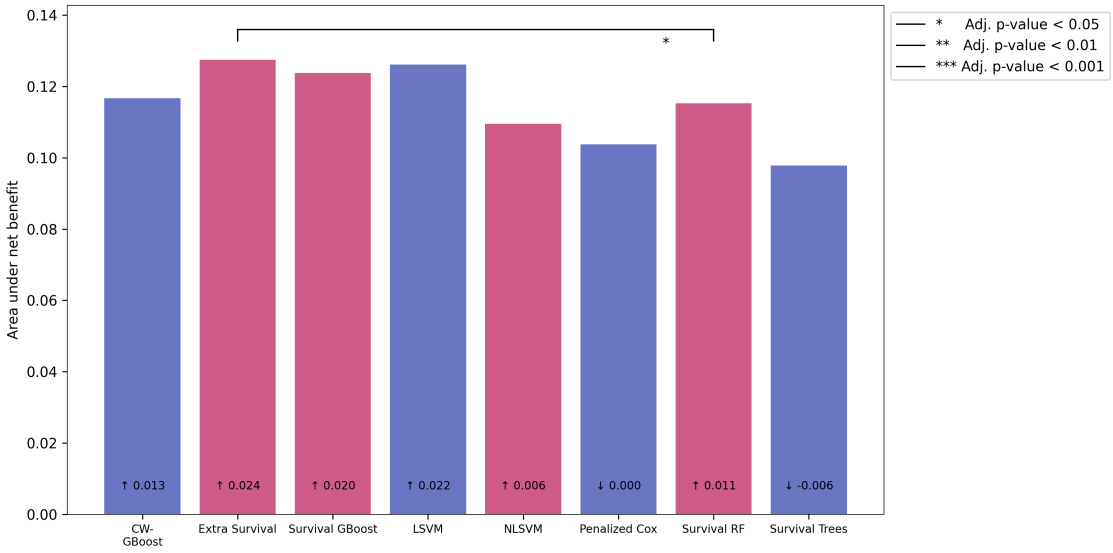
Figure 5.5 illustrates that the *optimized comprehensive* MF classification model trained by XGBoost shows the largest AUNBC, followed by GBoost and CatBoost. The CatBoost-*optimized* model also showed the highest cross-validated AUC and high hold-out AUC of 0.62 in the cross-cohort study. At the same time, the XGBoost-*optimized* model showed the highest hold-out AUC of 0.65. These models consistently show larger net benefit areas than the "treat all" strategy, which assumes that all patients receive the intervention regardless of their predicted risk. Among the top three models based on the

Figure 5.5 Bar plot of the area under the positive net benefit curve for the *comprehensive* MF classification models in cross-cohort analysis.



The bar plot shows the area under the positive net benefit for different cross-cohort *comprehensive* motor fluctuations classification models, with the lines above the bars indicating significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

Figure 5.6 Bar plot of the area under the positive net benefit curve for the *comprehensive* time-to-MF models in cross-cohort analysis.



The bar plot shows the area under the positive net benefit for different cross-cohort *comprehensive* time-to-MF models, with the lines above the bars indicating significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

AUNBC, GBoost shows superior calibration compared to XGBoost and CatBoost, as evidenced by Table 5.16 (left side).

For the *optimized refined* MF classification model, the AUNBC for AdaBoost has the largest net benefit area and highest hold-out AUC of 0.64, as illustrated in Figures B.14. AdaBoost also shows superior calibration slopes of 1.13. A comparison with the "treat all" strategy in this context further highlights the advantage of using predictive models to tailor interventions, with the potential to focus resources on patients who are most likely to benefit.

In the *comprehensive* time-to-MF model, the *optimized* models from Extra Survival Tree and LSVM achieved the highest hold-out C-index in the cross-cohort analysis, with values of 0.68 and 0.69, respectively. These models also show an excellent area under the net benefit curve in Figure 5.6. A comparable pattern is observed in the *refined* time-to-MF model, as illustrated in Figure B.15

Table 5.16 (right side) indicates that the *optimized comprehensive* time-to-MF model from LSVM is better calibrated than the Extra Survival Tree model, with a slope of 0.82, followed by the Extra Survival Tree model, with a slope of 0.62. For the *refined* time-to-MF model, NLSVM and CW-GBoost-*optimized* models are well-calibrated with a calibration slope of 0.68 and 0.66, respectively. A more accurate calibration analysis, in which the predicted probabilities more closely align with the actual outcomes, improves the reliability of the model's predictions.

5.5 Discussion

The prognosis of MF in PD presents a significant challenge due to intra-individual variability and the multifaceted factors influencing disease progression. Our study examines MF in PD through single and multi-cohort analyses. It addresses these complexities by testing the robustness and accuracy of interpretable MF prediction methods in a cross-cohort setting. We compared different ML and feature selection approaches within a two-level nested CV framework. Instead of focusing solely on single markers and one conceptual approach, we assessed multivariable ML classification and time-to-MF models. Additionally, by computing feature selection statistics across CV cycles and different cohorts, we ranked the relative importance of features to identify the most robust and generally applicable predictors. In the following, we discuss both the potential utility of the ML models as a prognostic tool and interpret the identified predictors in the context of the prior literature.

5.5.1 Comparative evaluation of predictive models

In previous studies, predictive modeling of MF in PD has achieved varying degrees of success. Most studies focused on assessing the influence of specific individual factors, e.g., a community-based study found that MF occurred in 29% of the overall PD sample and in 40% of those who had received levodopa, with disease duration and levodopa dose being significant predictors [181]. Another study confirmed longer disease duration as a significant predictor of MF and identified younger age at diagnosis as a further risk factor [182]. However, while the previous studies identified relevant statistical associations, they often focused on datasets from specific patient cohorts with limited sample sizes, which may reduce the statistical power and generalizability of some of the findings across diverse populations. The present study's comparative evaluation of different predictive models addresses these limitations by incorporating multivariable ML approaches and cross-cohort analyses to improve robustness and

reliability. Notably, using nested CV frameworks and including multiple cohorts (LuxPARK, PPMI, and ICEBERG) have facilitated a comprehensive assessment of the models' performance across different patient populations.

We applied various modeling approaches to predict MF, each with advantages and limitations. Tree-based algorithms such as AdaBoost, CatBoost, and XGBoost were used extensively due to their ability to handle complex interactions between variables and their robustness to overfitting through regularization techniques. GBoost showed a high AUNBC and better calibration, with calibration slopes closer to 1. Furthermore, support vector machines (LSVM and NLSVM) perform well in time-to-event analyses. These models are mainly advantageous in their ability to handle high-dimensional data and their flexibility in incorporating censoring information, which is important for accurate time-to-event analysis. Thus, future follow-up studies should ideally consider multiple approaches to avoid performance bottlenecks and identify the most suitable trade-off between the strengths and weaknesses of different modeling techniques.

The study's findings emphasize the benefits of integrating data from multiple cohorts, improving the robustness and generalizability. Furthermore, the higher predictive performance observed in the PPMI cohort for both MF classification and time-to-MF analysis highlights the important role of larger sample sizes and extended follow-up periods in developing reliable models. Despite introducing additional complexities due to our longitudinal cross-study analysis compared to prior cross-sectional, single-cohort analyses, the predictive performance for MF prognosis in this study is comparable to that reported in earlier research on MF in PD. For instance, earlier studies reported an AUC of 0.68 for dyskinesia prognosis [143], an accuracy of 72% for cross-sectional MF detection [181], and AUC values ranging from 60% to 82% for cross-sectional differentiation between ON and OFF medication states [183].

When evaluating the predictive models for MF classification and time-to-MF analysis in terms of their calibration and potential utility for clinical decision-making using a DCA, most models consistently show superior net benefit compared to simple strategies, such as "treat all" and "treat none", and the calibration analysis showed that the slope of predicted outcome categories matched well with the observed outcomes.

Despite these encouraging results, significant limitations and challenges still have to be addressed for the practical application of these predictive models in clinical settings. The hold-out performance metrics, although indicative of high predictive capability, still show variability across different cohorts, suggesting that individual patient characteristics and cohort-specific factors significantly influence model accuracy. Additionally, the reliance on features such as LEDD, disease duration, and non-motor symptoms underscores the complexity of predicting MF onset, as these factors can vary widely among patients [184]. Therefore, while the models could provide valuable insights for testing new therapeutic interventions in precision medicine trials, they will need to be further optimized and tailored to individual patient profiles to achieve the long-term goal of optimizing treatment strategies and improving clinical outcomes.

5.5.2 Differences between clinical features across cohorts

When interpreting the role of baseline clinical features as potential prognostic predictors of future MF development, we need to consider the marked variations in baseline characteristics between the three cohorts: LuxPARK, PPMI, and ICEBERG. The LuxPARK cohort displays a relatively long disease duration compared to other cohorts, indicative of more advanced disease stages and, consequently, a

potentially more significant influence of current clinical characteristics on MF development. In contrast, the ICEBERG cohort covers a population with a lower average body weight and BMI, which may influence motor symptom progression due to differences in lifestyle, physical activity, and dietary habits. Despite a similar disease duration to that observed in the ICEBERG cohort, the PPMI cohort showed significantly lower scores across MDS-UPDRS Parts I to III and the SCOPA-AUT. This indicates that factors other than disease duration may be implicated in the variations in symptom severity and non-motor symptoms. Therefore, to account for these significant differences between the cohorts, our interpretation of models and key predictive features for MF prognosis focuses on the results from the cross-study modeling analyses.

Overall, these findings emphasize the significance of multi-cohort analyses in capturing the heterogeneity of PD populations. Incorporating data from different cohorts enables the development of more robust and generalizable predictive models, ensuring they perform well across clinical settings and patient demographics.

5.5.3 Comparative evaluation of cross-study integration

This study aimed to assess the impact of cross-study normalization on the predictive performance of models for MF in PD. In multi-cohort analyses, cross-study normalization addresses the potential variability in data from different studies. The assessment of normalized versus unnormalized models revealed that, in general, cross-study normalization did not result in statistically significant improvements in hold-out performance for either the *comprehensive* or *refined* models. This consistency was observed in the leave-ICEBERG-out, leave-PPMI-out, and leave-LuxPARK-out analyses.

Implementing cross-study normalization methods in the cross-cohort study led to improvements in the hold-out C-index for the *optimized comprehensive* time-to-MF model. This indicates that normalizing study-specific biases and variations improves the model's capacity for generalization across patient cohorts. By reducing the influence of cohort-specific variations, normalization facilitated a better prediction capability of MF outcomes, particularly in the context of time-to-MF analysis.

The findings of this study emphasize the necessity of considering cross-study normalization to address inter-study variability and improve the reliability of predictive models in multi-cohort analyses. While normalization can occasionally improve model performance, its overall impact may be limited. Consequently, researchers must assess the necessity and efficacy of normalization on a case-by-case basis, considering the specific characteristics and variability of the data.

5.5.4 Interpretation of models and predictive features

The association between levodopa medication and MF has been well-documented in previous studies [134, 138], particularly in the advanced stages of the disease [34]. However, a recent trial provided conflicting results [185]. Despite the limitations imposed by the incomplete and inconsistent LEDD data across the two cohorts, the statistical analysis within the LuxPARK cohort yielded valuable insights. Our univariate analysis showed a significant association between levodopa use and MF, with higher LEDD ($\geq 400mg$) correlating with increased MF risk [139, 141]. However, considering multiple factors, our multivariate ML models did not identify a statistically significant difference in the risk of MF on baseline levodopa use. Furthermore, the KM curves did not reveal a significant impact. These results emphasize the value of multivariate models for capturing complex interactions that univariate analyses might miss.

Significant correlations between levodopa intake at baseline, disease duration, and H&Y stage of 0.47 and 0.34 (p = 5.4E-55 and 1.5E-27), respectively [34], suggest that levodopa may be indirectly associated with disease progression. It should be noted that the variable “levodopa" refers only to levodopa intake at baseline, not to LEDD. Moreover, the *comprehensive* and *refined* models showed comparable hold-out performance in cross-cohort analyses, with no statistically significant differences. This suggests that, despite the apparent influence of levodopa in univariate analyses, its predictive value may be overshadowed by other strongly correlated predictors. The results show that predictors associated with levodopa, such as disease duration, can maintain predictive capability when levodopa is excluded, indicating that the *refined* model performs well.

Patients with more extended disease duration and higher H&Y stages are more likely to experience MF, as previously documented in the literature [137]. This analysis verifies the notion that a longer disease duration [139, 141], which reflects the progressive nature of PD [3, 35], is significantly associated with a higher risk of MF [26]. This indicates that patients with more advanced PD at baseline tend to encounter instability in their motor functions [186]. As PD progresses, the therapeutic window of levodopa narrows [134], rendering it increasingly challenging to maintain stable motor control. Similarly, earlier age at PD diagnosis was associated with longer disease duration (Spearman correlation of -0.22, p = 3.4E-12), which is associated with a higher risk of MF (point-biserial correlation of 0.3, p = 6.7E-22). This is because these patients often experience a slower disease progression; therefore, a longer disease duration results in a higher risk of developing MF [1, 139].

Considering scores from commonly used PD symptom severity rating scales, higher baseline MDS-UPDRS Part II scores, which indicate greater severity of motor impairments and the presence of rigidity and freezing behaviors, also significantly predict the development of MF [143]. This aligns with prior expectations, as FoG is related to the OFF phase in most cases of PD, and ON/OFF phases define MF. In general, impairments in movement speed and gait instability, as well as axial symptoms, including postural instability, gait impairments, and rigidity, usually confined to the later stages of PD [17, 166], have already been associated with MF [17, 19, 166].

As PD progresses, patients frequently report an exacerbation of axial symptoms, including postural instability and gait impairment. This symptom can lead to an increased risk of falls and difficulties in performing activities of daily living [187]. Axial symptoms positively correlate with MF and are more prevalent in PD patients with *GBA* mutations [5, 48]. Furthermore, these symptoms correlate positively with several motor impairments, including bradykinesia, dyskinesia, and rigidity, heightening MF.

Another hallmark of PD is rigidity, which affects axial and limb muscles, leading to difficulties initiating and executing movements smoothly. There is a significant positive correlation between rigidity and tremor and documented evidence of a correlation between resting tremor and rigidity [19]. As rigidity progresses, it can lead to periods of increased stiffness and decreased mobility and is associated with an increased risk of MF onset.

Additionally, tremor and rigidity are significantly correlated with the stage of PD, as documented in Prashanth and Roy (2018) [17]. Notably, tremor, a symptom frequently observed in the initial stages of PD, negatively correlates with MF. This inverse relationship indicates that patients who experience pronounced tremors may be less likely to develop MF [139], which may be attributed to the milder disease progression in tremor-dominant patients [188].

Non-motor symptoms further complicate and exacerbate the burden of MF in PD. Higher scores on MDS-UPDRS Part I, which assesses non-motor symptoms such as fatigue, correlate positively with

an increased risk of MF. In addition, non-motor symptoms are less responsive to levodopa, and the MDS-UPDRS Part I assessment of non-motor aspects of daily living may be a more reliable measure of disease severity in treated patients [143]. Fatigue is a common nonmotor symptom in PD, affecting up to 73% of patients with MF. This symptom displays a state of physical and mental exhaustion that is disproportionate to the activity level, which reduces motivation and productivity. Consequently, patients may experience a decreased engagement in hobbies and activities. Such fluctuations in motor function negatively impact the ability to engage in these activities, affecting overall well-being and quality of life.

In addition, dystonia significantly contributes to the burden of MF in PD, a movement disorder often associated with LID, as documented in previous literature [140]. These disturbances frequently worsen during OFF, negatively affecting mobility, sleep quality, and overall comfort. Consequently, rapid eye movement sleep behavior disorder (RBD) and nighttime sleep problems positively correlate with MF. *GBA*-PD patients are at an increased risk of experiencing RBD and show a more significant burden of nonmotor symptoms, particularly sleep disturbances [47, 48]. At baseline, 70% of PD patients with MF experienced nighttime sleep problems. Previous literature also indicates that *LRRK2* mutations are associated with sleep disturbances [44]. Furthermore, these sleep disturbances not only exacerbate motor symptoms but also contribute to fluctuations in disease severity.

The progression of overall autonomic dysfunction, particularly in the urinary and gastrointestinal (GI) tracts, has positively correlated with MF. In the context of MF in PD, autonomic dysfunction, particularly GI issues, can impact the absorption of medications. Furthermore, levodopa transport and absorption barriers contribute to developing motor complications [30, 145]. This dysfunction is frequently documented before the onset of motor symptoms [144], resulting in various complications, including oral issues such as drooling and swallowing difficulties. They are positively correlated with MF. Drooling, often caused by swallowing difficulties, has negative effects, including difficulty eating [33].

Furthermore, pathogenic *GBA* mutations are associated with a rapid disease progression [5], and an increased risk of developing MF [47, 186], emphasizing the role of *GBA* mutations in heightening MF in PD. Conversely, while pathogenic *LRRK2* mutations are also associated with MF [189], the effect size is lower (HRs 2.0 and 3.8 for *LRRK2* and *GBA*). In addition, both *GBA* and *LRRK2* mutations have been linked to dyskinesia [190], a common PD complication linked to MF, highlighting the multifaceted impact of these genetic variants on the disease trajectory.

The Benton JLO test assesses visuospatial ability. The results of this test indicate that visuospatial ability is negatively associated with the likelihood of experiencing MF in PD. This finding suggests impaired visuospatial skills may predict motor decline [56]. This association is further supported by correlations with the H&Y stage and disease duration [45]. The study highlights the interrelationship between cognitive and motor domains in the progression of PD, with visuospatial function showing a significant correlation with motor symptoms such as bradykinesia and rigidity [2]. Although the Benton JLO test requires minimal motor skill [63], deficits often reflect cognitive impairments that impact motor functions. The Palermo et al. (2020) study observed a correlation between late-onset PD and lower cognitive performance [172]. However, this association was not significant in our cross-cohort analysis. In the context of genetic influences, individuals with *GBA* mutations show a notable progression of motor symptoms and rapid cognitive decline [5, 186], which correlates with pronounced impairments in visuospatial abilities [47]. Conversely, *LRRK2* mutations are associated with a lower risk of cognitive

impairment in the literature [44], although they may still influence visuospatial impairments related to MF. These findings highlight the complex relationship between genetics, cognitive function, and motor symptoms in PD.

Further investigation is required to clarify the relationship between body weight, BMI, and MF in PD. Although our study defined body weight and BMI as predictors, no statistically significant associations were found between these variables and the prevalence of MF. This finding suggests that, while there might be some associations between body weight or BMI and MF, our cohorts did not provide strong evidence to support these connections [143]. Further research with more diverse cohorts may be needed to clarify these potential associations.

5.5.5 Clinical utility and calibration

The comparison with the "treat all" and "treat none" strategies is a benchmark for evaluating the added value of these predictive models. The *optimized* models' consistently larger net benefit areas over these baseline strategies emphasize their advantageous decision-making potential. These models ensure interventions are more effectively targeted, potentially reducing unnecessary treatments and associated costs.

The *optimized* predictive models for MF classification and time-to-MF analysis show considerable clinical utility and reliable calibration, yielding valuable tools for personalized patient management. By outperforming the "treat all" strategy, these models offer a more pronounced approach to treatment decisions, enhancing the precision and effectiveness of interventions for patients at risk of MF. These findings highlight these models' potential to improve clinical outcomes and optimize the management of PD.

The predictive models developed in this study offer valuable applications for improving clinical trial protocols and enrollment strategies for MF-focused research in PD. First, these models can support risk stratification by identifying patients at higher risk of developing MF, allowing for the selection of enriched study populations for precision medicine trials. This approach may reduce sample size and trial duration, improving efficiency and cost-effectiveness. Second, the identified predictors can inform trial inclusion and exclusion criteria. For example, interventions aimed at preventing or delaying MF could focus on patients with characteristics associated with a higher risk of MF, such as those with PD-associated *GBA* and *LRRK2* variants, as highlighted by our models. Finally, these models may facilitate personalized follow-up schedules in clinical trials, with higher-risk participants undergoing more frequent monitoring to ensure early detection of MF onset.

5.6 Summary and conclusions

This study aimed to identify robust predictors of MF in PD by leveraging ML techniques and cross-cohort data integration. By analyzing data from three distinct longitudinal cohorts (LuxPARK, PPMI, and ICEBERG), the study showed the potential of ML models to predict MF with significant predictive performance. The models identified multiple clinical features such as disease duration, LEDD, and various motor and non-motor symptoms as significant predictors of MF.

As a key study finding, cross-cohort data integration increased the stability and generalizability of the predictive models, reducing cohort-specific biases and improving overall robustness. The decision

curve and calibration analyses confirmed the potential of these models for practical clinical decision-making, showing that the models provide a greater net benefit compared to simple "treat all" and "treat none" strategies. The PPMI cohort, with its larger sample size, showed superior predictive performance, emphasizing the importance of sample size and long follow-up durations in prognostic ML model development. However, the models with the highest stability and generalizability were obtained from the cross-cohort integrative analysis.

In the cross-cohort analyses, baseline LEDD showed its association with MF, likely due to the joint association of these variables with disease duration and progression. Additional significant predictors include longer disease duration, advanced H&Y stages, MDS-UPDRS Part I scores, and the presence of dyskinesia, FoG, axial symptoms, and rigidity. Conversely, tremors were inversely correlated with MF, likely due to their slower progression. GI tract dysfunction may impact the transport and absorption of levodopa, resulting in exacerbating MF.

Despite the promising results in prognostic model building using these features, several limitations still need to be addressed for future clinical translation. The variability in hold-out performance metrics across different cohorts indicates that individual patient characteristics and cohort-specific factors significantly influence model accuracy. The reliance on features such as baseline LEDD, disease duration, and non-motor symptoms underscores the multifacetedness and complexity of predicting MF onset, as the individual factors can vary widely among patients. Future research should further optimize these models across data from more diverse cohorts and tailor them to individual patient profiles to improve treatment strategies and clinical outcomes for PD patients.

Overall, this study highlights the potential of interpretable ML models in predicting MF in PD, emphasizing the benefits of cross-cohort data integration for improving model stability and generalizability. These findings may help lay the ground for designing precision medicine trials and developing personalizd management strategies for MF in PD patients, aiming to improve their quality of life and clinical outcomes.

5.7 Contribution statement

This study was a collaborative effort, with contributions from all authors in various aspects of the research and manuscript preparation.

Rebecca Loo Ting Jiin: Conducted the study as the first author, developed the methodology, conducted analysis, created visualizations, validated results, and drafted the original manuscript.

Enrico Glaab: He supervised the study as the corresponding author for the submitted manuscript, contributed to the study's correction, methodology, and investigation, guided the project, acquired funding, and reviewed and edited all sections of the manuscript.

Lukas Pavelka, Graziella Mangone, Fouad Khoury, Marie Vidailhet, Jean-Christophe Corvol, and Rejko Krüger: Contributed by reviewing and editing the manuscript, providing insights from clinical perspectives.

Chapter 6

Multi-cohort machine learning identifies predictors of cognitive impairment in Parkinson's disease

Cognitive impairment represents one of the non-motor symptoms of PD. It ranges from mild cognitive impairment (*PD-MCI*) to dementia and progressively affects various cognitive domains, including memory, attention, executive functions, and visuospatial abilities [62, 191]. These impairments can manifest early in the onset of PD [2, 192].

Among individuals diagnosed with PD for the first time, cognitive dysfunction is observed in between 20 to 50% of cases [4, 13, 57, 60], indicating that this is a significant clinical concern. As the disease progresses, the prevalence of cognitive decline increases, with 41.5% of PD patients experiencing cognitive decline within three years [60]. Approximately 5-10% of patients with *PD-MCI* transition to dementia annually [7]. The identification of predictive markers for cognitive decline can improve clinical management, as early intervention may prove effective in slowing the progression of cognitive dysfunction.

The assessment of objective cognitive impairment in PD typically uses standardized tests, such as the Montreal Cognitive Assessment (MoCA). A score below 26 on the MoCA typically indicates *PD-MCI* in individuals with PD. It is to assume that objective measures solely determine cognitive impairment in PD. Patient-reported cognitive decline (*PRCD*), also called subjective cognitive impairment, is measured based on patients' perception of their cognitive decline and may only sometimes align with objective assessments. Some patients report cognitive difficulties despite normal objective performance, while others show deficits on neuropsychological tests without subjective complaints. These two measures, though related [193], are only sometimes aligned. Some patients report cognitive difficulties despite performing at expected levels on objective cognitive assessments. Conversely, some individuals may not perceive cognitive decline despite neuropsychological test deficits.

Despite the expansion of research on cognitive impairment in PD, challenges remain. Previous studies investigating cognitive decline in PD have concentrated on identifying cohort-specific predictors within limited sample sizes [61, 191]. Although these studies have provided valuable insights, they often need more generalizability due to the potential for cohort bias. The MoCA has been used extensively for monitoring cognitive decline in patients with PD, with a score of less than 26 indicative of *PD-MCI*

[4, 57, 60]. Nevertheless, the relationship between objective cognitive measures and patient-reported outcomes still needs to be better understood. A more comprehensive understanding of cognitive decline in PD could be achieved by investigating objective and subjective measures of cognitive function.

Early identification of cognitive decline may facilitate the implementation of timely interventions that could decelerate the progression of cognitive decline [57]. These may include cognitive training, pharmacological adjustments, or lifestyle modifications, which may confer benefits in slowing cognitive decline [14]. In addition to its clinical implications, cognitive impairment in PD also has significant economic consequences. There is a clear association between cognitive decline and higher healthcare costs, driven by the increased demand for medical care and support [194].

Despite the growing body of research on cognitive impairment in PD, several challenges remain. Multiple studies have concentrated on cohort-specific predictors utilizing smaller sample sizes, which constrains the generalizability of their findings [5, 61, 191]. Furthermore, the relationship between clinical cognitive measures and patient-reported cognitive outcomes needs to be better understood, necessitating further investigation.

The application of ML techniques offers considerable promise for developing predictive models of cognitive impairment due to the fact that they can handle complex clinical data and reveal patterns that may otherwise be overlooked by traditional statistical methods [7, 116]. Nevertheless, the application of ML in cognitive impairment research has predominantly been confined to single-cohort studies, thereby hindering the development of more comprehensive and generalizable models across diverse patient populations.

Previous studies on cognitive impairment in PD have applied various data types and methodologies to examine the progression and risk factors associated with cognitive decline. Several studies have concentrated on the relationship between motor dysfunction and cognitive performance. For example, small-sample cohort studies using linear regression and Spearman's rho correlation have repeatedly shown significant correlations between motor deficits, such as balance and gait impairment, and cognitive decline, suggesting a close connection between these two domains in PD progression [2, 43, 61]. Furthermore, a longitudinal study involving PD patients and healthy controls revealed that progressive cortical thinning, particularly in areas associated with visuospatial functions, significantly contributes to cognitive decline over time [191].

Moreover, prior research has indicated that baseline cognitive function can predict the progression of motor disability in PD, thereby emphasizing the dynamic interaction between cognitive impairment and motor symptoms as the disease advances [56]. Larger cohort studies have used more robust statistical techniques to identify predictors of cognitive impairment. For instance, in a cohort of 294 PD patients from the PPMI study, baseline cognitive assessments and neuropsychiatric measures were shown to forecast the probability of cognitive decline over time [60]. Similarly, other studies utilizing MoCA scores from the PPMI cohort have reinforced the importance of early cognitive assessments by examining their correlation with cognitive performance [4, 195].

In recent years, ML approaches have emerged as a tool for analyzing complex interactions between clinical, genetic, and imaging data. To illustrate, Almgren et al. (2023) developed and validated multimodal ML models to predict continuous cognitive decline in PD patients, utilizing a range of data sources. The models showed improved accuracy in discerning patterns of cognitive decline in comparison to conventional methodologies, highlighting the potential of ML techniques to facilitate the early detection and intervention strategies for cognitive impairment in PD [14].

Cognitive impairment with various clinical features that contribute to the progression of dysfunction, including motor disability features, is closely linked to worsening cognitive outcomes [56]. Furthermore, studies have indicated that PD patients with more significant motor impairment, including postural instability, are at an increased risk of cognitive decline [196]. Furthermore, non-motor symptoms, such as apathy, have been shown to correlate with the emergence and advancement of cognitive impairment in patients with PD [46, 196].

Several risk factors have been identified as contributing to an increased risk of cognitive decline in patients with PD, particularly in those with more advanced disease stages. As the disease progresses, prolonged duration is also associated with an exacerbation of cognitive decline due to the accumulation of neurodegeneration [196]. Elder age is a well-established predictor of cognitive decline in patients with PD. Compared to younger patients with PD, older patients show a faster rate of cognitive decline [60]. Similarly, the severity of parkinsonism is associated with an increased risk of cognitive dysfunction [13], particularly in patients with advanced disease [45]. Additionally, gender differences have been observed, with women typically demonstrating superior performance in global cognition, memory, and language domains, while men tend to excel in visuospatial tasks [3].

Autonomic dysfunction, including gastrointestinal disturbances, has been identified as a predictor of cognitive decline in early PD [197]. Gastrointestinal symptoms, including constipation, have been shown to correlate with diminished cognitive performance and an accelerated progression to dementia [145], indicating a potential association between gut dysfunction and cognitive decline in PD.

This study aimed to address existing knowledge gaps by developing and validating ML models that used clinical data from multiple independent cohorts to predict both *PD-MCI* and *PRCD* in PD patients. By utilizing a cross-cohort approach, including leave-one-cohort-out validation, the study guaranteed that the predictors of cognitive impairment were applicable across diverse PD populations. The findings from this research can improve early detection strategies, facilitate a deeper comprehension of the factors that contribute to cognitive decline, and ultimately inform the development of more personalized interventions for PD patients at risk of cognitive impairment.

6.1 Rationale for the study

Cognitive impairment represents a prevalent non-motor symptom in PD. Despite the ongoing research in this field, there are still gaps in our understanding of the predictors of cognitive impairment and the progression of CI in PD. Many studies have focused on particular cohorts and used smaller sample sizes, challenging the generalization of their findings to different PD populations. Furthermore, the relationship between clinical cognitive assessments and patient-reported cognitive outcomes needs to be better understood, which challenges accurate diagnosis and management.

ML can help address these issues by analyzing complex clinical data and identifying patterns that traditional methods may overlook. However, most ML studies on cognitive impairment in PD have been limited to a single cohort, which limits the generalizability of their findings.

This study aims to address these challenges by utilizing ML models to predict both *PD-MCI* and *PRCD* in patients with PD across multiple independent cohorts. The objective of using a cross-cohort approach, including leave-one-cohort-out validation, is to develop more generalizable and reliable models. Generally, this research has the potential to improve the early identification of cognitive decline and facilitate the development of more tailored treatments for PD patients at risk of mild cognitive

impairment.

6.2 Objective of the current study

The main goal of this study is to develop and validate ML models that can predict *PD-MCI* and *PRCD* in patients with PD. The study aims to develop generalizable predictive models across diverse PD populations by incorporating clinical data from multiple independent cohorts. By utilizing a cross-cohort methodology, the study aims to improve early detection strategies for cognitive decline and to advance our comprehension of the factors associated with cognitive impairment. The specific objectives of this chapter are outlined as follows:

1. Integrate cross-cohort analysis and cross-study normalization:
Applying cross-study normalization techniques will ensure the analysis is conducted to account for the variation in measurement and findings regarding cognitive impairment across diverse PD populations.

2. Develop predictive models utilizing ML techniques:
The study aims to identify complex patterns and relationships within large clinical datasets from multiple cohorts that traditional statistical methods may overlook. The predictive models will be evaluated for their performance to ensure they provide reliable predictions for clinical use.

3. Identify key factors associated with cognitive impairment in PD patients:
The study will analyze the data to identify significant predictors, such as age, disease duration, and motor & non-motor symptoms. An understanding of these risk factors will not only improve early detection strategies but will also inform the development of target interventions tailored to the individual.

By focusing on these key areas, the study aims to contribute substantially to advancing our knowledge and enhancing our ability to manage cognitive impairment associated with PD. Ultimately, this should improve the quality of care provided for those affected.

6.3 Research methodology

This study uses research methodologies to analyze *PD-MCI* and *PRCD* in patients with PD, utilizing data from multiple independent cohorts. Section 3.1 outlines the information for the selected PD cohorts. The baseline variables were selected based on consistent availability across all three cohorts, ensuring comparability for cross-cohort analyses. The variables contain demographic data, assessment of disease severity, clinical evaluations of both motor and non-motor symptoms, and pathogenic gene mutations linked to PD. This approach enables building predictive models for both *PD-MCI* and *PRCD*, thereby enhancing the generalizability of the findings across different PD populations.

6.3.1 Inclusion criteria

This study used data from three independent PD cohorts. The LuxPARK, PPMI, and ICEBERG studies. The participants were selected based on two inclusion criteria:

(1) All participants were required to meet the diagnostic criteria for PD. The LuxPARK and ICEBERG cohorts were assessed according to the UK Parkinson's Disease Society Brain Bank (UKPDSBB) criteria [154]. However, the PPMI cohort was required to show at least two of the following characteristics: resting tremor, bradykinesia, or rigidity, with either resting tremor or bradykinesia being one of the symptoms; or a single asymmetric resting tremor or asymmetric bradykinesia [103].
(2) Participants were required to either have a clinically confirmed presence of *PD-MCI* or *PRCD* (*PD-MCI+* or *PRCD+*) within four years of the baseline visit or to have a confirmed absence of these cognitive symptoms (*PD-MCI-* or *PRCD-*) within the same period.

The MoCA score was used to define *PD-MCI*. Individuals with *PD-MCI+* were classified as having a MoCA score of less than 26, while those with *PD-MCI-* had a score of 26 or above. The assessment of *PRCD* was conducted using item 1.1 of the MDS-UPDRS Part I. The classification of *PRCD* was determined as follows: *PRCD+* was defined as a score > 1, while *PRCD-* was defined as a score ≤ 1. It is important to note that the analyses of *PD-MCI* and *PRCD* were conducted separately to facilitate the distinction between these two cognitive outcomes.

The study focuses on both single-cohort and multi-cohort analyses for the LuxPARK, PPMI, and ICEBERG cohorts. The data was examined for up to four years following the baseline visit for *PD-MCI* and *PRCD* classification to determine the relationship between baseline clinical characteristics and cognitive outcomes. The time-to-event analysis was extended to the point of conversion to *PD-MCI+*/*PRCD+* (events), or the end of the follow-up period known as censored.

Baseline characteristics represent the clinical assessment conducted at the baseline visit for each cohort. In particular, MDS-UPDRS Part III was conducted exclusively during the ON state for participants in the LuxPARK cohort. Consequently, the cohort-specific assessment protocol limited all analyses involving MDS-UPDRS Part III to solely ON state data.

The number of patients with PD who met the study's inclusion criteria for *PD-MCI* and *PRCD* is detailed in Tables 6.1 and 6.2, respectively. Table 6.1 presents the distribution of participants classified as *PD-MCI+* or *PD-MCI-* based on MoCA score across the three cohorts, while Table 6.2 provides a summary of the number of individuals categorized as *PRCD+* or *PRCD-* using the MDS-UPDRS Part I.

The multi-cohort approach is designed to facilitate the development and validation of ML models capable of identifying reliable predictors of cognitive impairment across diverse PD populations. Incorporating data from multiple independent cohorts represents a significant advancement in this field of study, as it addresses the limitations of previous research, which often relied on single-cohort studies lacking generalizability. The cross-cohort analysis improves the possibility of identifying reliable and clinically significant markers of cognitive decline in PD, thereby facilitating the development of more precise predictive models.

6.3.2 Machine learning framework

The ML framework for this study is designed to facilitate the development of predictive models for *PD-MCI* and *PRCD* in PD across multiple cohorts. The framework incorporates a data preprocessing pipeline described in Section 3.2, which ensures data quality and consistency across diverse cohorts and enables effective modeling. The following writing outlines the key components of the ML framework:

1. **Data aggregation**: As detailed in Section 3.2.1, the data aggregation process aggregates relevant variables to reduce data dimensionality while preserving information on each participant's clinical

Table 6.1 Number of patients meeting inclusion criteria for *PD-MCI* analysis.

Cohort	Inclusion criteria (1)	Inclusion criteria (2)	Events ***PD-MCI* Classification**	Events **Time-to-*PD-MCI***
LuxPARK	706	531	467 (87.9%)	471 (88.7%)
PPMI	1624	625	393 (62.9%)	462 (74.2%)
ICEBERG	162	117	56 (47.9%)	56 (47.9%)
Total	2492	1273	916 (72.0%)	989 (77.7%)

Number of patients who met the inclusion criteria and the distribution of the events. The "Events" columns show the total number and percentage of subjects who developed *PD-MCI* during the specified period. *PD-MCI* classification was performed up to 4 years of follow-up, and time-to-*PD-MCI* analysis was performed up to the last available follow-up visit for each patient.

Table 6.2 Number of patients meeting inclusion criteria for *PRCD* analysis.

Cohort	Inclusion criteria (1)	Inclusion criteria (2)	Events ***PRCD* Classification**	Events **Time-to-*PRCD***
LuxPARK	706	412	279 (67.7%)	291 (70.6%)
PPMI	1624	524	147 (28.1%)	285 (54.4%)
ICEBERG	162	117	61 (52.1%)	61 (52.1%)
Total	2492	1053	487 (46.2%)	637 (60.5%)

Number of patients who met the inclusion criteria and the distribution of the events. The "Events" columns show the total number and percentage of subjects who developed *PRCD* during the specified period. *PRCD* classification was performed up to 4 years of follow-up, and time-to-*PRCD* analysis was performed up to the last available follow-up visit for each patient.

features. This approach facilitates the interpretation of predictors by aggregating the data inputs, thereby focusing on clinically meaningful features.

