ABSTRACT

Title of Dissertation: EXPLORING THE POWER OF MACHINE LEARNING IN MEDICAL RESEARCH: A FOCUS ON MOVEMENT DISORDER DIAGNOSIS AND AGE-RELATED HEARING LOSS

Rana M. Khalil
Doctor of Philosophy, 2025

Dissertation Directed by: Professor Michael P. Cummings
Department of Biology

The medical research field is experiencing a remarkable evolution due to the application of data science and machine learning techniques and new developments in these fields. The accessibility of large datasets, enhanced computing capabilities, and advanced algorithms have opened up new possibilities to extract valuable insights, identify patterns, and develop predictive models from complex biomedical data. This integration has the potential to revolutionize medical research, resulting in enhanced diagnostic capabilities, personalized treatment approaches, and ultimately, improved patient care.

In this dissertation, I explore the impact of data science and machine learning in medical research, with a specific focus on the diagnosis of movement disorders and age-related hearing loss. In the first of these domain areas, I use data from a wearable sensor to accurately identify individuals with Parkinson's disease based on their movements during several motor tasks. I demonstrate that applying machine learning to wearable sensor data can achieve diagnostic

accuracy surpassing that of movement disorder experts in routine clinical settings for differentiating Parkinson's disease from controls and is comparable to expert clinicians for distinguishing Parkinson's disease from other parkinsonian disorders. I also found that repeating mobility tasks is unnecessary for improving diagnostic accuracy. I propose several steps to simplify mobility test protocols, which can save time and effort for both clinicians and participants without compromising accuracy. Specifically, using a single sensor, a single mobility task, and just one trial of each task for the classification tasks explored in this study can streamline the process. This approach facilitates the practical application of wearable sensors as a diagnostic tool in clinical settings.

In the second domain area, I study age-related hearing loss by constructing ensemble models to examine data from participants with diverse ages and varying degrees of hearing loss. By integrating audiometric, perceptual, electrophysiological, and cognitive data, I predict speech perception in challenging auditory conditions like noise, reverberation, and time compression. Leveraging machine learning techniques, my objective is to identify the variables that are highly predictive of demanding speech-perception conditions, thereby confirming existing associations and potentially uncovering novel ones. The findings underscore the critical role of audiometric thresholds, particularly within the 1–4 kHz range, and emphasize the utility of composite variables spanning multiple frequencies in accurately predicting speech perception. Furthermore, basic temporal processing ability demonstrates a moderate influence, whereas cognitive factors and extended high-frequency thresholds exhibit limited to negligible predictive capability in this context. Continued research and exploration of associations will contribute to a deeper understanding of the complex interplay between speech perception, aging, hearing loss, and cognition.

EXPLORING THE POWER OF MACHINE LEARNING IN MEDICAL RESEARCH: A FOCUS ON MOVEMENT DISORDER DIAGNOSIS AND AGE-RELATED HEARING LOSS

by

Rana M. Khalil

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2025

Advisory Committee:
Professor Michael P. Cummings, PhD, Chair/Advisor
Professor Mihai Pop, PhD
Professor Abhinav Shrivastava, PhD
Professor Rob Patro, PhD
Professor Rainer von Coelln, Dr. med.
Professor Matthew Goupell, PhD

© Copyright by
Rana M. Khalil
2025

Acknowledgments

I am deeply grateful to all the individuals who have contributed to this thesis and made my graduate journey a truly memorable and enriching experience.

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Michael P. Cummings, for his invaluable guidance, encouragement, and unwavering support throughout my PhD journey. His expertise, patience, and insightful advice have played a crucial role in shaping both my research and personal growth. I will always be grateful for the time and effort he invested in me, his open-door policy for help and guidance, and his continuous motivation to expand my knowledge and pursue intersecting projects. His support not only pushed me toward academic excellence but also emphasized the importance of a healthy work-life balance. I am truly grateful for the opportunity to learn from such an inspiring mentor.

I would also like to extend my sincere appreciation to Professor Rainer von Coelln, who served as the main collaborator and primary liaison with the Department of Neurology at the University of Maryland Medical School. His expertise in Parkinson's disease and movement disorders was invaluable to the main project of this thesis. He played a crucial role in designing the clinical study under which these data were collected, supervising data collection, personally contributing to data acquisition, and overseeing initial data processing. His clinical perspective and deep understanding of movement disorders provided essential context for this research. He has been deeply invested in my dissertation, actively participating in monthly project meetings,

offering insightful advice, and coauthoring manuscripts and papers.

Additionally, I would like to acknowledge the invaluable support and expertise of Professor Matthew Goupell from the Department of Hearing and Speech Sciences at the University of Maryland. He generously provided the data used for the second project of this thesis and played a key role in designing the study. His initiative in organizing project meetings, offering thoughtful advice, and coauthoring manuscripts has been greatly appreciated. His contributions have significantly enriched this work.

I would like to take this opportunity to acknowledge my colleagues in Cummings Lab, who have greatly enriched my graduate experience in various ways. Renee Ti Chou contributed to the second project of this thesis during my parental leave, providing me with the code and initial analysis for the project. Alexis S. Boleda offered invaluable assistance with different analyses required for my work, especially where our projects overlapped. My interactions with Yi Chen and Sondos Abdelhafeez have been very fruitful.

I owe my deepest thanks to my family for their unconditional love, patience, and encouragement. My parents, husband, and children have consistently stood by me, offering guidance and strength throughout my career, and helping me persevere through even the most challenging times. Words cannot fully express my gratitude for all they have done. Their unwavering belief in me has been my greatest source of strength, and their endless support and prayers have been invaluable. They have always been my pillars, and it is because of them that I have reached this point. Without their love and encouragement, none of this would have been possible.

Finally, I would like to acknowledge the financial support provided by the University of Maryland MPower Seed Grant, the Rosalyn Newman Foundation, the University of Maryland Claude D. Pepper Older Americans Independence Center, the University of Maryland Brain and

Behavior Institute Seed Grant, as well as the National Institute on Deafness and Other Communication Disorders and the National Institute on Aging of the National Institutes of Health. Their generous funding has been instrumental in supporting all the projects discussed in this thesis.

This dissertation is a reflection of the collective effort and support of all these incredible individuals, and I am truly thankful to each and every one of them. Thank you all and thank God!

Table of Contents

Acknowledgements ii

Table of Contents v

List of Tables ix

List of Figures xii

List of Abbreviations xvii

I Introduction and General Machine Learning Analysis 1

Chapter 1: Introduction 2

Chapter 2: Related Work 8
2.1 Machine Learning for Wearable Sensor Data in Parkinson's Disease 8
2.1.1 Diagnosis of Parkinson's Disease 8
2.1.2 Detecting Parkinson's Disease Motor Symptoms 12
2.1.3 Detecting Medication States in Parkinson's Disease Patients 19
2.2 Machine Learning for Human Hearing Data 21
2.2.1 Age-related Hearing Loss 21
2.2.2 Noise-induced Hearing Loss 25
2.2.3 Sudden Sensorineural Hearing Loss 27
2.2.4 Hearing-impaired Speech Perception 29
2.2.5 Clustering of Audiometric Profiles 32

II Movement Disorders 36

Chapter 3: Simplification of Mobility Tests and Data Processing to Increase Applicability of Wearable Sensors as Diagnostic Tool for Parkinson's Disease 37
3.1 Introduction 38
3.2 Materials and Methods 43
3.2.1 Participants 43
3.2.2 Clinical Evaluations 45
3.2.3 Assessment of Mobility 46

3.2.4 Device and Data Collection . . . 46
3.2.5 Data Extraction . . . 47
3.2.6 Segmentation . . . 48
3.2.7 Feature Engineering . . . 50
3.2.8 Feature Selection . . . 51
3.2.9 Machine Learning Model . . . 51
3.2.10 Group Feature Importance . . . 53
3.2.11 Statistical Analysis . . . 53
3.3 Results . . . 54
3.3.1 Classification of PD versus Control . . . 55
3.3.2 Strategies for Simplifying Mobility Testing and Its Associated Workload 58
3.4 Discussion . . . 62
3.5 Conclusions . . . 68
3.6 Funding . . . 68
3.7 Institutional Review . . . 68
3.8 Informed Consent . . . 69
3.9 Data Availability . . . 69
3.10 Acknowledgments . . . 69

Chapter 4: Machine Learning and Statistical Analyses of Sensor Data Reveal Variability Between Repeated Trials in Parkinson's Disease Mobility Assessments 70
4.1 Introduction . . . 71
4.2 Materials and Methods . . . 74
4.2.1 Participants . . . 74
4.2.2 Mobility Assessments . . . 75
4.2.3 Device and Data Acquisition . . . 76
4.2.4 Data Retrieval . . . 77
4.2.5 Segmentation . . . 77
4.2.6 Quantitative Measures . . . 78
4.2.7 Feature Selection . . . 79
4.2.8 Machine Learning Model . . . 79
4.2.9 Statistical Analysis . . . 80
4.3 Results . . . 81
4.3.1 Reliability of Task/Subtask Duration . . . 82
4.3.2 Reliability of Other Quantitative Measures . . . 86
4.3.3 Performance of Machine Learning Models . . . 89
4.4 Discussion . . . 91
4.5 Conclusions . . . 97
4.6 Funding . . . 98
4.7 Institutional Review . . . 98
4.8 Informed Consent . . . 98
4.9 Data Availability . . . 98
4.10 Acknowledgments . . . 99

Chapter 5: Applying Wearable Sensors and Machine Learning to the Diagnostic Challenge of Distinguishing Parkinson's Disease from Other Forms of Parkinsonism **100**
5.1 Introduction . . . 101
5.2 Materials and Methods . . . 106
5.2.1 Participants . . . 106
5.2.2 Experimental Procedure . . . 107
5.2.3 Feature Engineering . . . 108
5.2.4 Mutual Information-Based Feature Selection for EasyEnsemble (MIEE) . . . 109
5.2.5 Machine Learning Model . . . 110
5.2.6 Group Feature Importance . . . 111
5.2.7 Statistical Analysis . . . 111
5.3 Results . . . 112
5.3.1 Feature Importance Analysis . . . 113
5.3.2 Model Performance . . . 114
5.3.3 Misclassification Analysis . . . 116
5.3.4 Model Predictions for Challenging Cases . . . 118
5.3.5 The Performance of Alternative Models . . . 119
5.4 Discussion . . . 121
5.5 Conclusions . . . 126
5.6 Funding . . . 126
5.7 Institutional Review . . . 127
5.8 Informed Consent . . . 127
5.9 Data Availability . . . 127
5.10 Acknowledgments . . . 127

III Human Hearing **128**

Chapter 6: Using Machine Learning to Understand the Relationship Between Audiometric Data, Speech Perception, Temporal Processing, and Cognition **129**
6.1 Introduction . . . 130
6.2 Methods . . . 131
6.2.1 Participants . . . 131
6.2.2 Audiometric Tests . . . 132
6.2.3 Perception, Electrophysiological, and Cognitive Tests . . . 132
6.2.4 Machine Learning Analysis . . . 133
6.3 Results . . . 135
6.4 Discussion . . . 137
6.5 Acknowledgments . . . 139

IV Future Work **140**

Chapter 7: Future Work **141**

7.1 Movement Disorders . . . 141
7.1.1 Larger Sample Size . . . 141
7.1.2 Non-motor Features for PD Identification . . . 142
7.1.3 Longitudinal Analysis . . . 142
7.1.4 Impact of Confounding Factors on Classification Performance . . . 142
7.1.5 Unsupervised Analysis of Sensor-based Movement Data . . . 143
7.1.6 WearGait-PD Dataset . . . 144
7.1.7 Genetic Markers . . . 145
7.2 Human Hearing . . . 145
7.2.1 Inherent Biases in Audiological Thresholds and Standards . . . 145
7.2.2 Left-Right Ear Asymmetry . . . 147

V Appendix **149**

Appendix A: Supplementary Material for Simplification of Mobility Tests and Data Processing to Increase Applicability of Wearable Sensors as Diagnostic Tool for Parkinson's Disease **150**
A.1 Supplementary Methods . . . 163
A.1.1 SHAP Values and Individual Feature Importance . . . 163
A.2 Supplementary Results . . . 164
A.2.1 Classification of Moderate and Severe PD versus Control . . . 164
A.2.2 Feature-level Importance . . . 172
A.2.3 Single Feature-based Models: Total TUG/cogTUG Duration . . . 174

Appendix B: Supplementary Material for Machine Learning and Statistical Analyses of Sensor Data Reveal Variability Between Repeated Trials in Parkinson's Disease Mobility Assessments **177**

Appendix C: Supplementary Material for Applying Wearable Sensors and Machine Learning to the Diagnostic Challenge of Distinguishing Parkinson's Disease from Other Forms of Parkinsonism **189**
C.1 Supplementary Results . . . 190

Appendix D: Supplementary Material for Using Machine Learning to Understand the Relationship Between Audiometric Data, Speech Perception, Temporal Processing, and Cognition **192**
D.1 More Detailed Results . . . 192
D.2 R Code . . . 193

Bibliography **195**

List of Tables

3.1 Characteristics of the study cohort. . . . 55
3.2 Confusion matrices of the classifiers using controls and participants with Parkinson's disease (PD). Rows represent actual classes, and columns represent predictions. . . . 56
3.3 Classification results of models in distinguishing controls from all Parkinson's disease (PD) and mild-PD (H&Y $\leq$ 2) participants. CI: 95% confidence interval. 56
3.4 Clinical characteristics of correctly classified and misclassified controls and Parkinson's disease (PD) participants. . . . 59
3.5 Accuracy (%) of super-learner models using unsegmented tasks and a combination of segmented tasks and calculated kinesiological features to distinguish controls from PD participants. The numbers in parentheses represent the difference in accuracy between each model and the corresponding standard model, which used data from segmented tasks without incorporating kinesiological features (same row). . . . 60
3.6 Accuracy (%) of alternative models. TUG-only and cogTUG-only models used features derived from the first trial, the second trial, and the mean of corresponding features from both trials. TUG-duration and cogTUG-duration models used the durations of TUG and cogTUG tasks. The numbers in parentheses represent the difference in accuracy between the respective model and the corresponding all-task (standard) model (same row). . . . 61

4.1 Characteristics of the study cohort. . . . 82
4.2 Accuracy (percent) of the models. Trial 1 models used features derived only from the first trial; Trial 2 models used features derived only from the second trial; Trials 1&2 models used features derived from the first trial, the second trial, and the mean of corresponding features from both trials; and duration models used the mean durations from both trials. Numbers in parentheses represent the difference in accuracy between each model and the corresponding model that used only features from the first trial. . . . 91

5.1 Characteristics of the study cohort. UPDRS: Unified Parkinson's Disease Rating Scale. H&Y: modified Hoehn and Yahr. . . . 112
5.2 Classification results of all-tasks and TUG-only models using participants with Parkinson's disease (PD) and non-PD parkinsonism. CI: 95% confidence interval. 115
5.3 Confusion matrices of the classifiers using participants with Parkinson's disease (PD) and non-PD parkinsonism (Pksm). Rows represent actual classes and columns represent predictions. . . . 115

5.4 Clinical characteristics of participants with Parkinson's disease (PD) and non-PD parkinsonism correctly and incorrectly classified by the TUG-only model. 117
5.5 Classification results of alternative machine learning models using TUG-only features. RF: random forest; FS: feature selection; bacc: balanced accuracy; MI: mutual information; DT: decision tree; AIC: Akaike information criterion. 121
5.6 Classification results of sampling techniques using TUG-only features. FS: feature selection. 121

A.1 Number of sensors and mobility tasks used in a survey of literature from 2017 to 2024. 150
A.2 Frequency domain features. The power spectrum is obtained using the Discrete Fourier Transform (DFT) of the signal with a fast algorithm, the Fast Fourier Transform (FFT). Variables: f_j is the frequency of the spectrum at frequency bin j; P_j is the power spectrum at frequency bin j; and M is the length of the frequency bin. 152
A.3 Time domain features. Variables: x_n represents n^{th} sample of the signal; and N is the length of the signal. 153
A.4 Cross-time domain features. 157
A.5 Features selected by the feature selection approach for TUG-only and cogTUG-only models using features from the first trial, the second trial, and the mean of the corresponding features from both trials. 158
A.6 Confusion matrix of the classifier using controls and participants with moderate (H&Y 2.5, 3) and severe (H&Y 4) Parkinson's disease (PD). Rows represent actual class and columns represent predictions. 167
A.7 Classification results of models distinguishing moderate (H&Y = 2.5, 3) and severe (H&Y = 4) Parkinson's disease (PD) participants from controls. CI: 95% confidence interval. 168
A.8 Clinical characteristics of correctly classified and misclassified controls and moderate and severe Parkinson's disease (PD) participants. 170
A.9 Accuracy (%) of super learner models using the unsegmented tasks and combination of segmented tasks and calculated kinesiological features to distinguish controls from PD participants. The numbers in parentheses represent the difference in accuracy between each model and the corresponding standard model, which used data from segmented tasks without incorporating kinesiological features (same row). 171
A.10 Accuracy (%) of alternative models. TUG-only and cogTUG-only models used features derived from the first trial, the second trial, and the mean of corresponding features from both trials. TUG-duration and cogTUG-duration models used the duration of TUG and cogTUG tasks. The numbers in parentheses represent the difference in accuracy between the respective model and the corresponding all-tasks model (same row). 172
A.11 Comparing the duration of TUG sub-tasks between different groups of PD participants and controls (ctrl). 175
A.12 Comparing the duration of cogTUG sub-tasks between different groups of PD participants and controls (ctrl). 176

B.1 Sensor-derived quantitative parameters. All quantitative measures were computed for both the vertical and anteroposterior directions. 177
B.2 Intraclass correlation coefficient (ICC, 95% CI) of sensor-based total duration of each mobility task. CI: confidence interval. 180
B.3 Variance of sensor-based task duration (s^2). 180
B.4 Values of assessments measures. 181
B.5 Reliability of each of TUG, cogTUG, and 16-foot walk assessments. ICC intraclass correlation coefficient, CI confidence interval, SEM standard error of measurement agreement, MDC minimal detectable change. 182
B.6 Median of intraclass correlation coefficient (median, 95% confidence interval (CI)) across the quantitative measures. 183

C.1 The count of misclassified participants with Parkinson's disease (PD) that closely resemble each type of non-PD parkinsonism. UnspecPksm: unspecified parkinsonism. 189
C.2 Confusion matrices of the classifiers using Parkinson's disease (PD) and non-PD parkinsonism (Pksm) participants using features from TUG mobility task. Rows represent actual class and columns represent predictions. 190

D.1 Base models used to build the SuperLearner model and the parameters differing from the defaults. 192
D.2 Base models used to predict QuickSIN Speech-in-Noise, the average percent of variance explained by each model, and the average weighted contribution in the overall ensemble. 193
D.3 Base models used to predict sentence recognition with 60% time compression, the average percent of variance explained by each model, and the average weighted contribution in the overall ensemble. 194
D.4 Base models used to predict sentence recognition with 1.2-s reverberation time, the average percent of variance explained by each model, and the average weighted contribution in the overall ensemble. 194

List of Figures

3.1 Number of mobility tasks and sensors used in prior studies. The axes are in $\log_2$ scale. 41

3.2 Machine learning pipeline. The process started by collecting data from a wearable sensor on three linear and three angular axes. Filtration was then applied to reduce noise in signals using a Butterworth filter. The composite tasks were further processed by dividing them into segments corresponding to the different subtasks. Using the six signals recorded for each participant and task, a large set of time-domain and frequency-domain features was calculated. The data were then shuffled multiple times, and a cross-validation framework was applied in each iteration to generate train and test sets. Within each fold, forward feature selection was performed on the training set, and a super-learner model was built. This model used a wide array of base models and the final prediction was calculated by assigning a weight to the prediction of each base learner. Finally, the performance of the model was evaluated using accuracy, sensitivity, specificity, and F_1 score measures. Model interpretation was performed using SHapley Additive exPlanation (SHAP) analysis (see Supplementary Methods, Section A.1.1) to quantify the contributions of features to the final model prediction. AP: anteroposterior direction; V: vertical direction; ML: mediolateral direction. 48

3.3 Group feature importance of models distinguishing controls from (**a**) all Parkinson's disease (PD) and (**b**) mild-PD (H&Y $\leq$ 2) participants. Groups are ranked by their importance calculated by permuting features within each group simultaneously and reporting the mean decrease in accuracy between the original and permuted data. Under the null hypothesis that there is no association between the group of predictor variables and the model prediction, permutation should have no or little impact on predictive performance. More detailed individual feature-level importance scores and Shapley additive explanation (SHAP) values are illustrated in Supplementary Figures A.4 and A.6. 58

4.1 Intraclass correlation coefficient (ICC) (left) and median of the relative change (right) in total and subtask duration for TUG (**a**,**b**), cogTUG (**c**,**d**), and 16-foot walk (**e**,**f**). Vertical dashed lines represent thresholds for moderate (ICC = 0.5) and good (ICC = 0.75) reliability. 84

4.2 Left: Intraclass correlation coefficient (ICC) of total duration and other quantitative measures for TUG (**a**), cogTUG (**c**), and 16-foot walk (**e**). Gray points represent the ICC of the total duration and blue triangles represent the median of the ICC values across the other quantitative measures. Vertical dashed lines represent thresholds for moderate (ICC = 0.5) and good (ICC = 0.75) reliability. Median ICC values, along with their respective confidence intervals, across the quantitative measures for each task and participant group are presented in Supplementary Table B.6. Right: Relative change (RC) in total duration and other quantitative measures for TUG (**b**), cogTUG (**d**), and 16-foot walk (**f**). Gray points represent the median of the RC of the total duration across participants of each group. Blue triangles represent the median of the medians of the absolute RC values across participants and quantitative measures. 88

5.1 Study workflow. Pksm: non-PD parkinsonism; AUC-ROC: area under the receiver operating characteristic curve. 108
5.2 Group feature importance for the classification of Parkinson's disease and non-PD parkinsonism. The ranking of groups is determined based on their importance, which is computed by simultaneously permuting features within each group and measuring the average decrease in accuracy between the original and permuted data. Under the null hypothesis, which posits no association between the group of predictor variables and the model's prediction, permutation should result in no impact on predictive performance. 114

6.1 Variable importance for predicting QuickSIN Speech-in-Noise. 136
6.2 Variable importance for predicting sentence recognition with 60% time compression. 136
6.3 Variable importance for predicting sentence recognition with 1.2-s reverberation time. 137

A.1 **(a)** An example signal for the angular velocity around the roll axis (gx, black line) and the trapezoidal integration of the signal (red line) for the TUG task. Segments (1) and (2) represent the two turns. **(b)** The signal for the vertical acceleration (az, black line) and the trapezoidal integration of the signal (red line). Segments (1) and (2) represent the sit-to-stand and stand-to-sit parts. 159
A.2 Example results of our segmentation approach for the TUG task applied to all six channels. The components are (1) sit-to-stand, (2) first walk, (3) first turn, (4) second walk, (5) second turn, and (6) stand-to-sit. 159
A.3 Example results of our segmentation approach for the 32-foot walk task applied to all six channels. The components are (1) first walk, (2) first turn, (3) second walk, (4) second turn, (5) third walk, (6) third turn, and (7) fourth walk. 160

A.4 Feature importance scores (left) and SHAP values (right) of the features included in the all PD vs controls classification model. The first column conveys the task and signal channel. Features are ranked by their importance defined as the mean decrease in area under the curve (AUC) between the original and permuted models. SHAP values were calculated using the kernel SHAP method (see Supplementary Methods). Points in the SHAP plot represent participants from the test sets and SHAP values indicate the impact on model output. The color gradient represents the variable values normalized based on percentile ranks. DFT: Discrete Fourier Transform. 160
A.5 Correlation matrix of the features included in the all PD vs controls classification model. 161
A.6 Feature importance scores (left) and SHAP values (right) of the features included in the mild PD (H&Y values ≤ 2) vs controls classification model. The first column conveys the task and signal channel. Features are ranked by their importance defined as the mean decrease in area under the curve (AUC) between the original and permuted models. SHAP values were calculated using the kernel SHAP method (see Supplementary Methods). Points in the SHAP plot represent participants from the test sets and SHAP values indicate the impact on model output. The color gradient represents the variable values normalized based on percentile ranks. DFT: Discrete Fourier Transform. 161
A.7 Base model coefficients in final super learner classifier using all PD participants and controls. Models with zero coefficients are not shown. GBM: gradient boosting machines, DL: deep learning, DRF: distributed random forests, GLM: generalized linear model. 162
A.8 Base model coefficients in final super learner classifier using mild PD (H&Y $\leq$ 2) participants and controls. Models with zero coefficients are not shown. GBM: gradient boosting machines, DL: deep learning, DRF: distributed random forests, XGBoost: extreme gradient boosting. 162
A.9 Feature importance scores (left) and SHAP values (right) of the features included in the moderate PD (H&Y = 2.5, 3) vs controls classification model. The first column conveys the task and signal channel. Features are ranked by their importance defined as the mean decrease in AUC between the original and permuted models. SHAP values were calculated using the kernel SHAP method (see Supplementary Methods). Points in the SHAP plot represent participants from the test sets and SHAP values indicate the impact on model output. The color gradient represents the variable values normalized based on percentile ranks. DFT: Discrete Fourier Transform, LSP: Lomb-Scargle Periodogram, OARS: Older Americans Resource Survey-Instrumental Activities of Daily Living, ADL: activities of daily living, IADL: instrumental activities of daily living. 165

A.10 Feature importance scores (left) and SHAP values (right) of the features included in the severe PD (H&Y = 4) vs controls classification model. The first column conveys the task and signal channel. Features are ranked by their importance defined as the mean decrease in AUC between the original and permuted models. SHAP values were calculated using the kernel SHAP method (see Supplementary Methods). Points in the SHAP plot represent participants from the test sets and SHAP values indicate the impact on model output. The color gradient represents the variable values normalized based on percentile ranks. OARS: Older Americans Resource Survey-Instrumental Activities of Daily Living. 166
A.11 Base model coefficients in final super learner classifier using moderate PD (H&Y 2.5 and 3) and controls. Models with zero coefficients are not shown. DL: deep learning, DRF: distributed random forests, XGBoost: extreme gradient boosting. . 166
A.12 Base model coefficients in final super learner classifier using severe PD (H&Y 4) and controls. Models with zero coefficients are not shown. GBM: gradient boosting machines, DL: deep learning, DRF: distributed random forests, XGBoost: extreme gradient boosting. 167
A.13 Group feature importance of models distinguishing **(a)** moderate (H&Y = 2.5, 3) Parkinson's disease (PD) and **(b)** severe PD (H&Y = 4) participants from controls. Groups are ranked by their importance calculated by permuting features within each group simultaneously and reporting the mean decrease in accuracy between the original and permuted models. Under the null hypothesis that there is no association of the group of predictor variables and the model prediction, permutation should have no or little impact on predictive performance. More detailed individual feature-level importance scores and SHapley Additive exPlanations (SHAP) values are illustrated in Supplementary Figures A.9 and A.10. 169
A.14 Comparison of the total duration of **(a)** TUG and **(b)** cogTUG tasks between different groups of PD participants and controls. Different PD groups shows significantly longer durations when compared to the controls using two-tailed Student's t-test ($p < 0.0014$). 175

B.1 TUG intraclass correlation coefficient (ICC) **(a)** and median of the relative change **(b)** of each measure. Vertical dashed lines represent thresholds for moderate and good reliability. 184
B.2 cogTUG intraclass correlation coefficient (ICC) **(a)** and median of the relative change **(b)** of each measure. Vertical dashed lines represent thresholds for moderate and good reliability. 184
B.3 16-foot walk intraclass correlation coefficient (ICC) **(a)** and median of the relative change **(b)** of each measure. Vertical dashed lines represent thresholds for moderate and good reliability. 185
B.4 Bland-Altman plots of each measure for the TUG task. The dashed lines represent the mean of the differences and the 95% limits of agreement (LoA), defined as mean $\pm$ 1.96 SD. The blue and red shaded areas are the confidence intervals of the mean and LoA respectively. 186

B.5 Bland-Altman plots of each measure for the cogTUG task. The dashed lines represent the mean of the differences and the 95% limits of agreement (LoA), defined as mean ± 1.96 SD. The blue and red shaded areas are the confidence intervals of the mean and LoA respectively. 187
B.6 Bland-Altman plots of each measure for the 16-foot walk task. The dashed lines represent the mean of the differences and the 95% limits of agreement (LoA), defined as mean ± 1.96 SD. The blue and red shaded areas are the confidence intervals of the mean and LoA respectively. 188

List of Abbreviations

ADF	Assistive Devices and Falls
ADL	Activities of Daily Living
AIC	Akaike information criterion
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CBS	Corticobasal Syndrome
CI	Confidence Interval
CIRS-G	Cumulative Illness Rating Scale-Geriatrics
cogTUG	Cognitive Timed Up and Go
DBS	Deep Brain Stimulation
DFT	Discrete Fourier Transform
DIP	Drug-Induced Parkinsonism
DLB	Dementia with Lewy Bodies
DRF	Distributed Random Forests
EE	EasyEnsemble
EHF	Extended High Frequencies
ET	Essential Tremor
FFT	Fast Fourier Transform
GBM	Gradient Boosting Machines
GLM	Generalized Linear Model
GUI	Graphic User Interface
HL	Hearing Level
H&Y	Hoehn & Yahr
IADL	Instrumental Activities of Daily Living
ICC	Intraclass Correlation Coefficient
IMU	Inertial Measurement Unit
LoA	Limits of Agreement

LSP	Lomb–Scargle Periodogram
MDC	Minimal Detectable Change
MI	Mutual Information
ML	Machine learning
MoCA	Montreal Cognitive Assessment
MSA	Multiple System Atrophy
NCL	Neighborhood Cleaning Rule
OARS	Older Americans Resource and Services
PD	Parkinson's Disease
PSD	Power Spectral Density
PSP	Progressive Supranuclear Palsy
PTA	Pure Tone Average
RC	Relative Change
RF	Random Forest
SD	Standard Deviation
SEM	Standard Error of Measurement
SF	Standard Frequencies
SHAP	SHapley Additive exPlanation
SL	Super Learner
SMOTE	Synthetic Minority Oversampling Technique
TUG	Timed Up and Go
UPDRS	Unified Parkinson's Disease Rating Scale
XGBoost	Extreme Gradient Boosting

Part I

Introduction and General Machine Learning Analysis

Chapter 1: Introduction

The field of medical research is undergoing a paradigm shift with rapid advancements through the application of data science and machine learning (ML) [1–3]. The increasing availability of large datasets, improvements in computational power, and the development of sophisticated algorithms have paved the way for the application of these technologies in uncovering valuable insights, patterns, and predictive models from vast volumes of complex biomedical data [4–6]. This integration holds immense potential for transforming the landscape of medical research, leading to improved diagnostics, personalized treatments, and ultimately, enhancing patient care [7, 8].

Historically, medical research has relied on statistics and hypothesis-driven investigations, which typically involve controlled experiments and observational studies [9–11]. Although this conventional approach has yielded significant breakthroughs, it is often time-consuming, expensive, and limited in its ability to capture the intricate interactions of multiple variables that are inherent in complex biological systems. Furthermore, traditional statistical methods often face challenges when dealing with enormous datasets, intricate patterns, and non-linear relationships that are commonly encountered in the healthcare domain [12].

In recent years, data science and ML anlayses have emerged as promising alternatives to extract meaningful insights from diverse and extensive data sources within the medical research

field [13–15]. The accumulation of healthcare data has grown exponentially, propelled by electronic health records, genomics, wearable devices, medical imaging, and other sources. These vast repositories of information contain a wealth of knowledge that, when analyzed using advanced ML algorithms, can unlock invaluable discoveries, facilitate high-quality patient care, optimize operational efficiency, and improve overall healthcare outcomes.

Machine learning offers a novel approach to leverage the potential of complex and diverse healthcare data [16–19]. ML algorithms can handle structured and unstructured data, allowing researchers to extract knowledge from various sources, including patient records, clinical trials, medical literature, and real-time monitoring systems [20]. ML allows medical professionals to more deeply understand the intricate mechanisms underlying various health conditions, allowing informed decision-making and the identification of anomalies, patterns, and trends, while simultaneously reducing the potential for human error [21]. Moreover, ML techniques enable the development of predictive models that can aid in accurate prognoses, early detection of diseases, and personalized treatment recommendations [22–24]. The field of medical research holds significant potential for progress and innovation, driven by advancements in data science and ML.