2. **Cross-validation framework**: Data processing steps were implemented to manage missing values, categorical variables encoding and standardizing continuous variables to facilitate consistent multi-cohort analysis. Missing values were imputed within each fold of the cross-validation process, as outlined in Section 3.2.2, ensuring that the imputation was implemented independently for each training and testing fold. This approach served to minimize data leakage, thereby enhancing the robustness and reliability of the model. Subsequently, categorical variables were transformed into numerical representations by applying categorical encoding techniques, as detailed in Section 3.2.3. This conversion permitted the incorporation of categorical data into ML models without compromising data integrity, thereby maintaining the distinct characteristics of each variable. Subsequently, cross-study normalization techniques, as outlined in Section 3.2.4, were applied to continuous variables to reduce cohort-specific biases and ensure consistency across different cohorts in the multi-cohort analyses. The alignment of scales and distributions of continuous variables through normalization enabled cross-study comparisons and improved the generalizability of the predictive models developed in this study.

3. **Undersampling technique**: To address the issue of class imbalance in the outcome variables, an undersampling technique was applied to the training set, as detailed in Section 3.2.5. The objective of this method was to achieve a more balanced distribution of the outcome classes by reducing the size of the majority class. This ensured that the ML models trained on these data could more effectively learn patterns within minority class. By focusing the undersampling technique exclusively on the training set,

the evaluation of the model on the test set remained unaffected, thus enabling a fair assessment of the model's performance on imbalanced real-world data. This approach helped to prevent the model from showing bias towards the majority class, thereby enhancing the model's sensitivity and robustness in predicting both *PD-MCI* and *PRCD* outcomes.

4. **Nested cross-validation**: This study applied a nested cross-validation framework to optimize model performance by integrating feature selection with hyperparameter tuning, as outlined in Section 3.2.6. In each iteration of the nested process, feature selection techniques were applied to identify the most relevant and informative predictors for *PD-MCI* and *PRCD*. The model's interpretability and potential generalizability can be improved by focusing on the most significant predictors associated with cognitive impairment. Applying feature selection and hyperparameter tuning within a nested framework was important to prevent overfitting and improve the capability to generalize on unseen data. The application of feature selection prevented the model from being overwhelmed by less relevant predictors, while hyperparameter tuning optimized the model parameters to achieve the optimized predictive results. Moreover, these techniques supported the development of robust models with improved generalization across diverse PD cohorts, facilitating the identification of risk factors associated with *PD-MCI* and *PRCD* in PD.

A four-year follow-up period was applied to differentiate patients into groups of *PD-MCI+* vs. *PD-MCI-*, and *PRCD+* vs. *PRCD-*, using ML classification techniques. As detailed in Section 3.3, several classification algorithms were used within a cross-validation framework. This approach ensured the reliability of the model assessment and enabled the evaluation of predictive performance across cohorts. In addition to classification, a time-to-event analysis (see Section 3.4) was conducted to consider the duration until the diagnosis of *PD-MCI* (time-to-*PD-MCI*) or *PRCD* (time-to-*PRCD*). Furthermore, these approaches ensured a comprehensive assessment of both the probability and the time to cognitive impairment, thereby enhancing the potential for early identification and intervention in at-risk patients.

Model interpretability techniques, such as SHAP values, were implemented to improve the interpretability of the ML models, as detailed in Section 3.6. The SHAP value facilitates an in-depth comprehension of how each feature contributes to the model's predictions, quantifying the impact of individual predictors on the probability of *PD-MCI* and *PRCD* outcomes. By attributing prediction impact to each feature, SHAP values provide insights at the feature level, thus enabling the identification of key variables influencing cognitive impairment. This interpretability is to identify clinically relevant predictors and ensure the model's applicability in practical settings for PD patient care.

To further examine the risk associated with individual predictors in cognitive impairment outcomes, hazard ratios were derived by integrating SHAP values, as outlined in Section 3.6. This approach provides a metric that captures the contribution of each predictor to the model's predictions and reflects its influence over time on the likelihood of developing *PD-MCI* or *PRCD*. The resulting estimates facilitate the interpretation of risk factors, thereby enabling a more profound comprehension of the relative importance of predictors and their role in progressing to cognitive impairment in PD.

The predictive performance of the models was evaluated through two primary metrics, as outlined in Section 3.7. In the context of binary classification tasks, the AUC was used to evaluate the model's capacity to differentiate between patients showing mild cognitive impairment (*PD-MCI+*) and those without (*PD-MCI-*), as well as between reported mild cognitive impairment by the patients (*PRCD+*) and those patients did not report so (*PRCD-*). In the context of time-to-event analysis, the C-index was used

to assess the predictive capability of the models in estimating the time until the occurrence of events (*PD-MCI* and *PRCD*), taking into account the presence of censored data.

Furthermore, to assess the impact of cross-study normalization, each cohort compared the predictive capability and generalizability of non-normalized and normalized models. This comparative analysis aimed to confirm which cross-study normalization techniques improved the generalizability and robustness of the models across different PD cohorts.

The Bayesian signed-rank test was used to calculate cross-validated performance metrics to evaluate the various models' predictive performance. This statistical approach enabled a comparison of the performance of the models, with the probability of one model outperforming another being calculated. By applying this methodology, we could measure probabilities that indicated the likelihood of superiority for each model based on their respective performance metrics, such as AUC for binary classification and the C-index for time-to-event analysis. This approach provides the comparison of models and offers insights into the relative efficacy of predictive models, thereby aiding the selection of the model for clinical applications in identifying the possibility of cognitive impairment in PD patients.

Optimizing the model for each cohort analysis aims to ensure its reliability and generalizability. A stable model shows robustness by maintaining consistent performance across various subsets of the data, indicating that it is flexible to specific data points or characteristics within the cohorts. This reliability ensures that the model can generalize effectively to unseen data samples, a requirement for clinical applicability. By evaluating the model's performance across different cohorts and ensuring it consistently performs well, greater confidence can be placed in its predictive capabilities and ability to predict cognitive impairment in PD patients across diverse cohorts.

In the single cohort analyses, selected feature statistics were assessed across the 5-fold cross-validation process to identify consistent predictors of *PD-MCI* and *PRCD*. The objective was to evaluate the consistency of the selected predictors throughout the model training and evaluation iterations by calculating the percentage frequency of each feature across each fold. This analysis yielded insights into the consistency and variability of feature selection, showing which features were consistently selected across different folds and which showed variability in their selection. Understanding these patterns is to determine the most robust predictors associated with *PD-MCI* and *PRCD*, as discussed in Section 3.8. This approach improves our confidence in the identified features and informs future model development, ensuring that the most relevant clinical characteristics are prioritized in predictive models.

6.3.3 Statistical analysis

This study conducted a statistical analysis over four years to evaluate the relationships between predictors and outcomes, specifically *PD-MCI* and *PRCD*. A univariate analysis was conducted to assess the relationship between each predictor and the outcomes, specifically between *PD-MCI+* and *PD-MCI-* and between *PRCD+* and *PRCD-*. This analysis used hypothesis testing to test whether statistically significant differences existed in the outcomes and the baseline predictors. Furthermore, a correlation analysis was performed to investigate the interrelationships between the various predictors, offering insights into how these variables are associated, thus contributing to the overall interpretation of the data and supporting the study's findings on cognitive impairment in PD. The level of statistical significance was set at 5% ($p < 0.05$). The specific methodologies and techniques used for these statistical analyses are detailed in Section 3.9, ensuring reproducibility in the analysis approach.

6.3.4 Clinical utility analysis

This study performed clinical utility analyses, specifically decision curve analysis (DCA) and calibration analysis, as detailed in Section 3.10. The main objective of these analyses was to assess the applicability and efficacy of the predictive models in a clinical context.

Through the application of DCA, the predictive models' net clinical benefit was assessed compared to alternative strategies, such as the "treat all" approach. The AUNBC for the *optimized* models was calculated and subsequently compared to the AUNBC for the "treat all" strategy. A higher AUNBC indicates that the model provides a more significant net benefit than standard practices, thereby emphasizing its potential value in clinical decision-making. This comparison enabled us to assess whether our models could offer a clinically meaningful improvement. Bootstrapping hypothesis testing was used to derive p-values and assess the significance of the model's AUNBC. Adjusted p-values were calculated using the Benjamini-Hochberg method to account for multiple comparisons, ensuring the findings' robustness.

A calibration analysis was conducted to compare the predicted probabilities from the models with the actual observed outcomes for both *PD-MCI* and *PRCD* classifications. Furthermore, the predicted conversion probabilities were examined against the observed conversion probabilities, as measured by Kaplan-Meier (KM) estimates at the 4-year, for the time-to-*PD-MCI* and time-to-*PRCD* models. The calibration was evaluated by measuring the slope and MSE of the calibration curve, thereby providing insights into the degree of alignment between the predicted probabilities and the actual outcomes.

Clinical utility analysis supports the applicability of the model results while emphasizing their relevance in clinical practice. Ultimately, it improves patient care through the informed clinical decision-making process.

6.3.5 Code availability

R (v4.2.1) was used for data processing, normalization, and statistical analyses, while Python-3.8.6-GCCcore-10.2.0 was used for ML predictions. The open-source code is available in the GitLab repository under the MIT license at
https://gitlab.com/uniluxembourg/lcsb/bds/ml-cognitive-impairment.

6.4 Results

6.4.1 Individual cohort analyses

The predictive performance of the ML models for identifying *PD-MCI* and *PRCD* was initially evaluated in three cohorts: LuxPARK, PPMI, and ICEBERG.

PD-MCI classification (see Table 6.3): In the LuxPARK cohort, the model with the highest hold-out AUC achieved an average cross-validated AUC of 0.70 ± 0.07 and a hold-out AUC of 0.70, demonstrating good consistency between cross-validation and hold-out validation. The PPMI cohort demonstrated comparable performance, with an average cross-validated AUC of 0.70 ± 0.04 and a hold-out AUC of 0.69, indicating robust generalizability within this cohort. However, due to its smaller sample size, the ICEBERG cohort showed lower predictive performance, as evidenced by the results compared to the

LuxPARK and PPMI cohorts (see Figure 6.1).

Time-to-PD-MCI analysis (see Table 6.4): In the LuxPARK cohort, the model demonstrated a cross-validated C-index of 0.72±0.06 and a hold-out C-index of 0.63, indicating moderate predictive ability. The PPMI cohort showed a cross-validated C-index of 0.66±0.07 and a hold-out C-index of 0.64, which, although slightly lower than that observed in the LuxPARK cohort, still reflects a significant predictive capability for time-to-*PD-MCI*. The predictive capacity of the models in the ICEBERG cohort was once more constrained by the smaller sample size, which limited their efficacy (see Figure 6.2).

PRCD classification (see Table 6.5): In the LuxPARK cohort, the *optimized* model demonstrated an average cross-validated AUC of 0.69±0.11, with a hold-out AUC of 0.63, suggesting a moderate capacity for classifying PRCD cases. The PPMI cohort showed slightly superior performance, with an average cross-validated AUC of 0.71±0.08 and a hold-out AUC of 0.70. As observed in previous analyses, the smaller sample size in the ICEBERG cohort resulted in lower average cross-validated and hold-out AUC performance (see Figure 6.3).

Time-to-PRCD analysis (see Table 6.6): Regarding time-to-*PRCD* prediction, the PPMI cohort demonstrated the highest performance, with a promising average cross-validated C-index of 0.76±0.04 and the best hold-out C-index of 0.70. The LuxPARK cohort showed a lower best average cross-validated C-index of 0.69±0.08, yet a comparable best hold-out C-index of 0.71. As previously observed, the limited sample size in ICEBERG resulted in lower predictive performance than LuxPARK and PPMI for this analysis (see Figure 6.4).

Common predictors: The *PD-MCI* and *PRCD* analyses identified age at diagnosis of PD and baseline MoCA score as the most consistent predictors of future cognitive development (Table 6.7). In the case of *PD-MCI*, additional consistent predictors were identified as Benton Line Orientation (JLO) and cognitive impairment assessed within the MDS-UPDRS Part I at baseline (Table 6.7, left). In the *PRCD* analysis, predictor variables such as the MDS-UPDRS Part I and II total scores, SCOPA-AUT symptoms (particularly gastrointestinal and urinary symptoms), and disease duration since PD diagnosis were found to be most strongly associated with future patient-reported cognitive outcomes (Table 6.7, right).

Overall, these results indicate that while there is an overlap in the predictors of clinical measures of cognition (e.g., *PD-MCI*) and patient-reported cognitive outcomes, some of the identified associations are specific to the type of cognitive outcome being assessed.

Table 6.3 Predictive performance metrics for *PD-MCI* classification in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features
AdaBoost	**0.701 (0.129)**	**0.557**	**3 (9)**	0.697 (0.018)	0.663	1 (1)	**0.600 (0.103)**	**0.451**	**6 (11)**
CART	0.68 (0.098)	0.503	8 (14)	0.686 (0.033)	0.663	1 (1)	0.55 (0.046)	0.500	1 (1)
CatBoost	0.699 (0.106)	0.502	10 (15)	0.693 (0.057)	0.637	7 (12)	***0.581 (0.117)***	***0.674***	***5 (7)***
C4.5	0.637 (0.095)	0.456	4 (6)	**0.701 (0.044)**	**0.694**	**2 (3)**	0.534 (0.144)	0.527	4 (8)
FIGS	0.654 (0.053)	0.530	3 (7)	0.686 (0.033)	0.663	1 (1)	0.549 (0.091)	0.511	3 (6)
GOSDT-GUESSES	0.634 (0.077)	0.534	9 (18)	0.617 (0.055)	0.659	21 (38)	0.503 (0.113)	0.563	6 (13)
GBoost	0.657 (0.102)	0.589	8 (17)	***0.688 (0.032)***	***0.707***	***5 (20)***	0.51 (0.129)	0.479	4 (8)
HS	0.672 (0.111)	0.565	3 (6)	0.686 (0.033)	0.663	1 (1)	0.515 (0.083)	0.500	1 (4)
XGBoost	***0.697 (0.073)***	***0.702***	***5 (6)***	0.653 (0.083)	0.645	32 (50)	0.597 (0.175)	0.534	18 (25)

An overview of the *PD-MCI* prognostic classification's predictive performance statistics summarizes the *PD-MCI* prognostic classification's predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 6.4 Predictive performance metrics for time-to-*PD-MCI* in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**
CW-GBoost	**0.720 (0.060)**	**0.633**	**10 (24)**	0.612 (0.018)	0.601	10 (16)	0.57 (0.173)	0.558	5 (8)
Extra Survival	***0.676 (0.066)***	***0.634***	***94 (132)***	0.657 (0.047)	0.641	13 (13)	**0.644 (0.186)**	**0.542**	**45 (77)**
Survival GBoost	0.640 (0.084)	0.589	8 (13)	0.66 (0.036)	0.682	14 (23)	0.521 (0.067)	0.577	10 (21)
LSVM	0.630 (0.059)	0.548	19 (19)	***0.646 (0.064)***	***0.722***	***28 (28)***	0.625 (0.112)	0.591	11 (11)
NLSVM	0.606 (0.055)	0.611	174 (174)	0.627 (0.048)	0.682	36 (36)	0.564 (0.081)	0.521	77 (79)
Penalized Cox	0.652 (0.074)	0.597	3 (3)	0.631 (0.043)	0.677	1 (3)	***0.500 (0.042)***	***0.614***	***7 (7)***
Survival RF	0.694 (0.062)	0.590	4 (4)	**0.66 (0.068)**	**0.644**	**9 (11)**	0.603 (0.137)	0.495	62 (80)
Survival Trees	0.657 (0.060)	0.611	14 (24)	0.626 (0.045)	0.658	5 (11)	0.541 (0.136)	0.526	4 (5)

An overview of the time-to-*PD-MCI* predictive performance statistics summarizes the time-to-*PD-MCI* predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-index in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out C-index is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 6.5 Predictive performance metrics for *PRCD* classification in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features
AdaBoost	***0.643 (0.093)***	***0.631***	***7 (19)***	0.665 (0.081)	0.599	4 (5)	***0.653 (0.298)***	***0.615***	***1 (1)***
CART	0.632 (0.071)	0.599	7 (14)	0.638 (0.032)	0.570	1 (5)	0.677 (0.082)	0.538	6 (7)
CatBoost	0.693 (0.062)	0.604	13 (28)	**0.709 (0.081)**	**0.663**	**8 (16)**	0.703 (0.157)	0.596	7 (14)
C4.5	0.668 (0.025)	0.543	5 (15)	0.639 (0.036)	0.570	1 (6)	***0.628 (0.181)***	***0.615***	***2 (3)***
FIGS	0.667 (0.095)	0.518	8 (17)	0.649 (0.042)	0.570	1 (3)	***0.643 (0.177)***	***0.615***	***2 (4)***
GOSDT-GUESSES	0.616 (0.061)	0.570	19 (30)	0.619 (0.097)	0.606	14 (27)	**0.732 (0.195)**	**0.519**	**7 (7)**
GBoost	0.624 (0.076)	0.580	15 (26)	0.656 (0.039)	0.649	15 (23)	0.673 (0.229)	0.500	3 (5)
HS	0.625 (0.076)	0.482	1 (2)	0.649 (0.042)	0.570	1 (3)	***0.643 (0.177)***	***0.615***	***2 (4)***
XGBoost	**0.694 (0.114)**	**0.592**	**13 (15)**	***0.676 (0.044)***	***0.698***	***27 (35)***	0.578 (0.153)	0.500	19 (24)

An overview of the *PRCD* prognostic classification's predictive performance statistics summarizes the *PRCD* prognostic classification's predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 6.6 Predictive performance metrics for time-to-*PRCD* in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features
CW-GBoost	***0.647 (0.024)***	***0.707***	***14 (14)***	0.725 (0.061)	0.743	13 (26)	0.665 (0.126)	0.599	7 (10)
Extra Survival	0.648 (0.11)	0.685	158 (162)	0.725 (0.027)	0.730	14 (14)	0.773 (0.083)	0.595	91 (92)
Survival GBoost	0.664 (0.035)	0.679	26 (51)	**0.756 (0.036)**	**0.700**	**22 (50)**	0.714 (0.112)	0.392	21 (28)
LSVM	0.654 (0.073)	0.624	11 (11)	0.737 (0.021)	0.722	26 (26)	0.719 (0.218)	0.603	5 (5)
NLSVM	0.64 (0.064)	0.640	11 (11)	***0.742 (0.023)***	***0.762***	***31 (31)***	0.664 (0.233)	0.477	10 (10)
Penalized Cox	0.615 (0.109)	0.615	1 (3)	0.752 (0.032)	0.724	15 (48)	0.623 (0.241)	0.579	2 (2)
Survival RF	**0.689 (0.075)**	**0.66**	**99 (123)**	0.715 (0.041)	0.739	11 (12)	**0.811 (0.114)**	**0.549**	**42 (61)**
Survival Trees	0.661 (0.067)	0.608	17 (28)	0.693 (0.019)	0.582	7 (17)	***0.723 (0.249)***	***0.604***	***2 (2)***

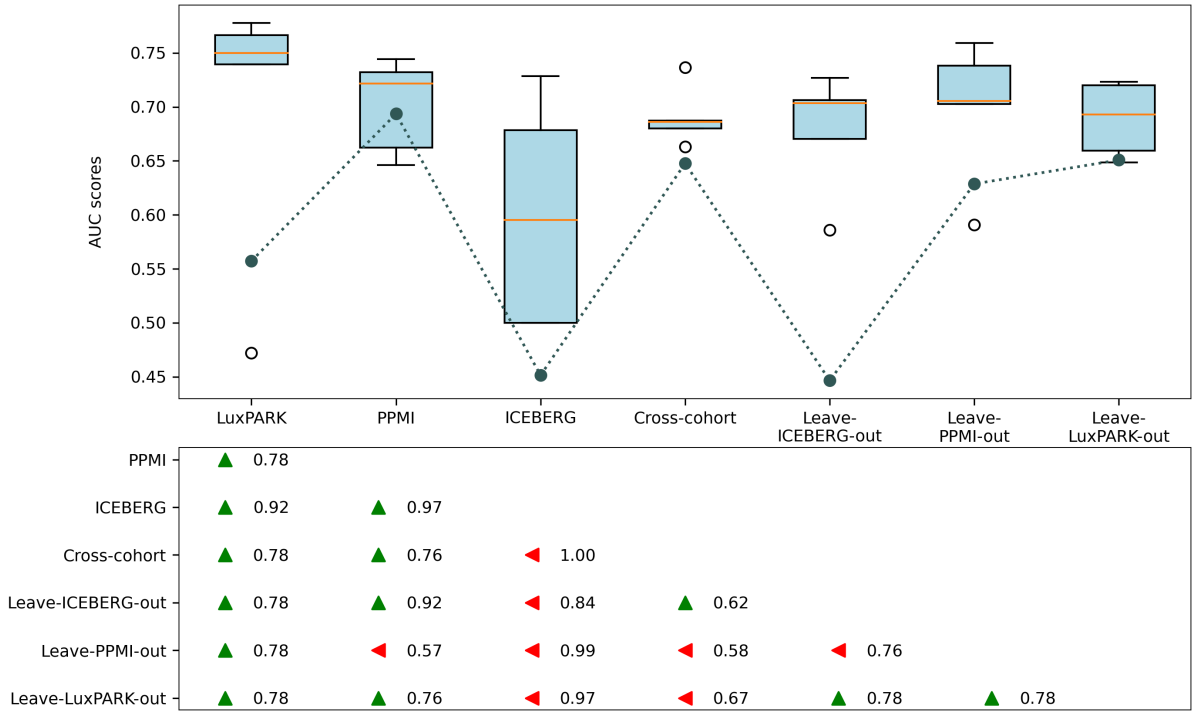
An overview of the time-to-PRCD predictive performance statistics summarizes the time-to-PRCD predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-index in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out C-index is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

6.4.2 Multi-cohort analyses

To address the limitations of single-cohort studies and develop more generalizable models, we conducted multi-cohort integrative analyses to predict *PD-MCI* and *PRCD* in PD.

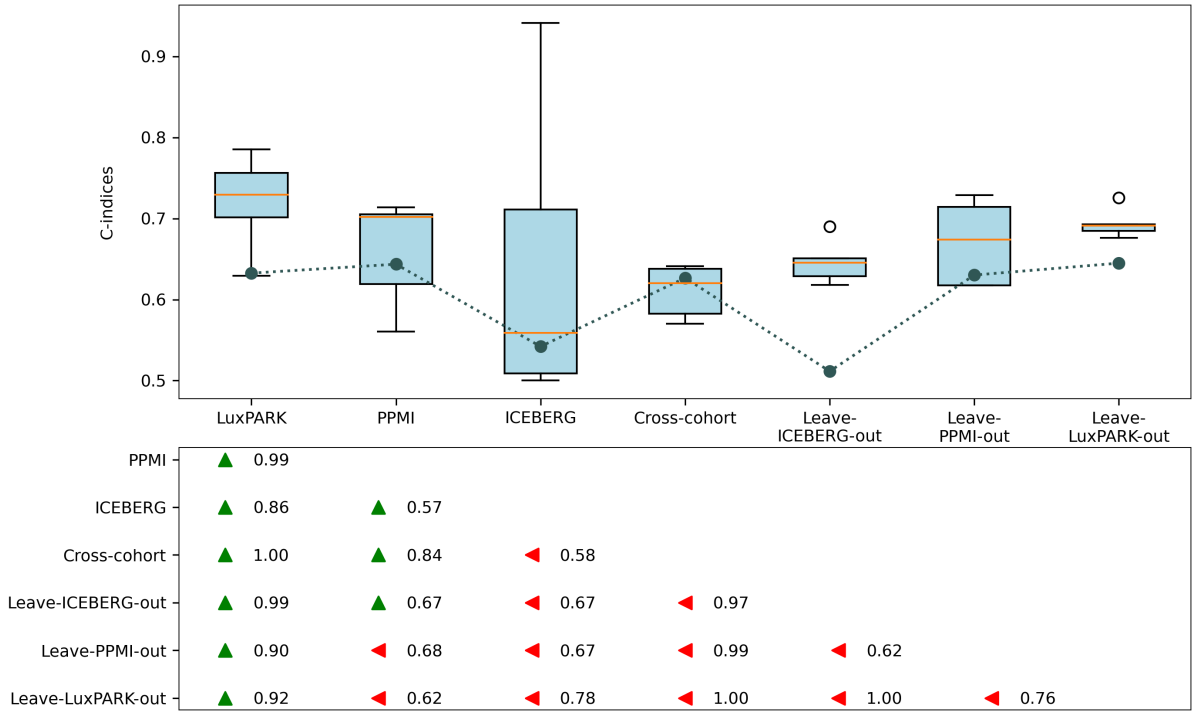
PD-MCI classification (see Table 6.8): In the cross-cohort analysis, the *optimized* model demonstrated an average cross-validated AUC of 0.69±0.03, with a hold-out AUC of 0.65. This performance was comparable to the optimized results observed in the single-cohort analyses, while also providing a model that is applicable across distinct cohorts. The *PD-MCI* classification models' hold-out predictive performance showed consistent trends across precision, recall, F-score, accuracy, balanced accuracy, and AUC scores (data is not shown in the thesis). This consistency highlights the robustness of the models in predicting *PD-MCI*. The leave-ICEBERG-out analysis indicated the presence of overfitting in the majority of models, as evidenced by reduced hold-out performance. This can be attributed to the smaller sample size of this cohort. The model with the highest performance (GBoost) achieved an average cross-validated AUC of 0.67±0.04 and a hold-out AUC of 0.60, indicating a higher generalization error than the cross-cohort analysis. In contrast, the leave-PPMI-out and leave-LuxPARK-out models showed comparable predictive performance to the cross-cohort analysis, with average cross-validated AUCs of 0.70±0.07 and 0.69±0.03 and hold-out AUCs of 0.63 and 0.65, respectively. These findings indicate that the predictive efficacy of multi-cohort models remains consistent even when a single cohort is excluded from the training process. However, the slight decline in performance when compared to

Figure 6.1 Comparison of cross-validated AUC scores for *PD-MCI* classification model.



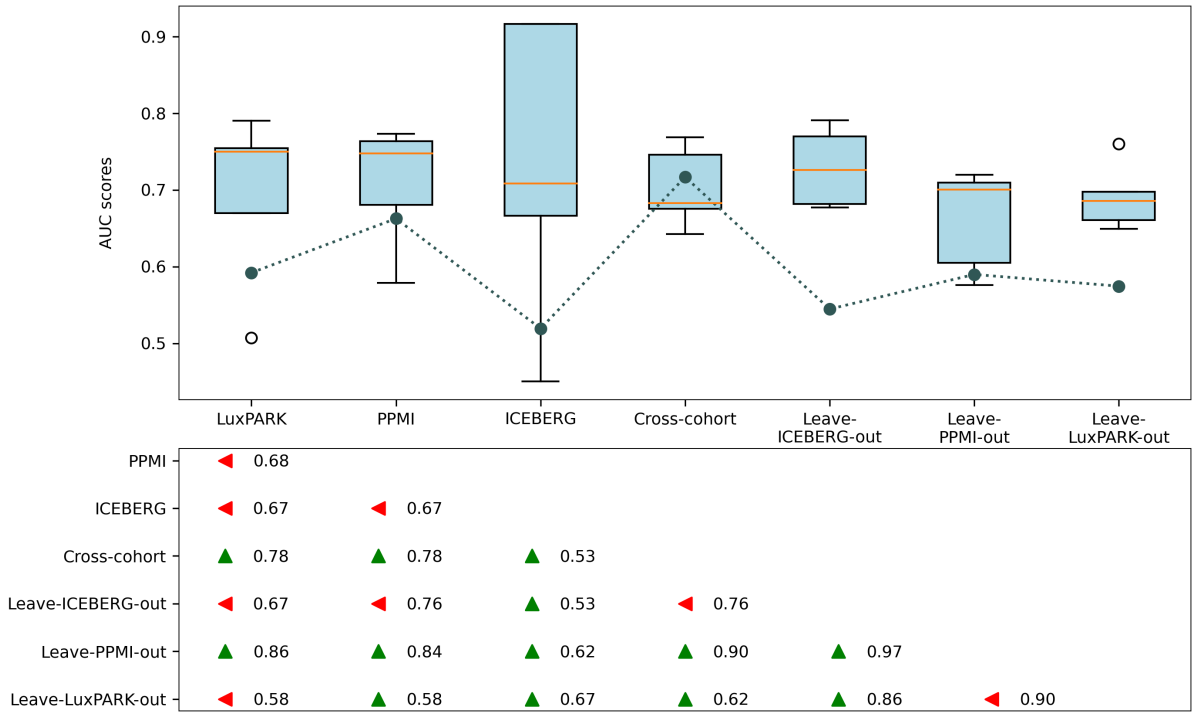
A comparison of cross-validated AUC scores and probabilities of better predictive performance for the *optimized PD-MCI* classification model in cross-cohort analyses. The upper part displays boxplots of the cross-validated AUC scores for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

Figure 6.2 Comparison of cross-validated C-indices for time-to-*PD-MCI* model.



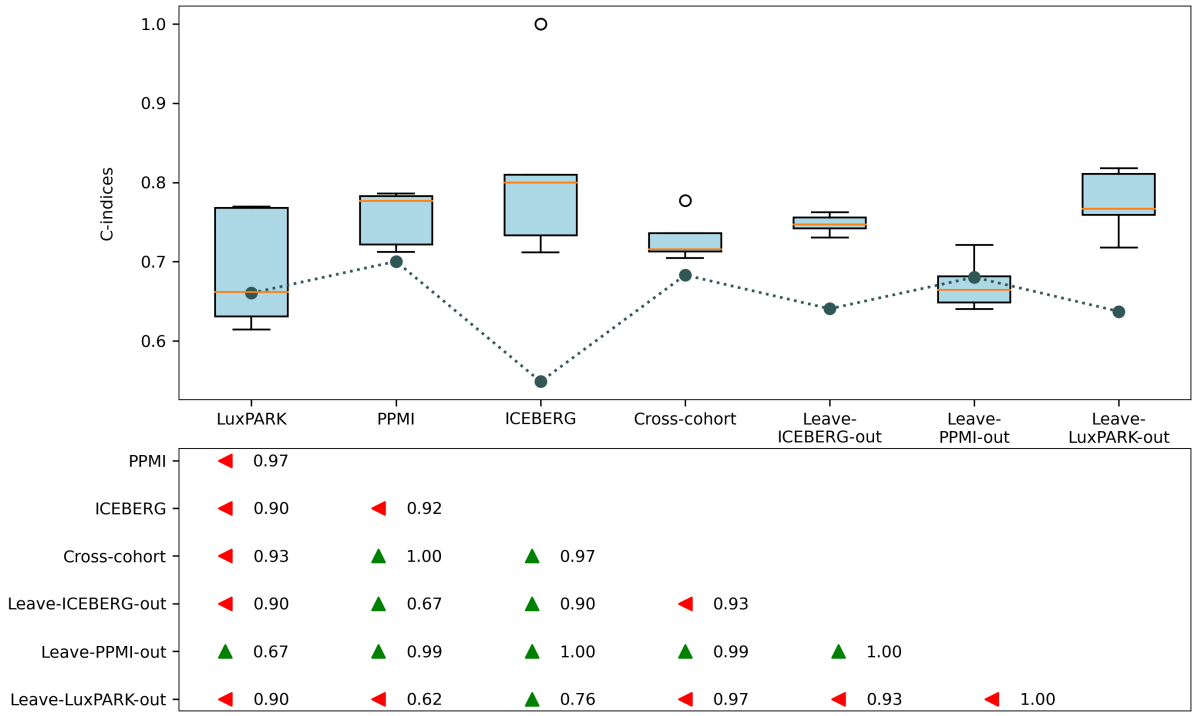
A comparison of cross-validated C-indices and probabilities of better predictive performance for the *optimized* time-to-*PD-MCI* model in cross-cohort analyses. The upper part displays boxplots of the cross-validated AUC scores for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

Figure 6.3 Comparison of cross-validated AUC scores for *PRCD* classification.



A comparison of cross-validated AUC scores and probabilities of better predictive performance for the *optimized PRCD* classification model across cohort analyses. The upper part displays boxplots of the cross-validated AUC scores for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

Figure 6.4 Comparison of cross-validated C-indices for time-to-*PRCD* model.



A comparison of cross-validated C-indices and probabilities of better predictive performance for the *optimized* time-to-PRCD model across cohort analyses. The upper part displays boxplots of the cross-validated C-indices for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

Table 6.7 Average percentage of predictors selected in 5-fold cross-validation for classification and time-to-event analyses across LuxPARK, PPMI, and ICEBERG cohorts.

Mild cognitive impairment (*PD-MCI*)			**Patient-reported cognitive decline (*PRCD*)**		
Predictors	**(1)**	**(2)**	**Predictors**	**(1)**	**(2)**
MDS-UPDRS II - Getting out of bed, car, or deep chair	-	80	SCOPA-AUT Gastrointestinal (GI)	53.3	93.3
MDS-UPDRS I - Urinary problems	-	73.3	MDS-UPDRS Part I score	46.7	100
Age at PD diagnosis	73.3	66.7	**Age at PD diagnosis**	40	100
MDS-UPDRS II - Saliva and drooling	40	66.7	Height (cm)	40	100
MDS-UPDRS II - Chewing and swallowing	-	53.3	Weight (kg)	46.7	93.3
MoCA score (adjusted for education)	60	33.3	Disease duration since PD diagnosis (years)	40	93.3
MDS-UPDRS I - Pain and other sensations	13.3	80	MDS-UPDRS Part II score	33.3	100
Benton Judgment of Line Orientation	26.7	66.7	SCOPA-AUT Sexual dysfunction	-	66.7
MDS-UPDRS III - Posture (ON)	26.7	66.7	SCOPA-AUT Urinary	26.7	100
Family history of PD	-	46.7	MDS-UPDRS II - Handwriting	40	80
MDS-UPDRS I - Depressed moods	-	46.7	**MoCA score (adjusted for education)**	-	60
MDS-UPDRS I - Cognitive impairment	20	66.7	MDS-UPDRS III - Global spontaneity of movement (ON)	-	60
MDS-UPDRS I - Fatigue	20	60	Tremor	-	60
MDS-UPDRS I - Sleep problems (night)	20	60	BMI (kg/m2)	40	73.3
MDS-UPDRS I - Constipation problems	20	60	MDS-UPDRS I - Apathy	26.7	86.7

Statistics on the average percentage of times predictors were selected during 5-fold cross-validation (CV) analyses. It compares data for mild cognitive impairment and patient-reported cognitive decline (1) classification and (2) time-to-event analyses across the LuxPARK, PPMI, and ICEBERG cohorts. The information presented includes the average percentage of times each feature was chosen in a 5-fold CV for single-cohort analyses in LuxPARK, PPMI, and ICEBERG for classification and time-to-event analyses across all cohorts. Features are listed in descending order based on their average selection percentages in classification and time-to-event analyses, with the top 15 features presented.

using data from all three cohorts suggests that cohort-specific information plays a role in enhancing model performance.

Time-to-PD-MCI analysis (see Table 6.9): The cross-cohort analysis yielded an average cross-validated C-index of 0.61±0.03 and a hold-out C-index of 0.63, indicating moderate predictive performance. These results are comparable to those of the LuxPARK and PPMI single-cohort analyses. In the leave-ICEBERG-out analysis, the overfitting phenomenon was observed in most models, as evidenced by the lower hold-out performance. However, the CW-GBoost model showed an average cross-validated C-index of 0.65±0.08 and a hold-out C-index of 0.63, comparable to the cross-cohort analysis results. The leave-PPMI-out and leave-LuxPARK-out analyses showed comparable predictive performance to the cross-cohort analysis, with average cross-validated C-indices of 0.67±0.05 and 0.69±0.02 and hold-out C-indices of 0.63 and 0.65, respectively. As with the *PD-MCI* classification analysis, these results demonstrate the robustness of the models to different cohort exclusions and illustrate the feasibility of developing predictive models that generalize across cohorts while maintaining performance levels comparable to those of single-cohort models.

PRCD classification (see Table 6.10): The cross-cohort analysis yielded an average cross-validated AUC of 0.70±0.05 and a hold-out AUC of 0.72, showing a slight superiority over the *optimized* single-cohort analyses. The hold-out predictive performance of the *PRCD* classification models showed consistent results for precision, recall, F-score, accuracy, balanced accuracy, and AUC scores (data is not shown in the thesis). This consistency highlights the reliability of the models, ensuring reliable prediction of *PRCD* in the study. In the leave-ICEBERG-out analysis, the majority of models demonstrated indications of overfitting, as evidenced by reduced hold-out AUCs. The *optimized* GOSDT-GUESSES model demonstrated an average cross-validated AUC of 0.65±0.04 and a hold-out AUC of 0.61, indicating a reduced capability for predictive accuracy compared to the cross-cohort model. The leave-LuxPARK-out analysis showed similar overfitting trends, with the *optimized* AdaBoost model achieving an average cross-validated AUC of 0.68±0.04 and a hold-out AUC of 0.63. The *optimized* FIGS model for leave-PPMI-out demonstrated comparable hold-out performance to the cross-cohort analysis, with a hold-out AUC of 0.71. However, the average cross-validated AUC was lower (0.61±0.05). In general, none of the leave-one-cohort-out models demonstrated superior cross-validated or hold-out predictive performance compared to the cross-cohort model.

Time-to-PRCD analysis (see Table 6.11): The cross-cohort analysis yielded an average cross-validated C-index of 0.73±0.04 and a hold-out C-index of 0.72, indicating that the performance was comparable to that of the most effective single-cohort PPMI model, despite the increased complexity of the cross-cohort prediction task. In the leave-ICEBERG-out analysis, the model achieved an average cross-validated C-index of 0.75±0.01 with a hold-out C-index of 0.64, indicating a slight decrease in generalization performance when ICEBERG was excluded from training despite a slight improvement in cross-validated performance compared to the cross-cohort analysis. The leave-PPMI-out and leave-LuxPARK-out analyses yielded comparable results, with average cross-validated C-indices of0.67±0.03 and 0.77±0.04 and hold-out C-indices of 0.67 and 0.64, respectively. While the predictive performance remained relatively robust for the leave-one-cohort-out models, these results suggest that excluding any cohort may slightly reduce overall robustness. Moreover, the cross-cohort model demonstrated comparable

performance to the *optimized* single-cohort model.