This dissertation aims to comprehensively explore the impact of data science and ML in medical research, with a particular focus on specific applications, such as movement disorder diagnosis and age-related hearing loss. First, for movement disorder diagnosis, I focused on Parkinson's disease (PD), the most common movement-related neurodegenerative disorder, affecting an estimated 7–10 million people worldwide [25]. Several risk factors, including age, gender, and race/ethnicity, have been associated with an increased likelihood of developing PD [26]. PD primarily impairs movement, causing symptoms such as tremors, rigidity, bradykinesia (slow-

ness of movement), and postural instability. However, its impact extends beyond motor function, leading to non-motor complications such as cognitive decline, sleep disturbances, and autonomic dysfunction, all of which significantly reduce quality of life. Given the progressive nature of PD and the challenges in distinguishing it from controls and other movement disorders [27], there is a critical need for improved diagnostic approaches. For this study, I analyzed data from participants with PD, other parkinsonian disorders, and healthy controls as they performed various motor tasks while wearing a sensor on their lower back in a clinical setting hallway. By employing ML techniques, I identified strategies to streamline and simplify both mobility testing and subsequent data analysis. This approach reduced the time and effort required by clinicians and participants while maintaining high predictive performance in diagnosing (or supporting the diagnosis of) PD. Additionally, I examined the reliability of mobility tasks when repeated twice, revealing differences in reliability between PD and control groups and within PD subgroups stratified by disease severity. I also addressed potential challenges of applying ML to datasets resembling real clinical scenarios, particularly for distinguishing PD from a heterogeneous group of participants with various non-PD parkinsonian disorders, and demonstrated effective strategies to overcome these challenges.

Second, to study age-related hearing loss, I developed ensemble models to analyze data from participants with varying ages and hearing loss. I utilized a combination of audiometric, perceptual, electrophysiological, and cognitive data to predict speech perception in challenging conditions such as noise, reverberation, and time compression. Employing ML techniques, my objective was to identify the variables that are predictive of demanding speech-perception conditions, and thereby validate known associations from previous studies and potentially uncover new associations. My results indicate that speech perception is best predicted using variables related

to audiometric thresholds, specifically focusing on the range between 1 and 4 kHz, including the use of new composite variables across frequencies. Basic temporal processing ability also plays a moderate role in predicting speech perception. However, cognitive factors and extended high-frequency thresholds have minimal to no predictive ability in this regard. Further research and exploration of associations will contribute to our understanding of the complex interplay between speech perception, aging, hearing loss, and cognition.

To achieve the stated research objectives, this dissertation is structured as follows:

Chapter 2 Describes different fundamental concepts related to applying machine learning in studying Parkinson's disease and human hearing.

Chapter 3 This chapter discusses my research focused on simplifying instrumented mobility testing to enhance the integration of wearable sensors for PD diagnosis in clinical settings. This work was conducted in collaboration with Professors Rainer von Coelln and Lisa M. Shulman at the University of Maryland School of Medicine as primary collaborators, along with additional contributors. My role included performing data analyses, interpreting results, and drafting the manuscript, which was published in the journal *Sensors* as:

Rana M. Khalil, Lisa M. Shulman, Ann L. Gruber-Baldini, Sunita Shakya, Rebecca Fenderson, Maxwell Van Hoven, Jeffrey M. Hausdorff, Rainer von Coelln, and Michael P. Cummings. Simplification of mobility tests and data processing to increase applicability of wearable sensors as diagnostic tools for Parkinson's disease. *Sensors*, 24(15):4983, 2024 [28].

Chapter 4 This chapter describes my research on evaluating the test–retest reliability of mobility tasks by examining machine learning models based on wearable sensor-derived measures and

statistical metrics. Collaborating again with Professors Rainer von Coelln and Lisa M. Shulman, along with other contributors, I conducted the analyses, interpreted the findings, and drafted the manuscript, which was published in the journal *Sensors* as:

Rana M. Khalil, Lisa M. Shulman, Ann L. Gruber-Baldini, Sunita Shakya, Jeffrey M. Hausdorff, Rainer von Coelln, and Michael P. Cummings. Machine learning and statistical analyses of sensor data reveal variability between repeated trials in Parkinson's disease mobility assessments. *Sensors*, 24(24):8096, 2024 [29].

Chapter 5 This chapter presents my research on leveraging machine learning and wearable sensors to distinguish PD from other parkinsonian disorders, addressing diagnostic challenges, and streamlining workflows. In collaboration with Professors Rainer von Coelln, Lisa M. Shulman, and other contributors, I was responsible for data analysis, result interpretation, and manuscript preparation. This work is published in the journal *Biomedicines* as:

Rana M. Khalil, Lisa M. Shulman, Ann L. Gruber-Baldini, Stephen G. Reich, Joseph M. Savitt, Jeffrey M. Hausdorff, Rainer von Coelln, and Michael P. Cummings. Applying wearable sensors and machine learning to the diagnostic challenge of distinguishing Parkinson's disease from other forms of parkinsonism. *Biomedicines*, 13(3):572, 2025 [30].

Chapter 6 Describes my research in applying machine learning to study the relationship between audiometric data, speech perception, temporal processing, and cognition. This was a collaborative work with Alexandra Papanicolaou, Dr. Bobby E. Gibbs, II, Professors Samira Anderson, Sandra Gordon-Salant, and Matthew J. Goupell from the Department of Hearing and Speech Sciences, University of Maryland. I was responsible for all data pre-processing, analyses,

initial interpretations, and writing the initial manuscript draft. The paper has been published in the IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP) as:

Rana M. Khalil, Alexandra Papanicolaou, Renee Ti Chou, Bobby E. Gibbs, Samira Anderson, Sandra Gordon-Salant, Michael P. Cummings, and Matthew J. Goupell. Using machine learning to understand the relationships between audiometric data, speech perception, temporal processing, and cognition. In ICASSP 2023 — 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5, 2023 [31].

Chapter 7 This chapter outlines potential directions for future research.

Appendix A Presents the supplementary material for Chapter 3.

Appendix B Presents the supplementary material for Chapter 4.

Appendix C Presents the supplementary material for Chapter 5.

Appendix D Presents the supplementary material for Chapter 6.

Chapter 2: Related Work

2.1 Machine Learning for Wearable Sensor Data in Parkinson's Disease

Wearable devices, equipped with sensors capable of capturing a wide range of movement data, offer a unique opportunity to gather rich and continuous information about individuals in their daily lives. However, the sheer volume and complexity of this data present significant challenges for traditional analytical methods. This is where machine learning plays a crucial role. In this section, we review relevant machine learning (ML) techniques applied to diagnose Parkinson's disease (PD) participants or monitor PD-related symptoms.

2.1.1 Diagnosis of Parkinson's Disease

Several studies have been conducted to explore different aspects of PD assessment and diagnosis using advanced technologies and machine learning algorithms. The studies differ in wearable sensor type and position, incorporated mobility tasks, signal processing approach, calculated features, and ML analyses. These studies provide valuable insights into the development of digital biomarkers, quantification of motor impairment, remote tracking of disease progression, and gait analysis for PD.

One noteworthy study employed deep learning techniques to identify digital biomarkers for

self-reported PD [32]. Deep convolutional neural networks were utilized to analyze smartphone-based gait data from a simple walking task, achieving an impressive area under the receiver-operating characteristic curve (AUCROC) of 0.86. To overcome spatial and temporal biases in real-world motion records, the researchers proposed a data augmentation technique. The study highlighted the significance of resting tremor as an indicator of Parkinsonian movement, enabling efficient PD screening in the general population.

In another study, the goal was to quantify motor impairment in PD by utilizing an instrumented Timed Up & Go test [33]. Reliable measures were sought to differentiate between healthy individuals and those with PD. Through feature selection and classification techniques, the researchers compared individual measures and identified the unique motor pattern of PD subjects. Chosen classifiers included the linear and quadratic discriminant analysis (LDA and QDA, respectively) and the Mahalanobis classifier (MC). The instrumented Timed Up & Go test was found to effectively characterize motor impairment in PD, offering a valuable tool for evaluation and potential monitoring of disease progression.

Arora et al. [34] conducted a pilot study exploring the potential of smartphones in detecting and monitoring PD symptoms. Participants used smartphones with specialized applications to assess various motor symptoms, including voice, posture, gait, finger tapping, and response time. By leveraging smartphone built-in sensors, the researchers aimed to capture and analyze these motor symptoms associated with PD. They achieved high accuracy rates, exceeding 96% in classifying PD participants from control subjects, using random forest (RF) classifiers.

Specific aspects of PD assessment and monitoring have also been investigated. In one study, RF feature classification was employed to quantify correlations between motor symptoms across different motor tasks and clinical conditions in PD [35]. Using data collected from a motion

capture suit, specific metrics characterizing each motor task were identified. The Embedded Gait Analysis using Intelligent Technology (eGaIT) was utilized in a separate study to objectively and automatically classify specific stages and motor signs in PD [36]. By combining sensor-derived signals with classifiers such as Linear LDA, AdaBoost, and Support Vector Machines (SVM), the eGaIT system achieved an overall classification rate of 81%, successfully distinguishing PD patients from controls. The system demonstrated increased accuracy with higher levels of motor impairment and advanced stages of PD.

Oung et al. [37] explored the use of wearable multimodal sensors, including accelerometers, gyroscopes, and magnetometers, to classify individuals with PD. The proposed SVM and Probabilistic Neural Network (PNN) classifiers exhibited high average accuracy, sensitivity, specificity, and receiver operating characteristic curve (ROC) values above 88%. The study investigated various feature extraction methods, comparing time domain and frequency domain analyses, and reported promising results with the proposed classifier.

Tien et al. [38] aimed to determine the most sensitive gait and mobility measures reflecting PD motor stages and identify the optimal sensor location for each stage. They found that upper-limb sensor measures were effective in discriminating between controls and early PD, trunk sensor measures were prominent in mid-stage PD, and stride timing and regularity were discriminative in more advanced stages. Sensor location and corresponding gait features significantly influenced the discriminatory value, as revealed by the random under-sampling boosting (RUSBoost) classification algorithm. Measures obtained from challenging walks were effective in detecting mobility impairments in the early stages but were less sensitive in the later stages of the disease.

Additionally, smartphone-based gait assessment and crowdsourced data were investigated

for inferring PD severity in a comparative study [39]. Various classification algorithms, including bagged trees, SVM, K-nearest neighbors (KNN), and RF were compared. The RF classifier showed the highest accuracy in distinguishing subjects from controls and assessing PD severity.

A comprehensive analysis assessing the postural behavior of early-mild PD subjects using accelerometers was proposed by Palmerini et al. [40]. The authors computed several measures and employed feature selection techniques and different classifiers to characterize early-mild PD and control subjects. The results demonstrated remarkably low misclassification rates, as low as 5%, using subsets of three features. Various commonly used classifiers, such as LDA, QDA, MC, logistic regression (LR), KNN, and SVM were integrated into the feature selection procedure, which was based on a wrapper approach and aimed to optimize predictive accuracy.

Several studies focused on the development and comparison of machine learning algorithms for PD diagnosis and assessment. LR, PNN, and SVM classifiers were compared for discriminating between normal and PD subjects [41]. This study focused on the effects of deep brain stimulation of the subthalamic nucleus (DBS-STN) on ground reaction force (GRF). The PNN classifier outperformed LR and SVM based on higher AUCROC and lower negative likelihood ratio values (NLR). Evaluating treatment effects, the study revealed that DBS-STN alone was more effective than medication alone, and the greatest improvements occurred with both treatments combined. Similarly, various classification methods including Neural Network (NN), Regression, and Decision tree (DT) were used for diagnosing PD [42]. The NN classifier achieved the highest classification accuracy with a score of 92.9% correct classification rate.

A comprehensive review of the literature identifying previous studies that utilized wearable sensors for PD assessment is provided in Section 3.1.

2.1.2 Detecting Parkinson's Disease Motor Symptoms

Motor symptoms such as bradykinesia, dyskinesia, freezing of gait (FoG), and tremors are key indicators of PD. Bradykinesia is characterized by a reduction in automatic movements such as blinking or swinging of arms while walking, or by trouble initiating intentional movements or just slowness of actions. Another movement problem, dyskinesia, involves involuntary, erratic, and sometimes jerky movements. FoG in PD is often likened to the sensation of having one's feet adhered to the ground, causing temporary immobility while the upper body remains mobile. Tremors are typically observed during periods of rest, are slow, and commonly originate in one hand, foot, or leg and eventually extend to involve both sides of the body. Researchers have dedicated their efforts to estimating the severity of PD symptoms and developing methods for continuous monitoring. For instance, one study focused on using wearable sensors to assess the severity of symptoms and motor fluctuations in PD patients [43]. A SVM classifier was implemented to estimate tremor, bradykinesia, and dyskinesia severity from accelerometer data features. The study analyzed the effect of various parameters on the estimation of clinical scores, including window duration for segmenting accelerometer data, SVM kernels, and data features derived from different motor tasks. The study achieved average estimation errors of 3.4% for tremor, 2.2% for bradykinesia, and 3.2% for dyskinesia. Another study investigated the data collected from wearable sensors that are most valuable for training symptom detection models in PD [44]. This study trained machine learning classifiers, including statistical ensembles (RF) and convolutional neural networks (CNN), to detect the presence of tremor or bradykinesia. The effects of the number and location of sensors, types of tasks, number of data collection sessions, and the number of individuals involved were evaluated. The results showed that a single wearable

sensor on the back of the hand was sufficient for detecting bradykinesia and tremor in the upper extremities. Both RF and CNN-based models were effective in detecting PD symptoms, with CNN models showing comparable performance to RF classifiers. Cole et al. [45] developed and evaluated dynamical machine-learning algorithms that track the presence and severity of tremors and dyskinesia in PD. These algorithms analyze signals from hybrid sensors combining 3-D accelerometers and surface-electromyographic modalities. The comparative analysis showed that dynamic neural networks (DNNs), dynamic support vector machines (DSVMs), and hidden Markov models (HMMs) performed equally well, maintaining error rates below 10% for dynamic tracking. Dynamic rather than static machine learning algorithms were used to leverage the time-varying nature of tremor and dyskinesia, enabling accurate detection of these disorders in the presence of unscripted and unconstrained voluntary movements.

Furthermore, several studies have specifically targeted individual PD symptoms. These studies have delved into various aspects of symptom assessment and monitoring. The sections below summarize some of these studies.

2.1.2.1 Bradykinesia

In a previous study, the researchers explored the application of deep learning in monitoring motor symptoms, with a specific focus on detecting bradykinesia [46]. They compared deep learning based on CNNs with standard machine learning pipelines. These pipelines typically involve the use of expert-defined features and algorithms such as boosting, DT, KNN, and SVM. The findings revealed that deep learning surpassed other state-of-the-art machine learning algorithms, achieving a classification rate improvement of at least 4.6%, indicating their potential for

PD assessment. Another relevant work introduced a novel machine-learning method for the assessment and quantification of bradykinesia using a waist-worn triaxial accelerometer [47]. The method combined an SVM classifier for gait detection, a specific signal processing technique to identify strides, and a characterization of these strides based on their frequency content. The dichotomic detection of bradykinesia based on frequency content achieves an accuracy of over 90%. Moreover, bradykinesia severity, as measured by the Unified Parkinson's Disease Rating Scale (UPDRS) scores, is approximated using a regression model with errors below 10%. Similarly, a methodology utilizing wireless, wearable accelerometers to automatically detect the severity of bradykinesia in patients with motor diseases was introduced in an another study [48]. By analyzing unstructured daily life activities, the novel method achieved symptom severity accuracies ranging from 70% to 86% based on UPDRS scores. This work presented a bradykinesia profiler that eliminated segment overlap and calculated the confidence of the resulting events. For segment classification, six classifiers were employed: KNN, Parzen classifier, Parzen density-based classifier, binary decision tree (BDT), backpropagation neural network (BPNN) classifier, and support vector classifier (SVC).

2.1.2.2 Dyskinesia

An objective and automatic procedure to evaluate the severity of levodopa-induced dyskinesia (LID) in PD patients was developed in a previous work [49]. The patients engaged in approximately 35 functional tasks while wearing triaxial accelerometers placed at six different body positions. A neural network was trained to assess LID severity in real-life situations using various accelerometer signal variables. The neural network achieved the accurate classification

of dyskinesia presence or absence with success rates of 93.7% for the arm, 99.7% for the trunk, and 97% for the leg. Analysis of the neural networks revealed novel variables that are relevant for assessing LID severity. Belgiovine et al. [50] presented a LID detection system that utilizes machine learning algorithms along with smartwatch data and a smartphone to detect LID in both upper-limbs and lower-limbs of PD patients. The study implemented SVM and DT in the LID detection system and found that on average, the SVM algorithm outperformed the DT algorithm and that the LID detection at the lower limbs exhibited better performance compared to LID detection at the upper limbs.

2.1.2.3 Freezing of Gait (FoG)

The aim of one study was to FoG events in patients with PD using signals from wearable sensors [51]. The methodology consists of four stages: replacing missing values, removing low-frequency components, calculating signal entropy, and testing four classification algorithms (Naïve Bayes, RFs, DTs, and Random Tree (RT)) for FoG detection. The evaluation involved various sensor configurations to determine the optimal set for detecting FoG episodes. The results showed that the proposed methodology achieved 96.1% accuracy in detecting FoG events using signals from all sensors and the Random Forests classification algorithm. What sets this methodology apart from others is its independence from experimental conditions and participants, reliance on motor-related signals from affected body parts, and suitability for long-term monitoring to assess disease progression. Another wearable accelerometer-based methodology was proposed in a different study [52]. This study aimed to achieve three objectives: quantifying gait parameters objectively using data from wearable accelerometers, comparing the estimated

gait parameters from the proposed algorithm with those obtained from the 3D motion capture system in terms of mean error rate and Pearson's correlation coefficient (PCC), and employing ML techniques for automatic discrimination between FoG and non-FoG patients. The proposed methodology, utilizing only accelerometer data, achieved good accuracy of approximately 88% for estimating spatiotemporal gait parameters using an SVM classifier, making it a cost-effective alternative for assessing FoG patients. Additionally, it demonstrated a high correlation between accelerometer-based and 3D motion capture system-based gait parameters for concurrent validity, with an error rate of less than 10%. Furthermore, it presented the first classification study using gait parameters for classification. Naghavi et al. [53] investigated the impact of data imbalance on the performance of classifiers in predicting the FoG. They applied three approaches: using ensemble classifiers, adding synthetic FoG samples to the training dataset for improved balance, and increasing the misclassification cost for the minority class (FoG). They utilized time series signals as inputs and extracted features from 2-second windows, with six successive windows feeding into the classifier. The results showed that the ensemble classifier, formed by SVMs, KNN, and Multi-Layer Perceptron using bagging techniques, achieved the highest performance (F_1 = 90.7) when synthetic FoG samples were added to the training set and the class cost was set twice as high as normal gait. The model successfully identified 97.4% of the onset FoG events, with 66.7% being accurately predicted before the actual onset of a potential FoG episode. Despite the limited frequency of events, this study demonstrated the potential of the algorithm to accurately predict gait events and provide preventive cueing. Another improved system was presented utilizing sensor signals and employing a patient-dependent model for FOG detection in PD patients [54]. The Long Short Term Memory (LSTM) network-based patient-dependent model was adopted for FoG detection, and a comparison was made with traditional machine learning

methods such as SVM. The results demonstrated the superiority of the LSTM model, achieving an average accuracy of 83.4% compared to 79.5% for SVM. In this study, the hybrid discrete wavelet transform – fast Fourier transform (DWT-FFT) features were used for the SVM classifier, while LSTM was directly applied to the network without feature extraction, reducing computational time. The researchers of an additional work proposed a wearable assistant, consisting of a smartphone and wearable accelerometers, designed for real-time detection of FoG [55]. The system utilizes machine learning techniques to automatically detect FoG episodes and provides rhythmic auditory cueing or vibrotactile feedback to prompt the patient to resume walking. In user-dependent settings, the final system achieved an average sensitivity and specificity of over 95% and a mean detection latency of 0.34 seconds. Their work aims to improve FoG detection performance using machine learning techniques and deploy the system on a smartphone as an unobtrusive and cost-effective wearable assistant. To achieve this, they evaluated various machine learning algorithms using a real FoG dataset and assessed their accuracy and detection latency. They conducted experiments to optimize sensor placement and sensory-data window size, considering wearability and computational costs. Based on these evaluations, they developed a system for online FoG detection using a smartphone and wearable sensors, providing auditory feedback upon FoG detection.

2.1.2.4 Tremor

Algorithms using wearable sensors to estimate total Parkinsonian tremor during various free body movements were developed in one study [56]. The researchers employed two methods: an ensemble model based on gradient tree boosting and a deep learning model based on LSTM

networks. The analysis revealed that the gradient tree boosting method exhibited a high correlation between the estimated and clinically assessed tremor subscores, whereas the LSTM-based method demonstrated a moderate correlation. Their proposed ML algorithm comprised feature extraction and regression models. The feature extraction process involved extracting a number of features from every 5-second segment of data, while the regression model estimated a tremor subscore for each 10-minute duration of the data. Additionally, they explored the potential of the gradient tree boosting model in estimating resting and action tremor. The model demonstrated a high correlation with the UPDRS-III assessment for resting tremor and a moderate correlation for action tremor. These results support the reliability of their developed tremor estimation method, combined with wearable sensors, in providing accurate assessments of tremor severity experienced by PD subjects throughout a typical day. Another study measured the tremor signals of patients with PD using a wrist-watch-type wearable device equipped with an accelerometer and a gyroscope [57]. Various machine-learning algorithms, including DT, SVM, discriminant analysis (DA), RF, and KNN, were explored to automatically score the severity of Parkinsonian tremor using selected features obtained through a pairwise correlation strategy. Compared to existing methods, the approach achieved high accuracy of 85.5% and the smallest margin of error in estimating the UPDRS, which is already used in clinical practice. An automated method for assessing both resting and action/postural tremor using accelerometers placed on various body segments of patients was suggested in a third study [58]. Features indicative of low-frequency movements were extracted to differentiate tremor from other PD symptoms. Selected features were incorporated into hidden Markov models (HMMs) for tremor severity and body action/posture recognition. HMMs were employed to incorporate the time-dependent nature of the symptom and improve classification accuracy. The results demonstrated the effectiveness of the proposed

method in quantifying tremor severity with 87% accuracy, discriminating resting from postural tremor, and distinguishing tremor from other Parkinsonian motor symptoms such as bradykinesia and Levodopa-induced dyskinesia during daily activities.

2.1.3 Detecting Medication States in Parkinson's Disease Patients

In PD, patients often experience fluctuations in their motor symptoms depending on whether they are in an "on" or "off" medication state. Levodopa, often combined with carbidopa, is a commonly prescribed medication for PD that helps replenish dopamine levels in the brain and alleviate motor symptoms during the "on" state. The "on" state refers to a period when the medication is effectively controlling the symptoms, resulting in improved motor function. During this time, individuals may experience reduced rigidity, tremors, and improved mobility. On the other hand, the "off" state occurs when the medication's effects wear off, leading to a return of symptoms and a decrease in motor control. This fluctuation can significantly impact the daily lives of Parkinson's patients and their ability to perform daily activities. A framework for detecting "on" and "off" states in PD patients by leveraging wearable accelerometer data obtained from patient diaries was developed by Aich et al. [59]. The algorithm utilized a combination of statistical features and spatiotemporal gait features as inputs to four different classifiers: RF, SVM, KNN, and Naïve Bayes. Among the classifiers, random forest demonstrated the highest accuracy, reaching 96.7%. Some key findings include the successful use of accelerometer signals in combination with machine learning techniques for detecting medication states. Additionally, the knee-based placement of wearable devices showed better performance compared to wrist- and ankle-based wearable sensors. A patient-specific algorithm for classifying medication states in PD patients

was presented in a separate analysis [60]. The approach involved extracting temporal and spectral features from ambulatory signals and applying a semi-supervised classification algorithm utilizing K-means and self-organizing tree map clustering methods. Two types of cluster labeling: hard and fuzzy labeling were introduced. The average accuracy obtained using K-means and fuzzy labeling on the trunk and more sensors readings from the affected leg was 76%. A correlation between the accuracy for individual patients and the severity of dyskinesia as well as the improvement of PD symptoms with medication was observed. In an additional study [61], the authors developed a data fusion approach to automatically identify different medication states of patients with PD by combining multi-dimensional data from body-worn inertial sensors. The proposed approach utilized signal processing algorithms such as time-frequency analysis and tensor decomposition to extract features representing spectral, temporal, and spatial behavior of body motion data. An SVM classifier was trained and tested achieving an overall average accuracy of 78%. The classifier kernel (linear vs. radial basis) and the cost parameter were tuned using grid search and 4-fold cross-validation on the training data set.

2.2 Machine Learning for Human Hearing Data

Machine learning techniques have been increasingly employed in the field of audiology to predict and analyze various aspects of hearing loss and related conditions. This section provides an overview of recent research efforts focusing on the application of machine learning algorithms in predicting age-related hearing loss, noise-induced hearing loss, sudden sensorineural hearing loss, predicting speech perception, and clustering of audiometric profiles.

2.2.1 Age-related Hearing Loss

Presbycusis, also known as age-related hearing loss (ARHL), is a gradual decline in hearing ability that occurs as a natural part of the aging process. Age-related hearing loss can have various causes, with the most common being changes that take place in the inner ear during aging. However, ARHL can also result from changes in the middle ear, complex alterations along the nerve pathways connecting the ear to the brain, the cumulative impact of prolonged exposure to loud noise, viral or bacterial infections, heart conditions or stroke, head injuries, tumors, or certain medications. Several studies have explored the use of machine learning for predicting age-related hearing loss. These studies utilize audiometric data and demographic information to develop predictive models. For instance, the researchers have employed different ML models to predict the presence and severity of ARHL based on audiometric thresholds, speech perception scores, and age-related factors. Hoppe et al. [62] employed two different model approaches, including a generalized linear model (GLM) and a random forest regression model (RFR), to investigate the relationship between pure-tone hearing loss, age, and speech perception. The models were applied to a large clinical dataset comprising 19,801 ears with varying degrees of hearing

loss. The aim was to estimate the age-related decline in speech recognition for different audiogram types. Results revealed that speech scores were influenced by the specific type of hearing loss and the individual's age group. Age effects were observed across all degrees of hearing loss, with speech recognition deteriorating by up to 25% throughout the lifespan, even when pure-tone thresholds remained constant. The most significant decline was observed in a specific type of audiogram during the fifth decade of life, with a reduction of 10% per decade. Interestingly, this age-related decline in speech recognition could not be solely attributed to elevated hearing thresholds measured by pure-tone audiometry. Additionally, using pure-tone thresholds and age as inputs and speech recognition as the target variable, the performance of the RFR model was compared to the GLM approach. Unlike the RFR, the GLM required predefined assumptions about the qualitative relationship between input and target variables, whereas the RFR did not rely on a predefined equation framework. In comparison to previous studies, the RFR model provided more detailed information regarding the time course and extent of degradation in speech recognition. Both the GLM and RFR models confirmed an age-related decline in speech recognition after accounting for pure-tone sensorineural hearing loss (PTSL). This finding indicates that the measurable age-related decline in speech recognition is influenced by factors such as age range, specific audiogram characteristics, and the specific application of speech audiometry. Other studies showed that other factors, including central processing and cognitive decline, also contribute to the deterioration of speech perception [63–65].

Ilyas et al. [66] presented a computer-aided method that utilizes auditory perception to predict and prevent hearing loss. The method involved three predictive models: human age estimation, prediction of hearing loss, and prediction of the level of hearing loss. The robustness of the predictive models was demonstrated, with the human age estimation model achieving a Mean

Absolute Error (MAE) value of 4.1 years, the hearing loss prediction model attaining an accuracy of 94%, and the level of hearing loss prediction model achieving an accuracy of 90%. The proposed system required real-time interaction, where the subject responds to auditory stimuli. Two tests were conducted: the first test involves the subject indicating when they stop hearing a sound generated from lower to higher frequencies, while the second test involves the subject indicating when they start hearing a sound generated from higher to lower frequencies. The corresponding audible frequencies and their mean were saved and provided as input to the different predictive models. The first predictive model focused on estimating the age of an individual based on their auditory responses compared the performance of four regression models including Regression Forests (RF), Support Vector Regression (SVR), Adaboost, and Artificial Neural Networks (ANNs). The best accuracy was achieved using RF. The second predictive model aimed to predict whether a person has hearing loss or not based on their auditory responses. The audible frequencies and the real age were provided as input to this model. Five classifiers were tested: RF, SVM, LDS, Adaboost, and ANNs. The RF classifier achieved the highest accuracy in this prediction task. The third predictive model focused on predicting the severity or level of hearing loss. Leave-One-Out-Cross-Validation (LOOCV) is performed using RF and SVM classifiers, considering six levels of hearing loss. The RF classifier achieved again the highest accuracy in predicting the level of hearing loss.

Another study aimed to examine the relationship between age and different audiometric phenotypes in age-related hearing loss [67]. Longitudinal audiometric data collected from 343 adults aged 50-93 years were analyzed to determine if metabolic phenotypes increase with age compared to sensory phenotypes. A QDA model was utilized to classify audiograms into four phenotypes: Older-Normal, Metabolic, Sensory, and Metabolic + Sensory. The results showed

that although hearing loss increased with age, the audiometric phenotypes remained stable for the majority of ears (61.5%) over an average of 5.5 years. For ears with changing phenotypes, most transitioned to a Metabolic or Metabolic + Sensory phenotype. Additionally, the study aimed to characterize the longitudinal declines in hearing sensitivity for each audiometric phenotype. Longitudinal analyses revealed that each phenotype exhibited a unique pattern of threshold changes. Generalized linear mixed models (GLMM) were used to estimate threshold changes per year for each phenotype at different frequencies. The results demonstrated that the Metabolic and Metabolic + Sensory phenotypes were more likely with increasing age. The average age of individuals classified within these phenotypes was older compared to other phenotypes. These findings support the notion that gradual age-related metabolic declines contribute significantly to age-related hearing loss.

Further analysis was performed to assess the accuracy of a machine learning algorithm in predicting a person's audiometric configuration using limited demographic information, hearing loss data, and self-reported difficulty hearing [68]. The National Health and Nutrition Examination Survey (NHANES) database (`https://www.cdc.gov/nchs/nhanes/index.htm`), consisting of 9,256 cases, was utilized to train and test three machine learning algorithms, namely RF, SVM, and KNN. The Wisconsin Age-Related Hearing Impairment Classification Scale (WARHICS) [69] was employed to classify audiometric data into eight categories. Among the tested algorithms, the RF machine learning algorithm demonstrated the best fit, utilizing a few variables such as the slope between 2 and 4 kHz, gender, age, military experience, and self-reported hearing ability. With this method, 54.8% of individuals were correctly classified, while 34.4% were predicted to have a milder hearing loss than measured, and 10.8% were predicted to have a more severe loss than measured. The objective of this study was to assess the accuracy

of predicting audiometric configurations using limited information that one might typically have when utilizing uncalibrated equipment in a person's home.

2.2.2 Noise-induced Hearing Loss

Machine learning has shown promise in predicting noise-induced hearing loss (NIHL), a common occupational and recreational health concern that is usually caused by exposure to excessively loud sounds and cannot be medically or surgically corrected. A pilot study aimed to demonstrate the feasibility of developing machine learning models to predict hearing impairment in individuals exposed to complex industrial noise [70]. Audiometric and noise exposure data were collected from a population of screened workers ($n = 1113$) in 17 factories in China. Four machine learning algorithms (SVM, neural network multilayer perceptron, RF, and adaptive boosting) were used to develop both classification and regression models for predicting NIHL. Age, exposure duration, equivalent A-weighted sound pressure level, and median kurtosis were used as the input for the four algorithms. The models achieved a prediction accuracy between 78.6% and 80.1%, indicating their usefulness in assessing hearing impairment from complex occupational noise exposures. The SVM model performed the best among the classification models, while the multilayer perceptron regression model had the best performance for the quantitative prediction of noise-induced hearing loss. Farhadian et al. [71] analyzed the potential of artificial neural networks (ANNs) and logistic regression (LR) techniques for estimating hearing impairment in industrial workers. A total of 210 workers from a steel factory in Iran were selected, and their occupational exposure histories and hearing loss thresholds were analyzed. Personal noise exposures were also measured. Five variables influencing hearing loss were used as input

features, and hearing loss thresholds were the target feature for prediction. Multilayer feedforward neural networks and LR models were developed. Among the studied workers, 74.2% had normal-hearing thresholds, 23.4% had slight hearing loss, and 2.4% had moderate hearing loss. The best neural networks achieved an accuracy of 88.6% and a kappa coefficient of 66.3 for predicting the grades of hearing loss. Logistic regression achieved an accuracy of 84.3% and a kappa coefficient of 51.3. ANNs provided more accurate predictions than LR. This prediction method can provide reliable and comprehensible information for occupational health experts. The study confirmed that ANNs are a suitable tool for analyzing complex phenomena like hearing loss, especially when multiple variables interact in ways that are not fully understood. Similarly, Aliabadi et al. [72] used ANNs to predict the hearing loss threshold among noise-exposed workers. Workers from a steel factory were selected ($n = 210$), and their occupational exposure histories and audiometric test results were collected. Five variables that can influence hearing loss were used to develop the prediction model. Multilayer feed-forward neural networks with different structures were created. The best neural network model, consisting of one hidden layer and ten neurons, accurately predicted the hearing loss threshold with an RMSE of 2.6 dB and an R^2 of 0.89. The results confirmed that ANNs provided more accurate predictions compared to multiple regressions. The effectiveness of applying Gradient Descent with Adaptive Momentum (GDAM) to predict noise-induced hearing loss was investigated in a separate study [73]. The researchers showed that GDAM exhibited good prediction results for both left hearing loss (LHL) and right hearing loss (RHL). The Mean Square Error (MSE) for LHL was calculated to be 2.18×10^{-3}, and for RHL, it was 2.30×10^{-3}. The overall accuracy achieved for predicting NIHL in human workers was approximately 99.4% for the left ear and 99% for the right ear. Although a high accuracy of the prediction models was reported, incomplete information on the model construction

created the risk of bias. A longitudinal, retrospective cohort analysis was presented in a study to analyze historical hearing sensitivity data and identify factors associated with a significant change in hearing sensitivity among U.S. Air Force aviation-related personnel [74]. The study included audiogram records for personnel on active duty in 2013 who did not have non-noise-related hearing loss. Potential predictor variables such as age, elapsed time in the cohort, gender, and Air Force Specialty Code were examined. Random forest analyses were conducted to identify relevant predictors, and mixed effects models were fitted for statistical inferences. The final dataset included 167,253 nonbaseline audiograms from 10,567 participants. The results showed that the observed change in hearing threshold was small during the first 20 years of an individual's career and before the age of 50. Age and elapsed time since the baseline audiogram were modestly associated with decreased hearing sensitivity. However, the association between age, elapsed time, and likelihood of a significant threshold shift (STS) was not statistically significant. Aircraft type, as determined by the Air Force Specialty Code, was not associated with any of the hearing-related outcomes studied, despite expectations of differences based on aircraft type. Furthermore, the factors assessed in this study explained less than one-quarter of the variability in hearing sensitivity, suggesting that other factors not considered in this study or the definition of the STS metric may account for the majority of changes in hearing sensitivity.