Model stability: Multi-cohort analyses demonstrated superior stability and robustness in predictive models compared to single-cohort analyses, as illustrated in Figures C.1-C.4 for both *PD-MCI* and *PRCD* analyses. This higher stability is attributed to integrating diverse patient populations, enabling models to better account for cohort-specific biases and produce more generalizable predictions. These advantages underscore the potential preference for multi-cohort approaches in clinical practice, where reliable predictions are important across varying demographic and clinical characteristics. However, the ICEBERG cohort showed reduced stability, likely attributable to its smaller sample size. This finding highlights the necessity of sufficient statistical power to guarantee model robustness, as restricted sample sizes can lead to performance variability and decrease the results' representativeness.

The comparison between multi-cohort and single-cohort analyses demonstrated the difficulties posed by cohort-specific variations and the substantial advantages of integrating data across cohorts. Multiple datasets allow for the implementation of multi-cohort approaches, which can address the challenges above and facilitate the development of more reliable and generalizable predictive models for cognitive impairment in PD.

Table 6.8 Predictive performance metrics for *PD-MCI* classification in multi-cohort analyses.

Algorithm	**Cross-cohort**			**Leave-ICEBERG-out**		
	Mean (SD)	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**
AdaBoost	0.678 (0.053)	0.670	2 (2)	0.673 (0.05)	0.530	3 (5)
CART	0.687 (0.055)	0.632	2 (2)	0.675 (0.047)	0.595	4 (8)
CatBoost	**0.691 (0.028)**	**0.648**	**5 (7)**	0.652 (0.023)	0.544	8 (9)
C4.5	0.62 (0.058)	0.551	4 (7)	0.661 (0.035)	0.489	3 (5)
FIGS	0.676 (0.043)	0.632	2 (2)	**0.679 (0.056)**	**0.446**	**1 (5)**
GOSDT-GUESSES	0.663 (0.075)	0.627	22 (34)	0.606 (0.061)	0.533	39 (65)
GBoost	0.688 (0.06)	0.621	14 (32)	***0.666 (0.035)***	***0.604***	***18 (35)***
HS	0.676 (0.043)	0.632	2 (2)	**0.679 (0.056)**	**0.446**	**1 (5)**
XGBoost	***0.672 (0.072)***	***0.671***	***48 (69)***	0.646 (0.019)	0.561	36 (63)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
AdaBoost	***0.699 (0.065)***	***0.629***	***2 (7)***	0.68 (0.029)	0.651	2 (7)
CART	0.668 (0.069)	0.523	1 (1)	0.677 (0.043)	0.653	2 (3)
CatBoost	0.683 (0.057)	0.602	5 (13)	***0.685 (0.051)***	***0.654***	***12 (26)***
C4.5	0.638 (0.042)	0.516	3 (6)	0.68 (0.027)	0.460	2 (3)
FIGS	0.681 (0.056)	0.523	1 (1)	**0.689 (0.034)**	**0.651**	**2 (2)**
GOSDT-GUESSES	0.658 (0.034)	0.489	17 (34)	0.621 (0.045)	0.588	16 (18)
GBoost	0.694 (0.05)	0.513	4 (9)	0.673 (0.047)	0.630	12 (39)
HS	0.681 (0.056)	0.523	1 (1)	**0.689 (0.034)**	**0.651**	**2 (2)**
XGBoost	0.694 (0.054)	0.553	39 (48)	0.67 (0.043)	0.650	17 (25)

An overview of the *PD-MCI* prognostic classification's predictive performance statistics summarizes the *optimized PD-MCI* prognostic classification's predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 6.9 Predictive performance metrics for time-to-*PD-MCI* in multi-cohort analyses.

Algorithm	Cross-cohort			Leave-ICEBERG-out		
	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features
CW-GBoost	0.598 (0.035)	0.637	4 (5)	***0.646 (0.083)***	***0.633***	***1 (2)***
Extra Survival	0.583 (0.03)	0.592	12 (12)	0.598 (0.028)	0.483	13 (13)
Survival GBoost	***0.577 (0.078)***	***0.651***	***11 (24)***	0.621 (0.037)	0.563	9 (18)
LSVM	0.59 (0.039)	0.570	29 (29)	0.61 (0.06)	0.500	68 (68)
NLSVM	0.589 (0.059)	0.615	40 (40)	0.628 (0.039)	0.506	53 (53)
Penalized Cox	**0.610 (0.032)**	**0.627**	**2 (3)**	**0.647 (0.027)**	**0.512**	**1 (1)**
Survival RF	0.562 (0.04)	0.630	6 (6)	0.62 (0.026)	0.483	11 (12)
Survival Trees	0.591 (0.029)	0.615	14 (28)	0.617 (0.024)	0.525	1 (1)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
CW-GBoost	***0.638 (0.037)***	***0.643***	***18 (27)***	0.691 (0.014)	0.643	12 (23)
Extra Survival	0.658 (0.063)	0.601	137 (138)	0.68 (0.008)	0.627	20 (20)
Survival GBoost	**0.67 (0.052)**	**0.63**	**27 (61)**	**0.694 (0.019)**	**0.645**	**25 (48)**
LSVM	0.65 (0.034)	0.559	33 (33)	0.687 (0.017)	0.648	59 (59)
NLSVM	0.628 (0.048)	0.560	53 (53)	***0.693 (0.023)***	***0.661***	***58 (58)***
Penalized Cox	0.617 (0.042)	0.610	1 (2)	***0.663 (0.019)***	***0.661***	***2 (2)***
Survival RF	0.665 (0.048)	0.641	8 (8)	0.681 (0.009)	0.658	22 (22)
Survival Trees	0.641 (0.046)	0.573	5 (16)	0.648 (0.019)	0.620	8 (15)

An overview of the *optimized* time-to-*PD-MCI* predictive performance statistics summarizes the *comprehensive* time-to-*PD-MCI* predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-index in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out C-index is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 6.10 Predictive performance metrics for *PRCD* classification in multi-cohort analyses.

Algorithm	Cross-cohort			Leave-ICEBERG-out		
	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features
AdaBoost	0.699 (0.044)	0.705	8 (16)	0.71 (0.027)	0.545	6 (13)
CART	0.683 (0.054)	0.663	7 (15)	0.704 (0.026)	0.536	1 (2)
CatBoost	***0.702 (0.053)***	***0.719***	***14 (23)***	0.716 (0.041)	0.527	7 (8)
C4.5	0.657 (0.04)	0.671	5 (13)	0.689 (0.054)	0.536	3 (5)
FIGS	0.682 (0.044)	0.686	3 (5)	0.71 (0.038)	0.482	4 (6)
GOSDT-GUESSES	0.652 (0.032)	0.639	30 (55)	***0.654 (0.037)***	***0.607***	***16 (16)***
GBoost	**0.703 (0.052)**	**0.717**	**10 (14)**	**0.729 (0.051)**	**0.545**	**15 (29)**
HS	0.682 (0.044)	0.686	3 (5)	0.71 (0.038)	0.482	4 (6)
XGBoost	0.678 (0.044)	0.697	35 (48)	0.702 (0.033)	0.509	48 (54)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
AdaBoost	**0.662 (0.067)**	**0.590**	**8 (15)**	***0.676 (0.042)***	***0.625***	***3 (7)***
CART	0.646 (0.055)	0.665	4 (6)	0.667 (0.04)	0.597	29 (50)
CatBoost	0.654 (0.044)	0.610	7 (12)	0.687 (0.033)	0.577	6 (15)
C4.5	0.604 (0.051)	0.559	8 (13)	0.685 (0.043)	0.560	2 (3)
FIGS	***0.614 (0.045)***	***0.705***	***4 (7)***	0.677 (0.03)	0.558	1 (1)
GOSDT-GUESSES	0.656 (0.052)	0.612	28 (42)	0.631 (0.051)	0.575	40 (64)
GBoost	0.643 (0.03)	0.583	12 (23)	**0.691 (0.043)**	**0.575**	**12 (28)**
HS	***0.614 (0.045)***	***0.705***	***4 (7)***	0.677 (0.03)	0.558	1 (1)
XGBoost	0.636 (0.053)	0.697	31 (65)	0.676 (0.025)	0.571	46 (50)

An overview of the *PRCD* prognostic classification's predictive performance statistics summarizes the *optimized PRCD* prognostic classification's predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 6.11 Predictive performance metrics for time-to-*PRCD* in multi-cohort analyses.

Algorithm	Cross-cohort			Leave-ICEBERG-out		
	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features
CW-GBoost	**0.729 (0.029)**	**0.683**	**14 (19)**	0.736 (0.025)	0.605	11 (18)
Extra Survival	0.726 (0.018)	0.700	164 (165)	0.734 (0.026)	0.588	164 (165)
Survival GBoost	0.722 (0.032)	0.704	12 (30)	0.744 (0.037)	0.585	14 (25)
LSVM	0.714 (0.032)	0.691	180 (180)	0.738 (0.019)	0.625	53 (53)
NLSVM	***0.729 (0.038)***	***0.718***	***48 (48)***	0.726 (0.029)	0.597	49 (49)
Penalized Cox	0.708 (0.02)	0.660	13 (29)	***0.748 (0.012)***	***0.641***	***17 (28)***
Survival RF	0.727 (0.019)	0.697	139 (149)	0.729 (0.033)	0.603	13 (13)
Survival Trees	0.674 (0.02)	0.657	7 (9)	0.692 (0.037)	0.659	8 (10)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
CW-GBoost	**0.671 (0.032)**	**0.680**	**5 (9)**	***0.755 (0.025)***	***0.644***	***12 (23)***
Extra Survival	0.64 (0.052)	0.675	18 (18)	0.753 (0.041)	0.616	129 (129)
Survival GBoost	0.669 (0.035)	0.676	10 (21)	0.76 (0.042)	0.633	24 (45)
LSVM	0.658 (0.032)	0.617	151 (151)	0.758 (0.033)	0.613	49 (49)
NLSVM	0.651 (0.029)	0.658	151 (151)	0.762 (0.037)	0.607	56 (56)
Penalized Cox	0.657 (0.04)	0.550	1 (32)	**0.774 (0.041)**	**0.637**	**83 (114)**
Survival RF	***0.659 (0.028)***	***0.698***	***68 (106)***	0.75 (0.05)	0.626	115 (120)
Survival Trees	0.64 (0.035)	0.510	19 (35)	0.691 (0.04)	0.602	3 (6)

An overview of the *optimized* time-to-*PRCD* predictive performance statistics summarizes the *comprehensive* time-to-*PRCD* predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-index in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out C-index is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

6.4.3 Comparative evaluation of cross-study normalization integration

To assess the impact of cross-study normalization on model performance, we conducted a comparative analysis of the hold-out prediction metrics of optimized models with and without normalization across a range of multi-cohort machine learning analyses.

In the hold-out analysis, cross-study normalization was found to significantly improve predictive performance for *PD-MCI* and *PRCD* classification and time-to-*PRCD* prediction (see Tables 6.12-6.13). In both classification tasks, models incorporating normalization demonstrated notably higher hold-out AUCs than unnormalized models. Similarly, the application of cross-study normalization resulted in a notable improvement in the hold-out C-index for time-to-*PRCD* prediction.

Notably, the advantages of cross-study normalization were pronounced in the leave-PPMI-out analysis, whereas other analyses demonstrated modest improvements. This discrepancy may be attributed to the distinctive value distributions of key predictors in the PPMI cohort (Table 6.14), where normalization effectively mitigated study-specific biases and normalized data variations across cohorts.

These findings highlight that the efficacy of cross-study normalization can depend on the type of study-specific biases in the data. While it demonstrated clear advantages in specific analyses, notably when the PPMI cohort was excluded from training, its impact was less consistent in other cohort integration analyses.

Thus, cross-study normalization requires a particular assessment to ascertain its appropriateness for disparate contexts. It is most effective when cohort-related biases are present, provided they are not overly complex and can be addressed with statistical adjustments. It is important to implement cross-study normalization to the specific characteristics and distributions of the data to maximize its utility and improve the robustness of predictive models.

6.4.4 Associations between clinical features and cognition outcome

A cross-cohort analysis was conducted to identify key predictors associated with cognitive impairment in PD using *PD-MCI* classification and time-to-*PD-MCI* prediction. The analysis yielded consistent results, indicating that the identified predictors are robust across different analytical approaches. The SHAP value plots (Figures 6.5-6.6) revealed that there were common predictors across these analyses.

Visuospatial ability, as measured by the Benton JLO, was a significant predictor for *PD-MCI* classification and time-to-*PD-MCI* prediction. The results consistently showed that superior performance on the JLO test was associated with a reduced risk of *PD-MCI* and a delayed onset of cognitive impairment. This finding highlights the potential value of visuospatial function as a cognitive health indicator for PD patients.

Age at diagnosis of PD was also identified as a key predictor in both the *PD-MCI* classification and time-to-*PD-MCI* models. A correlation was identified between older age at PD diagnosis and an increased risk of developing *PD-MCI*, as well as a shorter time to *PD-MCI* onset. This indicates that early PD onset may be associated with better preservation of cognitive abilities, whereas older age at diagnosis may be linked to accelerated cognitive decline.

Furthermore, body weight was identified as a factor in predicting *PD-MCI*. However, the relationship between body weight and cognitive outcome was not readily interpretable from the SHAP value plot. This indicates that body weight may interact with other predictors in the multifactorial model, thereby highlighting the intricate nature of the relationships observed in the data.

Table 6.12 Significance testing of hold-out predictive metrics between normalized and unnormalized models for *PD-MCI* and *PRCD* in multi-cohort analyses.

	Mild cognitive impairment (*PD-MCI*)		**Patient-reported cognitive decline (*PRCD*)**	
Cohort	**Normalized vs. Unnormalized**	**Cross-cohort normalization**	**Normalized vs. Unnormalized**	**Cross-cohort normalization**
Classification:				
Cross-cohort	0.400	Ratio-A	0.626	Standardize
Leave-ICEBERG-out	0.459	Ratio-A	0.683	Mean-centering
Leave-PPMI-out	0.021	Mean-centering	0.028	ComBat
Leave-LuxPARK-out	0.170	M-ComBat	0.892	ComBat
Time-to-event:				
Cross-cohort	0.051	M-ComBat	0.207	Mean-centering
Leave-ICEBERG-out	1.000	ComBat	0.037	Ratio-A
Leave-PPMI-out	0.121	ComBat	3.80E-05	Mean-centering
Leave-LuxPARK-out	0.065	ComBat	0.055	Standardize

A comparison of the statistical significance of the differences between the hold-out predictive performance metrics for the *optimized PD-MCI* and *PRCD* models across cohorts. The *p*-values for the significance of the difference were calculated using DeLong's test for classification and the one-shot nonparametric test for time-to-event analysis. A *p*-value < 0.05 indicates a significant difference in hold-out predictive performance between the two models. The normalization method used on the *optimized* model is indicated in the column "Normalization".

For *PRCD*, the SHAP values plot revealed similar key predictors to those found for *PD-MCI*, with age at diagnosis and Benton JLO scores emerging as significant predictors (Figures 6.7-6.8). A significant negative correlation was observed between age at diagnosis and both MoCA and Benton JLO scores (Table 6.17), indicating that older age at diagnosis is associated with poorer cognitive performance.

In the time-to-*PD-MCI* analysis, patients diagnosed with PD at age 53 or older had a nearly 2.4-fold higher risk of developing cognitive impairment compared to those at a younger age (Table 6.15 and Figure 6.9). Similarly, in the time-to-*PRCD* analysis, patients diagnosed at age 62 or older were 1.5 times more likely to report cognitive impairment (Table 6.16 and Figure 6.10).

Other clinical characteristics associated with an increased likelihood of *PRCD* included the MDS-UPDRS Part I score, disease duration, the presence of tremors, and sex. Males were more likely to report cognitive impairment than females.

Furthermore, patients with lower Modified Schwab & England Activities of Daily Living (ADL) scale were more likely to indicate that cognitive impairment affected their daily living. Patients with greater functional dependence may perceive or recognize cognitive decline more acutely, resulting in more prompt recognition of cognitive impairment and its impact on daily activities.

The presence of non-motor symptoms, as measured by the MDS-UPDRS Part I and SCOPA-AUT, was also identified as contributing to the risk of *PRCD*. This highlights the complex, multifactorial nature of cognitive impairment affecting daily living in patients with PD.

While the correlations between predictors and *PD-MCI* and *PRCD* outcomes were consistent, notable differences were observed (Table 6.17). For example, a positive correlation was observed between BMI and *PD-MCI*, yet no such correlation was shown in *PRCD* outcome. Thermoregulatory dysfunction

Table 6.13 Predictive performance metrics between normalized and unnormalized models for *PD-MCI* and *PRCD* in multi-cohort analyses.

	Cross-cohort			**Leave-ICEBERG-out**		
	Mean (SD)	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**
***PD-MCI* classification**						
Normalized	0.691 (0.028)	0.648	5 (7)	0.679 (0.056)	0.446	1 (5)
Unnormalized	0.669 (0.068)	0.617	7 (16)	0.666 (0.035)	0.604	18 (35)
***PRCD* classification**						
Normalized	0.703 (0.052)	0.717	10 (14)	0.71 (0.027)	0.545	6 (13)
Unnormalized	0.702 (0.053)	0.719	14 (23)	0.729 (0.051)	0.545	15 (29)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
***PD-MCI* classification**						
Normalized	0.699 (0.065)	0.629	2 (7)	0.689 (0.034)	0.651	2 (2)
Unnormalized	0.694 (0.05)	0.513	4 (9)	0.68 (0.042)	0.651	8 (16)
***PRCD* classification**						
Normalized	0.656 (0.052)	0.612	28 (42)	0.685 (0.051)	0.616	8 (17)
Unnormalized	0.662 (0.067)	0.59	8 (15)	0.691 (0.043)	0.575	12 (28)
	Cross-cohort			**Leave-ICEBERG-out**		
	Mean (SD)	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**
time-to-*PD-MCI* model						
Normalized	0.61 (0.032)	0.627	2 (3)	0.647 (0.027)	0.512	1 (1)
Unnormalized	0.594 (0.079)	0.662	1 (2)	0.631 (0.016)	0.512	1 (2)
time-to-*PRCD* model						
Normalized	0.729 (0.029)	0.683	14 (19)	0.744 (0.037)	0.585	14 (25)
Unnormalized	0.726 (0.018)	0.7	164 (165)	0.748 (0.012)	0.641	17 (28)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
time-to-*PD-MCI* model						
Normalized	0.67 (0.052)	0.63	27 (61)	0.694 (0.019)	0.645	25 (48)
Unnormalized	0.658 (0.063)	0.601	137 (138)	0.694 (0.019)	0.602	72 (110)
time-to-*PRCD* model						
Normalized	0.671 (0.032)	0.68	5 (9)	0.774 (0.041)	0.637	83 (114)
Unnormalized	0.658 (0.031)	0.605	11 (21)	0.762 (0.044)	0.655	19 (33)

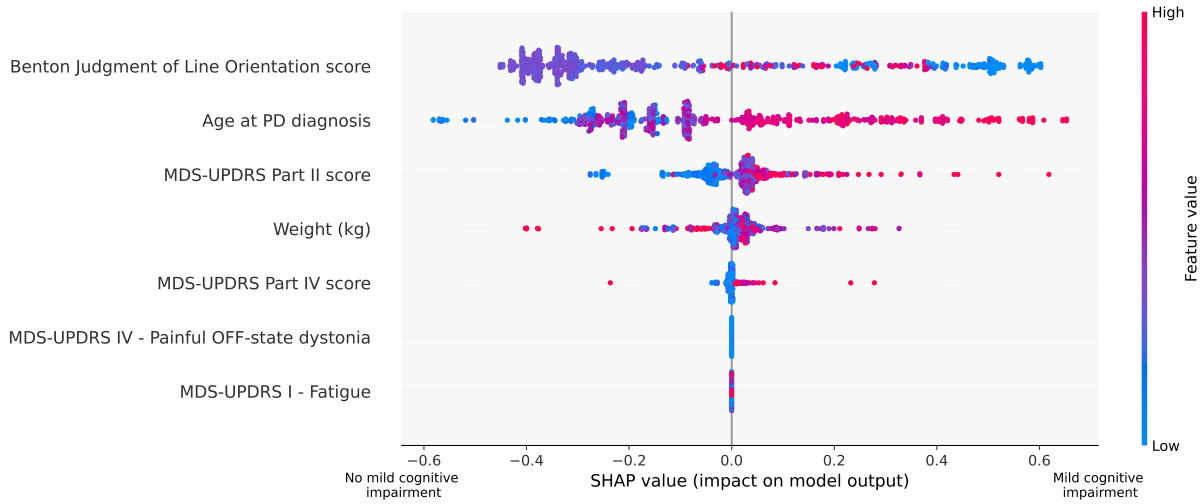
Assessment of the predictive performance of *PD-MCI* and *PRCD* prognostic models for cognitive impairment, including classification and time-to-event analyses. The evaluation includes cross-validated and hold-out AUC or C-index calculations for both normalized and unnormalized models and a detailed examination of the number of features used in each model. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table 6.14 Comparative analysis of baseline features mean differences across cohorts in *PD-MCI* and *PRCD* analyses.

Predictors	**LuxPARK vs. PPMI (*p*-values)**	**LuxPARK vs. ICEBERG (*p*-values)**	**PPMI vs. ICEBERG (*p*-values)**	***p*-values**
Mild cognitive impairment (*PD-MCI*)				
Age at PD diagnosis	1.72 (0.008)	0.71 (1.000)	-1.01 (0.858)	1.10E-02
Disease duration	3.29 (4.72E-32)	3.73 (1.55E-16)	0.44 (0.383)	1.40E-36
Weight (kg)	0.45 (1.000)	5.80 (0.002)	5.35 (0.005)	2.70E-03
Height (cm)	-2.90 (1.14E-05)	-2.04 (0.165)	0.86 (1.000)	2.05E-05
BMI (kg/m^2)	1.04 (2.20E-04)	2.67 (3.40E-08)	1.63 (0.002)	6.43E-09
MDS-UPDRS Part I score	0.82 (2.36E-15)	0.23 (0.108)	-0.59 (5.82E-05)	6.22E-15
MDS-UPDRS Part II score	8.18 (9.41E-107)	2.01 (1.000)	-6.16 (4.48E-33)	7.83E-115
MDS-UPDRS Part III (ON) score	5.77 (1.14E-37)	4.32 (5.56E-04)	-1.45 (4.91E-04)	5.82E-37
SCOPA-AUT total score	1.37 (8.65E-08)	0.48 (0.853)	-0.89 (0.101)	1.65E-07
Patient-reported cognitive decline (*PRCD*)				
Age at PD diagnosis	1.63 (0.05)	-0.35 (1.000)	-1.99 (0.14)	2.25E-02
Disease duration since PD diagnosis (years)	3.43 (5.71E-31)	3.99 (5.98E-19)	0.56 (0.26)	3.14E-36
Weight (kg)	0.37 (1.000)	5.49 (0.005)	5.12 (0.011)	5.90E-03
Height (cm)	-2.24 (0.003)	-1.75 (0.296)	0.48 (1.000)	4.40E-03
BMI (kg/m^2)	0.82 (0.013)	2.47 (7.55E-07)	1.65 (0.002)	9.30E-07
MDS-UPDRS Part I score	8.68 (2.27E-86)	2.42 (0.598)	-6.26 (1.16E-29)	1.41E-93
MDS-UPDRS Part II score	6.36 (6.04E-37)	4.88 (1.81E-05)	-1.48 (9.53E-04)	3.24E-36
MDS-UPDRS Part III (ON) score	19.49 (1.45E-35)	9.81 (3.32E-06)	-9.67 (6.73E-08)	7.00E-36
SCOPA-AUT total score	0.58 (8.54E-07)	0.74 (6.16E-06)	0.16 (0.344)	1.50E-08

A comparative analysis of the mean differences for baseline features across the LuxPARK, PPMI, and ICEBERG cohorts. The *p*-values indicated statistically significant differences in the average of the predictors between specific cohort pairs, providing insights into cohort-specific variations in predictor distributions in cognitive impairment analysis.

Figure 6.5 SHAP values plot for the *optimized PD-MCI* classification model in the cross-cohort analysis.



SHAP value plot displaying the top 15 predictors for the *optimized PD-MCI* model in cross-cohort prognostic classification. The plot shows the magnitude and direction (positive or negative) of each feature's influence on motor fluctuations prognosis status as output.

significantly correlated with *PRCD* but not with *PD-MCI*. This suggests that the association between thermoregulatory dysfunction and cognitive impairment may vary depending on the specific type of cognitive impairment.

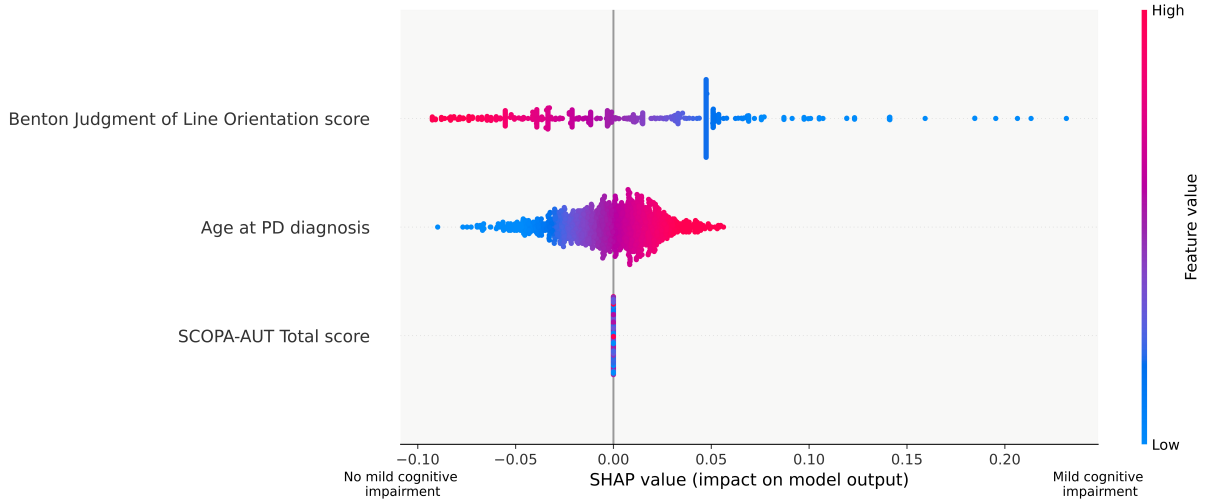
These findings provide valuable insights into the complex interplay of clinical features associated with objective and subjective cognitive impairment affecting daily living in PD patients. They highlight the necessity of considering the extensive range of factors in predicting cognitive decline in this population.

Table 6.15 Median conversion times and hazard ratios of the *optimized* time-to-*PD-MCI* model in the cross-cohort analysis.

Predictors	**Hazard Ratio (95% CI)**	**Median Conversion (95% CI)**	**Log-rank (*p*-values)**
Benton Judgment of Line Orientation score			
≥16	0.83 (0.66, 1)	4.09 (2.15, NR)	2.53E-04
<16		1.09 (1, 1.36)	
Age at PD diagnosis			
≥53	2.42 (1.16, 3.17)	1 (0.92, 1.13)	2.73E-10
<53		5.04 (3.53, NR)	

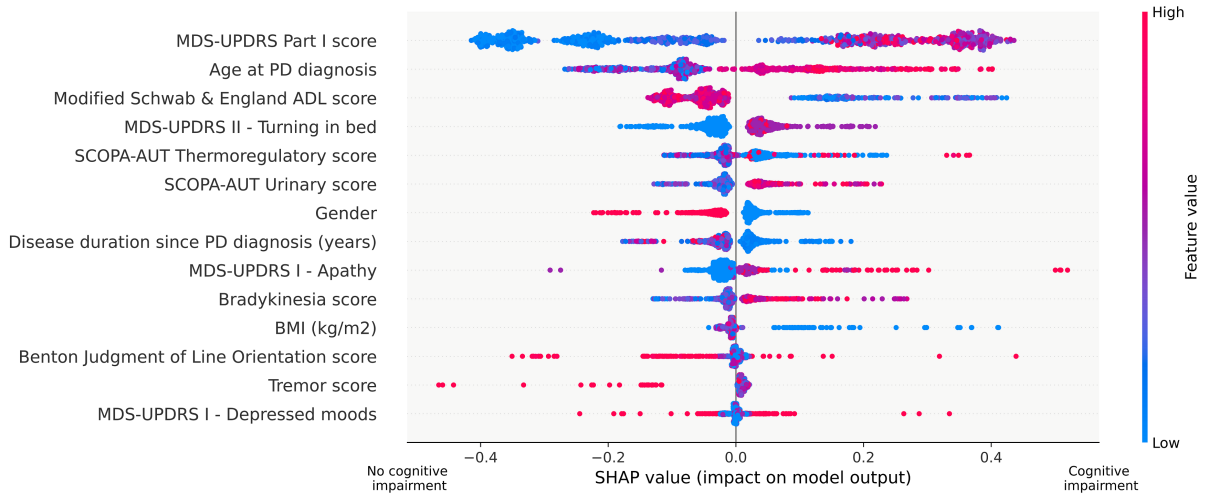
Summary of the hazard ratio (HR), median conversion time with 95% confidence interval (CI), and p-values from the log-rank test for the top 15 predictors identified in the time-to-*PD-MCI* model in the cross-cohort analysis. The HR provides insights into the risk associated with each predictor, while the median conversion time and log-rank test assess Kaplan-Meier (KM) curve differences between groups. "NR" (not reached) indicates that the MF event did not occur for some participants during the study period.

Figure 6.6 SHAP values plot for the *optimized* time-to-*PD-MCI* model in the cross-cohort analysis.



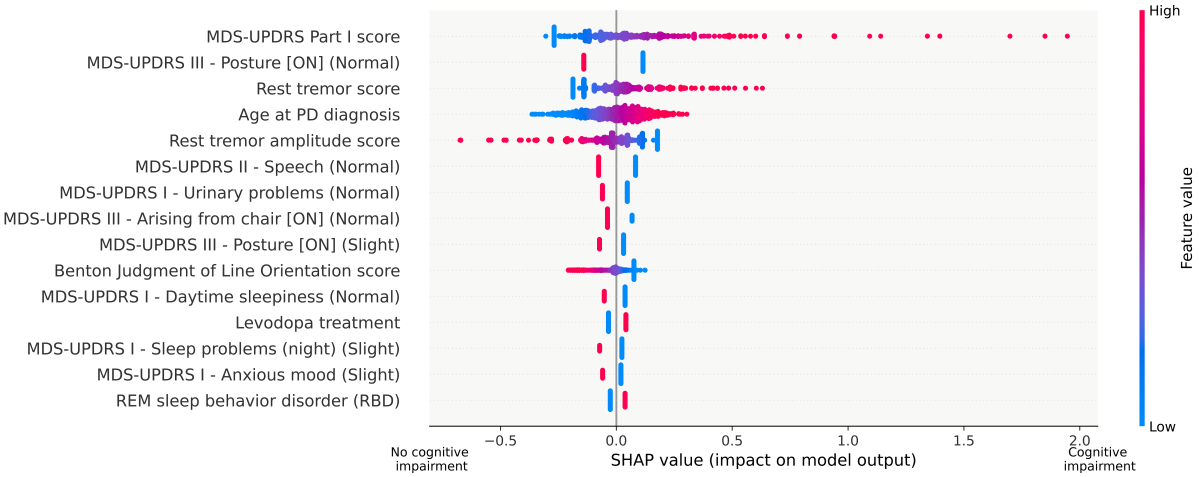
SHAP value plot displaying the top 15 predictors for the *optimized PD-MCI* model in cross-cohort time-to-*PD-MCI* analysis. The plot shows the magnitude and direction (positive or negative) of each feature's influence on time-to-*PD-MCI* as output.

Figure 6.7 SHAP values plot for the *optimized PRCD* classification model in the cross-cohort analysis.



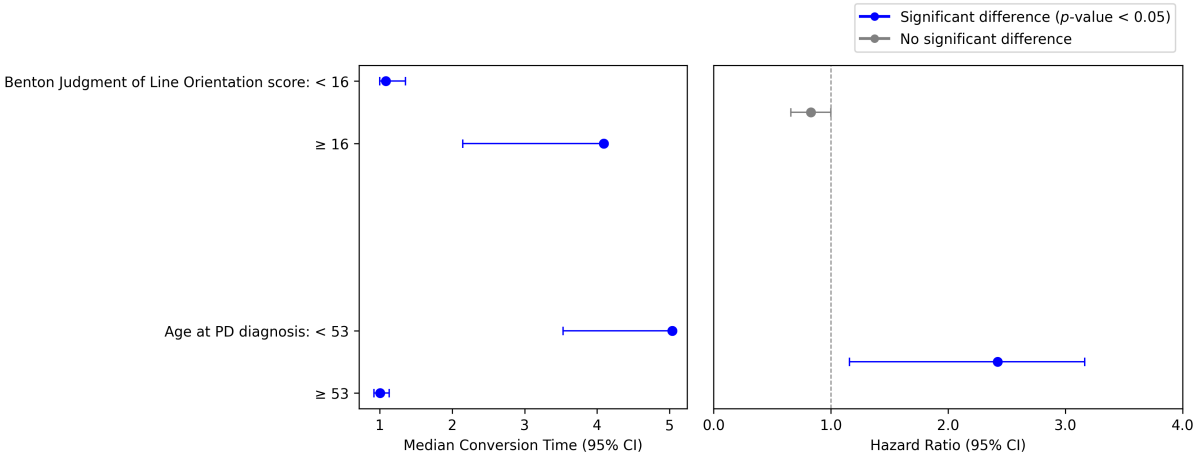
SHAP value plot displaying the top 15 predictors for the *optimized PRCD* model in cross-cohort prognostic classification. The plot shows the magnitude and direction (positive or negative) of each feature's influence on motor fluctuations prognosis status as output.

Figure 6.8 SHAP values plot for the *optimized* time-to-*PRCD* model in the cross-cohort analysis.



SHAP value plot displaying the top 15 predictors for the *optimized* time-to-*PRCD* model in cross-cohort analysis. The plot shows the magnitude and direction (positive or negative) of each feature's influence on time-to-*PRCD* as output.

Figure 6.9 Forest plot of median conversion times and hazard ratios for the *optimized* time-to-*PD-MCI* in cross-cohort analysis.



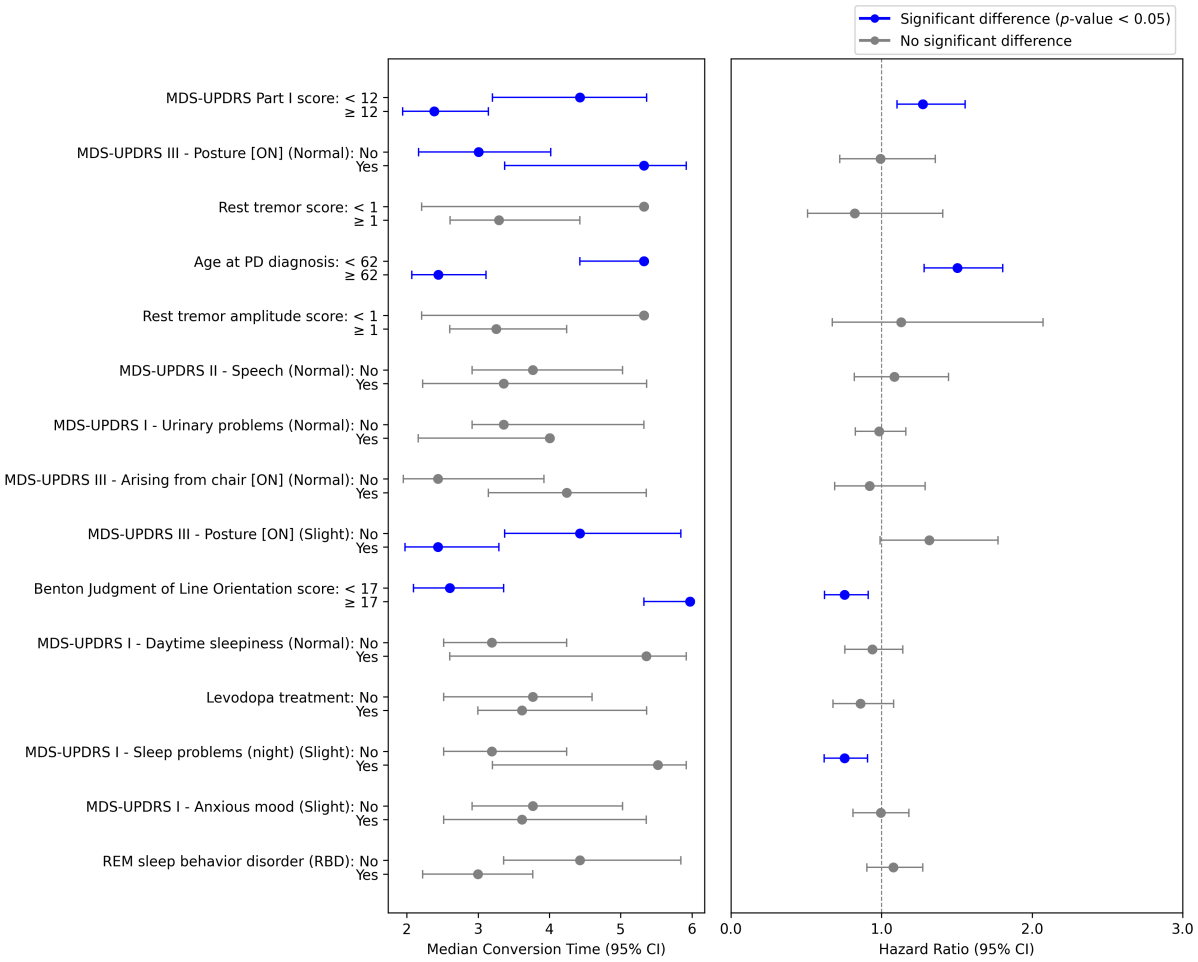
The plot illustrates how multiple clinical predictors affect *PD-MCI*. The left panel shows the median time to *PD-MCI* onset with 95% confidence intervals (CI), comparing subgroups for each predictor. The right panel displays corresponding hazard ratios (HR) with 95% CI, where HR > 1 indicates increased risk. Solid blue lines indicate statistically significant differences between groups (p-value < 0.05), while grey lines indicate non-significant differences.