2.2.3 Sudden Sensorineural Hearing Loss

Sudden sensorineural hearing loss (SSNHL), also known as sudden deafness, is a rapid and unexplained loss of hearing that occurs suddenly, either all at once or over a few days. It is typically caused by a malfunction in the sensory organs of the inner ear. It often affects only one ear.

Prediction models based on machine learning techniques have been explored for SSNHL. Bing et al. [75] developed predictive models for SSNHL using machine learning methods and determine the best performer for clinical application. A total of 1220 SSNHL patients were included in the study and four machine learning techniques, including deep belief network (DBN), LR, SVM, and multilayer perceptron (MLP), were employed. The original feature set consisted of 149 variables extracted from demographics, medical records, medications, pure tone audiometry, and laboratory test results. The authors used expert knowledge and univariate statistics to derive six subsets of features with subset-containment relations. The DBN model exhibited the best predictive ability when tested with the raw data set of 149 variables, achieving an accuracy of 77.6% and an AUC of 0.84. However, the performance of DBN decreased after feature pruning. On the other hand, LR, SVM, and MLP models demonstrated the opposite trend, with the best individual prediction powers achieved using only three variables. The ROC-AUC ranged from 0.79 to 0.81 and decreased as the size of input feature combinations increased. The authors claimed that while DBN can be a robust prediction tool for SSNHL when provided with enough features, LR is more practical for early prediction in routine clinical applications using three readily available variables. The researchers of another study proposed an ML-based model for predicting cochlear dead regions (DRs) in patients with various etiologies of hearing loss [76]. The analysis included 555 ears from 380 patients diagnosed with sensorineural hearing loss (SNHL). The presence of DRs was detected using a threshold-equalizing noise (TEN) test. Data on sex, age, affected ear, hearing loss etiology, word recognition scores (WRS), and pure-tone thresholds at each frequency were collected. The patients were categorized into six groups based on the cause of hearing loss. To develop the predictive model, two steps were employed. In the first step, a classification tree (CT) was constructed using recursive partitioning and regression. In the second

step, LR and RF were used to build binary classification models. The prediction models achieved positive predictive values (PPVs) ranging from 21.6% to 62.1%. While the LR model had a lower PPV compared to the CT or RF model, it exhibited a sensitivity of 59.7% as a function of frequency. This suggests that the LR model could serve as a screening tool with high sensitivity for detecting most DRs. The WRS, etiology types, and hearing thresholds at each frequency were found to be informative factors in all three models. Possible predictive factors for determining the presence of DRs were suggested based on the results. The effectiveness of five machine learning models (adaptive boosting, KNN, MLP, RF, and SVM) for predicting hearing prognosis in patients with SSNHL after 1 month of treatment was assessed in a comparative study [77]. Data from 227 patients (recovery: 106, no recovery: 121) were analyzed. Statistical hypothesis tests were conducted to compare patients who did or did not recover in order to determine risk factors. Variable selection was performed using an RF model based on mean decreases in the Gini index and accuracy. Among the models tested, the SVM model achieved the highest accuracy of 75.4%, followed by the RF model with an accuracy of 73.9%. The AdaBoost and MLP models had accuracies of 72.5%, while the KNN model had an accuracy of 65.2%. After variable selection based on RF, all models except RF showed improved performance. The SVM model with selected variables was found to be the most helpful for clinicians in predicting the hearing prognosis of patients with SSNHL,

2.2.4 Hearing-impaired Speech Perception

Speech perception tests provide a measure of how well listeners understand speech in a controlled environment. Machine learning approaches have been applied to predict speech perception

abilities, which play a crucial role in everyday communication. For example, a prediction model utilizing a Random Forest-based machine learning approach to estimate the Speech Discrimination Score (SDS) of a patient based on Pure Tone Average (PTA) thresholds was proposed by Kim et al. [78]. A large-scale dataset from 12,697 subjects, who underwent both PTA and speech discrimination tests, was collected to design and evaluate the prediction model. The accuracy of predicting SDS levels was reported as 95.1% for the left ear and 96.6% for the right ear. Comparison with other commonly used machine learning algorithms such as SVM and Multilayer Perceptron revealed the effectiveness of the RF-based approach. Age groups were also predicted to detect sudden changes in hearing at younger ages. The proposed scheme accurately predicted the age groups of patients, with better prediction performance for the elder population compared to the age groups between 0-49. This is because age-related hearing loss is less common in younger age groups. An interesting observation was made from the classification results for the 80-90 and 90-100 age bins. In the left ear, there were nearly twice as many misclassified cases compared to the right ear. While most cases were well-classified with high accuracy, there were some parts of the confusion matrix where a significant number of misclassified cases existed, exhibiting different patterns for the left and right ears. These findings carry significant implications for the development of a feasible screening tool aimed at detecting malingering in tests related to hearing loss. Haro et al. [79] proposed a computational model that combines a biophysical cochlear-nerve (auditory-nerve) spike generator with a deep neural network (DNN) classifier to simulate human speech-in-noise perception. The model carried out a digit classification task, mimicking human behavior by using only acoustic sound pressure as input. The goals were to test the feasibility of using a spiking cochlear model to replicate human-like speech recognition, examine the impact of cochlear model degradations on the digit-in-noise classifier performance, and explore

the effects of neural plasticity on cochlear degradation adaptation. The model achieved human-like performance across various signal-to-noise ratios (SNRs) under normal-hearing cochlear settings (normal auditory function and hearing thresholds), reaching 50% digit recognition accuracy at -20.7 dB SNR. This performance was comparable to eight normal-hearing participants. The authors also simulated medial olivocochlear reflex (MOCR) and auditory nerve fiber loss, which resulted in decreased digit recognition accuracy at lower SNRs. Their findings support the hypothesis that cochlear synaptopathy, a phenomenon that impairs speech comprehension in noisy environments without affecting the clinical pure-tone threshold audiogram, primarily affects communication in background noise rather than quiet environments. Despite implementing adaptation through the DNN, the models were unable to fully recover normal-hearing performance at the lowest SNRs, indicating limitations in performance recovery following peripheral damage. The results demonstrated that the model achieved human-like performance and provided insights into the effects of peripheral hearing loss on speech-in-noise perception. Another study aimed to identify key factors that distinguish individuals with clinically normal or "near-normal" hearing in terms of their speech-in-noise perception and to develop a regression model for predicting speech-in-noise difficulties in this population [80]. Additionally, the study sought to evaluate the potential effectiveness of the regression model as a diagnostic criterion for clinical use. The speech-in-noise perception was assessed using a composite speech-in-noise score (CSS) derived from three different speech-in-noise measures. Two subgroups were created based on the CSS, representing individuals with the lowest and highest scores, respectively. These groups were compared in terms of hearing thresholds, temporal perception, noise exposure, attention, and working memory. A multiple linear regression model was developed using variables with significant differences between the two groups. Extended high-frequency (EHF) hearing

and working memory emerged as significant predictors of the CSS. A second regression model was then constructed using these two predictors. The resulting regression formula was evaluated using Monte Carlo cross-validation on the complete dataset, which accounted for 41% of the total variance. The area under the receiver operating characteristics curve was 0.76, and the root mean square error was 0.60. These findings indicate that both peripheral (auditory) and central (cognitive) factors contribute to speech-in-noise difficulties in normal-hearing adults. Lenatti et al. [81] analyzed the performance of multivariate ML models in a speech-in-noise hearing screening test and examine the contribution of measured features to hearing loss detection using explainability techniques. Seven ML techniques were trained and evaluated on a dataset of 215 tested ears, including those with and without hearing loss. The dataset included six input features extracted from speech-in-noise testing and one output feature indicating the presence or absence of hearing loss determined by pure-tone average. Random forest demonstrated the best performance, followed by DT, SVM, LR, and gradient boosting. Post hoc explainability analysis revealed that age, number and percentage of correct responses, and average reaction time were the most relevant features in predicting hearing loss, while total test time had the least relevance. The classification performance was high in the balanced dataset used in the study (0.86 sensitivity, 0.85 specificity), but further validation on a larger representative population is necessary to evaluate the algorithms' performance in real-world settings.

2.2.5 Clustering of Audiometric Profiles

In addition to prediction tasks, machine learning has been utilized for clustering audiometric profiles to identify subgroups or patterns within a population. Unsupervised learning

algorithms have been employed to group individuals based on similarities in their audiometric data. This clustering analysis can aid in the characterization of different types or stages of hearing loss and provide insights into the underlying mechanisms. To better understand the diverse profiles of hearing loss, Parthasarathy et al. [82] employed a Gaussian Mixture Model (GMM), which allowed for the clustering of audiograms without making assumptions about frequency relationships, interaural symmetry, or etiology. The GMM analysis identified ten distinct types of audiograms, characterized by varying degrees of high-frequency loss, flat loss, mixed loss, and notched profiles. These included patients with normal thresholds, notch-type audiograms, and audiograms with varying degrees of high-frequency hearing loss at deflection points of 8 kHz, 4 kHz, 2 kHz, and 1 kHz. Additionally, there was a flat-type audiogram with loss across all frequencies, as well as audiogram types with varying degrees of low-frequency loss in addition to sloping high-frequency loss. A separate GMM clustering of audiograms from the NHANES dataset revealed six similar types, albeit without the more extreme hearing loss configurations observed in the patient cohort. These detected classes exhibited predictable relationships with age and sex. The prevalence of normal audiograms decreased with age, while the prevalence of hearing loss increased. The mixed types tended to be associated with older age, while normal and notch types were more prevalent among younger individuals. Traditionally, certain audiogram shapes, such as the 4000 Hz notch type and steeply sloping high-frequency loss type, have been linked to noise overexposure and sensorineural hearing loss. In this study, these shapes also showed a predominance of males, likely reflecting historical differences in occupational noise exposure between men and women. The researchers of another study conducted an examination of "presbycusis phenotypes" in a database consisting of 960 subjects ranging from 18 to 92 years old [83]. Each subject underwent 30 measures of peripheral hearing sensitivity. Cluster

and principal component analyses (PCA) were employed to examine measures of central tendency and dispersion with age and sex in the data. The goal was to identify natural classifications that could help distinguish different presbycusis sub-types and their underlying bases. Surprisingly, the analysis revealed that the hearing phenotypes did not naturally form discrete classes of presbycusis. Instead, PCA identified two components corresponding to the overall degree and configuration of loss. By applying the K-means algorithm, eight clusters were identified. Upon inspection, it was clear that the clusters formed have fluid and arbitrary boundaries, yet they group similar phenotypes together. Although there are no categorical boundaries that delineate presbycusis phenotypes, this analysis still allows for the identification of sub-types, particularly in the extremes of the data distributions.

A data-driven analysis was conducted on subjective data related to hearing disabilities and handicaps to explore "benefit patterns" resulting from rehabilitation in different audiometric groups [84]. The study aimed to achieve three objectives: identify the benefit patterns associated with different audiometric groups, determine the priorities for hearing rehabilitation by targeting specific hearing difficulties and handicaps in need of improvement, and explore the feasibility of implementing personalized post-fitting rehabilitation strategies. The methodology employed involved unsupervised and supervised learning and comprised five main steps. The first step involved dimensionality reduction using factor analysis. In the second step, the processed dataset was divided into four subsets representing listeners from four distinct audiometric groups. Archetypal analysis was then applied in the third step to identify the archetypal patterns of benefit. In the fourth step, listeners were clustered into benefit profiles. Finally, the importance of each questionnaire item as a predictor was inferred in the fifth step. A total of 572 hearing-aid users completed questionnaires assessing hearing difficulties and hearing handicaps. Through the

data-driven approach, four distinct benefit profiles (optimal, near-optimal, near-suboptimal, and suboptimal) were identified, with each profile varying across the different audiometric groups. Among the groups with a low degree of high-frequency hearing loss (HLHF), priority was given to rehabilitating hearing handicaps, while the groups with HLHF exceeding 50 dB HL prioritized improvements in speech understanding.

Part II

Movement Disorders

Chapter 3: Simplification of Mobility Tests and Data Processing to Increase Applicability of Wearable Sensors as Diagnostic Tool for Parkinson's Disease

The work of this chapter has been published in the journal Sensors as: Rana M. Khalil, Lisa M. Shulman, Ann L. Gruber-Baldini, Sunita Shakya, Rebecca Fenderson, Maxwell Van Hoven, Jeffrey M. Hausdorff, Rainer von Coelln, and Michael P. Cummings. Simplification of mobility tests and data processing to increase applicability of wearable sensors as diagnostic tools for Parkinson's disease. *Sensors*, 24(15):4983, 2024 [28].

Abstract

Quantitative mobility analysis using wearable sensors, while promising as a diagnostic tool for Parkinson's disease (PD), is not commonly applied in clinical settings. Major obstacles include uncertainty regarding the best protocol for instrumented mobility testing and subsequent data processing, as well as the added workload and complexity of this multi-step process. To simplify sensor-based mobility testing in diagnosing PD, we analyzed data from 262 PD participants and 50 controls performing several motor tasks wearing a sensor on their lower back containing a triaxial accelerometer and a triaxial gyroscope. Using ensembles of heterogeneous machine learning models incorporating a range of classifiers trained on a set of sensor features,

we show that our models effectively differentiate between participants with PD and controls, both for mixed-stage PD (92.6% accuracy) and a group selected for mild PD only (89.4% accuracy). Omitting algorithmic segmentation of complex mobility tasks decreased the diagnostic accuracy of our models, as did the inclusion of kinesiological features. Feature importance analysis revealed that Timed Up and Go (TUG) tasks to contribute the highest-yield predictive features, with only minor decreases in accuracy for models based on cognitive TUG as a single mobility task. Our machine learning approach facilitates major simplification of instrumented mobility testing without compromising predictive performance.

3.1 Introduction

Parkinson's disease (PD) is the most common neurodegenerative movement disorder and affects an estimated 7–10 million people worldwide [25]. It is characterized by a range of motor symptoms, including tremors [85, 86], bradykinesia [87, 88], rigidity [87, 89], a slow shuffling gait with small steps [90–92], freezing of gait [93, 94], and imbalance [90, 95]. Among all these motor symptoms, gait impairment and imbalance most substantially impact the quality of life of PD patients, ultimately resulting in disability and loss of independence [96, 97].

Currently, PD diagnosis is based on clinical assessment, including history and neurological examination, which is inherently subjective in nature and depends on the level of experience of the healthcare provider. Consequently, initial clinical diagnosis by movement disorders experts has limited accuracy (79.6%, increasing to 83.9% at follow-up) and limited overall validity [27]. The need for objective, low-cost, reliable, and valid biomarkers that can accurately identify patients with PD was identified as one of the top priorities in PD research by a National Institutes

of Health/National Institute of Neurological Disorders and Stroke (NIH/NINDS) panel [98].