Table 6.16 Median conversion times and hazard ratios of the *optimized* time-to-*PRCD* model in the cross-cohort analysis.

Predictors	**Hazard Ratio (95% CI)**	**Median Conversion (95% CI)**	**Log-rank (*p*-values)**
MDS-UPDRS Part I score			
$\geq$12	1.25 (1.01, 2.04)	2.38 (1.94, 3.15)	6.74E-03
<12		4.43 (3.2, 5.36)	
Age at PD diagnosis			
$\geq$62	2.05 (1.5, 2.87)	2.44 (2.07, 3.11)	1.87E-05
<62		5.33 (4.43, NR)	
Modified Schwab & England ADL score			
$\geq$78	0.53 (0.31, 0.98)	4.6 (3.61, 5.52)	1.04E-05
<78		2.09 (1.19, 2.38)	
MDS-UPDRS Part II score			
$\geq$12	1.01 (0.87, 1.5)	2.6 (2.07, 3.2)	6.39E-03
<12		4.6 (3.29, 5.84)	
SCOPA-AUT Gastrointestinal (GI) score			
$\geq$7	1.25 (0.96, 2.34)	2.06 (1.13, 2.43)	7.50E-06
<7		4.28 (3.36, 5.36)	
MDS-UPDRS I - Daytime sleepiness (Mild)			
Yes	1 (1, 1.3)	2.92 (2.16, 4)	8.75E-02
No		4.44 (3.2, 5.36)	
MDS-UPDRS Part III score (ON)			
$\geq$22	1.34 (1, 1.87)	2.68 (2.15, 3.77)	1.42E-04
<22		5.35 (4.76, NR)	
Dyskinesia score			
$\geq$1	1 (0.76, 1.37)	5.33 (2.22, NR)	4.01E-01
<1		3.61 (3, 4.44)	
Rigidity lower extremities score			
$\geq$4	1.16 (0.99, 1.52)	5.02 (1.95, NR)	2.84E-01
<4		3.37 (3, 4.44)	
MDS-UPDRS I - Pain and other sensations (Mild)			
Yes	1 (1, 1.23)	2.6 (1.32, 3.37)	7.24E-02
No		4.02 (3.01, 5.35)	
BMI (kg/m^2)			
$\geq$25	1.07 (0.86, 1.36)	3.19 (2.43, 4.76)	2.82E-01
<25		4.24 (3.11, 5.84)	
SCOPA-AUT Total score			
$\geq$16	1.36 (1, 2.3)	2.21 (1.32, 3)	1.44E-03
<16		4.44 (3.36, 5.52)	

Summary of the hazard ratio (HR), median conversion time with 95% confidence interval (CI), and p-values from the log-rank test for the top 15 predictors identified in the time-to-*PRCD* model in the cross-cohort analysis. The HR provides insights into the risk associated with each predictor, while the median conversion time and log-rank test assess Kaplan-Meier (KM) curve differences between groups. "NR" (not reached) indicates that the MF event did not occur for some participants during the study period.

Figure 6.10 Forest plot of median conversion times and hazard ratios for the *optimized* time-to-*PRCD* in cross-cohort analysis.



The plot illustrates how multiple clinical predictors affect *PRCD*. The left panel shows the median time to *PRCD* onset with 95% confidence intervals (CI), comparing subgroups for each predictor. The right panel displays corresponding hazard ratios (HR) with 95% CI, where HR > 1 indicates increased risk. Solid blue lines indicate statistically significant differences between groups (p-value < 0.05), while grey lines indicate non-significant differences.

Table 6.17 Correlation between predictors and *PD-MCI/PRCD* outcomes in cross-cohort analysis.

	Mild cognitive impairment (*PD-MCI*)		**Patient-reported cognitive decline (*PRCD*)**	
Predictors	**Correlation**	***p*-values**	**Correlation**	***p*-values**
Age at PD diagnosis	0.31	2.31E-30	0.24	8.33E-15
Disease duration since PD diagnosis (years)	0.14	1.36E-06	0.19	2.84E-10
Gender	0.04	2.20E-03	0.04	3.94E-03
Levodopa treatment	0.21	1.30E-13	0.25	4.58E-16
Weight (kg)	0.06	4.44E-02	0.02	5.44E-01
Height (cm)	-0.11	1.89E-04	-0.05	9.76E-02
BMI (kg/m^2)	0.14	3.51E-07	0.05	8.49E-02
Hoehn & Yahr stage	0.20	1.70E-12	0.29	2.65E-22
MOCA score (adjusted for education)	-0.47	7.91E-41	-0.34	1.86E-15
Benton Judgment of Line Orientation (JLO)	-0.26	4.00E-19	-0.30	3.88E-21
REM sleep behavior disorder (RBD)	0.11	1.74E-04	0.20	8.05E-10
Initial motor symptom - Resting tremor	-0.06	2.43E-02	-0.13	6.26E-05
Initial motor symptom - Rigidity or bradykinesia	-0.09	1.21E-03	-0.10	9.71E-04
Axial symptoms	0.26	2.43E-20	0.37	1.10E-35
Selective axial symptoms	0.24	3.08E-18	0.33	2.66E-28
Freezing of gait	0.14	3.95E-07	0.19	4.67E-10
Tremor	0.08	2.23E-02	0.08	3.91E-02
Rest tremor	0.09	1.30E-02	0.04	2.70E-01
Rest tremor amplitude	0.11	3.53E-03	0.05	1.82E-01
Rigidity upper extremities	0.17	2.80E-06	0.23	3.67E-09
Rigidity lower extremities	0.16	1.07E-05	0.19	8.32E-07
Total rigidity	0.20	2.57E-08	0.25	2.30E-10
Bradykinesia	0.26	8.53E-13	0.35	2.98E-20
Dyskinesia	0.03	3.77E-01	0.04	2.99E-01
Motor fluctuations	0.08	1.70E-02	0.09	1.94E-02
MDS-UPDRS Part I score	0.21	1.38E-14	0.53	6.85E-77
MDS-UPDRS Part II score	0.21	1.46E-14	0.40	1.46E-41
MDS-UPDRS Part III score (ON)	0.28	4.75E-15	0.35	8.32E-20
SCOPA-AUT Gastrointestinal (GI)	0.22	1.85E-15	0.35	9.77E-32
SCOPA-AUT Urinary	0.12	1.90E-05	0.29	5.41E-22
SCOPA-AUT Cardiovascular	0.09	1.33E-03	0.26	1.41E-17
SCOPA-AUT Thermoregulatory	0.02	5.22E-01	0.18	8.58E-09
SCOPA-AUT Sexual dysfunction	0.04	2.61E-01	0.11	1.29E-02
SCOPA-AUT Total score	0.21	3.98E-14	0.43	6.27E-47
Family history of PD	-0.08	1.09E-02	-0.13	2.56E-05
Pathogenic LRRK2 variant	-0.14	8.55E-06	-0.22	2.53E-11
Pathogenic GBA variant	-0.02	5.76E-01	-0.01	8.12E-01

The correlation of predictors with *PD-MCI* and *PRCD* outcomes was measured using the point biserial correlation for continuous or ordinal predictors and the Matthews correlation coefficient (MCC) for the binary predictor.

Table 6.18 Correlation analysis results for *PD-MCI* predictors in cross-cohort analysis.

Predictors	Age at PD diagnosis	MoCA score	Disease duration	MDS-UPDRS Part I	MDS-UPDRS Part II	MDS-UPDRS Part III (ON)	SCOPA-AUT total score
MoCA score	-0.31 (9.1E-17)						
Benton JLO	-0.13 (8.3E-06)	0.28 (7.5E-12)					
Disease duration	-0.19 (1.5E-11)	-0.15 (9.8E-05)	-0.16 (5.9E-08)				
MDS-UPDRS Part I	0.03 (3.0E-01)	-0.14 (3.0E-04)	-0.28 (2.6E-21)	0.32 (7.2E-31)			
MDS-UPDRS Part II	0.02 (3.9E-01)	-0.23 (5.4E-10)	-0.21 (2.9E-12)	0.36 (4.0E-39)	0.57 (3.2E-108)		
MDS-UPDRS Part III (ON)	0.17 (2.1E-06)	-0.37 (8.9E-22)	-0.34 (1.2E-18)	0.23 (2.5E-10)	0.39 (7.0E-28)	0.51 (2.9E-49)	
SCOPA-AUT total score	0.15 (9.8E-08)	-0.14 (2.6E-04)	-0.25 (2.8E-17)	0.36 (3.5E-39)	0.68 (3.9E-164)	0.55 (1.3E-96)	0.39 (1.5E-26)

Correlation analysis of *PD-MCI* predictors used Spearman correlation for two continuous/ordinal variables, point biserial correlation for continuous/ordinal and binary variables, and Matthews correlation coefficient (MCC) for two binary variables. The correlation coefficients are presented with the p-values in brackets.

6.4.5 Decision curve and calibration analysis

A DCA and a calibration analysis were conducted to evaluate the developed models' reliability and potential clinical utility for predicting *PD-MCI* and *PRCD* outcomes.

In the context of *PD-MCI* classification, the *optimized* AdaBoost model showed a high area under the net benefit curve (AUNBC=0.23), thereby indicating that the model offers considerable value in guiding decisions regarding potential preventive or early therapeutic interventions targeting cognitive decline. The calibration analysis for this model yielded a calibration slope of 1.13 (see Figure 6.11 and Table 6.19), reflecting a strong alignment between predicted probabilities and observed outcomes. This indicates that the model's predictions for *PD-MCI* are well-calibrated, offering clinicians reliable risk estimates.

In the time-to-*PD-MCI* analysis, the *optimized* penalized Cox model also demonstrated a high AUNBC of 0.22 (Figure 6.12). The calibration slope for this model was 1.30, indicating a slight tendency to overestimate risk at higher probability levels. Nevertheless, the calibration demonstrated reasonable concordance between predicted and observed time-to-*PD-MCI* events.

In terms of *PRCD* classification, the FIGS model showed the most optimal calibration slope (0.74), thereby indicating a superior degree of precision in aligning the predicted and observed probabilities compared to the other models. Nevertheless, alternative models attained a superior AUNBC (0.08, as illustrated in Figure 6.13).

The time-to-*PRCD* analysis demonstrated significant challenges with model calibration across all approaches. Although the model achieved a satisfactory AUNBC (0.10, Figure 6.14), its calibration slopes were observed to be relatively low. This indicates that although the models can differentiate between high- and low-risk patients, they cannot accurately predict the absolute risk of *PRCD* onset. The discrepancy between the good discrimination indicated by the AUNBC and the suboptimal calibration emphasizes the challenges in achieving reliable model predictions for this particular outcome.

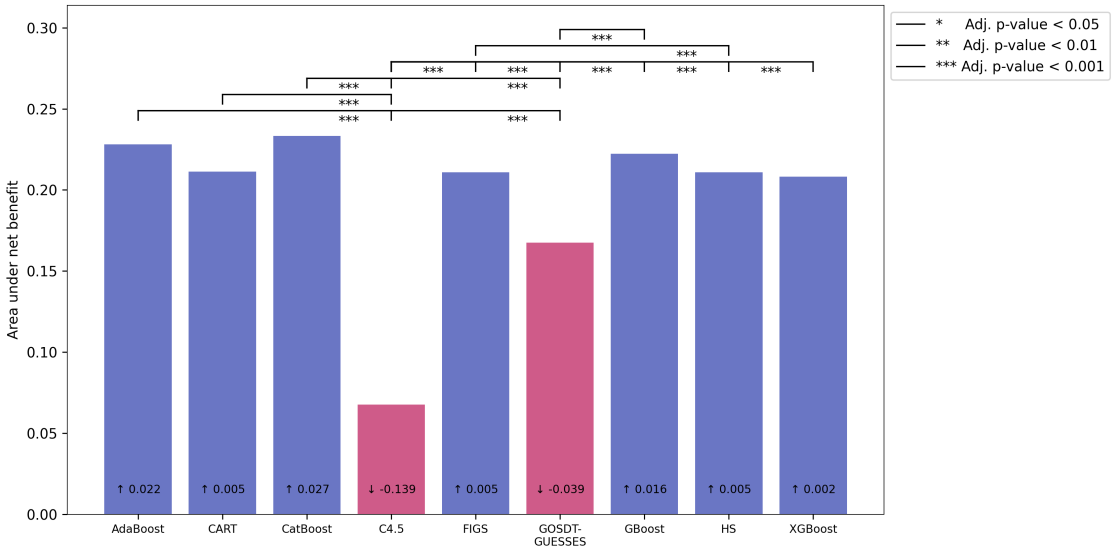
These findings collectively highlight the necessity of considering discrimination and calibration when evaluating model performance. While the models demonstrated high AUNBC values overall, the disparate calibration outcomes across different outcomes highlight the necessity for careful interpretation of model predictions in clinical practice. In particular, the findings indicate that the models developed for classifying and predicting *PD-MCI* could offer valuable assistance in clinical decision-making. In contrast, the *PRCD* models, particularly those used for time-to-event prediction, may require further refinement to improve their calibration and reliability for future clinical applications.

6.5 Discussion

This study used a comprehensive multi-cohort approach to identify shared predictors of *PD-MCI* and *PRCD* in PD. By integrating data from diverse cohorts, we demonstrated that specific clinical features, such as age at diagnosis, consistently predict both outcomes. This integrative strategy, which incorporates diverse input data and two alternative cognitive impairment outcomes, improves the generalizability of findings and provides robust insights into cognitive decline across different PD populations.

A key strength of the study is its cross-cohort analysis, which reduces the potential for cohort-specific biases often present in single-cohort studies. This approach yielded more stable predictive models and extended the applicability of findings to diverse PD populations. The application of cross-

Figure 6.11 Bar plot of the area under the positive net benefit curve for the *optimized PD-MCI* classification models in cross-cohort analysis.



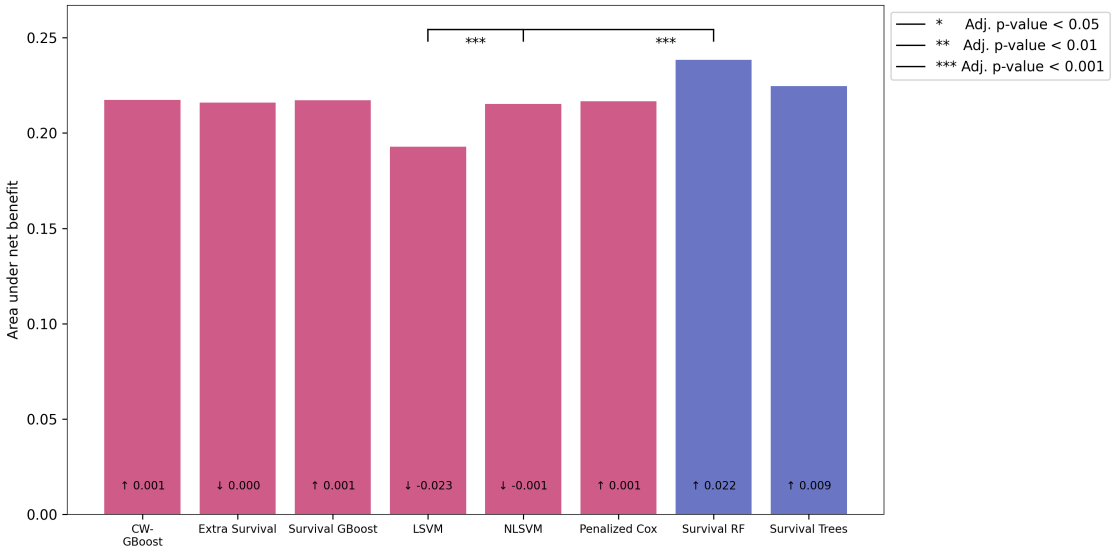
The bar plot shows the area under the positive net benefit for different cross-cohort *optimized PD-MCI* classification models, with the lines above the bars indicating significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

study normalization, particularly within the leave-one-cohort-out validation framework, improves predictive performance, although its effectiveness was observed to vary across different settings. This variability highlights the need to explore normalization strategies further to ensure robust predictions across diverse clinical contexts.

The age at diagnosis was identified as a significant predictor in the *PD-MCI* and *PRCD* analyses. Patients who were older at the time of onset were found to face a markedly higher risk of cognitive decline compared to younger individuals, consistent with previous research indicating a correlation between late-onset PD and more pronounced cognitive dysfunction [172]. Time-to-event analysis revealed a significant association between age at diagnosis and the risk of cognitive impairment in both *PD-MCI* and *PRCD*. Patients with late-onset PD were almost twice as likely to develop cognitive impairment as those with earlier disease onset. This association may reflect the increased vulnerability of age-related brain networks to neurodegenerative processes [198]. Nevertheless, early- and late-onset PD show altered functional connectivity within brain networks [199], particularly in domains associated with cognitive functions such as attention, executive function, and memory.

Furthermore, our study identified significant sex-based differences in *PRCD*, with women demonstrating superior global cognitive performance and a lower prevalence of cognitive impairments during clinical assessments [3, 200]. The Benton JLO test significantly predicted cognitive impairment, particularly visuospatial deficits. This finding aligns with previous studies highlighting the role of visuospatial

Figure 6.12 Bar plot of the area under the positive net benefit curve for the *optimized* time-to-*PD-MCI* models in cross-cohort analysis.



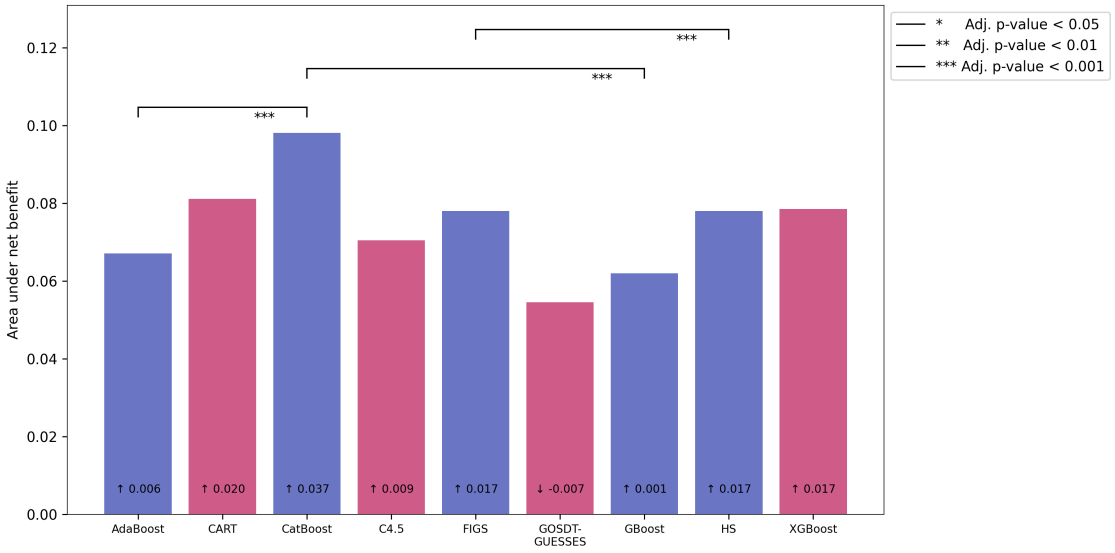
The bar plot shows the area under the positive net benefit for different cross-cohort *optimized* time-to-*PD-MCI* models, with the lines above the bars indicating significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

function in the progression of cognitive decline in PD [2, 63, 191]. Women generally outperformed men on global cognition [201, 202], whereas men performed better on visuospatial tasks [201, 203]. Hormonal differences between men and women are hypothesized to affect brain function and may contribute to the observed sex differences in cognition [202, 204]. This highlights the importance of considering gender-specific cognitive strengths and weaknesses when assessing and managing cognitive impairment in PD.

Non-motor symptoms, including autonomic dysfunction assessed by SCOPA-AUT, correlated with *PRCD*, highlighting the multifactorial nature of the cognitive decline in PD [14]. Among these, gastrointestinal symptoms such as constipation have been explicitly associated with cognitive impairment [197, 205]. Moreover, the established association between autonomic dysfunction and PD progression [206] emphasizes the complex interplay of systems impacted by PD pathology, thereby contributing to the complexity of cognitive decline.

Moreover, sleep disturbances may also impact cognitive impairment, as they can exacerbate cognitive difficulties, such as impaired memory processing, which is important for maintaining cognitive function [207]. In our analyses, sleep problems at night, as assessed in the MDS-UPDRS Part I, were identified as one of the predictors with a significant hazard ratio in the time-to-*PRCD* model. Patients with sleep disorders frequently report increased difficulties with attention, memory, and problem-solving [208]. Improving sleep quality through targeted interventions, such as cognitive behavioral therapy for

Figure 6.13 Bar plots of the area under the positive net benefit curve for the *optimized PRCD* classification models in cross-cohort analysis.



The bar plot shows the area under the positive net benefit for different cross-cohort *optimized PRCD* classification models, with the lines above the bars indicating significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

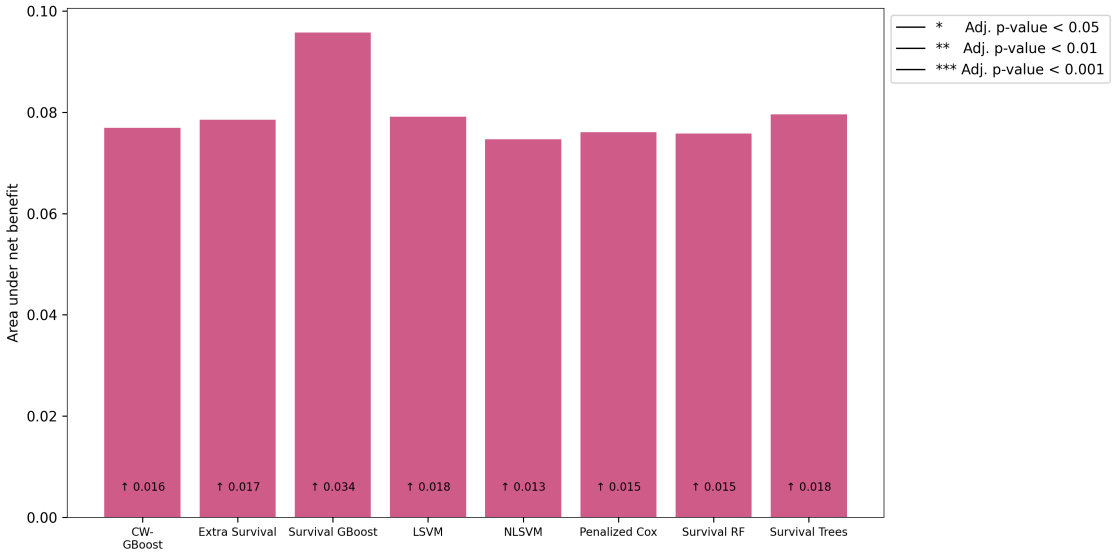
insomnia [209], may mitigate perceived cognitive impairment and improve overall cognitive health in PD.

Our study's application of ML models facilitated the evaluation of predictive performance and the discovery of distinctive clinical features associated with cognitive impairment in greater depth than traditional statistical approaches. Using advanced analytical techniques across multiple cohorts, we identified robust predictors with potential applications in precision medicine trials, paving the way for more targeted and personalized interventions.

The results of our study have considerable implications for clinical practice and future research. Identifying consistent predictors across cohorts provides clinicians with practical tools for the early detection of patients at high risk of cognitive decline, enabling the implementation of timely interventions and more personalized management strategies. Furthermore, the observed sex differences in *PRCD* highlight the necessity of developing sex-specific approaches to cognitive assessment and tailored management strategies for PD, ensuring more equitable and effective care.

It is important to acknowledge the limitations of our study, which may influence the generalizability of our findings. Although a multi-cohort approach was used, the generalizability of the findings may be limited by the characteristics of the included cohorts. The variability in predictive performance observed across cohorts can be attributed to differences in sample sizes and disparities in patient profiles. For example, the LuxPARK cohort included participants with a significantly average age at diagnosis and

Figure 6.14 Bar plot of the area under the positive net benefit curve for the *optimized* time-to-*PRCD* classification models in cross-cohort analysis.



The bar plot shows the area under the positive net benefit for different cross-cohort *optimized* time-to-*PRCD* models, with the lines above the bars indicating significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

longer disease duration than the PPMI and ICEBERG cohorts. In contrast, participants in the ICEBERG cohort had significantly lower average body weight and BMI than those in the LuxPARK and PPMI cohorts. Moreover, the PPMI cohort showed less severe disease progression, as indicated by lower overall MDS-UPDRS Parts I to III and SCOPA-AUT scores compared to LuxPARK.

These discrepancies in baseline characteristics emphasize the necessity of considering cohort-specific variables when interpreting the predictive efficacy of *PD-MCI* and *PRCD* models, highlighting the challenge of developing applicable predictive tools for cognitive impairment in PD. Moreover, while the models showed satisfactory predictive performance despite the challenges of cross-cohort prediction, their clinical utility requires further optimization and validation in prospective studies to ensure robust and reliable application in diverse clinical settings.

Overall, this study highlights the association of non-motor symptoms and sex-related differences on the progression of cognitive decline in PD. By using a multi-cohort approach on two outcome measures, as well as advanced ML techniques, we were able to identify robust predictors of cognitive impairment, thereby enhancing our understanding of its underlying mechanisms. These findings provide a foundation for improved risk stratification and the development of personalized interventions for individuals with PD. It is recommended that future research prioritize validating these predictors in more diverse clinical settings and investigating their potential to inform treatment decisions and optimize patient management strategies.

Table 6.19 Calibration analysis for *PD-MCI* and *PRCD* analyses.

Classification					Time-to-event				
Algorithm	***PD-MCI***		***PRCD***		**Algorithm**	***PD-MCI***		***PRCD***	
	Slope	**MSE**	**Slope**	**MSE**		**Slope**	**MSE**	**Slope**	**MSE**
AdaBoost	1.13	0.23	4.17	0.23	CW-GBoost	0.51	0.23	0.24	0.05
CART	0.53	0.25	0.57	0.23	Extra Survival	0.29	0.28	0.3	0.04
CatBoost	0.62	0.24	0.67	0.19	Survival GBoost	0.26	0.27	0.22	0.1
C4.5	0.15	0.49	0.3	0.35	LSVM	0.06	0.04	0.36	0.03
FIGS	0.51	0.26	0.74	0.21	NLSVM	0.1	0.03	0.39	0.03
GOSDT-GUESSES	0.25	0.37	0.27	0.33	Penalized Cox	1.3	0.22	0.28	0.03
GBoost	0.8	0.24	2.09	0.23	Survival RF	0.06	0.05	0.28	0.04
HS	0.51	0.26	0.74	0.21	Survival Trees	0.1	0.34	0.2	0.08
XGBoost	0.45	0.28	0.55	0.24					

Calibration analysis for *PD-MCI* and *PRCD* models in both classification and time-to-event for cross-cohort analysis with the calibration slope and mean square error (MSE).

6.6 Summary and conclusions

This study aimed to identify the predictors of cognitive decline in PD by utilizing objective and subjective cognitive impairment measures. Our findings emphasize the significance of non-motor symptoms, particularly autonomic dysfunction, in the progression of cognitive decline and clarify the existence of notable sex-based discrepancies in cognitive outcomes. Our findings indicate that older age at diagnosis is associated with an increased risk of cognitive impairment, which is consistent with previous research that has linked late-onset PD with more severe cognitive dysfunction. Furthermore, visuospatial deficits were identified as a key predictor of cognitive decline, thereby reinforcing their significance in PD-related cognitive dysfunction.

Applying ML models across multiple cohorts enabled the identification of a range of clinical features that predict cognitive impairment. These findings highlight the potential of ML to improve predictive accuracy and enable more precise risk stratification for cognitive decline in PD. The models demonstrated reliable performance in identifying high-risk patients, thereby offering the potential for early interventions and more personalized treatment strategies in clinical practice.

However, the study also identified cohort-specific factors impacting the model's predictive performance. When interpreting the results and developing predictive tools with universal applicability, it is important to consider the impact of patient characteristics, including age at diagnosis, disease duration, and symptom severity. Further optimization and prospective validation of these models are required to improve their clinical utility and generalizability.

In conclusion, our study provides valuable insights into the predictors of cognitive decline in PD and emphasizes the importance of considering non-motor symptoms and sex differences in clinical assessments. ML methodologies have facilitated a more profound comprehension of cognitive decline in PD, thereby paving the way for future investigations into risk prediction, early diagnosis, and tailored treatment regimens.

6.7 Contribution statement

This study was a collaborative effort, with contributions from all authors in various aspects of the research and manuscript preparation.

Rebecca Loo Ting Jiin: Conducted the study as the first author, developed the methodology, conducted analysis, created visualizations, validated results, and drafted the original manuscript.

Enrico Glaab: He supervised the study as the corresponding author for the submitted manuscript, contributed to the study's correction, methodology, and investigation, guided the project, acquired funding, and reviewed and edited all sections of the manuscript.

Lukas Pavelka, Graziella Mangone, Fouad Khoury, Marie Vidailhet, and Jean-Christophe Corvol: Contributed by reviewing and editing the manuscript, providing insights from clinical perspectives.

Chapter 7

Conclusions and perspectives

This chapter provides a comprehensive summary of the key findings from the analyses of levodopa-induced dyskinesia (*LID*), motor fluctuations (*MF*), mild cognitive impairment (*PD-MCI*), and patient-reported cognitive decline (*PRCD*) in Parkinson's disease (PD). The primary objective of this research was to identify predictors of these complications across multiple cohorts, using a combination of clinical features and advanced machine learning techniques to ensure robust and generalizable insights.

LID is a well-documented complication of long-term levodopa therapy in PD, often manifesting as involuntary movements that emerge after prolonged treatment. *MF*, another complication in PD, has been hypothesized in previous research to be associated with extended levodopa use. Prior studies have analyzed *LID* and *MF* independently, providing valuable insights into their respective predictors. In this study, we adopt an independent analytical approach for *LID* and *MF* to improve our understanding of their respective associations and determine the distinctiveness or overlap of the factors associated with each. The correlation between *LID* and *MF* highlights the complex interrelationship between these two PD complications.

In addition, both mild cognitive impairment in PD (*PD-MCI*) and patient-reported cognitive decline (*PRCD*) were analyzed independently to explore their distinct predictors and potential overlap. *PD-MCI*, measured using standardized cognitive tools such as the MoCA, represents an objective assessment of cognitive impairment. In contrast, *PRCD* reflects the subjective cognitive concerns reported by patients, often assessed through the MDS-UPDRS Part I, which captures cognitive difficulties as experienced in daily life. By analyzing these outcomes independently, we aim to gain a deeper understanding of the factors associated with both objective and subjective cognitive dysfunction in PD and assess whether the same predictors are involved in both types of cognitive decline.

The age at which PD diagnosis is observed represents a key predictor across both motor and cognitive complications examined in this study. However, the observed association differs between these domains. Earlier onset of PD is significantly correlated with an increased likelihood of developing motor complications, including *LID* and *MF*. This may be influenced by the longer disease duration observed in younger-onset patients, which provides a prolonged window of opportunity for the emergence of motor complications. In contrast, later-onset PD is associated with an increased risk of cognitive impairments, including *PD-MCI* and *PRCD*, which may reflect age-related vulnerabilities to cognitive decline. These contrasting patterns show how age at PD diagnosis influences the course of PD, with younger patients predominantly experiencing motor complications and challenges, while older individuals are more likely

to develop cognitive decline. These findings emphasize the importance of developing management strategies that address the distinct risks faced by different patient groups in PD.

Disease duration, a key marker of PD progression, not only correlates with *LID* and *MF* but is also associated with other predictors, including levodopa equivalent daily dose (LEDD), motor impairments, and non-motor symptoms. Furthermore, an association between disease duration and *PRCD* was identified, whereby patients with longer disease durations are more likely to report cognitive decline that impacts their daily functioning, highlighting the impact of disease progression on both motor and cognitive domains in PD. Although previous studies have indicated a correlation between *LID* and *MF* with prolonged levodopa treatment, this study produced contradictory results in *MF* analysis. Specifically, although levodopa medication intake was identified as one of the top predictors for *MF*, its direct association with *MF* was not statistically significant. This finding aligned with the results of a recent clinical trial, which indicated that *MF* may be more closely associated with disease progression than with levodopa therapy.

Motor dysfunctions, including axial symptoms, freezing of gait, and rigidity, were positively correlated with *LID* and *MF*. These findings suggest that these motor symptoms serve as important clinical markers for predicting the occurrence of these complications. Moreover, motor and non-motor disabilities, as assessed by the MDS-UPDRS and SCOPA-AUT, provide a comprehensive evaluation of motor, non-motor, and autonomic dysfunction. The observed associations between these assessments and both motor complications (*LID* and *MF*), as well as subjective cognitive decline (*PRCD*), highlight the multifactorial nature of PD, where motor and non-motor symptoms often coexist and contribute to the overall disease burden.

In contrast, resting tremor was associated with a reduced risk of *LID* and *MF*, likely reflective of the relatively mild disease progression typically observed in tremor-dominant patients. These patients tend to experience a lower risk of motor complications. Furthermore, resting tremor was associated with *PD-MCI* but not with *PRCD*, suggesting that while tremor may be related to cognitive function, its association differs from other forms of cognitive decline observed in PD. Visuospatial function, a cognitive domain involved in spatial attention and navigation, was linked not only with *LID* and *MF* but also with cognitive decline and was found to be associated with the progression of motor disabilities. These findings highlight the interconnectivity of motor and cognitive impairments in PD and their potential influence on disease progression.

Assessing non-motor symptoms is also important in predicting the emergence of *LID* and *MF*. Gastrointestinal, urinary, and thermoregulatory symptoms were associated with increased PD severity and a higher risk of *LID*, highlighting the complex relationship between motor and non-motor dysfunctions in disease progression. Moreover, non-motor symptoms play a role in cognitive decline in patients with PD. Notably, constipation was associated with *PD-MCI*, while gastrointestinal, urinary, and thermoregulatory symptoms were associated with *PRCD*, highlighting the impact of non-motor symptoms on the disease's motor and cognitive aspects.

Furthermore, body weight was identified as a predictor, with lower body weight being associated with a higher risk of *LID*. This inverse relationship may be attributed to pharmacokinetic differences or metabolic influence on levodopa metabolism. In the cross-cohort analysis, BMI and body weight were positively associated with *PD-MCI* but not *PRCD*. This finding highlights the distinct relationships between motor and cognitive outcomes in PD.

Despite common predictors between motor complications (*LID* and *MF*) and cognitive impairment

(*PD-MCI* and *PRCD*), including disease duration and non-motor symptoms, no significant correlation was identified between *LID* and cognitive outcomes. A weak correlation was observed between *MF* and cognitive outcomes. These findings highlight PD complications' complex and multifaceted nature and the overlapping yet distinct predictors that influence motor and cognitive outcomes. These findings highlight the necessity for targeted therapeutic approaches that consider specific predictors when assessing the risk of different outcomes. Such approaches must address motor and cognitive symptoms while accounting for their unique and shared determinants.

Genetic factors contribute to a comprehensive understanding of the variability of *LID* and *MF* risks. *GBA* mutations were associated with an earlier onset of both *LID* and *MF*, which is likely reflective of the more aggressive disease progression commonly observed in these patients. In contrast, pathogenic *LRRK2* mutations were also linked to *MF*, although with a smaller effect size than *GBA* mutations, indicating that the genetic influence on these complications is not identical.

Furthermore, sex-based differences were identified in the progression of *PRCD*, with women demonstrating superior performance on cognitive tests and reporting less cognitive impairment that interferes with daily functioning. This suggests the existence of potential sex-based differences in both the cognitive manifestations and perceptions of cognitive decline in PD.

Incorporating multi-cohort analysis, which integrates data from different cohorts and expands the sample size, substantially improves the stability and generalizability of the predictive model. Pooling data from diverse patient populations reduces variability and increases the robustness of the model, thereby ensuring more reliable predictions across different groups. The results demonstrated that, despite the added challenges of cross-cohort analysis, such as potential cohort-specific biases and differences in clinical feature characteristics, the predictive performance was comparable to that of single-cohort analyses. This illustrates the capacity of cross-cohort methodologies to preserve predictive capability, even when there is an increase in complexity.

Moreover, cross-study normalization was a practical approach for improving predictive performance, particularly when cohort variability could impact the analysis. This normalization technique often led to a notable improvement in model performance by accounting for inter-cohort differences, resulting in more consistent and reliable outcomes. However, cross-study normalization did not consistently improve performance across all cases, suggesting that its effectiveness may depend on the cohorts' specific characteristics and data patterns.

Furthermore, cross-cohort analysis helps to reduce cohort-specific biases that frequently emerge in single-cohort studies due to sample homogeneity or other study-specific factors. By incorporating data from multiple cohorts, this approach helps to ensure that the findings are independent of the distinctive characteristics of a single cohort, thereby improving the generalizability of the results. Consequently, cross-cohort analysis offers a more comprehensive understanding of predictive factors, applicable across a broader range of patient populations.

7.1 Limitations

This thesis offers valuable insights into the predictors associated with levodopa-induced dyskinesia (*LID*), motor fluctuations (*MF*), mild cognitive impairment in PD (*PD-MCI*), and patient-reported cognitive decline (*PRCD*) in Parkinson's disease (PD) by applying state-of-the-art machine learning (ML) techniques and a multi-cohort approach. However, it is important to acknowledge several limitations

inherent to the study. These limitations stem from methodological challenges, PD's complexities, and the included cohorts' characteristics. Addressing these limitations is important for interpreting the findings and guiding future research that builds upon this work. The key limitations of the thesis are discussed below.