Wearable sensors containing accelerometers and gyroscopes show great promise in objectively detecting the presence of key PD motor symptoms in a convenient and accessible manner [99–103]. The widespread availability of wearable sensors makes it possible to quantitatively assess mobility without costly equipment or a dedicated gait laboratory [104]. Used in instrumented mobility testing, these sensors provide highly granular data on gait, balance, and other motor functions to identify features and patterns indicative of PD [105, 106]. Ultimately, the expectation is for wearable sensors to support earlier detection and treatment of PD, leading to better outcomes for patients [100, 107]. However, critical challenges have yet to be overcome in order to turn sensors from research instruments into clinical tools widely used to support the diagnosis of PD in clinical practice [108, 109].

Studies on the use of wearable sensors to diagnose PD vary widely in terms of which and how many sensors to use and where to place them [34, 37, 45, 110–114]. There is also a large variation in which and how many instrumented mobility tasks participants are required to perform [35, 37, 39, 115]. Moreover, the methods employed to process sensor data and derive relevant features are complex and vastly different across studies [33, 37, 116–118].

We conducted a search of the literature from 2017 to 2024 to identify previous studies that used wearable sensors for PD assessment. This involved reviewing three prior review articles [112–114] and performing a search of the PubMed database with the keywords "(wearable OR IMU) AND Parkinson", which yielded 72 relevant articles. Our survey revealed that the number of sensors used in these studies ranged from 1 to 18, and the number of tasks ranged from 1 to 14, occurring in various combinations (Figure 3.1). We categorized these studies into four groups based on their objectives. The first group included studies focusing on PD diagnosis. There

were 21 studies in this group, using between 1 and 18 sensors and from 1 to 12 tasks [119–139]. Among these, six studies used a single sensor placed on the lower trunk [124,132,135], foot [125], wrist [130], or torso [136]. Additionally, 11 studies used data from a single mobility task—either simple gait tasks [119, 122, 127, 129, 133, 135, 137–139] or more complex tasks [128, 132]—with only two of these studies also using a single sensor [132, 135]. The second group, comprising 24 studies, focused on quantifying/monitoring the severity of PD symptoms [44, 56, 110, 140–160], with 4 using a single sensor [140–143] and 7 using a single mobility task [142, 144–149]. Only one study in this group used both a single sensor and a single mobility task [142]. The third group, consisting of 15 studies, aimed to detect freezing of gait in PD [161–175], with 8 using a single sensor [161–166, 172, 173] and 2 using a single mobility task [167, 168]. The last group included 12 studies that performed quantitative characterization of gait or identified imbalance in PD [111, 176–186], with 3 using a single sensor [176–178] and 5 using a single mobility task [111, 179–181, 186]. None of the studies from the third or fourth groups used both a single sensor and a single mobility task. The detailed results of the surveyed studies are presented in Supplementary Table A.1.

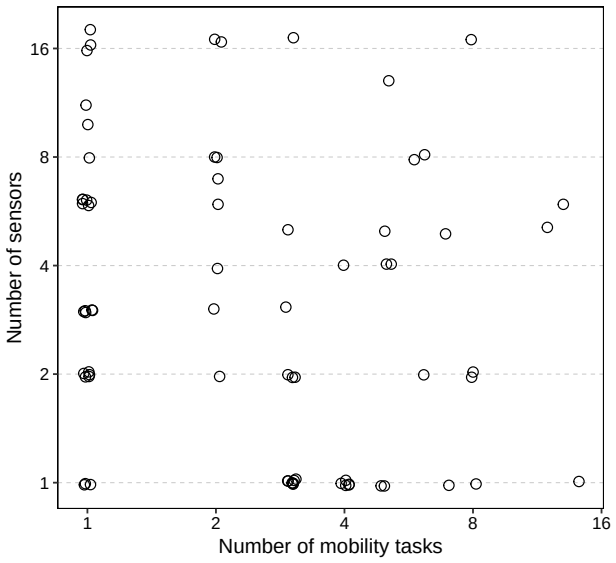


Figure 3.1: Number of mobility tasks and sensors used in prior studies. The axes are in $\log_2$ scale.

The reviewed papers exhibit a variety of approaches using wearable sensors for PD assessment, focusing on different tasks and parameters. Some studies emphasize the importance of comprehensive data collection to train symptom detection models [44], objectively measure gait parameters [119, 122, 124, 126–128, 145, 160, 170, 176, 177, 180, 186], monitor daily life activities [146, 170], or detect early stages of the disease [111, 124, 133, 138, 174]. The choice of the number of sensors and their placement is often driven by the specific objectives of each study. For instance, capturing hand motor tests [120, 140, 154], assessing balance impairments [111, 123, 142, 160, 182], and measuring dyskinesia severity [153] require specific sensor setups. Sensors are commonly placed on the wrist for tremor and bradykinesia monitoring [56, 130, 140, 144, 150, 152, 154, 156, 173]; on the foot or shoe to measure gait parameters and detect freezing gait and gait phases [119, 125, 129, 145, 162, 171, 179, 180]; on the waist, hips, torso, or lower back for general gait and balance analysis [124, 126, 135, 136, 164, 165, 172, 178];

and on the lower limbs to measure several key aspects of PD [137, 158, 168, 175, 185, 186]. Some studies integrate multiple sensor locations to capture a broader range of symptoms, such as axial rigidity and postural instability, providing comprehensive data [110, 121, 131, 138, 139, 146, 148, 149, 151, 155, 170, 184]. Overall, the reviewed studies aim to balance detailed data capture with practical application, ensuring accurate and relevant information for their specific goals.

Increasing the number of tasks (abscissa of Figure 3.1) is associated with more time required from both the patient and clinician, while an increasing number of sensors (ordinate) indicates greater complexity in terms of maintaining, storing, and placing sensors, as well as processing the data. Therefore, the goal of this project is rigorous simplification of the number of sensors, mobility testing, and data processing protocols based on a systematic and data-driven analysis of comprehensive sensor-based mobility data.

Machine learning offers a powerful and flexible analytical approach to accomplish this goal in a large-scale, high-dimensional, and complex dataset collected by wearable sensors [32, 36, 43, 45, 52]. Furthermore, machine learning can identify and capture hidden patterns and relationships that would be hard to detect using standard observation and statistical methods [12].

Although previous studies have successfully used machine learning and wearable sensors for PD diagnosis [121, 125, 126, 128, 129, 131, 133, 135–139], wearable sensors are still not commonly applied in clinical settings due to the complexity and variability of proposed protocols for mobility testing and data processing. Our study objective is to employ machine learning techniques to identify strategies that streamline and simplify the process, reducing the time and effort required by both participants and clinicians while maintaining the predictive performance of models to diagnose (or support the diagnosis of) PD based on instrumented mobility testing. This will ease the application of wearable sensors as a diagnostic tool in clinical practice.

As an initial step toward the simplification of mobility testing, we used a single sensor positioned on the lower backs of participants. Then, we show that sensor data collected from participants performing a single mobility task in a conventional clinic hallway can effectively differentiate individuals with PD from controls. Moreover, we demonstrate that manual segmentation and complex calculations of kinesiological features, which have been prevalent in prior studies [33, 35, 40, 110, 115, 187], may be unnecessary. Instead, algorithmic segmentation, followed by the calculation of a small set of sensor-derived features from a single sensor and a single mobility task, offers substantial simplification of both mobility testing and data processing without significantly compromising predictive accuracy. Feature importance analysis revealed that timed up and go (TUG) tasks, which include walking, turning, and sit-to-stand and stand-to-sit transitions, provided more valuable information to our models than simpler mobility tasks such as gait and sway assessments. These results suggest that substantial simplification in mobility testing with a wearable sensor and subsequent data analysis is possible without loss of diagnostic accuracy for PD, facilitating the integration of wearable sensors into clinical practice.

3.2 Materials and Methods

3.2.1 Participants

Our study on sensor-based mobility analysis in parkinsonism ("Towards Next-Generation Phenotyping in Parkinson Disease: Quantitative Analysis of Gait and Balance Using a Portable Biosensor Device") at the University of Maryland Movement and Memory Disorders Center (UM-MMDC) enrolled a total of 368 participants between October 2015 and March 2020, including 50 control participants and 318 participants with parkinsonism. Of those, 293 were diagnosed

with idiopathic PD by a UM-MMDC movement disorder specialist, applying the UK Parkinson's Disease Society Brain Bank Clinical Diagnostic Criteria [188]. All participants with parkinsonism were also enrolled in the Health Outcomes Measurement (HOME) study at UM-MMDC. The HOME study is a naturalistic cohort study that collects patient- and clinician-reported data during routine office visits. In addition to HOME study participation, all 318 participants with parkinsonism plus 50 non-parkinsonian age-matched control participants were separately enrolled in the mobility analysis in the parkinsonism study. Patients were excluded from the mobility analysis study if they were unable to stand or walk without assistance (Hoehn & Yahr stage 5); presented with other conditions unrelated to parkinsonism that, in the judgment of the treating physician, significantly interfere with gait and balance (e.g., advanced lower extremity arthritis or severe lower extremity neuropathy); or failed to provide signed consent. A total of 24 of the study participants had previously undergone deep brain stimulation (DBS) surgery (20 to the subthalamic nucleus (STN), 3 to the globus pallidus internus (GPi), and 1 to the thalamic ventral intermediate (Vim) nucleus). All participants signed informed consent. A person accompanying a patient (spouse/significant other, sibling, or friend) was approached for enrollment as a control participant if they were of similar age to the patient and did not carry a diagnosis of parkinsonism or another condition significantly affecting their gait and balance. All study procedures for both the HOME study and the mobility analysis study were approved by the University of Maryland Institutional Review Board.

Of 318 participants with parkinsonism, 293 were diagnosed with idiopathic PD. The remaining 25 participants were diagnosed with other forms of parkinsonism. Participants with non-PD forms of parkinsonism were excluded from data analysis for this study. Of 293 participants with a diagnosis of idiopathic PD, one was unable to complete the study procedures and

was excluded. Of the remaining 292 PD participants, 30 were excluded from our analysis due to missing data (2 due to technical problems with the wearable sensor and 28 due to a lack of important clinical data). None of the 50 control participants were excluded.

3.2.2 Clinical Evaluations

Participants were enrolled during routine outpatient visits at the UM-MMDC. All PD participants underwent full clinical assessment according to the HOME study protocol, as previously described [96, 189–191]. In brief, collected general data included basic demographics, Cumulative Illness Rating Scale-Geriatrics (CIRS-G) [192], the Older Americans Resource and Services (OARS) Disability Subscale [193] (activities of daily living (ADLs) and instrumental activities of daily living (IADLs)), a questionnaire on the use of assistive devices and falls (ADF), PROMIS Profile 29 [194], and the Montreal Cognitive Assessment (MoCA) [195]. PD-specific data included PD duration (time since symptom onset and time since diagnosis), PD severity (Total and Motor (Part III) Unified Parkinson's Disease Rating Scale (UPDRS) [196], and modified H&Y staging [97, 197]). The HOME study is a naturalistic study, and patients are enrolled during routine clinic visits and on their regular medications. For 87% of assessments performed for this study, the patients rated themselves as being in a medication ON state. Patients who had previously undergone DBS surgery were examined in the stimulation ON state, both for clinical assessments and for the instrumented mobility testing.

3.2.3 Assessment of Mobility

Mobility testing with a wearable sensor was performed using a previously published protocol [115, 198, 199]. We adopted the mobility testing protocol from the Memory and Aging Project and the Minority Aging Research Study at the RUSH Alzheimer's Disease Center at Rush University [200] based on the previous successful use of mobility data from this study in parkinsonism. In brief, participants performed mobility tasks with a single wearable sensor strapped to their lower back. For this study, the following data from five of those mobility tasks were analyzed: 1. a 32-foot walk, where participants walk eight feet back and forth twice without stopping, including three turns; 2. standing with eyes open, where participants stand with their feet shoulder-width apart for 20 seconds with their eyes open; 3. standing with eyes closed, where participants stand with their feet shoulder-width apart for 20 seconds with eyes closed; 4. two trials of the timed up and go test (TUG), where participants stand up from a seated position in a chair, walk eight feet to a line on the floor, turn, walk back to the chair, make a second turn, and sit down again; and 5. two trials of cognitive TUG (cogTUG), where participants perform the TUG test while counting backwards from 100 in steps of 3.

3.2.4 Device and Data Collection

Preceding the mobility testing procedure, a small (106.6 $\times$ 58 $\times$ 11.5 mm) lightweight (55 g) sensor device (Dynaport MT, McRoberts B.V., The Hague, The Netherlands, technical specifications: `https://www.mcroberts.nl/products/movetest/`, accessed on 17 June 2024) was strapped to the participant's lower back with a neoprene belt around the waist. This device contains a triaxial accelerometer (range: $\pm$ 8 g; resolution: 1 mg) and a triaxial

gyroscope (range: ± 2000 dps; resolution: 70 dps) with 100 Hz sampling of three-dimensional acceleration (vertical (az), mediolateral (ay), and anteroposterior (ax)) and rotation (yaw (gz), pitch (gy), and roll (gx)) of the lower trunk. Recording by the sensor was controlled (started and stopped) remotely via wireless Bluetooth signals from a notebook computer using a handheld controller. To facilitate post hoc segmentation of individual mobility tasks, digital markers were placed to label the beginning and end of each mobility task. At the end of each test session, the sensor was removed from the participant and connected to an encrypted notebook computer for data transfer. Raw data files were then transferred to a secure data server.

3.2.5 Data Extraction

A custom-made, runtime-based Matlab graphic user interface (GUI) was used to view the accelerometry data. The 32-foot walk, TUG, and cogTUG were identified by the typical accelerometry pattern of each of these tasks and by the digital markers recorded at the time of mobility testing. The two 20 seconds segments corresponding to standing with eyes open and closed, respectively, were identified solely based on the digital markers recorded at the time of mobility testing. The raw accelerometry data for each individual mobility task were saved as separate files for subsequent data processing.

Figure 3.2 presents an overview of the machine learning pipeline implemented in this study. The following sections describe the steps of the workflow in detail.

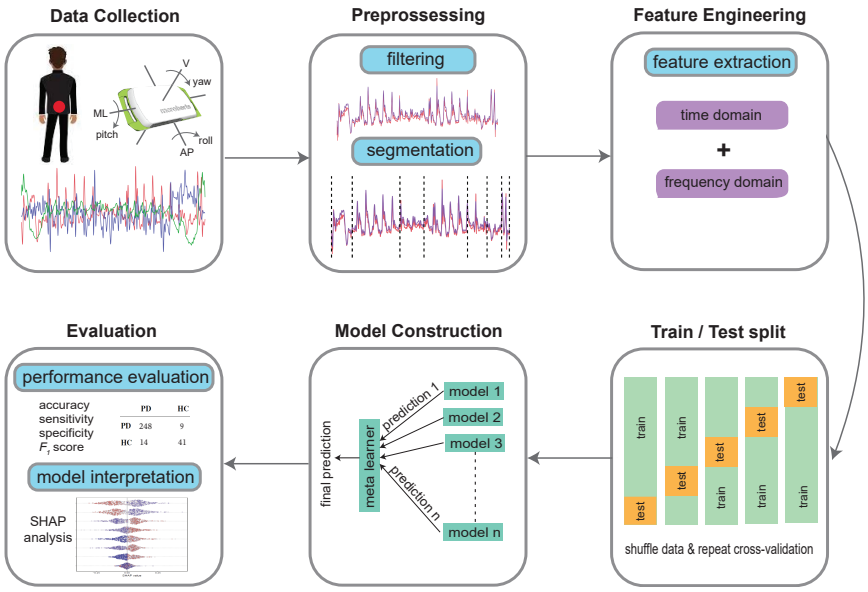


Figure 3.2: Machine learning pipeline. The process started by collecting data from a wearable sensor on three linear and three angular axes. Filtration was then applied to reduce noise in signals using a Butterworth filter. The composite tasks were further processed by dividing them into segments corresponding to the different subtasks. Using the six signals recorded for each participant and task, a large set of time-domain and frequency-domain features was calculated. The data were then shuffled multiple times, and a cross-validation framework was applied in each iteration to generate train and test sets. Within each fold, forward feature selection was performed on the training set, and a super-learner model was built. This model used a wide array of base models and the final prediction was calculated by assigning a weight to the prediction of each base learner. Finally, the performance of the model was evaluated using accuracy, sensitivity, specificity, and F_1 score measures. Model interpretation was performed using SHapley Additive exPlanation (SHAP) analysis (see Supplementary Methods, Section A.1.1) to quantify the contributions of features to the final model prediction. AP: anteroposterior direction; V: vertical direction; ML: mediolateral direction.