1. Cohort-specific biases and generalizability
Despite using multiple cohorts to reduce cohort-specific biases, the generalizability of the findings remains a challenge. The cohorts included in this thesis demonstrate variations in demographic and clinical characteristics, including sample size, disease severity, and age at disease onset. For example, discrepancies in disease duration and body mass index (BMI) were observed across cohorts, which may have influenced the patterns and predictors identified in the analyses. These differences may restrict the generalizability of the findings to populations outside the studied cohorts, particularly those with different demographic or clinical profiles. Consequently, while the multi-cohort approach improves the robustness of the findings in comparison to single-cohort studies, it is possible that the diversity of PD presentations globally may not be fully captured. This limitation highlights the need for the results to be validated in more diverse cohorts to ensure the identified predictors are robust across unrepresented or underrepresented patient populations.

2. Potential influence of unmeasured variables
The availability of variables across the different cohorts limited the analyses presented in this thesis. In cross-cohort analyses, certain variables may be excluded when these variables are only present in a specific cohort, resulting in the exclusion of these variables from the analysis across the cohort, such as medications, comorbid conditions, physical activity levels, and clinical health metrics. As a result, valuable information may be excluded, potentially reducing the models' predictive power and limiting the findings' scope. This limitation reflects the challenges of conducting cross-cohort analyses, where differences in data availability across studies can reduce the robustness and generalizability of the results, leaving out key predictors that could contribute to a deeper understanding of the mechanisms underlying PD complications. Therefore, the findings may not fully capture the complexity of contributing factors, which may limit their application in clinical decision-making or research contexts where these excluded variables are of particular relevance. It is recommended that future studies incorporate more factors to provide a more comprehensive understanding of the predictors involved in PD complications.

3. Focus on clinical features
The analyses presented in this study primarily focused on examining clinical features as potential predictors of complications associated with PD, with limited integration of multimodal data, including imaging and biomarkers, which could provide further insights. Although clinical features provide valuable insights into the progression of PD, including additional data types could offer a more comprehensive understanding of the underlying pathophysiology of PD complications. For example, neuroimaging data may reveal structural or functional alterations in the brain, while biomarkers could provide objective measures of disease progression or treatment response. The lack of such data in this study may restrict the depth of the findings and inhibit a more detailed comprehension of the complexities of PD. Analyzing clinical features alone may limit the comprehensiveness of predictive models, potentially excluding key predictors that could improve the accuracy and relevance of the results.

4. Heterogeneity in outcome definitions
Despite the consistent application of standardized tools across cohorts, intraindividual variability in outcome assessments may still occur due to inherent differences in clinical environments and cohort-specific practices [197, 210, 211]. Despite the use of uniform tools, variations in the measurement of symptoms or outcomes during patient evaluations may occur, reflecting differences in the individual patient's condition at the time of assessment. Factors such as varying patient populations or communication can influence how data is collected and recorded, potentially contributing to variations in measurement. Such variability in outcome assessments may introduce noise into the analysis, impacting the reliability of the prediction models. Discrepancies in assessment practices could reduce the consistency of identified predictors and their applicability to diverse populations. Therefore, it is recommended that these potential differences be addressed in future research through robust statistical methods or sensitivity to ensure more reliable and generalizable predictive findings.

7.2 Future works

This study has provided insights into the predictors of levodopa-induced dyskinesia, motor fluctuations, and cognitive impairment in PD. Nevertheless, several directions for further investigation could improve the robustness and applicability of the findings. Despite the progress that has been made, limitations such as cohort-specific biases, incomplete variable coverage, and the focus on clinical features highlight the need for future research to extend these analyses. Therefore, it is important to address these gaps to improve predictive models and ensure their broader clinical applicability, thereby advancing our understanding of PD complications and their management.

1. Expanding cohort diversity and enhancing generalizability and predictive power
The expansion of the number of cohorts in cross-cohort analysis has the potential to improve the generalizability and predictive capability of the findings. The incorporation of three cohorts has enabled this study to effectively address cohort-specific biases, thereby enhancing the stability and reliability of the model. Future research may benefit from incorporating additional cohorts with diverse demographic and clinical characteristics to evaluate how increasing cohort diversity improves predictive performance. Furthermore, incorporating additional cohorts may help mitigate overfitting in leave-one-cohort-out validation, thereby enhancing the robustness and external validity of the identified predictors. Such an expansion would facilitate a more comprehensive understanding of PD complications and ensure the findings apply to a broader range of patient populations.

2. Integration of additional variables
To gain a more comprehensive understanding of the complex complications of PD, future research should integrate a broader range of variables into predictive models. The current study primarily relied on clinical features available across the included cohorts. However, the exclusion of variables, such as medication regimens, comorbidities, physical activity levels, and other clinical health metrics, may have limited the scope of the findings. Integrating these additional variables would facilitate the development of more comprehensive models capable of capturing the complex interplay of factors influencing PD complications. Nevertheless, this goal can only be attained through a unified effort to standardize data

collection protocols across all cohorts. Inconsistencies in measuring and recording variables can limit their utility and increase variability in cross-cohort analyses. Standardizing the data collection, such as clinical and lifestyle data, has the potential to improve the quality of predictive models and increase their applicability across diverse PD populations. By addressing these challenges, future studies can construct robust predictive tools, offering valuable insights into the mechanisms of PD complications and informing more effective interventions.

3. Leveraging multimodal and longitudinal data to improve prediction

The advancement of predictive models for PD complications requires a shift in focus beyond solely clinical features to incorporate multimodal data. Integrating neuroimaging and biomarker data with clinical assessments may facilitate a more comprehensive and multidimensional understanding of the underlying pathophysiology of PD. Neuroimaging modalities, such as magnetic resonance imaging (MRI) or positron emission tomography (PET) scans, can provide insights into structural or functional alterations in the brain, including atrophy or alterations in connectivity. Biomarkers, such as cerebrospinal fluid proteins or blood-based markers, may reveal biochemical changes indicative of disease progression and treatment response. Combining these data types with clinical variables enables researchers to construct more detailed and accurate models that capture the interplay between the disease's biological, structural, and symptomatic aspects. Furthermore, integrating multiple data types can facilitate techniques like multi-view learning [212, 213] or data fusion algorithms [214, 215] to process and combine heterogeneous data sources. These methods facilitate the identification of complex, cross-modal patterns and associations that may be overlooked when analyzing each data type independently. For instance, correlating particular imaging characteristics with biomarker concentrations could reveal novel predictors of complications. Joint modeling of multimodal data may improve the robustness and generalizability of the findings. Moreover, longitudinal multimodal data could facilitate a dynamic understanding of disease development. Monitoring alterations in biomarkers, imaging, and clinical characteristics over time could improve predictive models by integrating disease trajectories, thereby facilitating the prediction of complications and their timing and progression. This approach could prove invaluable in identifying early indicators of complications, thereby enabling timely and personalized interventions tailored to the patient's evolving needs. By embracing multimodal and dynamic data integration, future research can develop more accurate predictive models that reflect PD progression's complex, multifaceted nature and associated complications.

4. Addressing heterogeneity in outcomes and predictors

To address the heterogeneity in outcome definitions and predictors, robust analytical approaches can ensure the consistency and reliability of findings despite the variations in assessment practices that may occur. Future research may benefit from using advanced statistical methodologies to accommodate intra-individual variability [216]. One promising approach is the application of hierarchical Bayesian models, which can explicitly account for cohort-specific differences by incorporating varying effects for each cohort [217]. This approach accommodates differences in individual and population, potentially mitigating noise introduced by these variations, thereby enhancing the stability and reliability of predictive models. Furthermore, mixed-effects models offer another option for addressing variability by introducing random effects that model intra- and inter-individual differences [218]. These models are beneficial for accounting for individual variability within cohorts and ensuring that predictions remain

robust. Researchers can evaluate the impact of heterogeneity on model performance and ensure that the findings remain consistent and reliable even when data variability is introduced. Incorporating these advanced analytical techniques will enable future studies to effectively address the challenges posed by heterogeneity in outcome definitions and predictors, thereby facilitating the development of more robust and generalizable insights into PD complications.

References

[1] H. You et al. “Molecular basis of dopamine replacement therapy and its side effects in Parkinson’s disease”. In: *Cell and Tissue Research* 373 (2018), pp. 111–135. DOI: 10.1007/s00441-018-2813-2.

[2] Y.X. Wang et al. “Associations between cognitive impairment and motor dysfunction in Parkinson’s disease”. In: *Brain and Behavior* 7.6 (2017), e00719. DOI: 10.1002/brb3.719.

[3] M. Picillo et al. “Sex-related longitudinal change of motor, non-motor, and biological features in early Parkinson’s disease”. In: *Journal of Parkinson’s Disease* 12.1 (2022), pp. 421–436. DOI: 10.3233/JPD-212892.

[4] B.E. Yeager et al. “Salience network and cognitive impairment in Parkinson’s disease”. In: *medRxiv Preprint* (2023), pp. 2023–10. DOI: 10.2139/ssrn.4608191.

[5] J. Ren et al. “Comparing the effects of GBA variants and onset age on clinical features and progression in Parkinson’s disease”. In: *CNS Neuroscience & Therapeutics* (2024). DOI: 10.1111/cns.14387.

[6] P. Filzmoser and K. Nordhausen. “Robust linear regression for high-dimensional data: An overview”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 13.4 (2021), e1524. DOI: 10.1002/wics.1524.

[7] M. Wang et al. “Dementia risk prediction in individuals with mild cognitive impairment: A comparison of Cox regression and machine learning models”. In: *BMC Medical Research Methodology* 22 (2022), p. 284. DOI: 10.1186/s12874-022-01754-y.

[8] A. Spooner et al. “A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction”. In: *Scientific Reports* 10.1 (2020), p. 20410. DOI: 10.1038/s41598-020-77220-w.

[9] C. Janiesch, P. Zschech, and K. Heinrich. “Machine learning and deep learning”. In: *Electron Markets* 31 (2021), pp. 685–695. DOI: 10.1007/s12525-021-00475-2.

[10] H.S.R. Rajula et al. “Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment”. In: *Medicina* 56.9 (2020), p. 455. DOI: 10.3390/medicina56090455.

[11] P. Choudhury, R.T. Allen, and M.G. Endres. “Machine learning for pattern discovery in management research”. In: *Strategic Management Journal* 42.1 (2021), pp. 30–57. DOI: 10.1002/smj.3215.

[12] X. Shu and Y. Ye. "Knowledge discovery: Methods from data mining and machine learning". In: *Social Science Research* 110 (2023), p. 102817. DOI: 10.1016/j.ssresearch.2022.102817.

[13] J. Harvey et al. "Machine learning-based prediction of cognitive outcomes in de novo Parkinson's disease". In: *npj Parkinson's Disease* 8.1 (2022), p. 150. DOI: 10.1038/s41531-022-00409-5.

[14] H. Almgren et al. "Machine learning-based prediction of longitudinal cognitive decline in early Parkinson's disease using multimodal features". In: *Scientific Reports* 13.1 (2023), p. 13193. DOI: 10.1038/s41598-023-37644-6.

[15] Z. Zhang et al. "Effect of onset age on the levodopa threshold dosage for dyskinesia in Parkinson's disease". In: *Neurological Sciences* (2022), pp. 1–10. DOI: 10.1007/s10072-021-05694-1.

[16] D. Shen et al. "Bilateral globus pallidus interna deep brain stimulation in Parkinson's disease: Therapeutic effects and motor outcomes prediction in a short-term follow-up". In: *Frontiers in Human Neuroscience* 16 (2023), p. 1023917. DOI: 10.3389/fnhum.2022.1023917.

[17] R. Prashanth and S.D. Roy. "Novel and improved stage estimation in Parkinson's disease using clinical scales and machine learning". In: *Neurocomputing* 305 (2018), pp. 78–103. DOI: 10.1016/j.neucom.2018.04.049.

[18] D. Nyholm et al. "Large differences in levodopa dose requirement in Parkinson's disease: Men use higher doses than women". In: *European Journal of Neurology* 17.2 (2010), pp. 260–266. DOI: 10.1111/j.1468-1331.2009.02866.x.

[19] F.S. Kanellos et al. "Clinical evaluation in Parkinson's disease: Is the golden standard shiny enough?" In: *Sensors* 23.8 (2023), p. 3807. DOI: 10.3390/s23083807.

[20] R. Hornung et al. "Improving cross-study prediction through addon batch effect adjustment or addon normalization". In: *Bioinformatics* 33.3 (2017), pp. 397–404. DOI: 10.1093/bioinformatics/btw650.

[21] S. Decherchi et al. "Opportunities and challenges for machine learning in rare diseases". In: *Frontiers in Medicine* 8 (2021), p. 747612. DOI: 10.3389/fmed.2021.747612.

[22] D.K. Kwon et al. "Levodopa-induced dyskinesia in Parkinson's disease: Pathogenesis and emerging treatment strategies". In: *Cells* 11.23 (2022), p. 3736. DOI: 10.3390/cells11233736.

[23] B. Thanvi, N. Lo, and T. Robinson. "Levodopa-induced dyskinesia in Parkinson's disease: Clinical features pathogenesis, prevention and treatment". In: *Postgraduate Medical Journal* 83.980 (2007), pp. 384–388. DOI: 10.1136/pgmj.2006.054759.

[24] J. Wu et al. "The impact of levodopa therapy-induced complications on quality of life in Parkinson's disease patients in Singapore". In: *Scientific reports* 9.1 (2019), p. 9248. DOI: 10.1038/s41598-019-45110-5.

[25] Z. Chen, G. Li, and J. Liu. "Autonomic dysfunction in Parkinson's disease: Implications for pathophysiology, diagnosis, and treatment". In: *Neurobiology of Disease* 134 (2020), p. 104700. DOI: 10.1016/j.nbd.2019.104700.

[26] C.A. Hansen et al. "Levodopa-induced dyskinesia: A historical review of Parkinson's disease, dopamine, and modern advancements in research and treatment". In: *Journal of Neurology* 269.6 (2022), pp. 2892–2909. DOI: 10.1007/s00415-022-10963-w.

[27] M. Porta et al. “Pre-dopa deep brain stimulation: Is early deep brain stimulation able to modify the natural course of Parkinson’s disease?” In: *Frontiers in Neuroscience* 14 (2020), p. 492. DOI: 10.3389/fnins.2020.00492.

[28] C.S. Lee. “Levodopa-induced dyskinesia: Mechanisms and management”. In: *British Columbia Medical Journal* 43.4 (2001), pp. 206–209.

[29] R. Bandopadhyay et al. “Molecular mechanisms and therapeutic strategies for levodopa-induced dyskinesia in Parkinson’s disease: A perspective through preclinical and clinical evidence”. In: *Frontiers in Pharmacology* 13 (2022), p. 805388. DOI: 10.3389/fphar.2022.805388.

[30] V. Leta et al. “Gastrointestinal barriers to levodopa transport and absorption in Parkinson’s disease”. In: *European Journal of Neurology* 30.5 (2023), pp. 1465–1480. DOI: 10.1111/ene.15734.

[31] C.H. Williams-Gray and P.F. Worth. “Parkinson’s disease and related conditions”. In: *Medicine* 51.9 (2023), pp. 645–651. DOI: 10.1016/j.mpmed.2023.06.004.

[32] M.G. Cersosimo and E.E. Benarroch. “Neural control of the gastrointestinal tract: Implications for Parkinson disease”. In: *Movement disorders official journal of the Movement Disorder Society* 23.8 (2008), pp. 1065–1075. DOI: 10.1002/mds.22051.

[33] A. Fasano et al. “Gastrointestinal dysfunction in Parkinson’s disease”. In: *The Lancet Neurology* 14.6 (2015), pp. 625–639. DOI: 10.1016/S1474-4422(15)00007-1.

[34] J.T.B. Keun et al. “Dietary approaches to improve efficacy and control side effects of levodopa therapy in Parkinson’s Disease: A systematic review”. In: *Advances in Nutrition* 12.6 (2021), pp. 2265–2287. DOI: 10.1093/advances/nmab060.

[35] L. di Biase et al. “Levodopa-induced dyskinesias in Parkinson’s disease: An overview on pathophysiology, clinical manifestations, therapy management strategies and future directions”. In: *Journal of Clinical Medicine* 12.13 (2023), p. 4427. DOI: 10.3390/jcm12134427.

[36] C. Rusch et al. “To restrict or not to restrict? Practical considerations for optimizing dietary protein interactions on levodopa absorption in Parkinson’s disease”. In: *npj Parkinson’s disease* 9 (2023), p. 98. DOI: 10.1038/s41531-023-00541-w.

[37] E. Knight et al. “The role of diet and dietary patterns in Parkinson’s disease”. In: *Nutrients* 14.21 (2022), p. 4472. DOI: 10.3390/nu14214472.

[38] M.ó. Breasail et al. “Parkinson’s disease: the nutrition perspective”. In: *Proceedings of the Nutrition Society* 81.1 (2022), pp. 12–26. DOI: 10.1017/S0029665121003645.

[39] I. van der Berg et al. “Dietary interventions in Parkinson’s disease”. In: *Journal of Parkinson’s disease* 14.1 (2024), pp. 1–16. DOI: 10.3233/JPD-230366.

[40] D. Kwon et al. “Diet quality and Parkinson’s disease: Potential strategies for non-motor symptom management”. In: *Parkinsonism & Related Disorders* 115 (2023), p. 105816. DOI: 10.1016/j.parkreldis.2023.105816.

[41] A.O. Omotosho et al. “Parkinson’s disease: Are gut microbes involved?” In: *Brain and Behavior* 13.8 (2023), e3130. DOI: 10.1002/brb3.3130.

[42] S. Navailles and P. De Deurwaerdère. "Contribution of serotonergic transmission to the motor and cognitive effects of high-frequency stimulation of the subthalamic nucleus or levodopa in Parkinson's disease". In: *Molecular Neurobiology* 45 (2012), pp. 173–185. DOI: 10.1007/s12035-011-8230-0.

[43] L.A. King et al. "Do cognitive measures and brain circuitry predict outcomes of exercise in Parkinson's disease: A randomized clinical trial". In: *BMC Neurology* 15 (2015), pp. 1–8. DOI: 10.1186/s12883-015-0474-2.

[44] Sauerbier A. et al. "New concepts in the pathogenesis and presentation of Parkinson's disease". In: *Clinical Medicine* 16.4 (2016), pp. 365–370. DOI: 10.7861/clinmedicine.16-4-365.

[45] G.P. Crucian and M.S. Okun. "Visual-spatial ability in Parkinson's disease". In: *Frontiers in Bioscience* 8 (2003), pp. 992–997. DOI: 10.2741/1171.

[46] Z. Yildiz et al. "Relationship between apathy and cognitive functions in Parkinson's disease". In: *Psychological Applications and Trends* (2023), pp. 661–665. DOI: 10.36315/2023inpact145.

[47] G. Devigili et al. "Unraveling autonomic dysfunction in GBA-related Parkinson's disease". In: *Movement Disorders Clinical Practice* 10.11 (2023), pp. 1620–1638. DOI: 10.1002/mdc3.13892.

[48] Y. Zhou et al. "Mutational spectrum and clinical features of GBA1 variants in a Chinese cohort with Parkinson's disease". In: *npj Parkinson's Disease* 9.1 (2023), p. 129. DOI: 10.1038/s41531-023-00571-4.

[49] A. Shcherbak, E. Kovalenko, and A. Somov. "Detection and classification of early stages of Parkinson's disease through wearable sensors and machine learning". In: *IEEE Transactions on Instrumentation and Measurement* 72 (2023), pp. 1–9. DOI: 10.1109/TIM.2023.3284944.

[50] J.M. Templeton, C. Poellabauer, and S. Schneider. "Classification of Parkinson's disease and its stages using machine learning". In: *Scientific Reports* 12 (2022), p. 14036. DOI: 10.1038/s41598-022-18015-z.

[51] S. Aich et al. "A supervised machine learning approach to detect the on/off state in Parkinson's disease using wearable based gait signals". In: *Diagnostics* 10.6 (2020), p. 421. DOI: 10.3390/diagnostics10060421.

[52] C. Sotirakis et al. "Identification of motor progression in Parkinson's disease using wearable sensors and machine learning". In: *npj Parkinson's disease* 9 (2023), p. 142. DOI: 10.1038/s41531-023-00581-2.

[53] B. Yang et al. "The amplitude of low-frequency fluctuation predicts levodopa treatment response in patients with Parkinson's disease". In: *Parkinsonism & Related Disorders* 92 (2021), pp. 26–32. DOI: 10.1016/j.parkreldis.2021.10.003.

[54] R. Djaldetti et al. "Levodopa responsiveness in Parkinson's disease: harnessing real-life experience with machine-learning analysis". In: *Journal of Neural Transmission* 129 (2022), pp. 1289–1297. DOI: 10.1007/s00702-022-02540-2.

[55] J. He et al. "Instrumented timed up and go test and machine learning-based levodopa response evaluation: a pilot study". In: *Journal of NeuroEngineering and Rehabilitation* 21 (2024), p. 163. DOI: 10.1186/s12984-024-01452-4.

[56] S.J. Chung et al. “Baseline cognitive profile is closely associated with long-term motor prognosis in newly diagnosed Parkinson’s disease”. In: *Journal of Neurology* 268 (2021), pp. 4203–4212. DOI: 10.1007/s00415-021-10529-2.

[57] N. Kandiah et al. “Montreal cognitive assessment for the screening and prediction of cognitive decline in early Parkinson’s disease”. In: *Parkinsonism & related disorders* 20.11 (2014), pp. 1145–1148. DOI: 10.1016/j.parkreldis.2014.08.002.

[58] A. Manson, P. Stirpe, and A. Schrag. “Levodopa-induced dyskinesias clinical features, incidence, risk factors, management and impact on quality of life”. In: *Journal of Parkinson’s disease* 2.3 (2012), pp. 189–198. DOI: 10.3233/JPD-2012-120103.

[59] R. Saunders-Pullman et al. “Progression in the LRRK2-associated Parkinson disease population”. In: *JAMA Neurology* 75.3 (2018), pp. 312–319. DOI: 10.1001/jamaneurol.2017.4019.

[60] H. Wilson et al. “Predict cognitive decline with clinical markers in Parkinson’s disease (PRECODE-1)”. In: *Journal of Neural Transmission* 127 (2020), pp. 51–59. DOI: 10.1007/s00702-019-02125-6.

[61] M. Ikeda, H. Kataoka, and S. Ueno. “Can levodopa prevent cognitive decline in patients with Parkinson’s disease?” In: *American Journal of Neurodegenerative Disease* 6.2 (2017), pp. 9–14.

[62] A. Luca et al. “Cognitive impairment and levodopa induced dyskinesia in Parkinson’s disease: A longitudinal study from the PACOS cohort”. In: *Scientific Reports* 11.1 (2021), p. 867. DOI: 10.1038/s41598-020-79110-7.

[63] J. Ciafone et al. “The neuropsychological profile of mild cognitive impairment in Lewy body dementias”. In: *Journal of the International Neuropsychological Society* 26.2 (2020), pp. 210–225. DOI: 10.1017/S1355617719001103.

[64] M.T.M. Prenger et al. “Social symptoms of Parkinson’s disease”. In: *Parkinson’s Disease* (2020), p. 8846544. DOI: 10.1155/2020/8846544.

[65] E.C. DeMarco, N. Al-Hammadi, and L. Hinyard. “Exploring treatment for depression in Parkinson’s patients: A cross-sectional analysis”. In: *International Journal of Environmental Research and Public Health* 18.16 (2021), p. 8596. DOI: 10.3390/ijerph18168596.

[66] S. Lukas, S. Friederike, and H. Wiebke. “Management of sleep disturbances in Parkinson’s disease”. In: *Journal of Parkinson’s Disease* 12.7 (2022), pp. 2029–2058. DOI: 10.3233/JPD-212749.

[67] E.N. Minakawa. “Bidirectional relationship between sleep disturbances and Parkinson’s disease”. In: *Frontiers in Neurology* 13 (2022), p. 927994. DOI: 10.3389/fneur.2022.927994.

[68] Z. Xu, K.N. Anderson, and N. Pavese. “Longitudinal studies of sleep disturbances in Parkinson’s disease”. In: *Current Neurology and Neuroscience* 22 (2022), pp. 635–655. DOI: 10.1007/s11910-022-01223-5.

[69] Y. Zhang et al. “Multiple comorbid sleep disorders adversely affect quality of life in Parkinson’s disease patients”. In: *npj Parkinson’s Disease* 6 (2020), p. 25. DOI: 10.1038/s41531-020-00126-x.

[70] D. Santos-García et al. “Sleep problems are related to a worse quality of life and a greater non-motor symptoms burden in Parkinson’s disease”. In: *Journal of Geriatric Psychiatry and Neurology* 34.6 (2021), pp. 642–658. DOI: 10.1177/0891988720964250.

[71] R.A. Berk. "Classification and Regression Trees (CART)". In: *Statistical Learning from a Regression Perspective*. Springer Texts in Statistics. Cham: Springer, 2016. DOI: 10.1007/978-3-319-44048-4_3.

[72] Y.S. Tan et al. "Fast interpretable greedy-tree sums (FIGS)". In: *arXiv preprint arXiv:2201.11931* (2022). DOI: 10.48550/arXiv.2201.11931.

[73] Y. Freund, R. Schapire, and N. Abe. "A short introduction to boosting". In: *Journal-Japanese Society for Artificial Intelligence* 14.5 (1999), pp. 771–780.

[74] J.H. Friedman. "Greedy function approximation: A gradient boosting machine". In: *Annals of Statistics* (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451.

[75] T. Chen and C. Guestrin. "XGBoost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794. DOI: 10.1145/2939672.2939785.

[76] L. Prokhorenkova et al. "CatBoost: Unbiased boosting with categorical features". In: *Advances in Neural Information Processing Systems* (2018), pp. 6639–6649.

[77] X. Hu, C. Rudin, and M. Seltzer. "Optimal sparse decision trees". In: *Advances in Neural Information Processing Systems* 32 (2019).

[78] J. Lin et al. "Generalized and scalable optimal sparse decision trees". In: *International Conference on Machine Learning* (2020), pp. 6150–6160.

[79] H. McTavish et al. "Fast sparse decision tree optimization via reference ensembles". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.9 (2022), pp. 9601–9613.

[80] A. Agarwal et al. "Hierarchical Shrinkage: Improving the accuracy and interpretability of tree-based methods". In: *arXiv preprint arXiv:2202.00858* (2022). DOI: 10.48550/arXiv.2202.00858.

[81] D.R. Cox. "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

[82] D. Bertsimas et al. "Optimal survival trees". In: *Machine Learning* 111.8 (2022), pp. 2951–3023. DOI: 10.1007/s10994-021-06117-0.

[83] M.Y. Park and T. Hastie. "L1-regularization path algorithm for generalized linear models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69.4 (2007), pp. 659–677. DOI: 10.1111/j.1467-9868.2007.00607.x.

[84] N. Simon et al. "Regularization paths for Cox's proportional hazards model via coordinate descent". In: *Journal of Statistical Software* 39.5 (2011), p. 1. DOI: 10.18637/jss.v039.i05.

[85] H. Ishwaran et al. "Random survival forests". In: *The Annals of Applied Statistics* 2.3 (2008), pp. 841–860. DOI: 10.1214/08-AOAS169.

[86] A. De Rose and A. Pallara. "Survival trees: An alternative non-parametric multivariate technique for life history analysis". In: *European Journal of Population* 13 (1997), pp. 223–241. DOI: 10.1023/A:1005844818027.

[87] P. Geurts, D. Ernst, and L. Wehenkel. "Extremely randomized trees". In: *Machine Learning* 63 (2006), pp. 3–42. DOI: 10.1007/s10994-006-6226-1.

[88] I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. "A review of survival trees". In: *Statistics Surveys* 5 (2011), pp. 44–71. DOI: 10.1214/09-SS047.

[89] Y. Chen et al. "A gradient boosting algorithm for survival analysis via direct optimization of concordance index". In: *Computational and Mathematical Methods in Medicine* 2013 (2013). DOI: 10.1155/2013/873595.

[90] G. Karami et al. "Predicting overall survival time in glioblastoma patients using gradient boosting machines algorithm and recursive feature elimination technique". In: *Cancers* 13.19 (2021), p. 4976. DOI: 10.3390/cancers13194976.

[91] G.W. Ji et al. "Development and validation of a gradient boosting machine to predict prognosis after liver resection for intrahepatic cholangiocarcinoma". In: *BMC Cancer* 22.1 (2022), p. 258. DOI: 10.1186/s12885-022-09352-3.

[92] K. He et al. "Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates". In: *Bioinformatics* 32.1 (2016), pp. 50–57. DOI: 10.1093/bioinformatics/btv517.

[93] W.E. Johnson, C. Li, and A. Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127. DOI: 10.1093/biostatistics/kxj037.

[94] J.A. Lee, K.K. Dobbin, and J. Ahn. "Covariance adjustment for batch effect in gene expression data". In: *Statistics in Medicine* 33.15 (2014), pp. 2681–2695. DOI: 10.1002/sim.6157.

[95] C. Lazar et al. "Batch effect removal methods for microarray gene expression data integration: A survey". In: *Briefings in bioinformatics* 14.4 (2013), pp. 469–490. DOI: 10.1093/bib/bbs037.

[96] R. Hornung, A.L. Boulesteix, and D. Causeur. "Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment". In: *BMC Bioinformatics* 17 (2016), pp. 1–19. DOI: 10.1186/s12859-015-0870-z.

[97] C.K. Stein et al. "Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat". In: *BMC Bioinformatics* 16.1 (2015), pp. 1–9. DOI: 10.1186/s12859-015-0478-3.

[98] J. Luo et al. "A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data". In: *The Pharmacogenomics Journal* 10.4 (2010), pp. 278–291. DOI: 10.1038/tpj.2010.57.

[99] J.T. Leek and J.D. Storey. "Capturing heterogeneity in gene expression studies by surrogate variable analysis". In: *PLoS Genetics* 3.9 (2007), e161. DOI: 10.1371/journal.pgen.0030161.

[100] H.S. Parker, H.C. Bravo, and J.T. Leek. "Removing batch effects for prediction problems with frozen surrogate variable analysis". In: *PeerJ* 2 (2014), e561. DOI: 10.7717/peerj.561.

[101] S.E. Reese et al. "A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis". In: *Bioinformatics* 29.22 (2013), pp. 2877–2883. DOI: 10.1093/bioinformatics/btt480.

[102] L. Pavelka et al. "Luxmbourg Parkinson's study -comprehensive baseline analysis of Parkinson's disease and atypical parkinsonism". In: *Frontiers in Neurology* 14 (2023), p. 1330321. DOI: 10.3389/fneur.2023.1330321.

[103] K. Marek et al. “The Parkinson’s progression markers initiative (PPMI) - establishing a PD biomarker cohort”. In: *Annals of Clinical and Translational Neurology* 5.12 (2018), pp. 1460–1477. DOI: 10.1002/acn3.644.

[104] K. Marek et al. “The Parkinson Progression Marker Initiative (PPMI)”. In: *Progress in Neurobiology* 95.4 (2011), pp. 629–635. DOI: 10.1016/j.pneurobio.2011.09.005.

[105] P. Dodet et al. “Sleep disorders in Parkinson’s disease, an early and multiple problem”. In: *npj Parkinson’s Disease* 10 (2024), p. 46. DOI: 10.1038/s41531-024-00642-0.

[106] M. Bologna et al. “Evolving concepts on bradykinesia”. In: *Brain* 143.3 (2020), pp. 727–750. DOI: 10.1093/brain/awz344.

[107] S. van Buuren and K. Groothuis-Oudshoorn. “mice: Multivariate imputation by chained equations in R”. In: *Journal of Statistical Software* 45 (2011), pp. 1–67. DOI: 10.18637/jss.v045.i03.

[108] M.J. Azur et al. “Multiple imputation by chained equations: what is it and how does it work?” In: *International Journal of Methods in Psychiatric Research* 20.1 (2011), pp. 40–49. DOI: 10.1002/mpr.329.

[109] B.C. Jaeger, N.J. Tierney, and N.R. Simon. “When to impute? Imputation before and during cross-validation”. In: *arXiv preprint arXiv:2010.00718* (2020). DOI: 10.48550/arXiv.2010.00718.

[110] M.K. Dahouda and I. Joe. “A deep-learned embedding technique for categorical features encoding”. In: *IEEE Access* 9 (2021), pp. 114381–114391. DOI: 10.1109/ACCESS.2021.3104357.

[111] B.M. Bolstad et al. “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”. In: *Bioinformatics* 19.2 (2003), pp. 185–193. DOI: 10.1093/bioinformatics/19.2.185.

[112] D. Kostka and R. Spang. “Microarray based diagnosis profits from better documentation of gene expression signatures”. In: *Plos Computational Biology* (2008), p. 22. DOI: 10.1371/journal.pcbi.0040022.

[113] M.M. Rahman and D. Davis. “Cluster based under-sampling for unbalanced cardiovascular data”. In: *Proceedings of the World Congress on Engineering* 3 (2013), pp. 3–5.

[114] M. Bach, A. Werner, and M. Palt. “The proposal of undersampling method for learning from imbalanced datasets”. In: *Procedia Computer Science* 159 (2019), pp. 125–134. DOI: 10.1016/j.procs.2019.09.167.

[115] A. Gramegna and P. Giudici. “Shapley feature selection”. In: *FinTech* 1.1 (2022), pp. 72–80. DOI: 10.3390/fintech1010006.

[116] S. Nirmalraj et al. “Permutation feature importance-based fusion techniques for diabetes prediction”. In: *Soft Computing* (2023), pp. 1–12. DOI: 10.1007/s00500-023-08041-y.

[117] Y. Freund and R.E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. DOI: 10.1006/jcss.1997.1504.

[118] J.R. Quinlan. “Improved use of continuous attributes in C4.5”. In: *Journal of Artificial Intelligence Research* 4 (1996), pp. 77–90. DOI: 10.1613/jair.279.

[119] S.M. Lundberg and S.I. Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[120] S. Sundrani and J. Lu. "Computing the hazard ratios associated with explanatory variables using machine learning models of survival data". In: *Clinical Cancer Informatics* 5 (2021), pp. 364–378. DOI: 10.1200/CCI.20.00172.

[121] C. Chen et al. "Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods". In: *PLoS One* 6.2 (2011), e17238. DOI: 10.1371/journal.pone.0017238.

[122] X. Sun and W. Xu. "Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristics curves". In: *IEEE Signal Processing Letters* 21.11 (2014), pp. 1389–1393. DOI: 10.1109/LSP.2014.2337313.

[123] L. Kang et al. "Comparing two correlated C indices with right-censored survival outcome: A one-shot nonparametric approach". In: *Statistics in Medicine* 34.4 (2015), pp. 685–703. DOI: 10.1002/sim.6370.

[124] A. Benavoli et al. "Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis". In: *Journal of Machine Learning Research* 18.77 (2017), pp. 1–36.

[125] G. Corani and A. Benavoli. "A Bayesian approach for comparing cross-validated algorithms on multiple data sets". In: *Machine Learning* 100.2 (2015), pp. 285–304. DOI: 10.1007/s10994-015-5486-z.

[126] A.J. Vickers, B. van Calster, and E.W. Steyerberg. "A simple, step-by-step guide to interpreting decision curve analysis". In: *Diagnostic and Prognostic Research* 3 (2019), p. 18. DOI: 10.1186/s41512-019-0064-7.

[127] D. Piovani et al. "Optimizing clinical decision making with decision curve analysis: Insights for clinical investigators". In: *Healthcare (Basel)* 11.16 (2023), p. 2244. DOI: 10.3390/healthcare11162244.

[128] Z. Zhang et al. "Decision curve analysis: a technical note". In: *Annals of translational medicine* 6.15 (2018), p. 308. DOI: 10.21037/atm.2018.07.02.

[129] P.C. Austin, F.E. Harrell Jr, and D. van Klaveren. "Graphical calibration curves and the integrated calibration index (ICI) for survival models". In: *Statistics in Medicine* 39.21 (2020), pp. 2714–2742. DOI: 10.1002/sim.8570.

[130] C. Leiva-Santana and M. Alvarez-Saùco. "Levodopa and cognitive disorders in Parkinson's disease". In: *Revista de Neurologia* 43.2 (2006), pp. 95–100.

[131] M. Hutny et al. "Current knowledge on the background pathophysiology and treatment of levodopa-induced dyskinesia - literature review". In: *Journal of Clinical Medicine* 10.19 (2021), p. 4377. DOI: 10.3390/jcm10194377.

[132] A. Tirozzi et al. "Analysis of genetic and non-genetic predictors of levodopa induced dyskinesia in Parkinson's disease". In: *Frontiers in Pharmacology* 12 (2021), p. 640603. DOI: 10.3389/fphar.2021.640603.

[133] E.V. Encarnacion and R.A. Hauser. "Levodopa-induced dyskinesias in Parkinson's disease: Etiology, impact on quality of life, and treatments". In: *European neurology* 60.2 (2008), pp. 57–66. DOI: 10.1159/000131893.

[134] S. Pandey and P. Srivanitchapoom. "Levodopa-induced dyskinesia: Clinical features, pathophysiology, and medical management". In: *Annals of Indian Academy of Neurology* 20.3 (2017), p. 190. DOI: 10.4103/aian.AIAN_239_17.

[135] E. Bezard, J.M. Brotchie, and C.E. Gross. "Pathophysiology of levodopa-induced dyskinesia: Potential for new therapies". In: *Nature Reviews Neuroscience* 2.8 (2001), pp. 577–588. DOI: 10.1038/35086062.

[136] C. Ahlrichs and M. Lawo. "Parkinson's disease motor symptoms in machine learning". In: *arXiv preprint arXiv:1312.3825* (2013). DOI: 10.48550/arXiv.1312.3825.

[137] S. Thanprasertsuk et al. "Levodopa-induced dyskinesia in early-onset Parkinson's disease (EOPD) associates with glucocerebrosidase mutation: A next-generation sequencing study in EOPD patients in Thailand". In: *PLoS One* 18.10 (2023), p. 0293515. DOI: 10.1371/journal.pone.0293516.