3.2.6 Segmentation

Some motor tasks performed in this study are composed of several subtasks. For instance, the TUG and cogTUG tasks consist of the following six components: sit to stand, two turns, two walks, and stand to sit. The 32-foot walk task has seven parts representing the four walks and three turns. We developed a segmentation procedure extending a previously described algorithm [201] to include more tasks and subtasks. Our algorithm starts by excluding the stationary

segments at the start and end of the signal. These stationary segments result if the recording starts before the participant starts the task or the recording stops after the participant terminates the task. The largest constant parts (with small slope change) common among the six channels and found at the start or the end of the signal were defined as stationary segments and were trimmed. Then, the turns of a task were detected by calculating the angular position from the angular velocity in the vertical axis using the trapezoidal integration method [202]. The model for turning fits a line with three horizontal segments and two segments with constant slopes. The transition points from zero to constant and from constant to zero slopes were used to mark the starts and ends of the turns (Supplementary Figure A.1a). This method of determining the onset and offset time points of 180° turns has been previously validated [203]. The same integration method was applied to the acceleration data in the vertical direction (az) to determine the sit-to-stand and stand-to-sit parts of the TUG and cogTUG tasks. Specifically, the start and the end of the peak of the fitted line represent the sit-to-stand part, and the start and the end of the valley represents the stand-to-sit part (Supplementary Figure A.1b). Finally, our algorithm marked the segments not mapped as turns in the 32-foot walk task and those between the sit-to-stand part and the first turn or between two turns in the TUG and cogTUG tasks as walking segments. Supplementary Figures A.2 and A.3 show the different components extracted for a participant performing the TUG and 32-foot walk tasks, respectively. Once we applied our segmentation procedure, the time-domain and frequency-domain features described in the next section were calculated separately for each subtask.

3.2.7 Feature Engineering

For each of the six channels recorded for each participant and mobility test, we started by applying a low-pass, 20 Hz, zero-lag, 4th-order Butterworth filter [204] to reduce noise in signals. Then, we derived the following sets of features from each signal: statistical measures consisting of 23 frequency-domain features (Supplementary Table A.2); 44 time-domain features (Supplementary Table A.3); the top-ten power amplitudes from Discrete Fourier Transform (DFT) and Lomb–Scargle Periodogram (LSP) [205, 206]; and three time-domain features calculated across the signals of each pair of the six channels (Supplementary Table A.4). These sensor-derived features were combined from several previous studies [37, 59, 118, 207, 208]. Kinesiological features were calculated/extracted using a second custom-made, runtime-based, mobility task-specific Matlab graphic user interface (GUI) as previously described [116, 117, 209]. Additionally, we included features not derived from the sensor data, including demographics and other variables (e.g., age, race, gender, height, OARS, and ADF scores) as predictive variables. Each participant was represented using a total of 26,515 features calculated from the six channels of the segmented subtasks of the 32-foot walk, standing-with-closed-eyes task, standing-with-open-eyes task, the segmented tasks of the TUG task (from the first trial and the second trial and the mean of corresponding features from both trials), and the segmented subtasks of the cogTUG task (from the first trial and the second trial and the mean of corresponding features from both trials). Preprocessing steps were applied to the data before running machine learning models, including removal of variables with constant or infinite values, imputation of missing values using the median of observed values, and normalization of each variable by subtracting its mean and dividing it by its standard deviation.

3.2.8 Feature Selection

To reduce the number of features, we applied forward feature selection. The goal was to select an optimal subset of features that results in a simpler model with similar or better performance. We started by ranking all features by their importance using a random forest model after randomly undersampling the larger group. Then, the elbow point of the feature importance plot was defined as the point that is farthest away from a straight line drawn from the first to the last point of the plot. This elbow point was used as a cut-off threshold to select the set of candidate features. If the size of this set was n_{thr}, then multiple models were trained using the top one, top two, or . . . , top n_{thr} important features. The best model was selected as the one that minimizes the total Akaike information criterion (AIC) [210], and the corresponding set of features was used in subsequent analyses. This variable reduction strategy results in fewer features compared to the use of principal component analysis, as reported in similar studies [115,211,212], and retains the original features.

3.2.9 Machine Learning Model

Prior to model training, we split the data into five groups. We used stratified sampling to preserve the class frequencies within each group. Then, a nested-loop framework was applied, where in each iteration of the inner loop, a five-fold cross-validation was implemented by choosing four groups for training and one for testing of the model. This ensured that each participant was included exactly once in the test data. The benefit of this technique is that all observations were used for both training and testing. Within each fold, the feature selection process described in Section 3.2.8 was applied exclusively to the training data. Selected features from the training

set were then applied to the corresponding test set to prevent data leakage and to ensure generalization [213]. In the outer loop, we created five replicates by randomly shuffling the whole dataset using different random number generator seeds to create different combinations of participants within each training and testing group. The final class of each participant was defined as the majority vote of the predicted classes from the five outer loops.

Prior to constructing each classifier, a balanced sampling process was applied to the training sets, where the mean sample sizes of the two classes were calculated and used to undersample the majority class and/or oversample the minority class. This mean-based sampling approach provides a balance between simplicity and computational efficiency. An ensemble method combining multiple machine learning algorithms was then applied to the training sets. This approach is known to achieve performance equal to or better than any of the base learners. The user of a super learner (SL) [214] was proposed as a stacked generalization using an ensemble of methods that generates an optimal model by creating a weighted linear convex combination of a set of base algorithms. The super-learner algorithm compares the performance of the user-chosen base algorithms by applying k-fold cross-validation to determine the weighted contribution of each base algorithm to the ensemble. We used the R interface H2O with five-fold cross-validation and a generalized linear model as a meta learner. For base models, we included a generalized linear model (GLM) [215], distributed random forests (DRFs) [216], gradient boosting machines (GBMs) [217], extreme gradient boosting (XGBoost) [218], and neural networks [219]. Hyperparameters of the base algorithms were selected using grid search criteria. The final ensemble incorporated 327 base models, including 11 GLM models, 100 DRF models, 100 GBM models, 16 XGBoost models, and 100 neural network models. The class prediction thresholds were selected based on the highest F_1 score.

Given the predicted class of each participant, the performance of the classification model was assessed using overall accuracy, area under the receiver operating characteristic curve (AUC-ROC), sensitivity, specificity, and F_1 score.

3.2.10 Group Feature Importance

We employed the concept of group feature importance to assess the information of various mobility tasks in predicting the outcome of interest. Group feature importance refers to the assessment of the collective information of a set of features within a specific category. All features derived from each mobility task were grouped together to form five separate groups corresponding to the five mobility tasks. We also created an additional group by combining features not derived from the sensor data, including demographics and other variables (e.g., age, race, gender, height, OARS, and ADF scores). To measure group importance, we adapted the permutation importance method for feature groups within the context of random forests, building upon a previously outlined approach [220]. For each feature group, the grouped variables were simultaneously permuted, and the resultant mean decrease in accuracy prior to and post permutation was computed.

3.2.11 Statistical Analysis

Data analysis was performed with R version 4.2.3 (2023-03-15) and RStudio. A two-tailed Student's t-test was applied to compare continuous variables, and Pearson's χ^2 test was used to compare categorical variables between two groups. Results are presented as mean $\pm$ s.d. for continuous variables. Actual p values for the statistical tests are reported.

For each performance metric of the models, we display both the computed values and their respective 95% two-sided confidence intervals (CIs). These intervals were derived via balanced bootstrap resampling with 10K replicates and were adjusted using bias correction and acceleration to better manage bias and skewness in the bootstrap distribution. The bootstrap replicates were generated using the boot function in R package boot version 1.3–30 [221, 222], and the boot.ci function from the same package was used to generate the CIs.

3.3 Results

We collected wearable sensor data from 262 participants with PD (87% were in a medication ON state) and 50 age-matched controls. We partitioned PD participants by their modified Hoehn & Yahr (H&Y) stage, defining mild PD as H&Y ≤ 2, moderate PD as H&Y 2.5 and 3, and severe PD as H&Y 4 (note that H&Y 5 patients were excluded from this study). Although our analysis encompasses four classifiers to distinguish control participants from all PD, mild PD, moderate PD, and severe PD groups, respectively, here, we focus on the results of our models for all PD participants and mild PD participants only. The all-PD classifier best mimics a routine clinical situation with patients in various stages of PD. The mild-PD classifier focuses on the group of PD participants that is most similar to control participants, making these two groups the most challenging to differentiate from one another. Results of the models for moderate- and severe-PD participants are detailed in the Supplementary Results (Section A.2.1). Table 3.1 provides the characteristics of the study cohort.

After the participants performed the five motor tasks (Section 3.2.3) wearing a Dynaport MT sensor strapped to their lower back, we followed the experimental design described in Fig-

ure 3.2 to predict the final class of each participant.

Table 3.1: Characteristics of the study cohort.

Feature	**Controls** ($n = 50$)	**All PD** ($n = 262$)	**Mild PD** ($n = 185$)
Age (years, $\bar{x} \pm SD$)	64.1 ± 9.8	66.9 ± 9.3	65.4 ± 9.1
Gender (%male)	38.0	62.0	58.4
Height (cm, $\bar{x} \pm SD$)	168.1 ± 10.9	172.2± 10.4	172.0 ± 10.3
UPDRS (total, $\bar{x} \pm SD$)	–	35.5 ± 17.1	29.4 ± 13.4
UPDRS (motor-part III, $\bar{x} \pm SD$)	–	22.2 ± 11.8	18.5± 9.7
Disease duration (years, $\bar{x} \pm SD$)	–	7.8 ± 6.5	6.6 ± 5.6
H&Y ($\bar{x} \pm SD$)	–	2.2 ± 0.62	1.9 ± 0.26
stage 1 (n)	–	12	12
stage 1.5 (n)	–	4	4
stage 2 (n)	–	169	169
stage 2.5 (n)	–	35	–
stage 3 (n)	–	25	–
stage 4 (n)	–	17	–

3.3.1 Classification of PD versus Control

By applying the forward feature selection approach, the total number of features was reduced from 25,705 to 9 for the all-PD classification and 14 for the mild-PD classification (Supplementary Figures A.4–A.6). Eight of the selected features for the all-PD classification were derived from one or both trials of the cogTUG task; three features were from the second cogTUG trial (cogTUG2), one feature was from the first cogTUG trial (cogTUG1), and four features were the mean of the corresponding features from the two cogTUG trials (mean cogTUG). One feature was derived from the open-eyes standing task. For the mild-PD classification, 13 of the 14 selected features were derived from the cogTUG task; 2 features were from cogTUG2, 1 feature was from cogTUG1, and 10 were from mean cogTUG. One feature was from the eyes-closed standing task. The selected cogTUG features for both classifiers were derived from the turn,

walk, and stand-to-sit components of the task, representing acceleration and rotational measures along the vertical, mediolateral, and anteroposterior axes. None of the selected features for both classifiers were from the 32-foot walk or TUG tasks.

Among the 327 base models included in the ensemble classifier, a gradient boosting machine model made the greatest weighted contribution to the final models (Supplementary Figures A.7 and A.8). The models trained on the selected features achieved an overall accuracy of 92.6% (confidence interval [CI] 88.8%, 94.9%) in distinguishing all-PD from control participants and 89.4% (CI 84.3%, 92.3%) in distinguishing those with mild PD from control participants. Confusion matrices (prediction summaries) and performance metrics for the classifiers are presented in Tables 3.2 and 3.3, respectively.

Table 3.2: Confusion matrices of the classifiers using controls and participants with Parkinson's disease (PD). Rows represent actual classes, and columns represent predictions.

(a) Controls vs. All PD			**(b) Controls vs. Mild PD (H&Y $\leq$ 2)**		
	Controls	PD		Controls	PD
Controls	41	9	Controls	36	14
PD	14	248	PD	11	174

Table 3.3: Classification results of models in distinguishing controls from all Parkinson's disease (PD) and mild-PD (H&Y $\leq$ 2) participants. CI: 95% confidence interval.

Controls vs. PD	**All PD**	**Mild PD**
Number of PD participants	262	185
Accuracy (CI) [%]	92.6 (88.8, 94.9)	89.4 (84.3, 92.3)
AUC-ROC (CI)	0.88 (0.83, 0.94)	0.83 (0.77, 0.90)
Sensitivity (CI)	0.95 (0.91, 0.97)	0.94 (0.90, 0.97)
Specificity (CI)	0.82 (0.69, 0.91)	0.72 (0.58, 0.83)
F_1 score (CI)	0.96 (0.93, 0.97)	0.93 (0.90, 0.96)

To further elucidate which mobility tasks were most relevant for the performance of our models, we conducted a group feature importance analysis (see Section 3.2.10). Figure 3.3 illus-

trates the group feature importance of the five mobility tasks and the non-sensor (demographic and clinical) features. The highest-yield group was the one with cogTUG-derived features for both models distinguishing controls from all-PD and mild-PD participants. A substantial decline of at least 64% in importance was observed when comparing the cogTUG group with the remaining groups for either model, underscoring the remarkable discriminatory power inherent in the cogTUG features. More detailed individual feature-level importance scores and Shapley additive explanation (SHAP) values were also calculated (Supplementary Figures A.4 and A.6, Supplementary Methods (Section A.1.1), Supplementary Results (Section A.2.2)).

To identify potentially confounding factors that may impact our models, we conducted a comparison of demographic and clinical characteristics between correctly classified and misclassified PD and control participants. A two-tailed Student's t-test was applied to compare continuous variables, and Pearson's χ^2 test was used to compare categorical variables between separate groups. The comparison results are summarized in Table 3.4. For both classifiers (all PD and mild PD), false positives (FPs, i.e., controls falsely classified as PD) were significantly older compared to the correctly classified controls (true negatives, TNs). False negatives (FNs, i.e., PD participants classified as controls) were younger compared to the correctly classified PD participants (true positives, TPs), but this difference was not statistically significant. These results suggest that older controls have movement characteristics that confuse the classifier, but there was no significant difference in age between the correctly and falsely classified PD participants. Also, the mean Unified Parkinson's Disease Rating Scale part III (UPDRS_PIII) motor score of the FNs (falsely classified PD participants) was significantly lower than that of TPs (correctly classified PD participants). In contrast to the UPDRS_PIII scores, misclassified PD participants were evenly distributed among the different H&Y stages. Similarly, there was no significant dif-

ference in the medication state, sex, MoCA scores, or CIRS-G scores between the correctly and falsely classified PD participants, suggesting that these factors do not have a confounding impact on model performance.

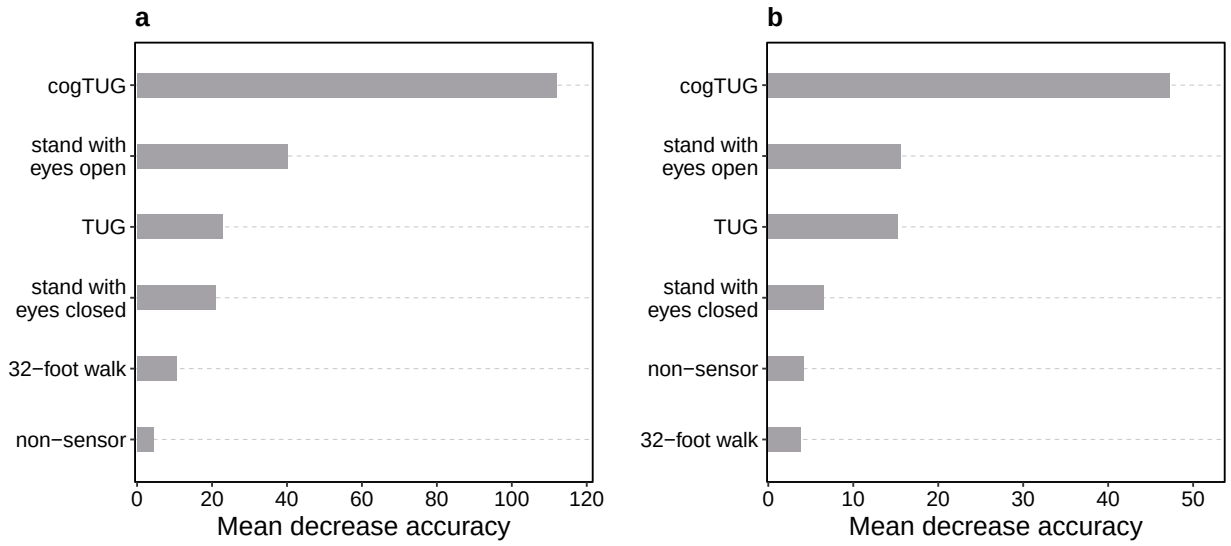


Figure 3.3: Group feature importance of models distinguishing controls from (**a**) all Parkinson's disease (PD) and (**b**) mild-PD (H&Y $\leq$ 2) participants. Groups are ranked by their importance calculated by permuting features within each group simultaneously and reporting the mean decrease in accuracy between the original and permuted data. Under the null hypothesis that there is no association between the group of predictor variables and the model prediction, permutation should have no or little impact on predictive performance. More detailed individual feature-level importance scores and Shapley additive explanation (SHAP) values are illustrated in Supplementary Figures A.4 and A.6.