[138] B.L. Santos-Lobato et al. "Diagnostic prediction model for levodopa-induced dyskinesia in Parkinson's disease". In: *Arquivos de Neuro-Psiquiatria* 78 (2020), pp. 206–216. DOI: 10.1590/0004-282X20190191.

[139] S. Redenšek et al. "Clinical-pharmacogenetic predictive models for time to occurrence of levodopa related motor complications in Parkinson's disease". In: *Frontiers in genetics* 10 (2019), p. 461. DOI: 10.3389/fgene.2019.00461.

[140] R.T. Athulya et al. "Predictors of levo-dopa induced dyskinesias in Parkinson's disease". In: *Annals of Indian Academy of Neurology* 23.1 (2020), p. 44. DOI: 10.4103/aian.AIAN_460_18.

[141] R. Cilia et al. "The modern pre-levodopa era of Parkinson's disease: Insights into motor complications from sub-Saharan Africa". In: *Brain* 10 (2014), pp. 2731–2742. DOI: 10.1093/brain/awu195.

[142] J.H. Jung et al. "White matter connectivity networks predict levodopa-induced dyskinesia in Parkinson's disease". In: *Journal of Neurology* 269.6 (2022), pp. 2948–2960. DOI: 10.1007/s00415-021-10883-1.

[143] M.J. Kelly et al. "Predictors of motor complications in early Parkinson's disease: A prospective cohort study". In: *Movement Disorders* 34.8 (2019), pp. 1174–1183. DOI: 10.1002/mds.27783.

[144] R.A. Travagli, K.N. Browning, and M. Camilleri. "Parkinson disease and the gut: new insights into pathogenesis and clinical relevance". In: *Nature Reviews Gastroenterology & Hepatology* 17.11 (2020), pp. 673–685. DOI: 10.1038/s41575-020-0339-z.

[145] A.H. Tan et al. "Gastrointestinal dysfunction in Parkinson's disease: Neuro-gastroenterology perspectives on a multifaceted problem". In: *Journal of Movement Disorders* 16.2 (2023), p. 138. DOI: 10.14802/jmd.22220.

[146] C.M. Shin et al. "DA-9701 on gastric motility in patients with Parkinson's disease: A randomized controlled trial". In: *Parkinsonism & Related Disorders* 54 (2018), pp. 84–89. DOI: 10.1016/j.parkreldis.2018.04.018.

[147] A. Poirier et al. “Gastrointestinal dysfunctions in Parkinson’s disease: Symptoms and treatments”. In: *Parkinson’s Disease* (2016), p. 6762528. DOI: 10.1155/2016/6762528.

[148] W. Agnieszka, P. Pawel, and K. Malgorzata. “How to optimize the effectiveness and safety of Parkinson’s disease therapy? - A systematic reviews of drugs interactions with food and dietary supplements”. In: *Current Neuropharmacology* 20.7 (2022), pp. 1427–1447. DOI: 10.2174/1570159X19666211116142806.

[149] M.B. Guebila and I. Thiele. “Model-based dietary optimization for late-stage, levodopa-treated, Parkinson’s disease patients”. In: *npj Systems Biology and Applications* 2 (2016), p. 16013. DOI: 10.1038/npjsba.2016.13.

[150] A. Lenka et al. “Practical pearls to improve the efficacy and tolerability of levodopa in Parkinson’s disease”. In: *Expert Review of Neurotherapeutics* 22.6 (2022), pp. 489–498. DOI: 10.1080/14737175.2022.2091436.

[151] R. Ana, G. Cristina, and G.Y. Justo. “Vitamin B6 deficiency in patients with Parkinson disease treated with levodopa/carbidopa”. In: *Clinical Neuropharmacology* 43.5 (2020), pp. 151–157. DOI: 10.1097/WNF.0000000000000408.

[152] K. Lizárraga and A.E. Lang. “Vitamin B6 and B12, levodopa, and their complex interactions in patients with Parkinson’s disease”. In: *Brain* 145.9 (2022), pp. 77–78. DOI: 10.1093/brain/awac225.

[153] A.L.P. de Menezes et al. “Vitamin B6 deficiency in patients with Parkinson disease treated with levodopa/carbidopa”. In: *Journal of Parkinson’s Disease* 14.7 (2024), pp. 1353–1368. DOI: 10.3233/JPD-240.

[154] W.R. Gibb and A.J. Lees. “The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson’s disease”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 51.6 (1988), pp. 745–752. DOI: 10.1136/jnnp.51.6.745.

[155] K.A. Severson et al. “Discovery of Parkinson’s disease states and disease progression modelling: a longitudinal data study using machine learning”. In: *The Lancet Digital Health* 3.9 (2021), e555–564. DOI: 10.1016/S2589-7500(21)00101-1.

[156] M.H. Li et al. “Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation”. In: *Journal of NeuroEngineering and Rehabilitation* 15 (2018), p. 97. DOI: 10.1186/s12984-018-0446-z.

[157] G.A.S. Ferreira et al. “On the classification of tremor signals into dyskinesia, Parkinsonian tremor, and essential tremor by using machine learning techniques”. In: *Biomedical Signal Processing and Control* 73 (2022), p. 103430. DOI: 10.1016/j.bspc.2021.103430.

[158] D.A.B. Leal et al. “Prediction of dyskinesia in Parkinson’s disease patients using machine learning algorithms”. In: *Scientific Reports* 13 (2023), p. 22426. DOI: 10.1038/s41598-023-49617-w.

[159] I.D. Dinov et al. “Predictive big data analytics: A study of Parkinson’s disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations”. In: *PLoS ONE* 11.8 (2016), e0157077. DOI: 10.1371/journal.pone.0157077.

[160] T.N. Tran et al. “Levodopa-induced dyskinesia: clinical features, incidence, and risk factors”. In: *Journal of Neural Transmission* 125 (2018), pp. 1109–1117. DOI: 10.1007/s00702-018-1900-6.

[161] A. Bjornestad et al. “Risk and course of motor complications in a population-based incident Parkinson’s disease cohort”. In: *Parkinsonism & Related Disorders* 22 (2016), pp. 48–53. DOI: 10.1016/j.parkreldis.2015.11.007.

[162] R. Ou et al. “Association between positive history of essential tremor and disease progression in patients with Parkinson’s disease”. In: *Scientific Reports* 10 (2020), p. 21749. DOI: 10.1038/s41598-020-78794-1.

[163] Y. Zhong et al. “Association of motor subtype and tremor type with Parkinson’s disease progression: An exploratory longitudinal analysis”. In: *Journal of Parkinson’s Disease* (2024). DOI: 10.1177/1877718X241305715.

[164] E. Goubault et al. “Remnants of cardinal symptoms of Parkinson’s disease, not dyskinesia, are problematic for dyskinetic patients performing activities of daily living”. In: *Frontiers in Neurology* 10 (2019), p. 256. DOI: 10.3389/fneur.2019.00256.

[165] R.B. Postuma et al. “MDS clinical diagnostic criteria for Parkinson’s disease”. In: *Movement disorders* 30.12 (2015), pp. 1591–1601. DOI: 10.1002/mds.26424.

[166] A. Fasano and B.R. Bloem. “Gait disorders”. In: *CONTINUUM: Lifelong Learning in Neurology* 19.5 (2013), pp. 1344–1382. DOI: 10.1212/01.CON.0000436159.33447.69.

[167] A. Josiah et al. “The effects of gait impairment with and without freezing of gait in Parkinson’s disease”. In: *Parkinsonism & Related Disorders* 18.3 (2012), pp. 239–242. DOI: 10.1016/j.parkreldis.2011.10.008.

[168] S. Azmin et al. “Nonmotor symptoms in a Malaysian Parkinson’s disease population”. In: *Parkinson’s Disease* (2014). DOI: 10.1155/2014/472157.

[169] Y. Nakayama, M. Washio, and M. Mori. “Oral health conditions in patients with Parkinson’s disease”. In: *Journal of Epidemiology* 14.5 (2004), pp. 143–150. DOI: 10.2188/jea.14.143.

[170] N.E. Allen et al. “The association between Parkinson’s disease motor impairments and pain”. In: *Pain Medicine* 17.3 (2015), pp. 456–462. DOI: 10.1111/pme.12898.

[171] J. Jankovic and E.K. Tan. “Parkinson’s disease: etiopathogenesis and treatment”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 91.8 (2020), pp. 795–808. DOI: 10.1136/jnnp-2019-322338.

[172] G. Palermo et al. “Dopamine transporter, age, and motor complications in Parkinson’s disease: A clinical and single-photon emission computed tomography study”. In: *Movement Disorders* 35.6 (2020), pp. 1028–1036. DOI: 10.1002/mds.28008.

[173] S. Gerke, T. Minssen, and G. Cohen. “Ethical and legal challenges of artificial intelligence-driven healthcare”. In: *Artificial Intelligence in Healthcare* (2020), pp. 295–336. DOI: 10.1016/B978-0-12-818438-7.00012-5.

[174] H. Fröhlich et al. “Leveraging the potential of digital technology for better individualized treatment of Parkinson’s disease”. In: *Frontiers in Neurology* 13 (2022), p. 788427. DOI: 10.3389/fneur.2022.788427.

[175] A. Harrie et al. “Cognitive correlates of dual tasking costs on the timed up and go test in Parkinson disease”. In: *Clinical Parkinsonism & Related Disorders* 7 (2022), p. 100158. DOI: 10.1016/j.prdoa.2022.100158.

[176] M. Muleiro Álvarez et al. “A comprehensive approach to Parkinson’s disease: Addressing its molecular, clinical, and therapeutic aspects”. In: *International Journal of Molecular Sciences* 25.13 (2024), p. 7183. DOI: 10.3390/ijms25137183.

[177] G. DeMichele et al. “Parkinson’s disease patients carrying GBA gene mutations”. In: *Acta Neurologica Belgica* 123 (2023), pp. 221–226. DOI: 10.1007/s13760-022-02165-y.

[178] R.F. Pfeiffer, S.H. Isaacson, and R. Pahwa. “Clinical implications of gastric complications on levodopa treatment in Parkinson’s disease”. In: *Parkinsonism & Related Disorders* 76 (2020), pp. 63–71. DOI: 10.1016/j.parkreldis.2020.05.001.

[179] N. Dag and H. Per M. “Effects of *Helicobacter pylori* on levodopa pharmacokinetics”. In: *Journal of Parkinson’s Disease* 11.1 (2021), pp. 61–69. DOI: 10.3233/JPD-202298.

[180] C.Q. Chu et al. “Dietary patterns affect Parkinson’s disease via the microbiota-gut-brain axis”. In: *Trends in Food Science & Technology* 116 (2021), pp. 90–101. DOI: 10.1016/j.tifs.2021.07.004.

[181] A. Schrag and N. Quinn. “Dyskinesias and motor fluctuations in Parkinson’s disease. A community-based study”. In: *Brain* 123.11 (2000), pp. 2297–2305. DOI: 10.1093/brain/123.11.2297.

[182] J. Jankovic. “Motor fluctuations and dyskinesias in Parkinson’s disease: Clinical manifestations”. In: *Movement Disorder Society* 20.11 (2005), pp. 11–16. DOI: 10.1002/mds.20458.

[183] B.D. Wissel et al. “Tablet-based application for objective measurement of motor fluctuations in Parkinson disease”. In: *Digital Biomarkers* 1.2 (2018), pp. 126–135. DOI: 10.1159/000485468.

[184] R. Krüger et al. “Classification of advanced stages of Parkinson’s disease: translation into stratified treatments”. In: *Journal of Neural Transmission* 124.8 (2017), pp. 1015–1027. DOI: 10.1007/s00702-017-1707-x.

[185] H.L. Frequin et al. “Long-term follow-up of the LEAP study: Early versus delayed levodopa in early Parkinson’s disease”. In: *Movement Disorders* 39.6 (2024), pp. 975–982. DOI: 10.1002/mds.29796.

[186] H. Kin et al. “Motor complications in Parkinson’s disease: 13 years follow-up of the CamPaIGN cohort”. In: *Movement Disorders* 35.1 (2020), pp. 185–190. DOI: 10.1002/mds.27882.

[187] A. Zampogna et al. “Axial impairment and falls in Parkinson’s disease: 15 years of subthalamic deep brain stimulation”. In: *npj Parkinson’s Disease* 8 (2022), p. 121. DOI: 10.1038/s41531-022-00383-y.

[188] T.H. Lian et al. “Tremor-dominant in Parkinson disease: The relevance to iron metabolism and inflammation”. In: *Frontiers in Neuroscience* 13 (2019), p. 255. DOI: 10.3389/fnins.2019.00255.

[189] T. Tezuka et al. “Pathophysiological evaluation of the LRRK2 G2385R risk variant for Parkinson’s disease”. In: *npj Parkinson’s Disease* 8 (2022), p. 97. DOI: 10.1038/s41531-022-00367-y.

[190] Y.L. Sosero et al. “Dopamine pathway and Parkinson’s risk variants are associated with levodopa-induced dyskinesia”. In: *Movement Disorders* 0.0 (2024), pp. 1–2. DOI: 10.1002/mds.29796.

[191] A.I. Garcia-Diaz et al. “Cortical thinning correlates of changes in visuospatial and visuoperceptual performance in Parkinson’s disease: A 4-year follow-up”. In: *Parkinsonism & Related Disorders* 46 (2018), pp. 62–68. DOI: 10.1016/j.parkreldis.2017.11.003.

[192] M. Poletti and U. Bonuccelli. "Acute and chronic cognitive effects of levodopa and dopamine agonists on patients with Parkinson's disease: A review". In: *Therapeutic Advances in Psychopharmacology* 3.2 (2013), pp. 101–113. DOI: 10.1177/2045125312470130.

[193] K.A. Mills et al. "Cognitive impairment in Parkinson's disease: Association between patient-reported and clinically measured outcomes". In: *Parkinsonism & Related Disorders* 33 (2016), pp. 107–114. DOI: 10.1016/j.parkreldis.2016.09.025.

[194] J.M. Chandler et al. "Characteristics of Parkinson's disease in patients with and without cognitive impairment". In: *Journal of Parkinson's Disease* 11.3 (2021), pp. 1381–1392. DOI: 10.3233/JPD-202190.

[195] G. Gramotnev, D.K. Gramotnev, and A. Gramotnev. "Parkinson's disease prognostic scores for progression of cognitive decline". In: *Scientific Reports* 9.1 (2019), p. 17485. DOI: 10.1038/s41598-019-54029-w.

[196] C. Baiano et al. "Prevalence and clinical aspects of mild cognitive impairment in Parkinson's disease: A meta-analysis". In: *Movement Disorders* 35.1 (2020), pp. 45–54. DOI: 10.1002/mds.27902.

[197] J.D. Jones et al. "Gastrointestinal symptoms are predictive of trajectories of cognitive functioning in de novo Parkinson's disease". In: *Parkinsonism & Related Disorders* 72 (2020), pp. 7–12. DOI: 10.1016/j.parkreldis.2020.01.009.

[198] Y. Xiao et al. "Different associated factors of subjective cognitive complaints in patients with early- and late-onset Parkinson's disease". In: *Frontiers in Neurology* 12 (2021), p. 12. DOI: 10.3389/fneur.2021.749471.

[199] F. Zhou et al. "Abnormal intra- and inter-network functional connectivity of brain networks in early-onset Parkinson's disease and late-onset Parkinson's disease". In: *Frontiers in Aging Neuroscience* 15 (2023), p. 1132723. DOI: 10.3389/fnagi.2023.1132723.

[200] I. Beheshti, S. Booth, and J.H. Ko. "Differences in brain aging between sexes in Parkinson's disease". In: *npj Parkinson's Disease* 10 (2024), p. 35. DOI: 10.1038/s41531-024-00646-w.

[201] P. Chiara et al. "Cognitive function in Parkinson's disease: The influence of gender". In: *Basal Ganglia* 3.2 (2013), pp. 131–135. DOI: 10.1016/j.baga.2012.10.004.

[202] T.H. Reekes et al. "Sex specific cognitive differences in Parkinson disease". In: *npj Parkinson's disease* 6 (2020), p. 7. DOI: 10.1038/s41531-020-0109-1.

[203] B. Cholerton et al. "Sex differences in progression to mild cognitive impairment and dementia in Parkinson's disease". In: *Parkinsonism & Related Disorders* 50 (2018), pp. 29–36. DOI: j.parkreldis.2018.02.007.

[204] M.C. Bakeberg et al. "Differential effects of sex on longitudinal patterns of cognitive decline in Parkinson's disease". In: *Journal of Neurology* 268 (2021), pp. 1903–1912. DOI: 10.1007/s00415-020-10367-8.

[205] S.H. Kang, J. Lee, and S.B. Koh. "Constipation is associated with mild cognitive impairment in patients with de novo Parkinson's disease". In: *Journal of Movement Disorders* 15.1 (2021), pp. 38–42. DOI: 10.14802/jmd.21074.

[206] E.K. Tur and E. Gözke. "Autonomic symptoms in early-stage Parkinson's patients and their relationship with cognition and disease parameters". In: *Anatolian Current Medical Journal* 5.4 (2023), pp. 498–502. DOI: 10.38053/acmj.1355855.

[207] A.V. Nagy et al. "Cognitive impairment in REM-sleep behaviour disorder and individuals at risk of Parkinson's disease". In: *Parkinsonism & Related Disorders* 109 (2023), p. 105312. DOI: 10.1016/j.parkreldis.2023.105312.

[208] G. Maggi et al. "Sleep disorders and cognitive dysfunctions in Parkinson's disease: A meta-analytic study". In: *Neuropsychology Review* 31 (2021), pp. 643–682. DOI: 10.1007/s11065-020-09473-1.

[209] S. Thangaleela et al. "Neurological insights into sleep disorders in Parkinson's disease". In: *Brain Sciences* 13.8 (2023), p. 1202. DOI: 10.3390/brainsci13081202.

[210] M. Malek-Ahmadi et al. "Trajectory and variability characterization of the Montreal cognitive assessment in older adults". In: *Aging Clinical and Experimental Research* 30 (2018), pp. 993–998. DOI: 10.1007/s40520-017-0865-x.

[211] L.J.W. Evers et al. "Measuring Parkinson's disease over time: The real-world within-subject reliability of the MDS-UPDRS". In: *Movement Disorders* 34.10 (2019), pp. 1480–1487. DOI: 10.1002/mds.27790.

[212] N.D. Nguyen and D. Wang. "Multiview learning for understanding functional multiomics". In: *PLoS Computational Biology* 16.4 (2020), e1007677. DOI: 10.1371/journal.pcbi.1007677.

[213] J. Zhao et al. "Multi-view learning overview: Recent progress and new challenges". In: *Information Fusion* 38 (2017), pp. 43–54. DOI: 10.1016/j.inffus.2017.02.007.

[214] J. Liu et al. "Identification of potential Parkinson's disease drugs based on multi-source data fusion and convolutional neural network". In: *Molecules* 27.15 (2022), p. 4780. DOI: 10.3390/molecules27154780.

[215] C.M.T. Karthigeyan and C. Rani. "Optimizing Parkinson's disease diagnosis with multimodal data fusion techniques". In: *Information Technology and Control* 53.1 (2024), pp. 262–279. DOI: 10.5755/j01.itc.53.1.34718.

[216] D.R. Williams, S.R. Martin, and P. Rast. "Putting the individual into reliability: Bayesian testing of homogeneous within-person variance in hierarchical models". In: *Behavior Research Methods* 54 (2022), pp. 1272–1290. DOI: 10.3758/s13428-021-01646-x.

[217] K. Shao, B.C. Allen, and M.W. Wheeler. "Bayesian Hierarchical structure for quantifying population variability to inform probabilistic health risk assessments". In: *Risk Analysis* 37.10 (2017), pp. 1865–1878. DOI: 10.1111/risa.12751.

[218] N.J. Dingemanse and N.A. Dochtermann. "Quantifying individual variation in behaviour: mixed-effect modelling approaches". In: *Journal of Animal Ecology* 82.1 (2013), pp. 39–54. DOI: 10.1111/1365-2656.12013.

Appendix A

Levodopa-induced dyskinesia in Parkinson's disease: Insights from cross-cohort prognostic analysis using machine learning

A.1 Model performance metrics for dyskinesia prognosis across cohort analyses

Table A.1 Predictive performance metrics for *refined* LID classification model in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**	**Mean (SD)**	**Hold-out AUC**	**Number of features**
AdaBoost	0.661 (0.107)	0.617	10 (20)	***0.613 (0.051)***	***0.671***	***4 (8)***	0.490 (0.087)	0.512	1 (2)
CART	0.628 (0.047)	0.515	2 (14)	0.597 (0.033)	0.642	7 (18)	0.480 (0.130)	0.203	1 (3)
CatBoost	0.686 (0.038)	0.553	6 (9)	**0.639 (0.091)**	**0.659**	**9 (19)**	0.500 (0.131)	0.468	9 (19)
C4.5	0.647 (0.073)	0.525	3 (6)	0.635 (0.081)	0.666	4 (11)	**0.558 (0.122)**	**0.515**	**2 (4)**
FIGS	0.638 (0.054)	0.693	8 (14)	0.621 (0.045)	0.630	2 (6)	0.503 (0.080)	0.512	1 (2)
GOSDT-GUESS	0.688 (0.044)	0.673	19 (25)	0.605 (0.091)	0.559	17 (31)	***0.518 (0.158)***	***0.718***	***6 (6)***
GBoost	***0.637 (0.044)***	***0.686***	***18 (31)***	0.612 (0.057)	0.655	6 (18)	0.517 (0.097)	0.510	5 (9)
HS	0.638 (0.054)	0.682	9 (15)	0.621 (0.045)	0.630	2 (6)	0.453 (0.179)	0.505	6 (14)
XGBoost	**0.706 (0.051)**	**0.655**	**29 (47)**	0.624 (0.064)	0.644	11 (12)	0.466 (0.115)	0.513	19 (25)

An overview of the *refined* LID prognostic classification's predictive performance statistics summarizes the LID prognostic classification's predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table A.2 Predictive performance metrics for *refined* time-to-LID in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**
CW-GBoost	0.682 (0.053)	0.662	10 (25)	0.669 (0.028)	0.633	9 (13)	0.570 (0.119)	0.454	3 (9)
Extra Survival	**0.701 (0.051)**	**0.648**	**91 (130)**	0.665 (0.053)	0.640	110 (113)	0.546 (0.076)	0.649	8 (8)
Survival GBoost	0.693 (0.084)	0.643	10 (17)	0.654 (0.049)	0.641	15 (30)	0.542 (0.182)	0.314	7 (7)
LSVM	0.664 (0.075)	0.675	32 (32)	0.653 (0.030)	0.623	29 (29)	0.547 (0.132)	0.452	9 (9)
NLSVM	0.637 (0.083)	0.666	19 (19)	0.658 (0.028)	0.615	19 (19)	0.546 (0.167)	0.446	9 (10)
Penalized Cox	0.685 (0.046)	0.532	1 (5)	**0.688 (0.048)**	**0.662**	**32 (59)**	0.535 (0.087)	0.517	4 (4)
Survival RF	***0.678 (0.091)***	***0.713***	***87 (118)***	0.670 (0.056)	0.638	53 (85)	**0.586 (0.108)**	**0.572**	**9 (17)**
Survival Trees	0.640 (0.095)	0.640	6 (8)	0.634 (0.077)	0.614	3 (12)	0.557 (0.095)	0.498	1 (3)

An overview of the *refined* time-to-LID predictive performance statistics summarizes the time-to-LID predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-indices in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out C-index is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table A.3 Predictive performance metrics for *refined* LID classification in multi-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features
AdaBoost	**0.688 (0.043)**	**0.639**	**5 (9)**	0.679 (0.028)	0.548	2 (3)	0.640 (0.056)	0.517	4 (9)
CART	0.674 (0.036)	0.618	2 (4)	0.667 (0.044)	0.519	10 (19)	0.607 (0.027)	0.574	6 (17)
CatBoost	***0.678 (0.046)***	***0.665***	***16 (29)***	**0.692 (0.058)**	**0.522**	**9 (19)**	**0.679 (0.033)**	**0.599**	**9 (20)**
C4.5	0.639 (0.045)	0.585	5 (8)	0.643 (0.018)	0.527	4 (9)	0.623 (0.113)	0.564	4 (7)
FIGS	0.663 (0.058)	0.636	3 (9)	0.673 (0.036)	0.488	5 (12)	0.624 (0.057)	0.531	9 (20)
GOSDT-GUESS	0.619 (0.057)	0.555	37 (53)	***0.639 (0.026)***	***0.574***	***38 (59)***	0.615 (0.062)	0.488	32 (45)
GBoost	0.654 (0.030)	0.612	17 (27)	**0.692 (0.021)**	**0.534**	**6 (15)**	0.630 (0.078)	0.550	10 (20)
HS	0.663 (0.058)	0.636	3 (9)	0.673 (0.036)	0.488	5 (12)	0.624 (0.057)	0.531	9 (20)
XGBoost	0.653 (0.050)	0.612	36 (62)	0.675 (0.023)	0.548	11 (15)	0.669 (0.045)	0.557	58 (84)

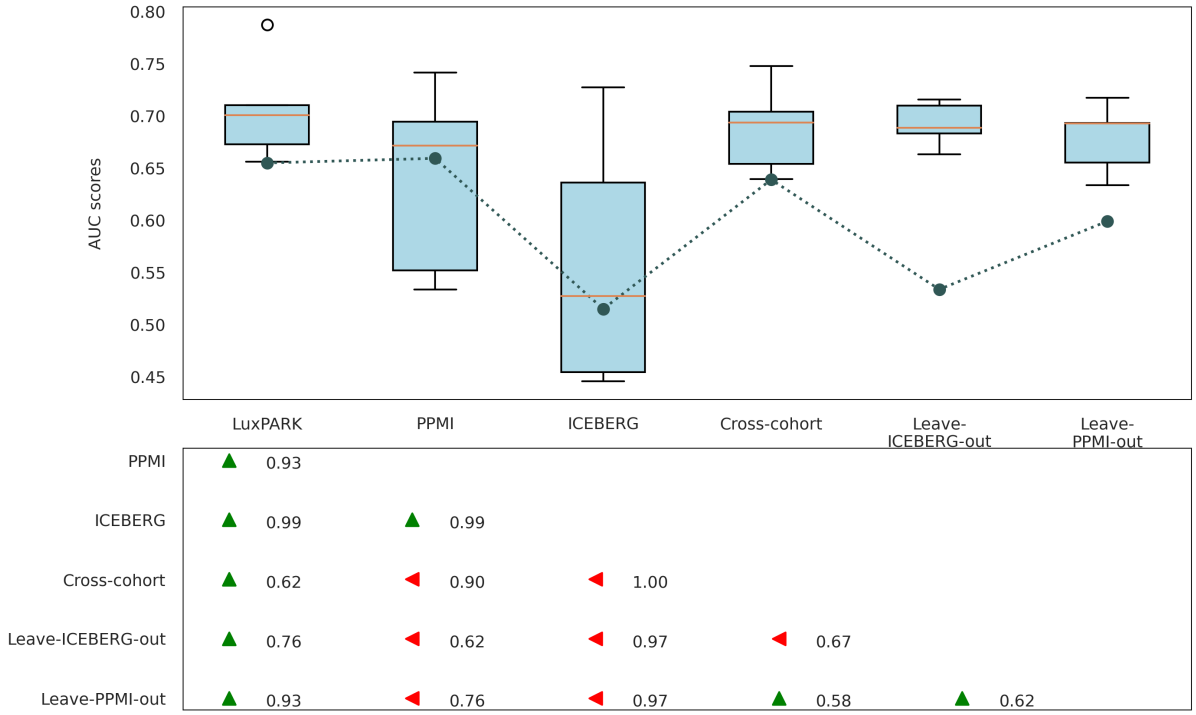
An overview of the *refined* LID prognostic classification's predictive performance statistics summarizes the LID prognostic classification's predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table A.4 Predictive performance metrics for *refined* time-to-LID in multi-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**
CW-GBoost	0.697 (0.055)	0.654	18 (37)	0.681 (0.037)	0.536	18 (31)	***0.674 (0.025)***	***0.637***	***6 (9)***
Extra Survival	0.693 (0.053)	0.676	156 (158)	0.694 (0.040)	0.533	159 (162)	0.675 (0.027)	0.632	131 (135)
Survival GBoost	0.692 (0.047)	0.631	10 (16)	0.680 (0.054)	0.603	147 (149)	0.682 (0.031)	0.616	10 (24)
LSVM	0.686 (0.053)	0.640	39 (39)	0.671 (0.059)	0.575	58 (58)	0.682 (0.023)	0.603	147 (147)
NLSVM	0.690 (0.082)	0.668	47 (47)	***0.665 (0.020)***	***0.646***	***58 (58)***	0.679 (0.029)	0.633	46 (46)
Penalized Cox	0.673 (0.074)	0.669	30 (36)	**0.695 (0.061)**	**0.531**	**50 (80)**	0.679 (0.068)	0.551	1 (67)
Survival RF	***0.715 (0.054)***	***0.685***	***106 (131)***	0.687 (0.052)	0.559	53 (97)	**0.687 (0.037)**	**0.634**	**95 (119)**
Survival Trees	0.644 (0.071)	0.558	7 (14)	0.636 (0.034)	0.549	11 (16)	0.649 (0.055)	0.527	11 (19)

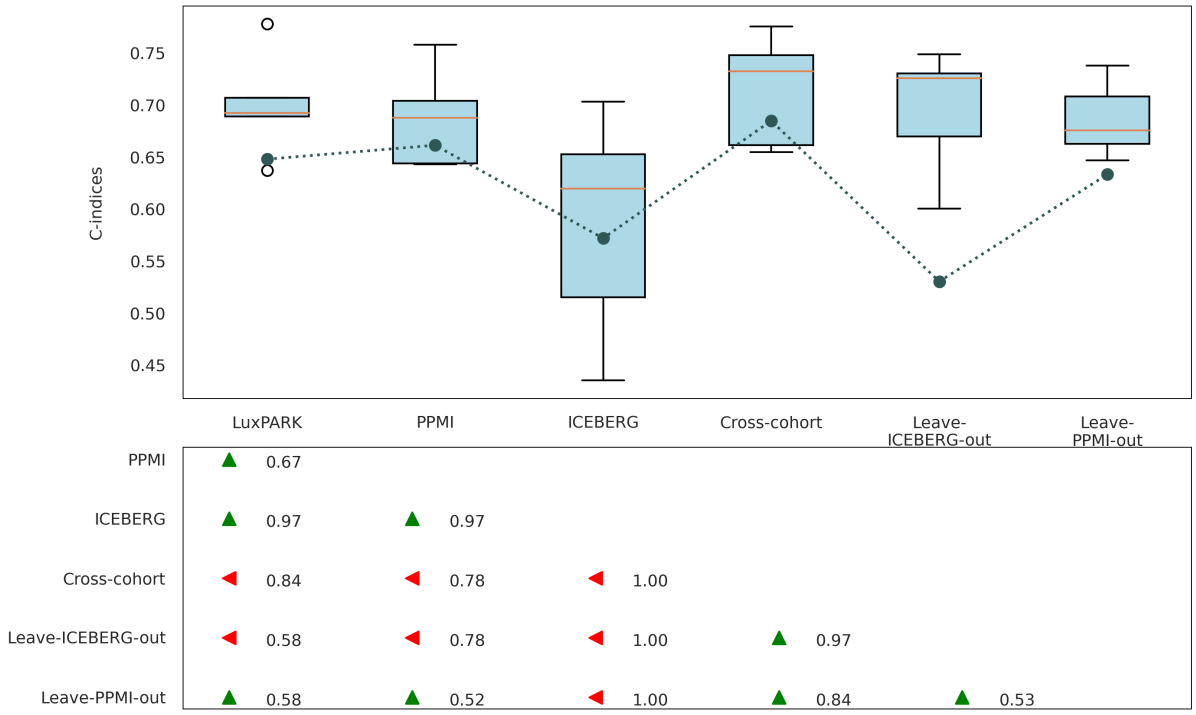
An overview of the *refined* time-to-LID predictive performance statistics summarizes the time-to-LID predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-indices in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out C-index is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Figure A.1 Comparison of cross-validated AUC scores for *refined* LID classification models.



A comparison of cross-validated AUC scores and probabilities of better predictive performance for the *optimized refined* LID classification model across cohort analyses. The upper part displays boxplots of the cross-validated AUC scores for each cohort, with the line indicating the hold-out AUC scores for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

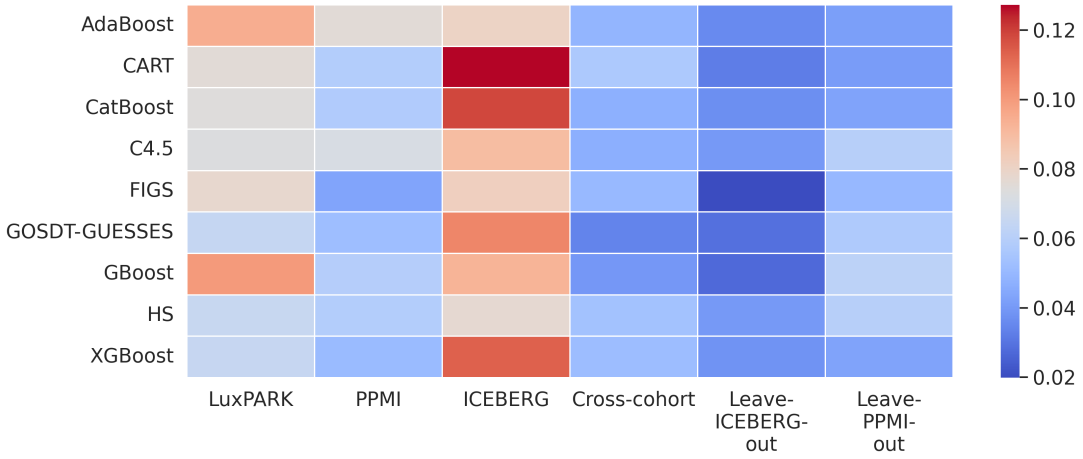
Figure A.2 Comparison of cross-validated C-indices for *refined* time-to-LID models.



A comparison of cross-validated C-indices and probabilities of better predictive performance for the *optimized refined* time-to-LID model across cohort analyses. The upper part displays boxplots of the cross-validated C-indices for each cohort, with the line indicating the hold-out C-indices for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

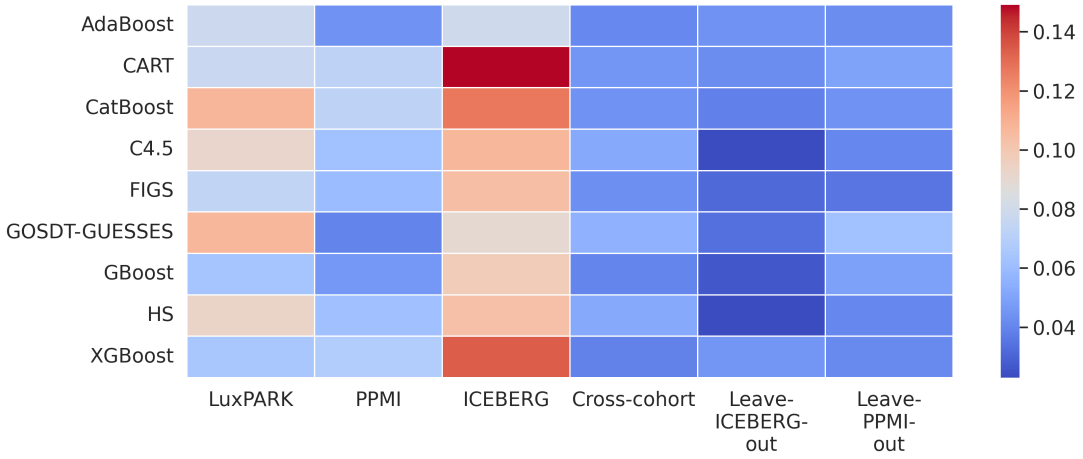
A.2 Stability of the optimized predictive models for predicting the risk of dyskinesia

Figure A.3 Stability analysis of *optimized comprehensive* LID classification models.



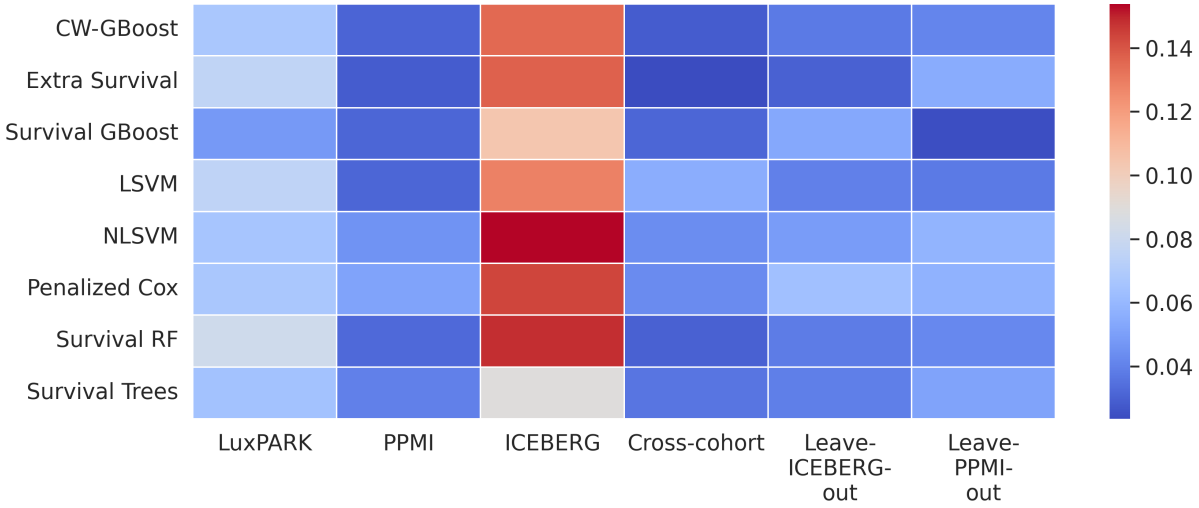
Stability analysis of *optimized comprehensive* predictive models for LID classification in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the AUC values obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

Figure A.4 Stability analysis of *optimized refined* LID classification models.



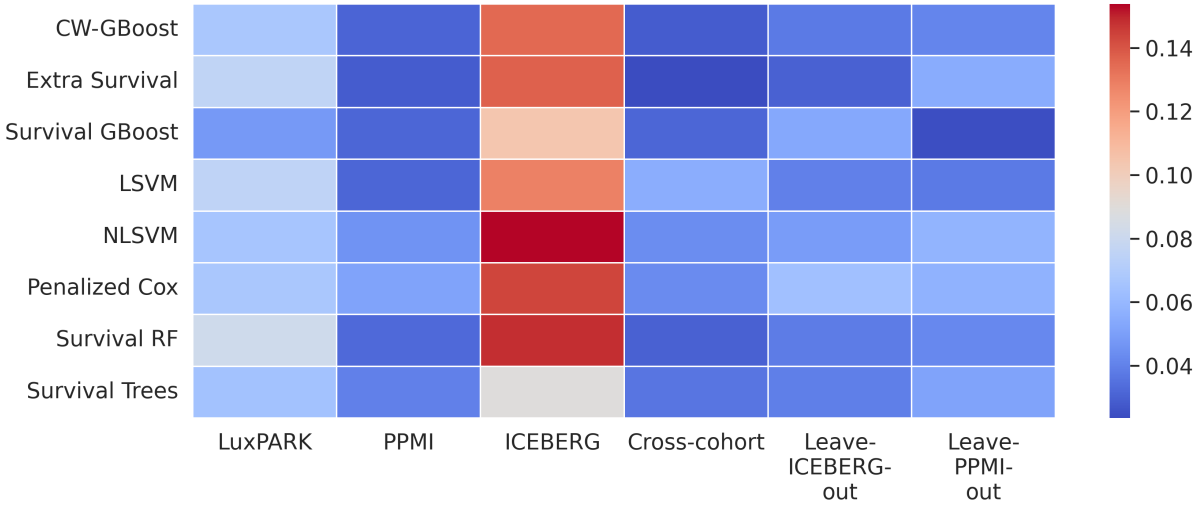
Stability analysis of *optimized refined* predictive models for LID classification in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the AUC values obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

Figure A.5 Stability analysis of *optimized comprehensive* time-to-LID models.



Stability analysis of *optimized comprehensive* time-to-LID models in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the C-indices obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

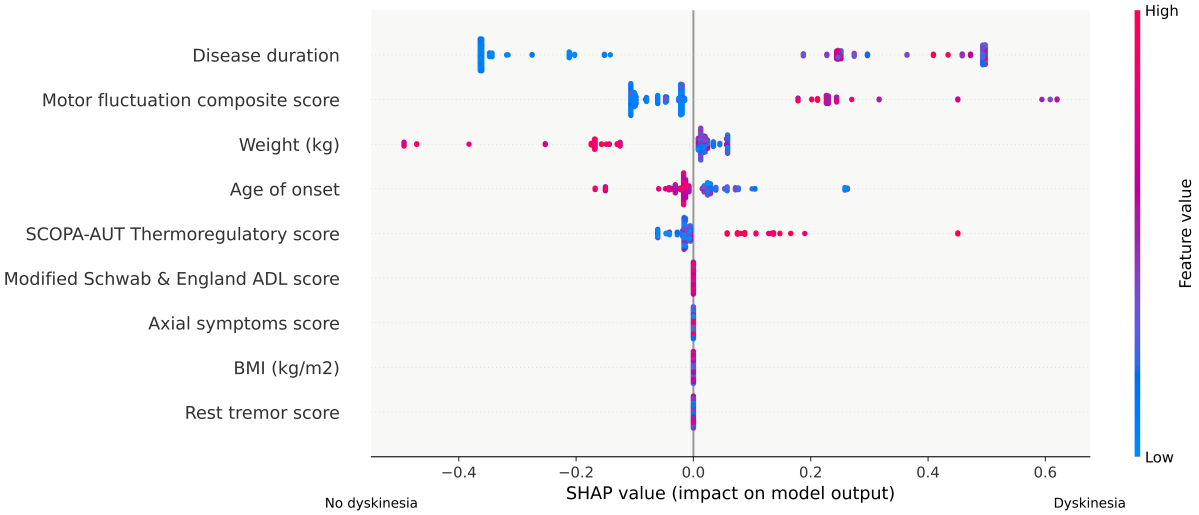
Figure A.6 Stability analysis of *optimized refined* time-to-LID models.



Stability analysis of *optimized refined* time-to-LID models in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the C-indices obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

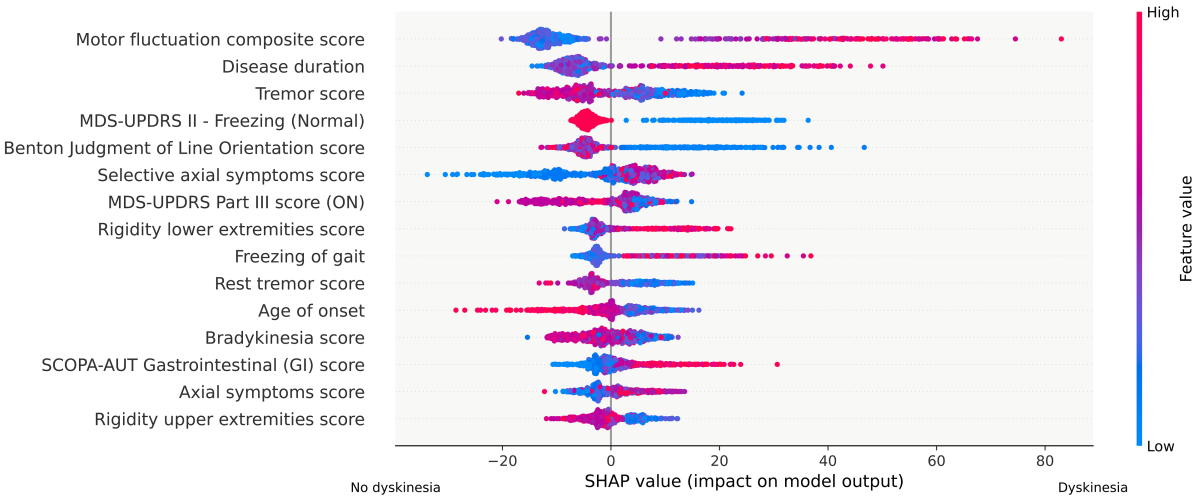
A.3 SHAP value analysis of optimized models with top 15 predictors for cross-cohort analysis

Figure A.7 SHAP values plot for the *optimized refined* LID classification model in cross-cohort analysis.



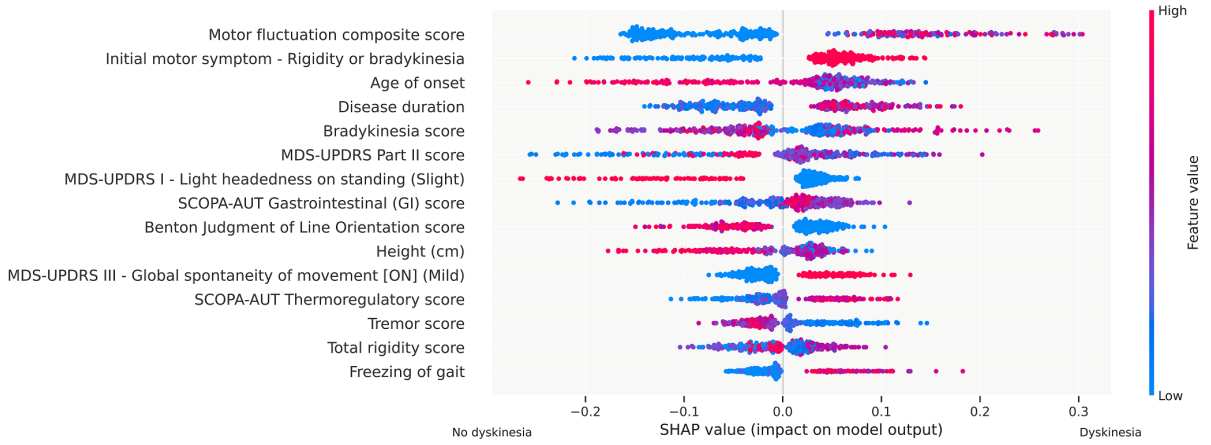
SHAP value plot displaying the top predictors for the *optimized refined* model in cross-cohort LID prognostic classification. The plot shows the magnitude and direction (positive or negative) of each feature's influence on LID prognosis status as output.

Figure A.8 SHAP values plot for the *optimized refined* time-to-LID model in cross-cohort analysis.



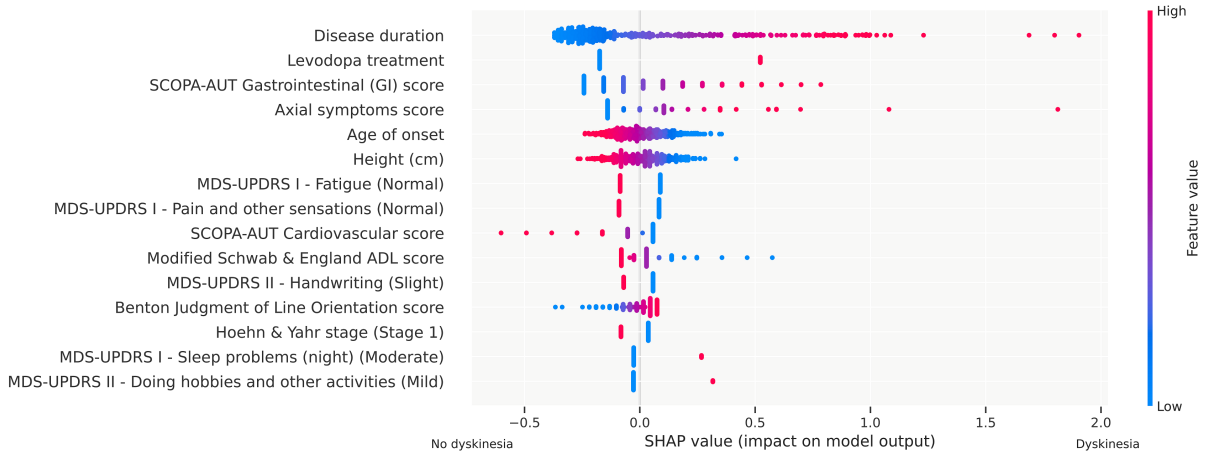
SHAP value plot displaying the top 15 predictors for the *optimized refined* model in cross-cohort time-to-LID analysis. The plot shows the magnitude and direction (positive or negative) of each feature's influence on time-to-LID as output.

Figure A.9 SHAP values plot for the *optimized comprehensive* LID classification model in LuxPARK.



SHAP value plot displaying the top 15 predictors for the *optimized comprehensive* LID classification in LuxPARK. The plot shows the magnitude and direction (positive or negative) of each feature's influence on LID prognosis status as output.

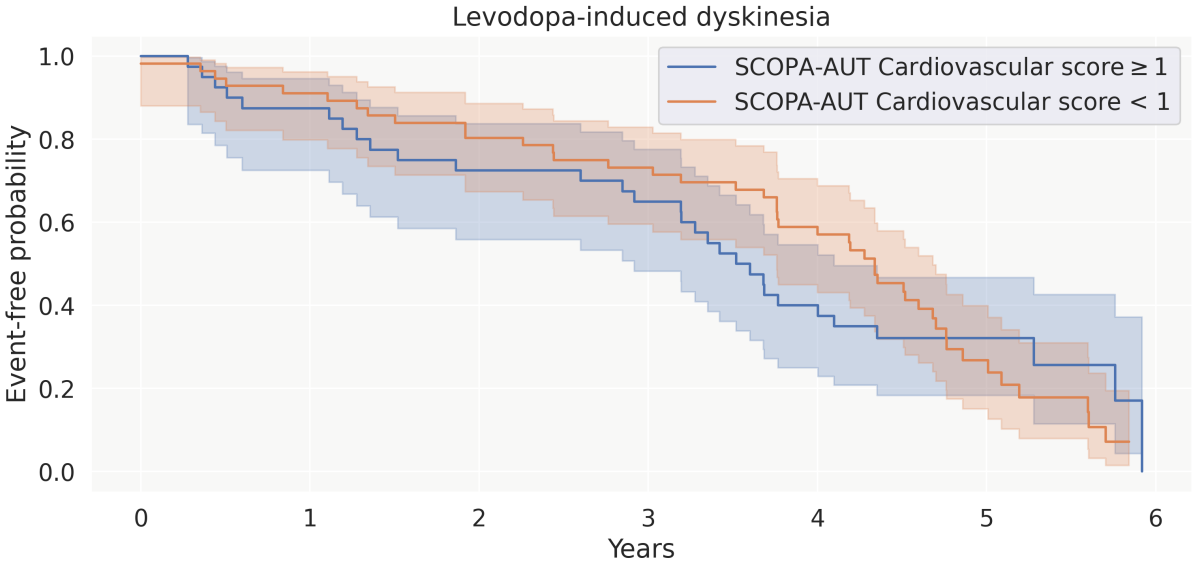
Figure A.10 SHAP values plot for the *optimized comprehensive* time-to-LID model in PPMI.



SHAP value plot displaying the top 15 predictors for the *optimized comprehensive* time-to-LID model in PPMI. The plot shows the magnitude and direction (positive or negative) of each feature's influence on time-to-LID as output.

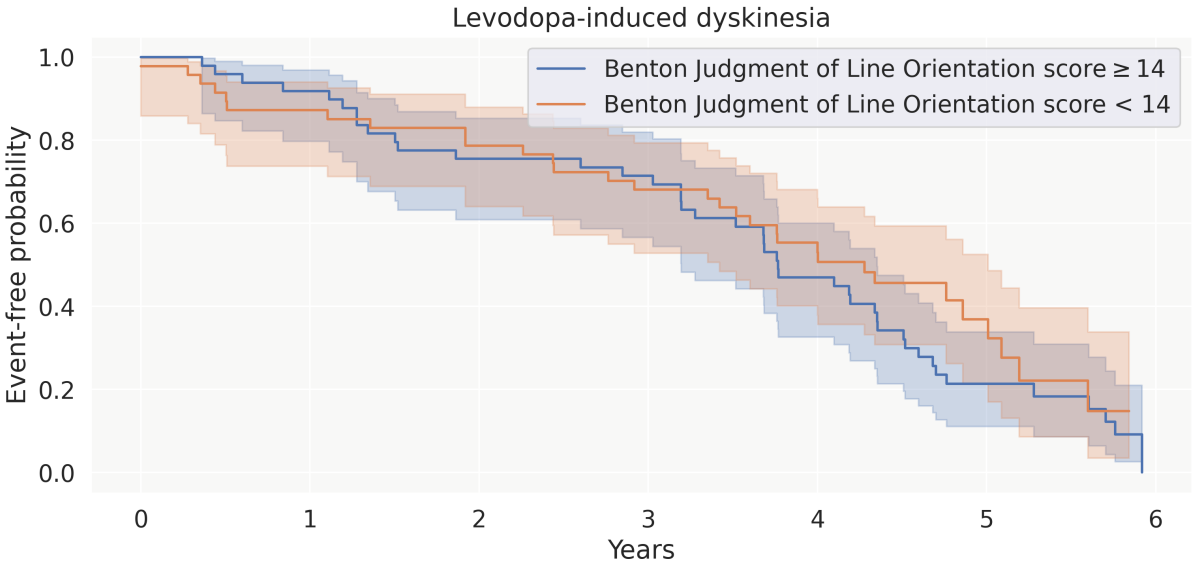
A.4 Kaplan-Meier analysis of predictors for time-to-LID

Figure A.11 Kaplan-Meier plot for SCOPA-AUT cardiovascular in PPMI.



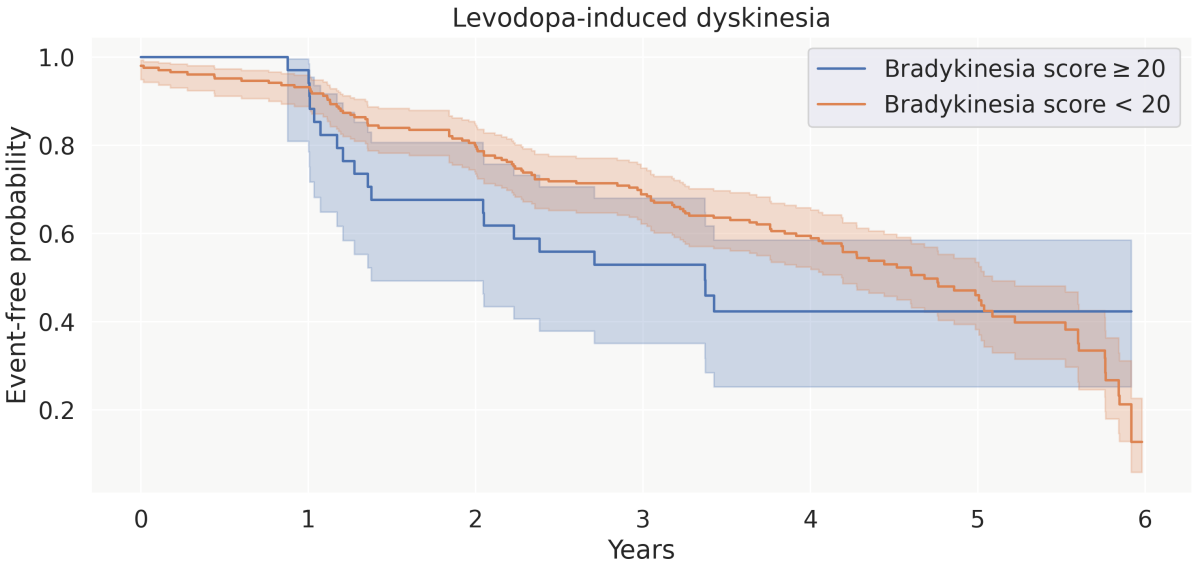
The Kaplan-Meier plot for the SCOPA-AUT cardiovascular feature about the onset of LID illustrates the time-to-LID distributions between the subgroups categorized based on cardiovascular scores in PPMI analysis (≥ 14 vs. < 14).

Figure A.12 Kaplan-Meier plot for Benton Judgment of Line Orientation in PPMI.



The Kaplan-Meier plot for the Benton Judgment of Line Orientation feature about the onset of LID illustrates the time-to-LID distributions between the subgroups categorized based on visuospatial scores in PPMI analysis ($\geq$ 14 vs. $<$ 14).

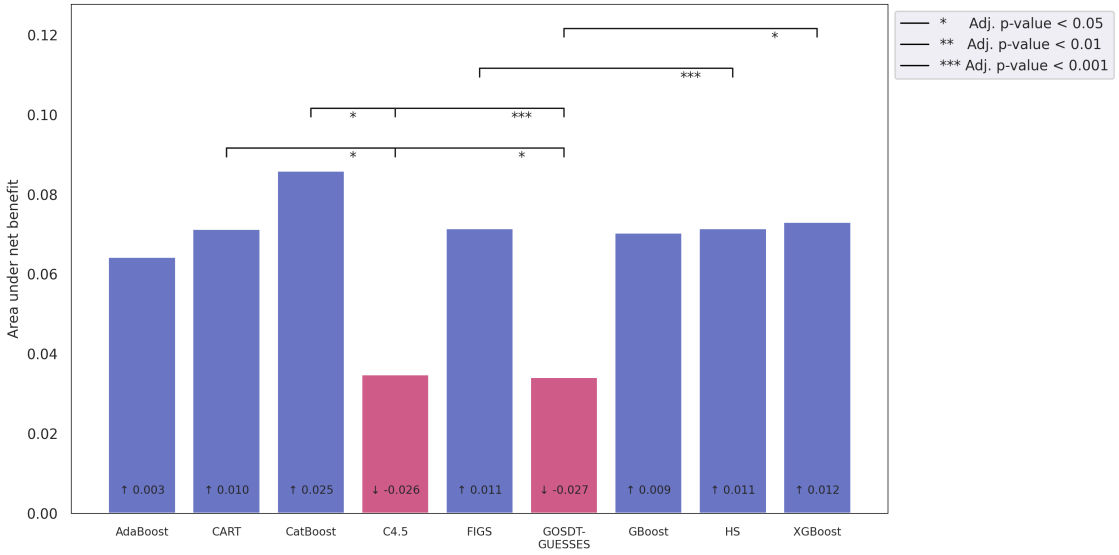
Figure A.13 Kaplan-Meier plot for bradykinesia in cross-cohort analysis.



The Kaplan-Meier plot for the bradykinesia feature about the onset of LID illustrates the time-to-LID distributions between the subgroups categorized based on bradykinesia scores ($\geq$ 20 vs. $<$ 20) in cross-cohort analysis.

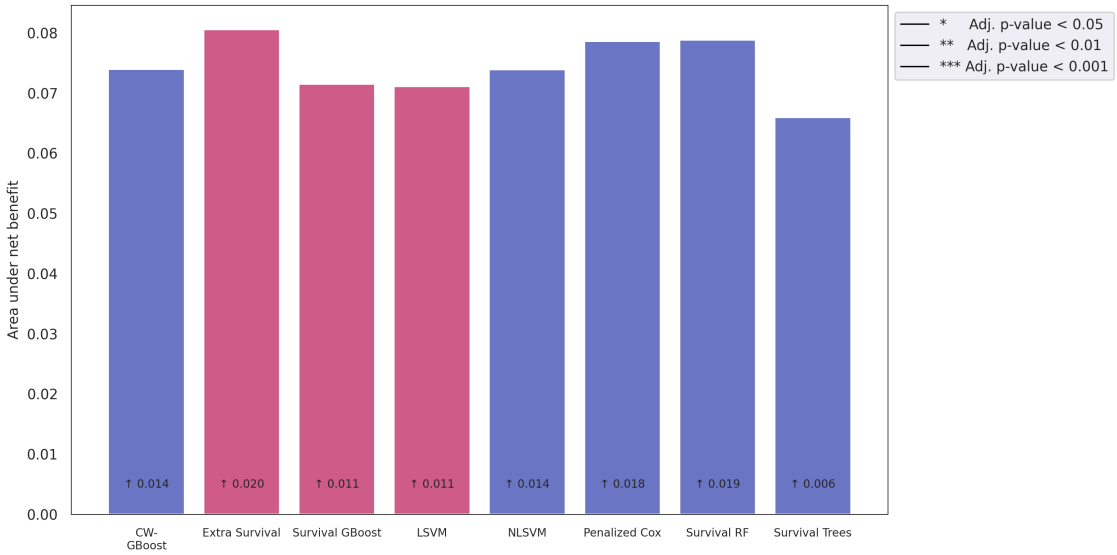
A.5 Evaluation of LID predictive models using decision curve analysis and calibration analysis

Figure A.14 Bar plot of the area under the positive net benefit curve for the *optimized refined* LID classification models in cross-cohort analysis.



Bar plot showing the area under the positive net benefit for different models in cross-cohort *refined* LID classification, with lines indicating significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

Figure A.15 Bar plot of the area under the positive net benefit curve for the *optimized refined* time-to-LID models in cross-cohort analysis.



Bar plot showing the area under the positive net benefit for different models in cross-cohort *refined* time-to-LID, with lines indicating significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

Appendix B

Interpretable machine learning for cross-cohort prediction of motor fluctuations in Parkinson's disease

B.1 Model performance metrics for motor fluctuations prognosis across cohort analyses

Table B.1 Predictive performance metrics for *refined* MF classification model in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features
AdaBoost	0.572 (0.100)	0.494	5 (6)	***0.682 (0.065)***	***0.661***	***6 (11)***	0.567 (0.103)	0.333	5 (7)
CART	0.532 (0.081)	0.542	8 (9)	0.676 (0.073)	0.606	2 (7)	0.566 (0.096)	0.592	3 (8)
CatBoost	0.549 (0.082)	0.526	7 (8)	0.694 (0.050)	0.648	6 (8)	***0.629 (0.127)***	***0.642***	***15 (25)***
C4.5	0.553 (0.082)	0.535	5 (5)	0.679 (0.046)	0.606	3 (5)	**0.715 (0.170)**	**0.503**	**8 (13)**
FIGS	0.546 (0.056)	0.541	1 (1)	0.683 (0.063)	0.637	4 (7)	0.651 (0.103)	0.558	7 (10)
GOSDT-GUESSES	0.553 (0.113)	0.551	12 (16)	0.626 (0.039)	0.625	14 (15)	0.560 (0.123)	0.464	7 (9)
GBoost	***0.625 (0.074)***	***0.600***	***5 (7)***	0.688 (0.042)	0.592	6 (12)	0.545 (0.149)	0.506	7 (12)
HS	0.525 (0.088)	0.541	1 (1)	0.683 (0.063)	0.637	4 (7)	0.639 (0.079)	0.531	6 (9)
XGBoost	0.577 (0.097)	0.533	43 (57)	**0.700 (0.048)**	**0.631**	**32 (43)**	0.476 (0.212)	0.553	22 (28)

An overview of the *refined* MF prognostic classification's predictive performance statistics summarizes the *comprehensive* MF prognostic classification's predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table B.2 Predictive performance metrics for *refined* time-to-MF model in single-cohort analyses.

Algorithm	LuxPARK			PPMI			ICEBERG		
	Mean (SD)	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**	**Mean (SD)**	**Hold-out C-index**	**Number of features**
CW-GBoost	0.514 (0.131)	0.654	8 (13)	0.691 (0.045)	0.704	16 (31)	0.577 (0.063)	0.525	3 (11)
Extra Survival	***0.555 (0.114)***	***0.665***	***132 (135)***	**0.692 (0.053)**	**0.704**	**78 (107)**	0.498 (0.135)	0.665	80 (94)
Survival GBoost	0.525 (0.116)	0.602	17 (21)	0.676 (0.047)	0.671	11 (16)	0.565 (0.073)	0.617	40 (60)
LSVM	0.475 (0.162)	0.577	26 (26)	0.682 (0.061)	0.707	28 (28)	0.49 (0.159)	0.647	11 (11)
NLSVM	0.506 (0.202)	0.583	18 (18)	0.683 (0.031)	0.685	37 (37)	0.496 (0.112)	0.554	15 (15)
Penalized Cox	0.466 (0.106)	0.602	1 (3)	0.683 (0.039)	0.701	25 (28)	***0.509 (0.083)***	***0.674***	***42 (56)***
Survival RF	0.543 (0.092)	0.628	7 (9)	***0.672 (0.037)***	***0.715***	***95 (104)***	**0.577 (0.057)**	**0.517**	**11 (11)**
Survival Trees	0.527 (0.171)	0.657	6 (12)	0.623 (0.046)	0.637	6 (7)	0.572 (0.071)	0.453	7 (11)

An overview of the *refined* time-to-MF predictive performance statistics summarizes the *comprehensive* time-to-MF predictive performance statistics in single cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-indices in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out C-index is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table B.3 Predictive performance metrics for *refined* MF classification model in multi-cohort analyses.

Algorithm	Cross-cohort			Leave-ICEBERG-out		
	Mean (SD)	Hold-out AUC	Number of features	Mean (SD)	Hold-out AUC	Number of features
AdaBoost	***0.619 (0.03)***	***0.640***	***2 (4)***	0.638 (0.032)	0.528	3 (7)
CART	0.614 (0.057)	0.584	3 (6)	0.622 (0.052)	0.528	2 (4)
CatBoost	0.617 (0.037)	0.582	8 (9)	**0.650 (0.036)**	**0.586**	**17 (28)**
C4.5	0.605 (0.037)	0.577	4 (10)	0.607 (0.041)	0.476	2 (6)
FIGS	0.614 (0.042)	0.629	1 (3)	0.622 (0.038)	0.618	5 (12)
GOSDT-GUESSES	0.599 (0.025)	0.520	30 (45)	0.606 (0.052)	0.569	35 (58)
GBoost	**0.625 (0.014)**	**0.591**	**7 (17)**	0.633 (0.043)	0.558	20 (42)
HS	0.614 (0.042)	0.629	1 (3)	***0.622 (0.038)***	***0.618***	***5 (12)***
XGBoost	0.595 (0.042)	0.598	14 (15)	0.636 (0.051)	0.520	41 (69)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
AdaBoost	0.625 (0.057)	0.635	9 (20)	0.640 (0.046)	0.579	3 (4)
CART	0.615 (0.060)	0.631	9 (29)	0.628 (0.048)	0.569	4 (10)
CatBoost	0.616 (0.054)	0.603	11 (25)	**0.651 (0.022)**	**0.581**	**5 (7)**
C4.5	0.582 (0.034)	0.658	7 (12)	0.611 (0.032)	0.609	5 (9)
FIGS	0.620 (0.057)	0.660	4 (9)	0.650 (0.058)	0.534	4 (11)
GOSDT-GUESSES	0.609 (0.067)	0.582	25 (39)	0.622 (0.03)	0.532	19 (19)
GBoost	**0.635 (0.064)**	**0.647**	**19 (33)**	***0.647 (0.042)***	***0.613***	***3 (5)***
HS	***0.620 (0.057)***	***0.660***	***4 (9)***	0.650 (0.058)	0.534	4 (11)
XGBoost	0.633 (0.036)	0.622	47 (53)	0.638 (0.019)	0.532	56 (80)

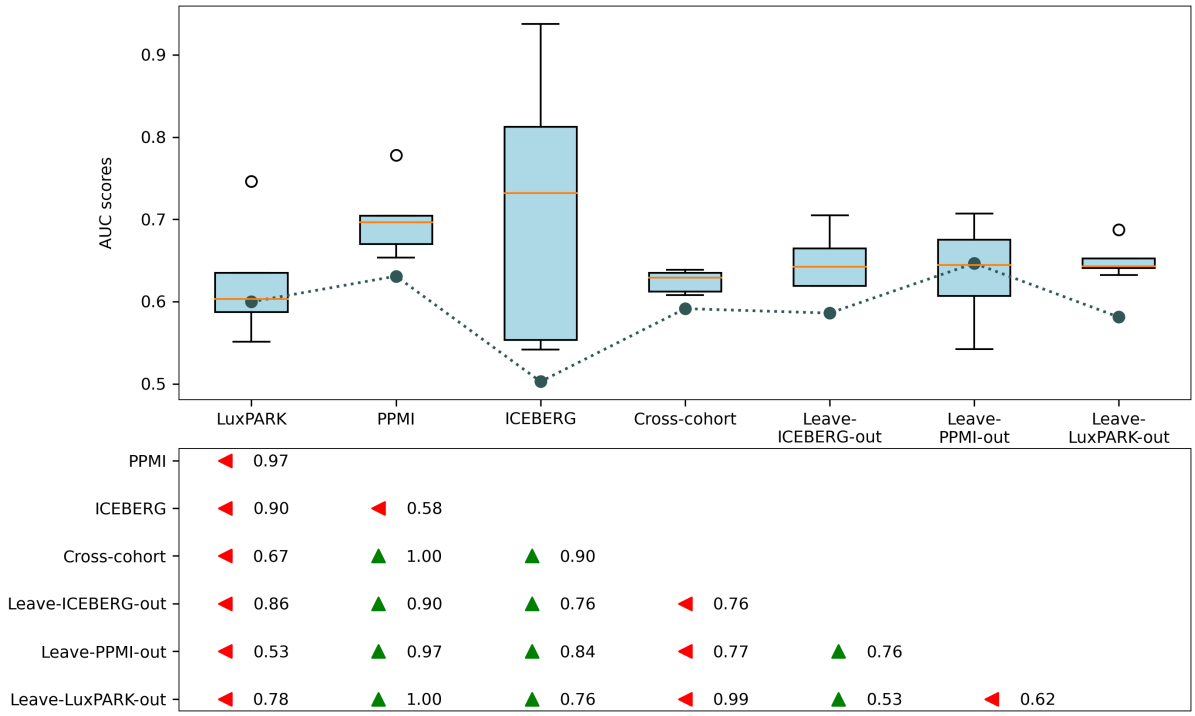
An overview of the *refined* MF prognostic classification's predictive performance statistics summarizes the *comprehensive* MF prognostic classification's predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out AUC values and the corresponding number of features used in each *optimized* model. Models with the highest average AUC scores in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out AUC score is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Table B.4 Predictive performance metrics for *refined* time-to-MF model in multi-cohort analyses.

Algorithm	Cross-cohort			Leave-ICEBERG-out		
	Mean (SD)	Hold-out C-index	Number of features	Mean (SD)	Hold-out C-index	Number of features
CW-GBoost	0.643 (0.02)	0.640	7 (23)	0.683 (0.042)	0.607	10 (24)
Extra Survival	***0.637 (0.057)***	***0.687***	***160 (165)***	0.680 (0.010)	0.581	95 (142)
Survival GBoost	**0.646 (0.021)**	**0.633**	**14 (26)**	***0.662 (0.091)***	***0.624***	***14 (26)***
LSVM	0.636 (0.030)	0.667	60 (60)	0.673 (0.038)	0.598	49 (49)
NLSVM	0.630 (0.047)	0.673	45 (45)	0.669 (0.031)	0.617	52 (52)
Penalized Cox	0.590 (0.082)	0.614	2 (3)	0.673 (0.048)	0.500	1 (3)
Survival RF	0.637 (0.027)	0.670	28 (62)	**0.687 (0.031)**	**0.572**	**50 (91)**
Survival Trees	0.603 (0.039)	0.593	16 (17)	0.614 (0.073)	0.534	6 (14)
	Leave-PPMI-out			**Leave-LuxPARK-out**		
CW-GBoost	0.609 (0.034)	0.668	3 (7)	0.730 (0.035)	0.629	16 (29)
Extra Survival	0.618 (0.036)	0.665	19 (19)	***0.716 (0.024)***	***0.636***	***30 (63)***
Survival GBoost	0.613 (0.034)	0.582	36 (70)	0.704 (0.022)	0.632	16 (20)
LSVM	**0.623 (0.033)**	**0.661**	**45 (45)**	0.722 (0.033)	0.616	36 (36)
NLSVM	0.614 (0.045)	0.636	50 (50)	0.725 (0.031)	0.583	53 (53)
Penalized Cox	0.607 (0.042)	0.590	2 (4)	**0.738 (0.043)**	**0.585**	**86 (119)**
Survival RF	***0.605 (0.040)***	***0.687***	***101 (120)***	0.712 (0.042)	0.628	106 (112)
Survival Trees	0.562 (0.058)	0.541	25 (47)	0.664 (0.016)	0.588	12 (17)

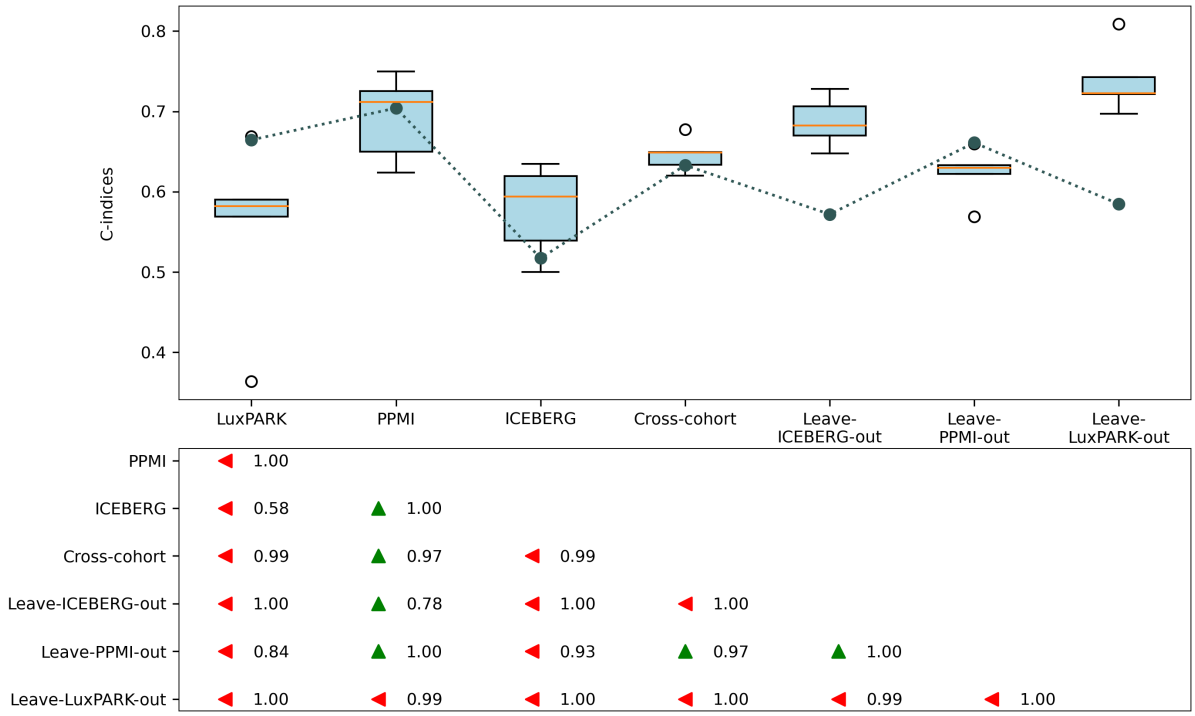
An overview of the *refined* time-to-MF predictive performance statistics summarizes the *comprehensive* time-to-MF predictive performance statistics in multi-cohort analyses. The *optimized* models are listed with cross-validated and hold-out C-indices and the corresponding number of features used in each *optimized* model. Models with the highest average C-indices in the cross-validation of the cohort analyses are highlighted in bold. The model with the highest hold-out C-index is indicated in *italics*. The column labeled 'Number of features' displays the number of candidate features selected during nested cross-validation. The number in front of the brackets indicates the number of selected predictive features in the cross-validation determined through permutation importance analysis.

Figure B.1 Comparison of cross-validated AUC scores for *refined* MF classification models.



A comparison of cross-validated AUC scores and probabilities of better predictive performance for the *optimized refined* MF classification model across cohort analyses. The upper part displays boxplots of the cross-validated AUC scores for each cohort, with the line indicating the hold-out AUC scores for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

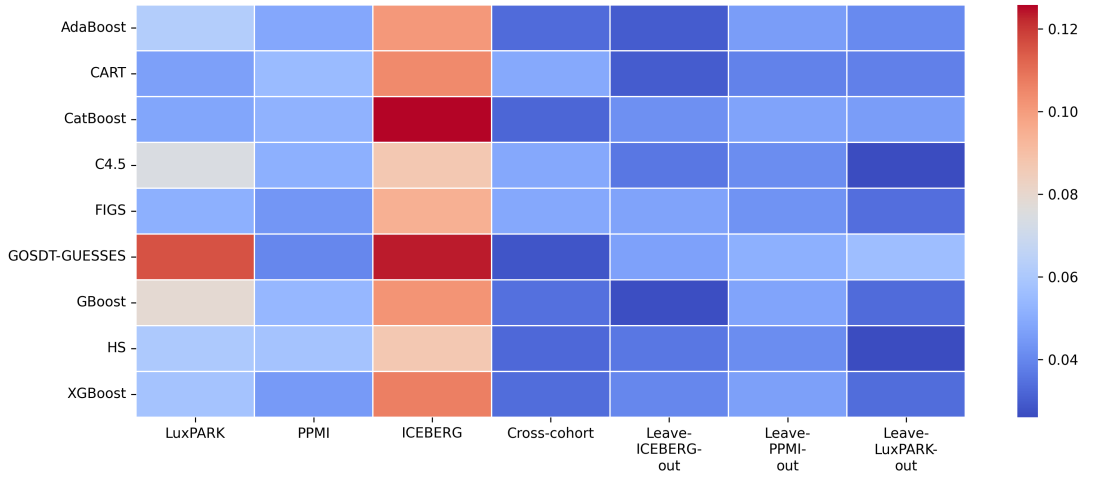
Figure B.2 Comparison of cross-validated C-indices for *refined* time-to-MF models.



A comparison of cross-validated C-indices and probabilities of better predictive performance for the *optimized refined* time-to-MF model across cohort analyses. The upper part displays boxplots of the cross-validated C-indices for each cohort, with the line indicating the hold-out C-indices for each cohort, while the lower part shows the probabilities of one cohort's predictive performance being better than another's. The arrows indicate higher probabilities of predictive performance. They point towards the cohort with the higher probability of better performance for the *optimized* model.

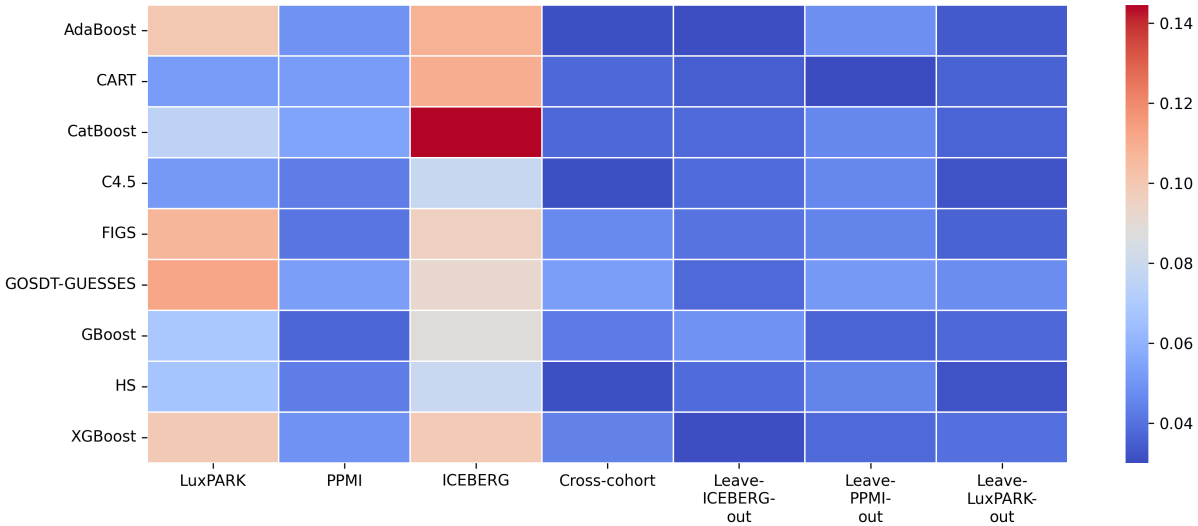
B.2 Stability of the optimized predictive models for predicting the risk of motor fluctuations

Figure B.3 Stability analysis of *optimized comprehensive* MF classification models.



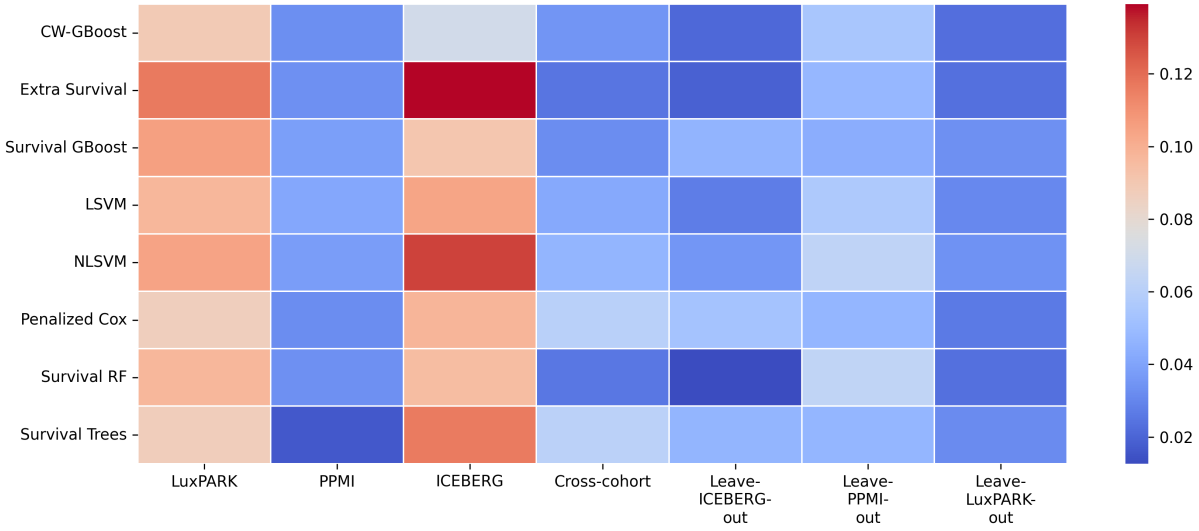
Stability analysis of *comprehensive optimized* predictive models for MF classification in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the AUC values obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

Figure B.4 Stability analysis of *optimized refined* MF classification models.



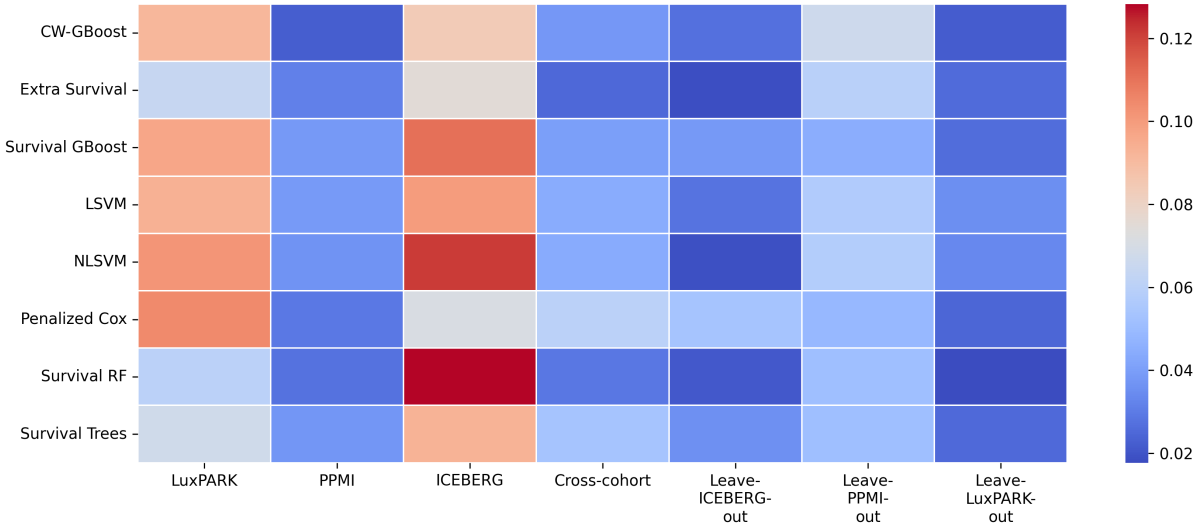
Stability analysis of *refined optimized* predictive models for MF classification in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the AUC values obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

Figure B.5 Stability analysis of *optimized comprehensive* time-to-MF models.



Stability analysis of *comprehensive optimized* time-to-MF models in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the C-indices obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

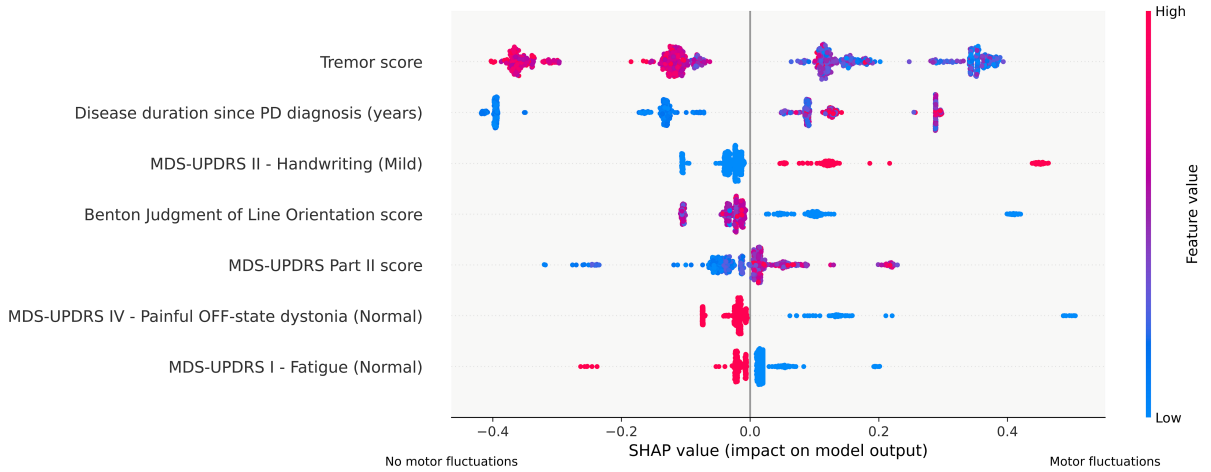
Figure B.6 Stability analysis of *optimized refined* time-to-MF models.



Stability analysis of *refined optimized* time-to-MF models in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the C-indices obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

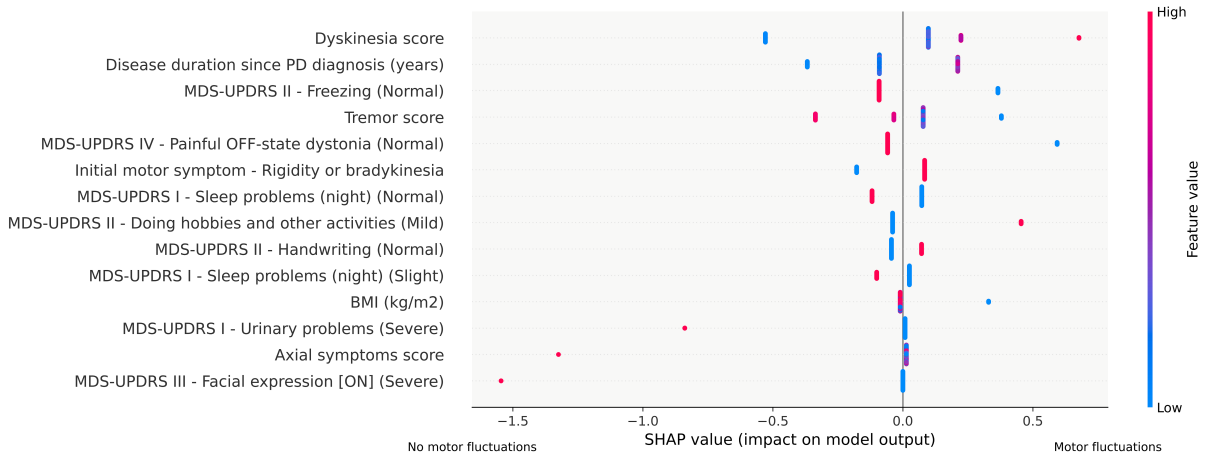
B.3 SHAP value analysis of optimized models with top 15 predictors for cross-cohort analysis

Figure B.7 SHAP values plot for the *optimized refined* MF classification model in cross-cohort analysis.



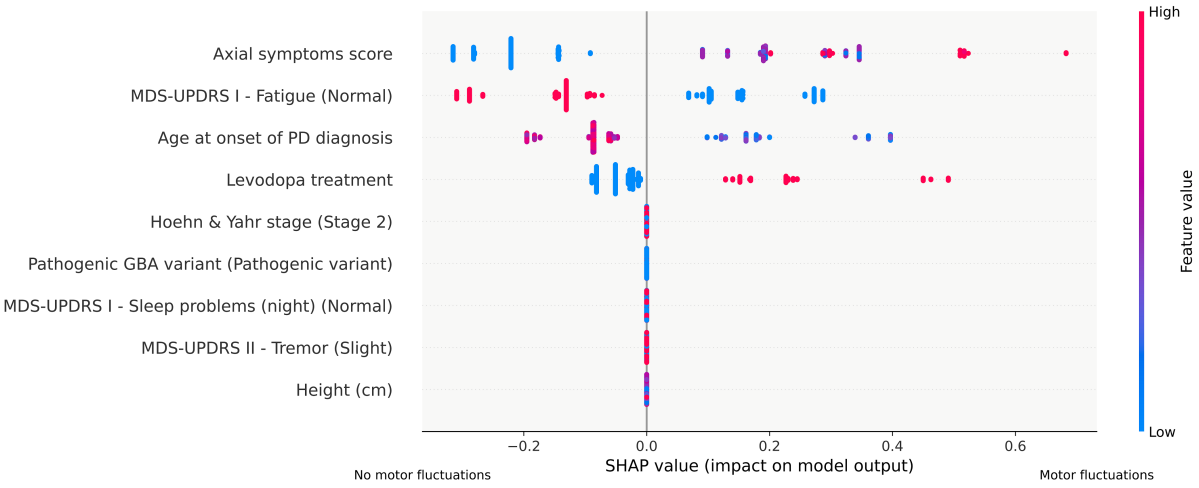
SHAP value plot displaying the top 15 predictors for the *optimized refined* model in cross-cohort motor fluctuations prognostic classification. The plot shows the magnitude and direction (positive or negative) of each feature's influence on LID prognosis status as output.

Figure B.8 SHAP values plot for the *optimized refined* time-to-MF model in cross-cohort analysis.



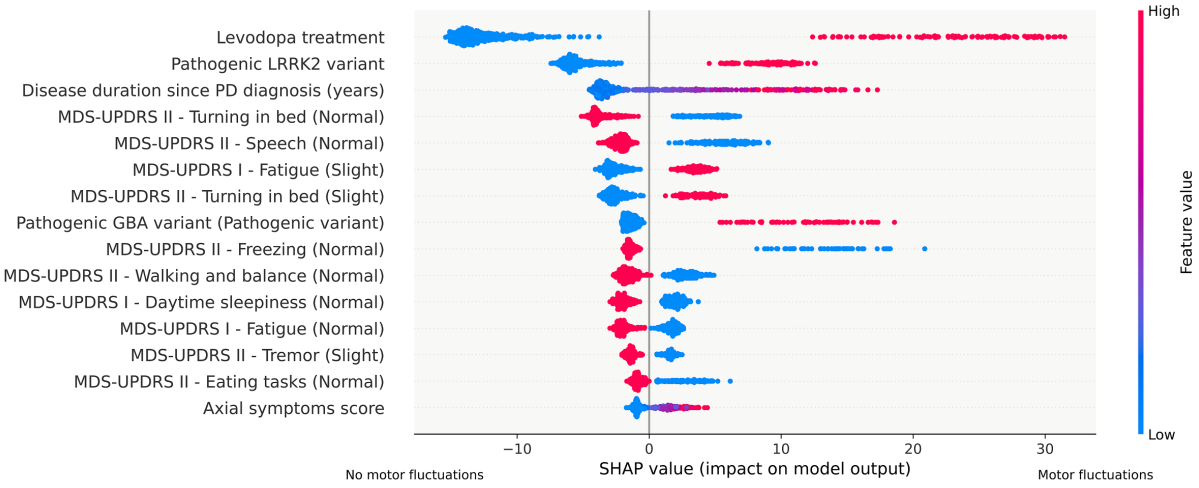
SHAP value plot displaying the top 15 predictors for the *optimized refined* model in cross-cohort time-to-MF analysis. The plot shows the magnitude and direction (positive or negative) of each feature's influence on time-to-MF as output.

Figure B.9 SHAP values plot for the *optimized comprehensive* MF classification model in PPMI.



SHAP value plot displaying the top predictors for the *optimized comprehensive* motor fluctuations classification in PPMI. The plot shows the magnitude and direction (positive or negative) of each feature's influence on motor fluctuations prognosis status as output.

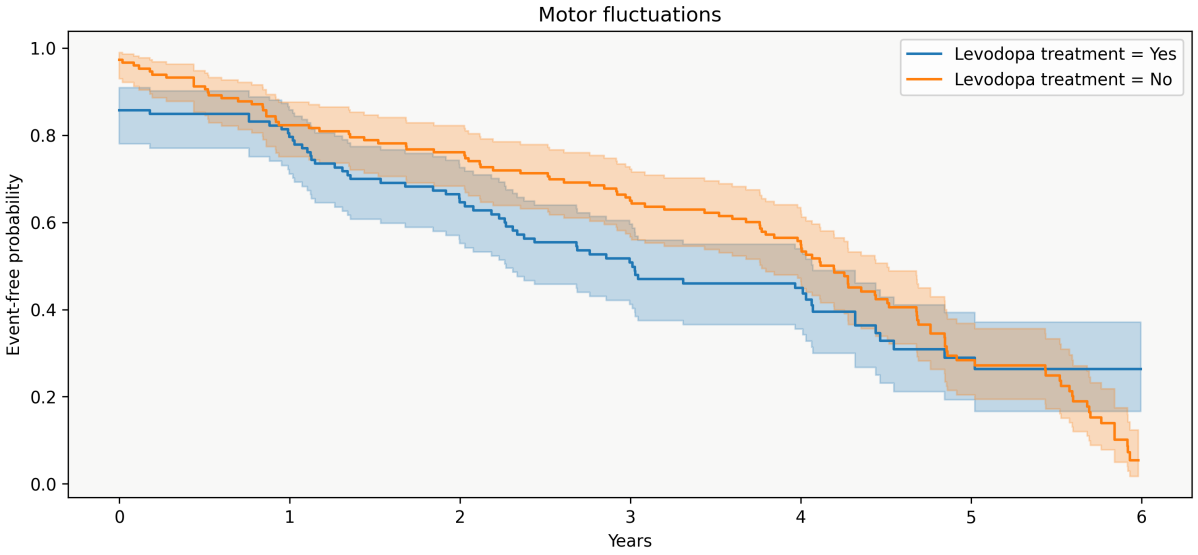
Figure B.10 SHAP values plot for the *optimized comprehensive* time-to-MF model in PPMI.



SHAP value plot displaying the top 15 predictors for the *optimized comprehensive* time-to-MF in PPMI. The plot shows the magnitude and direction (positive or negative) of each feature's influence on time-to-MF as output.

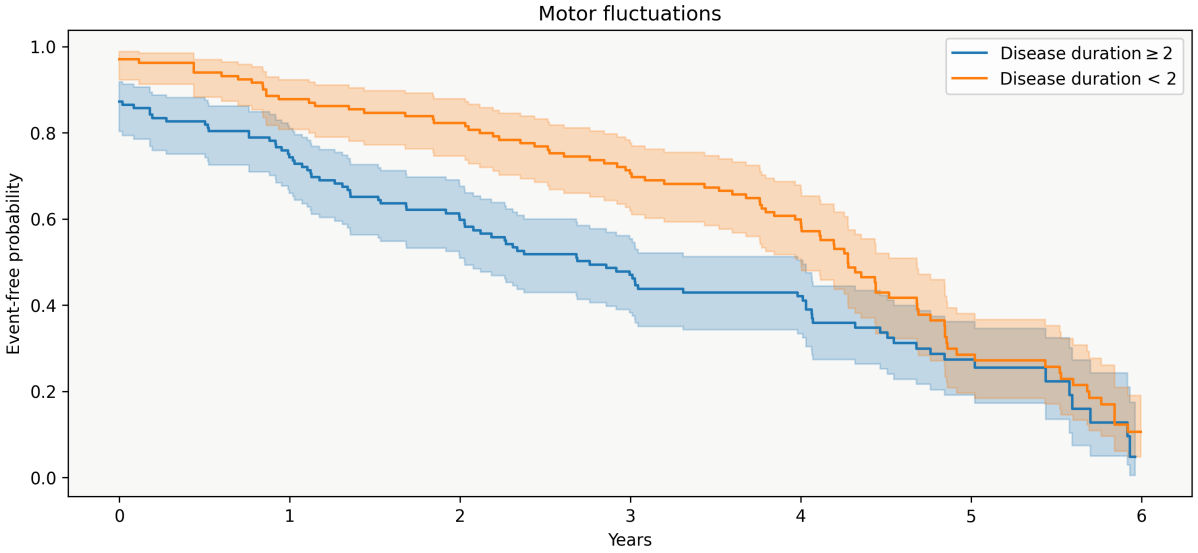
B.4 Kaplan-Meier analysis of predictors for time-to-MF

Figure B.11 Kaplan-Meier plot for levodopa medication intake in cross-cohort analysis.



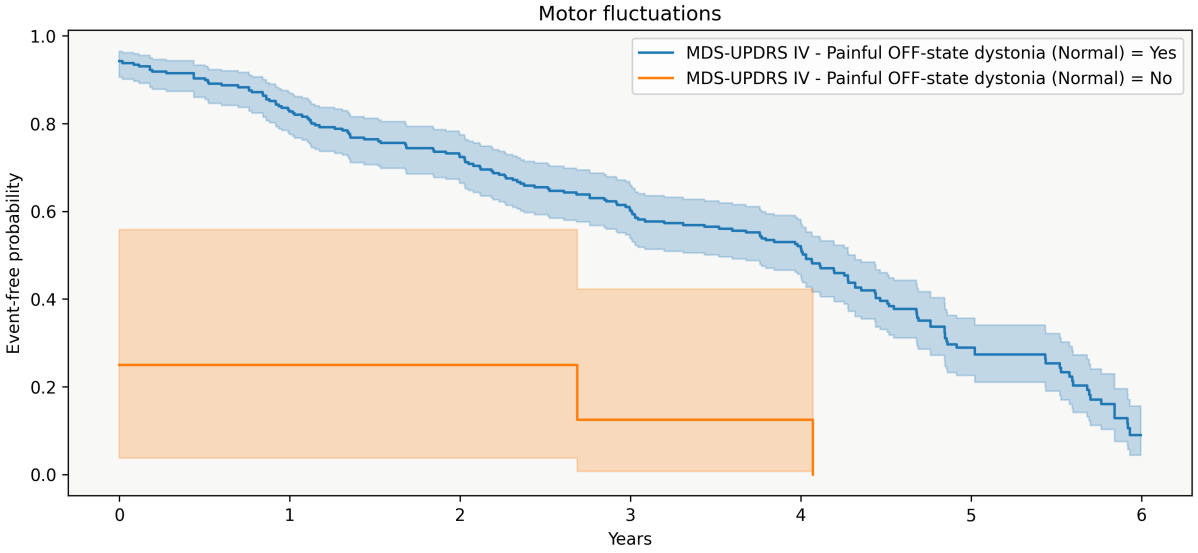
The Kaplan-Meier plot for the levodopa treatment about the onset of motor fluctuations illustrates the time-to-MF distributions between the subgroups categorized based on levodopa treatment in cross-cohort analysis.

Figure B.12 Kaplan-Meier plot for disease duration in cross-cohort analysis.



The Kaplan-Meier plot for the disease duration about the onset of motor fluctuations illustrates the time-to-MF distributions between the subgroups categorized based on disease duration in cross-cohort analysis.

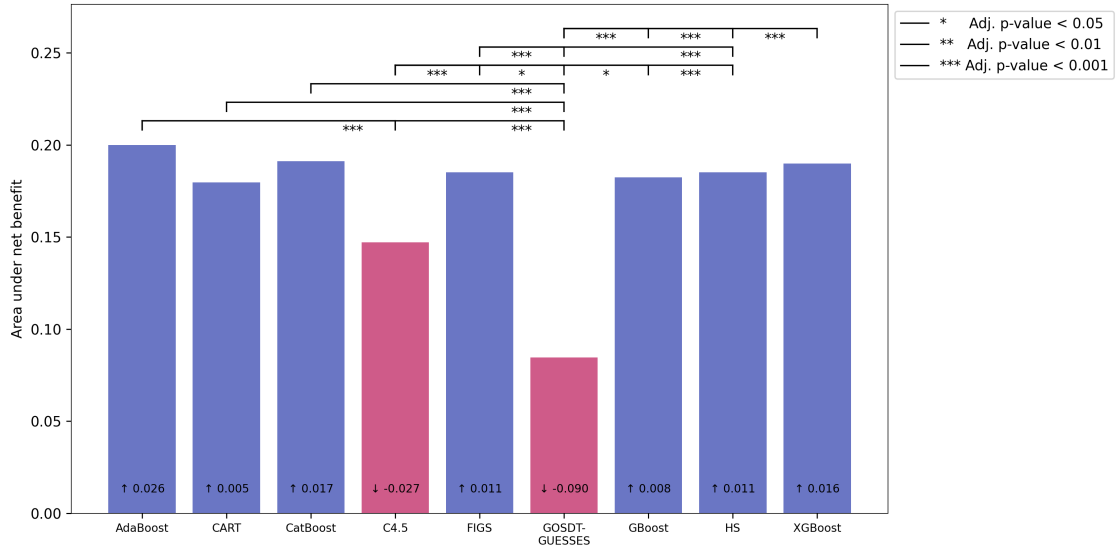
Figure B.13 Kaplan-Meier plot for MDS-UPDRS IV - painful off-state dystonia in cross-cohort analysis.



The Kaplan-Meier plot for the MDS-UPDRS IV - painful Off-state dystonia about the onset of motor fluctuations illustrates the time-to-MF distributions between the subgroups categorized based on the presence of dystonia in cross-cohort analysis.

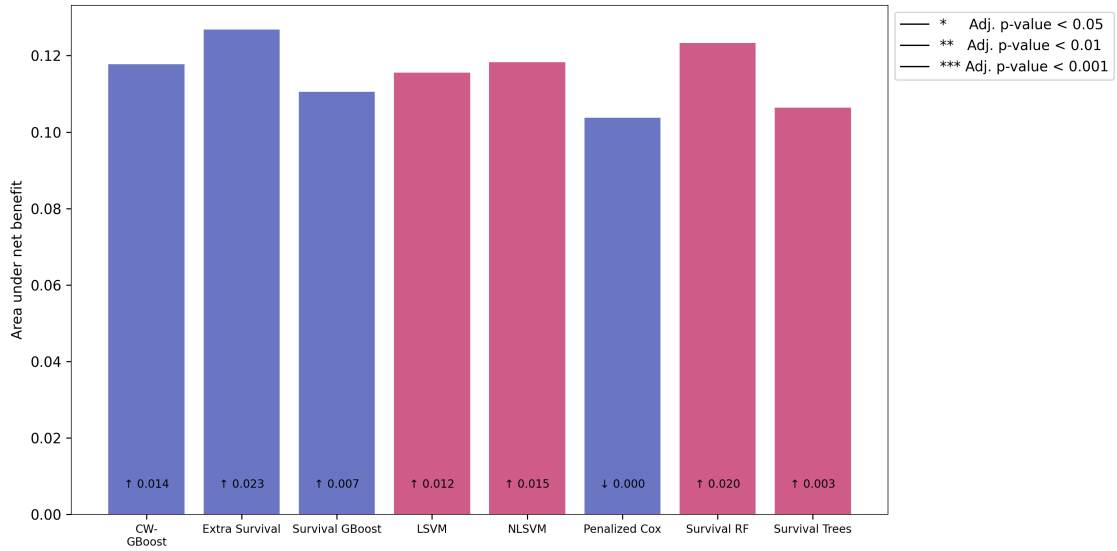
B.5 Evaluation of motor fluctuations predictive models using decision curve analysis and calibration analysis

Figure B.14 Bar plot of the area under the positive net benefit curve for the *optimized refined* MF classification models in cross-cohort analysis.



The bar plot shows the area under the positive net benefit for different cross-cohort *refined* motor fluctuations classification models. The lines indicate significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

Figure B.15 Bar plot of the area under the positive net benefit curve for the *optimized refined* time-to-MF models in cross-cohort analysis.



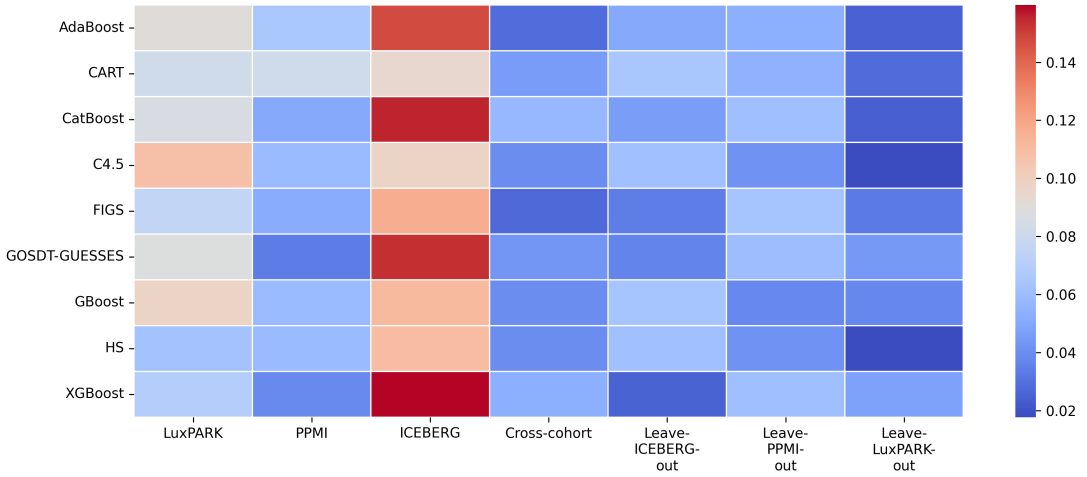
The bar plot shows the area under the positive net benefit for different cross-cohort *refined* time-to-MF. The lines indicate significant differences in the net benefit area across models. Blue bars represent models with a larger positive net benefit area than the negative net benefit, while red bars indicate the opposite. The numbers within the bars show the difference in the net benefit area between each model and the 'all intervention' baseline. Upward arrows (↑) signify a larger area than 'all intervention', while downward arrows (↓) indicate a smaller area.

Appendix C

Multi-cohort machine learning identifies predictors of cognitive impairment in Parkinson's disease

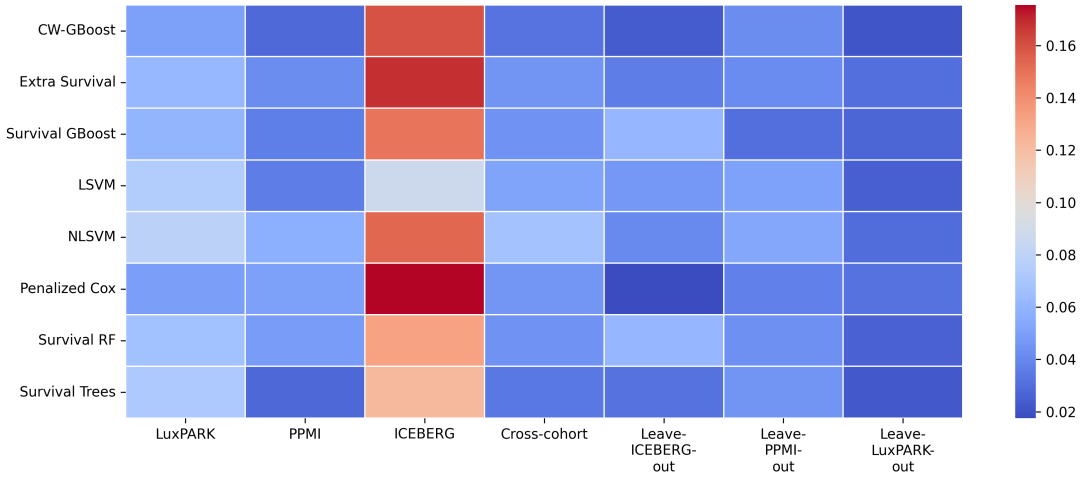
C.1 Stability of the optimized predictive models for predicting the risk of cognitive impairment

Figure C.1 Stability analysis of *optimized PD-MCI* classification models.



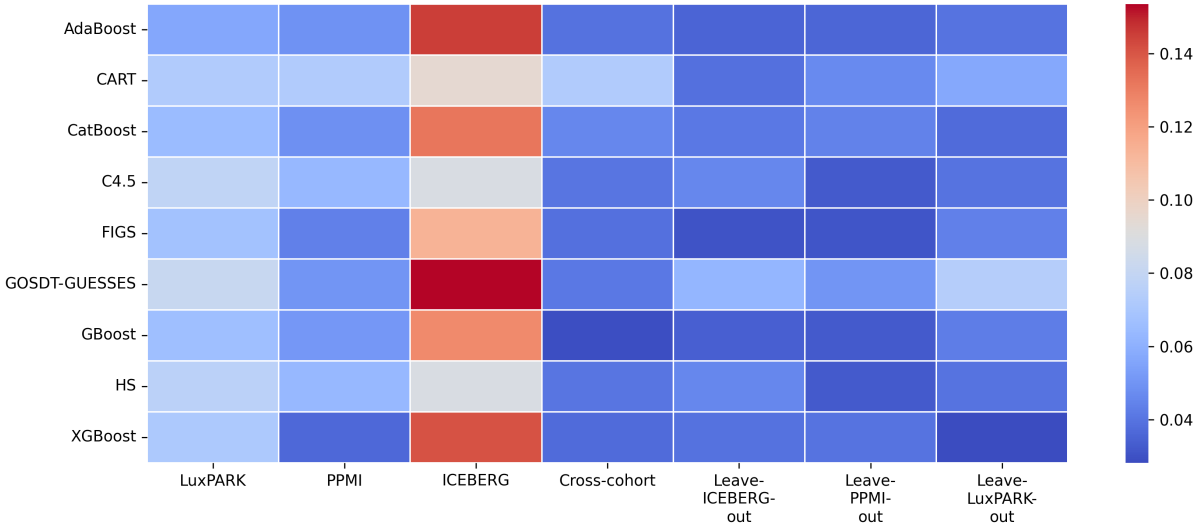
Stability analysis of *optimized* predictive models for *PD-MCI* classification in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the AUC values obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

Figure C.2 Stability analysis of *optimized* time-to-*PD-MCI* models.



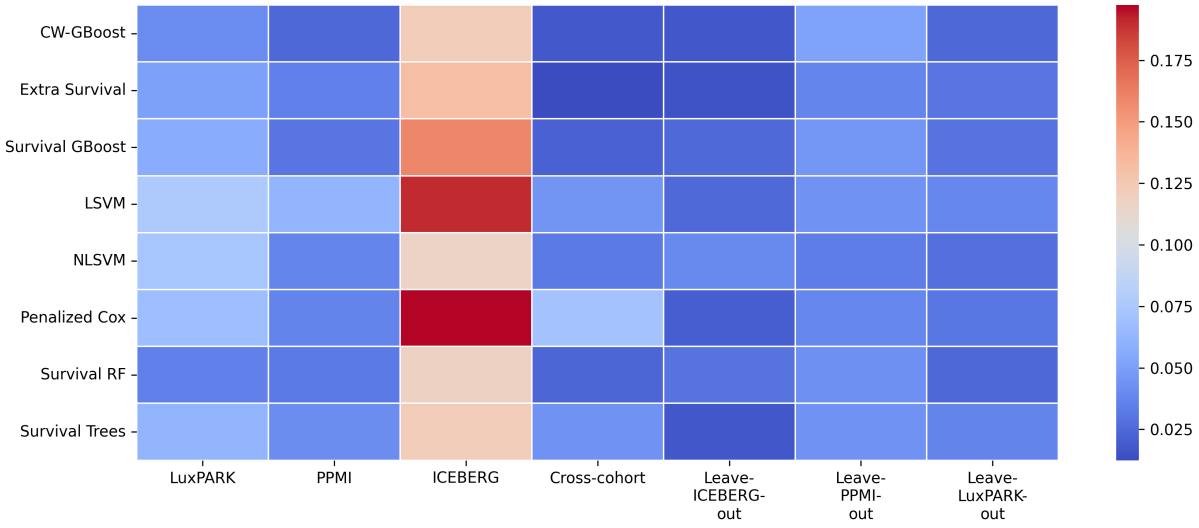
Stability analysis of *optimized* time-to-*PD-MCI* models in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the C-indices obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

Figure C.3 Stability analysis of *optimized PRCD* classification models.



Stability analysis of *optimized* predictive models for *PRCD* classification in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the AUC values obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.

Figure C.4 Stability analysis of *optimized* time-to-*PRCD* models.



Stability analysis of *optimized* time-to-*PRCD* models in PD across different algorithms and cohort studies. The stability of the model is evaluated by calculating the SD of the C-indices obtained from the nested CV process, which corresponds to the optimal hyperparameters. A lower SD indicates a higher stability of the predictive models, reflecting a consistent performance across diverse cohorts and ML techniques.