3.3.2 Strategies for Simplifying Mobility Testing and Its Associated Workload

One of the major obstacles preventing widespread use of wearable sensors for PD diagnosis is the increased burden in terms of time and effort in performing sensor-based mobility testing and subsequent data processing and analysis. To overcome these obstacles, we first focused on possible simplification of the data analysis process. As detailed above, our data processing pipeline includes an algorithmic segmentation approach (Section 3.2.6) to calculate a set of sensor-derived frequency-domain and time-domain features from the resulting task segments,

Table 3.4: Clinical characteristics of correctly classified and misclassified controls and Parkinson's disease (PD) participants.

		Controls			PD		
		False Positive	**True Negative**	***p***	**False Negative**	**True Positive**	***p***
All PD							
	UPDRS_PIII	–	–	–	15.8 ± 8.3	22.5 ± 11.8	0.006
	MoCA	–	–	–	28.6 ± 2.5	27.3 ± 3.0	0.075
	CIRS-G	–	–	–	4.7 ± 4.3	4.9 ± 3.4	0.88
	Age	70.9 ± 7.3	62.6 ± 7.3	0.006	65.9 ± 6.8	66.9 ± 9.3	0.30
	Sex (% male)	11.1	36.0	0.10	50.0	63.3	0.33
	Medication state (% ON)	–	–	–	85.7	70.6	0.44
	H&Y (n)			–			0.47
	1				1	11	
	1.5				1	3	
	2				10	159	
	2.5				1	34	
	3				1	24	
	4				0	17	
Mild PD							
	UPDRS_PIII	–	–	–	13.7 ± 7.4	18.7 ± 9.6	0.03
	MoCA	–	–	–	28.1 ± 1.9	27.5 ± 2.8	0.36
	CIRS-G	–	–	–	4.9 ± 4.4	4.4 ± 3.1	0.72
	Age	71.5 ± 6.3	61.3 ± 6.3	4.0×10^{-5}	65.2 ± 9.1	69.2 ± 7.5	0.12
	Sex (% male)	42.9	36.1	0.71	45.5	60.9	0.32
	Medication state (% ON)	–	–	–	88.2	71.3	0.36
	H&Y (n)			–			0.38
	1				1	11	
	1.5				1	3	
	2				9	160	

referred to as "standard models" below. To evaluate the effectiveness of the automated segmentation method and whether segmentation is necessary at all, we constructed alternative models using features derived from data of the unsegmented tasks. These models decreased the accuracy in distinguishing PD from control participants by 3.2% and 5.1% for all-PD and mild-PD classifications, respectively, compared to our standard models (Table 3.5, second column vs. first column).

We also developed another set of models incorporating calculated kinesiological features (see Section 3.2.7), in addition to the time- and frequency-domain features from the algorithmic

segmentation tasks. These models also reduced the accuracy by 4.8% and 0.50% compared to our standard models (Table 3.5, third column vs. first column). These findings show that algorithmic task segmentation increases diagnostic accuracy, but complex calculation of kinesiological features does not. These time-consuming and labor-intensive additional steps required to calculate the kinesiological features can therefore be eliminated without compromising the predictive performance of our diagnostic models.

Table 3.5: Accuracy (%) of super-learner models using unsegmented tasks and a combination of segmented tasks and calculated kinesiological features to distinguish controls from PD participants. The numbers in parentheses represent the difference in accuracy between each model and the corresponding standard model, which used data from segmented tasks without incorporating kinesiological features (same row).

	With Segmented Tasks and No Kinesiological Features	**With Unsegmented Tasks and No Kinesiological Features**	**With Segmented Tasks and Kinesiological Features**
All PD vs. controls	92.6	89.4 (−3.2)	87.8 (−4.8)
Mild PD (H&Y $\leq$ 2) vs. controls	89.4	84.3 (−5.1)	88.9 (−0.50)

As a second step towards simplification of the overall workflow, we focused on the tasks of mobility testing itself, by assessing the predictive performance achievable with fewer mobility tasks. Given the outsized importance of cogTUG features in the context of our models based on data from all available mobility tasks (Figure 3.3), we examined the performance of two drastically simplified models, using TUG- or cogTUG-derived features only, respectively, in addition to the non-sensor (demographic and clinical) features. We applied the feature selection approach to these single-task models (see Supplementary Table A.5 for the selected features). For the all-PD classifier, the TUG-only model and cogTUG-only model decreased the accuracy by 2.5% and 3.2%, respectively (Table 3.6, columns 2 and 3, row 1). For participants with mild PD, applying the cogTUG-only model decreased the accuracy by 2.4%, whereas the accuracy dropped

more substantially (by 11.7%) using the TUG-only model compared to the standard model based on data from all tasks (Table 3.6, columns 2 and 3, row 2). These results suggest that sensor data from a single appropriately chosen mobility task may be sufficient to achieve satisfactory performance in differentiating PD and control participants. Specifically, a dual task paradigm (cognitive plus mobility task) like the cogTUG task achieves superior diagnostic performance to a mobility task alone (like the TUG task) in identifying participants in the early stages of PD.

Table 3.6: Accuracy (%) of alternative models. TUG-only and cogTUG-only models used features derived from the first trial, the second trial, and the mean of corresponding features from both trials. TUG-duration and cogTUG-duration models used the durations of TUG and cogTUG tasks. The numbers in parentheses represent the difference in accuracy between the respective model and the corresponding all-task (standard) model (same row).

	All Tasks	**TUG-Only**	**cogTUG-Only**	**TUG-Duration**	**cogTUG-Duration**
All PD vs. controls	92.6	90.1 (−2.5)	89.4 (−3.2)	83.3 (−9.3)	84.0 (−8.6)
Mild PD (H&Y $\leq$ 2) vs. controls	89.4	77.9 (−11.7)	87.2 (−2.4)	78.7 (−10.7)	77.0 (−12.4)

Our observation that data from additional mobility tasks only offer a limited increase in diagnostic accuracy led us to consider whether further radical simplification would be possible without compromising the performance of our models. Therefore, we tested whether the total duration of TUG or cogTUG tasks as a single feature may be sufficient to build classifiers with predictive performance similar to that of our more comprehensive models described above. Even though we found a significant time difference between PD and control participants for both TUG and cogTUG tasks (see Supplementary Results, Section A.2.3), the corresponding logistic regression models using total duration as a single predictor variable decreased the accuracy in distinguishing PD from control participants by at least 8.6% compared to our standard models (Table 3.6, columns 4 and 5 compared to column 1). Based on these results, we conclude that total TUG/cogTUG duration alone is not sufficient to capitalize on the discriminatory potential

of these complex tasks.

3.4 Discussion

Our study demonstrates the effectiveness of analyzing wearable sensor data to distinguish PD participants from controls using the proposed machine learning pipeline for five mobility tasks. By leveraging a large dataset comprising participants in different stages of PD and controls performing five mobility tasks, along with a carefully selected subset of sensor-derived features and suitable machine learning models, we achieved an accuracy of 92.6% in identifying participants across different PD stages, surpassing the accuracy of 81% previously reported for clinical diagnosis by movement disorder experts [27]. We focused on our modeling results for all-PD participants and for mild-PD participants specifically. The all-PD classifier represents a typical clinical scenario with patients at various stages, while the mild-PD classifier targets the group most similar to control participants, where differentiation is most challenging [223] and where wearable sensors could be most clinically relevant.

Previously published studies have reported accuracies ranging from 73.5% to 96% in discriminating between PD and control participants [33, 35–37, 39, 40, 42, 121, 128, 131, 135, 137, 139]. Some studies have utilized the AUC-ROC as a performance measure, with values ranging from 0.86 to 0.99 [32, 41, 126], compared to an AUC of 0.86 for our model. Additional studies have reported sensitivity and specificity metrics, with sensitives ranging from 40% to 96.2% and specificity ranging from 82% to 96.9% [34, 38, 125, 136], as compared to 95% sensitivity and 82% specificity in our case. Major differences between studies in terms of the number of participants, PD severity, tasks conducted, types and quantity of sensors employed, and reported

performance metrics likely account for these differences in diagnostic performance and make direct comparison very challenging.

Analysis of group and individual feature importance revealed that the most important mobility tasks for classifying PD participants are the TUG and cogTUG tasks (Figure 3.3). Both tasks involve various everyday activities, including walking, turning, and sit-to-stand and stand-to-sit transitions. Our results demonstrate that more complex mobility tasks with multiple mobility subtasks provide diagnostically superior information compared to simpler motor tasks. Models based on the cogTUG task as a classic cognitive–motor dual task, in particular, showed superior performance in differentiating between participants with mild PD and controls (see Table 3.6). Simpler tasks, such as standing with eyes open, standing with eyes closed, and the 32-foot walk were much less important. Previous studies have reported a strong correlation between these simpler motor assessments and PD [110, 224–227]. In the context of our study and with the included mobility tasks and features, we demonstrate that simpler motor tasks are not as informative as more complex ones and, thus, can be omitted from the battery of our mobility tests with very minor loss of predictive performance. This is consistent with the results of a previous study that utilized multiple sensors and a higher number of sensor-derived features [228]. However, our results show that further simplification is possible using a single sensor and a reduced set of features, even with a single mobility task.

In our machine learning framework, we implemented automatic segmentation of the composite tasks into their subtask components, coupled with the calculation of many sensor-derived features combined from prior studies. Segmentation of complex tasks into subtasks (e.g., sit to stand, walk, and turn for the TUG and cogTUG tasks) increased the diagnostic accuracy of our models compared to models based on unsegmented data (see Table 3.5). This automatic

segmentation, as opposed to the use of manual annotations in other studies [33, 187], not only expedites the processing but also minimizes subjective biases as a source of errors. Additionally, we employed ensembles comprising disparate machine learning classifiers. This approach distinguishes our study from others that primarily depend on individual models [33,38,40,42,110,111], thereby improving predictive accuracy and robustness. Weighted ensembles have demonstrated performance equal to or better than that of any constituent individual learner [214, 229], as they leverage the diverse properties and strengths of various algorithms. Although we observed a modest performance difference of approximately 1% between the ensembles and the models with the highest weight in each ensemble, the process of finding the optimal model requires searching through multiple models and their respective hyperparameters.

Many kinesiological variables can be calculated based on signal processing and analysis of accelerometric and gyroscopic data from a wearable sensor using a developed software. This provides quantitative movement characterization in kinesiological terms, as demonstrated in previous studies [40,115,119,230]. However, these variables require additional manual input for the developed software, and our analysis demonstrates that these kinesiological features are neither necessary nor more helpful than analysis based on sensor-derived features alone in achieving better model performance (see Table 3.5). This observation suggests that properties of movement beyond those corresponding to conventional kinesiological variables were captured in our feature set derived from data from a wearable sensor and that those properties were instrumental in elevating the predictive performance of our models. Eliminating the additional data processing steps required for the generation of kinesiological variables significantly reduces the workload associated with sensor-based mobility testing and facilitates the use of quantitative mobility testing in research (e.g., testing large numbers of participants) and clinical practice.

Finally, we attempted radical simplification of our models by examining the time to complete either the TUG or cogTUG task as the only feature in our models. Total task duration was significantly shorter for controls compared to PD participants (see Supplementary Results, Section A.2.3). However, using elapsed time alone as the basis for discriminating PD was markedly less accurate than models including additional sensor-derived features (see Table 3.6). Another study also showed moderate accuracy (AUC = 0.699) when discriminating between PD patients and controls using the total duration of a TUG test performed at a self-selected speed [231]. The decrease in model performance demonstrates a trade-off between simplicity of testing and data analysis on the one hand and predictive accuracy (and, hence, the value of mobility testing) on the other hand. It is worth noting that if a model based on total duration alone performed equally as well as our multi-feature-based models, one would conclude that timing the TUG (or cogTUG) task with a stopwatch is just as good as using a state-of-the-art wearable sensor. Our results show that is not, in fact, the case. Rather, our models clearly show the added benefit of using a highly granular mobility analysis provided by a wearable accelerometer/gyroscope.

We observed that increasing the diversity of sensor-derived features improves diagnostic accuracy. Our use of forward feature selection greatly reduced the number of features included in the model. Consequently, it is not the number of features per se but the nature of the features that determines the accuracy. Including multiple tasks gives us some additional informative features, as does segmenting tasks into subtasks. On the whole, the latter appears to be the best approach, as it requires reduced time and effort for both patients and clinicians, as the additional effort is handled algorithmically by a computer.

One of the strengths of our study is the large sample size and the wide spectrum of PD severity of our PD participants. We included 262 participants with PD and 50 control partici-

pants, compared to an n of 13 to 20 in many other previously published studies [34, 111, 232]. Our PD participants differed widely in terms of disease severity, ranging from mild (H&Y $\leq$ 2) to moderate (H&Y 2.5 and 3) and severe (H&Y 4). Therefore, participants in this study represent the full spectrum of ambulatory PD patients encountered in clinical practice, compared to previous studies that focused on a more limited range of disease severity [34, 111, 232]. Another strength is the simplicity of instrumented mobility testing using a single wearable sensor. This is in contrast to the multiple sensors employed in many other studies [35, 43, 110, 111] and facilitates clinical use and scalability by streamlining data collection and analysis.

Although our proposed set of features might not be easily interpretable by clinicians, it shows potential for distinguishing between PD and control participants. Given our primary aim of achieving the highest possible diagnostic accuracy and real-world applicability, it is essential to include features that contribute to this goal, regardless of their interpretability. Studies that emphasize interpretability might consider using features that are more clinically explainable and may apply prior correlation analysis, as feature importance measures can show bias towards correlated predictor variables [233, 234].

In addition to typical motor symptoms, PD is characterized by non-motor features including cognitive impairment and autonomic dysfunction (e.g., dizziness due to a drop in blood pressure in an upright position). Such non-motor features were not assessed in control participants in our study and, therefore, could not be included in our machine learning models, which is one of the limitations of our study. Incorporating such non-motor manifestations of PD might lead to even more accurate models to predict (or identify) PD participants and will be part of our future work. Additionally, while the H&Y scale is commonly used to assess overall PD severity, it is limited by its inter-rater variability due to its reliance on clinical judgment. This subjectivity

introduces potential inconsistencies in the partitioning of PD participants, raising the question of how appropriate the H&Y scale is as frame of reference for objective assessment of PD severity. Information regarding the levodopa equivalent daily dose (LEDD) of each PD participant is not available in the PD clinical database (HOME study) we used for this study. However, ON/OFF medication status, which was available, was not found to be an important feature in any of our machine learning analyses, which is evidence that LEDD and ON/OFF status are not consequential in discrimination of PD patients and controls.

Objective and granular measures of disease manifestation as provided by a wearable sensor are promising candidates as superior markers of disease severity and disease progression. However, longitudinal studies are necessary to test this hypothesis. The cross-sectional nature of the data presented here is another limitation of our study. Longitudinal analysis of a subgroup of our study cohort is currently ongoing. Another aspect worth exploring in more detail in future studies is the potential impact of confounding factors on classification performance. Our analysis of misclassified participants shows that certain variables that are likely to be associated with the level of medical comorbidity, such as age, could influence the sensor data and classification outcomes. Older controls were falsely classified as PD, and PD participants with relatively minor motor symptoms (low UPDRS_PIII motor score) were falsely classified as controls. The differentiation between only mildly affected PD patients and older (non-PD) adults is also among the biggest challenges in clinical routine, which is why, for clinicians, the use of a wearable sensor as a supportive biomarker is relevant and attractive in such situations. That was also the rationale for us focusing on models for mild PD versus controls, in addition to all PD versus controls, in this project. Future studies with an even larger sample size for those particular groups of participants may facilitate the development of more comprehensive models that incorporate movement

characteristics to overcome these limitations of our models and to further enhance the accuracy and reliability of sensor-based PD classification/identification.

3.5 Conclusions

In conclusion, our study demonstrates that diagnostic accuracy for PD exceeding that typically achieved by movement disorders experts in a routine clinical setting is possible with few or even just one complex mobility task instrumented with a single wearable sensor when appropriately analyzed using machine learning. Our results provide a framework for such simplification and the ultimate harmonization of existing approaches to facilitate the use of wearable sensors and machine learning-based PD diagnosis in clinical settings.

3.6 Funding

Funding was provided by a University of Maryland MPower Seed Grant Award (R.v.C. and M.P.C.), the Rosalyn Newman Foundation (L.M.S.), and the University of Maryland Claude D. Pepper Older Americans Independence Center (P30-AG028747; R.v.C.).

3.7 Institutional Review

All study procedures were approved by the University of Maryland, Baltimore (UMB), Institutional Review Board (IRB) on 10 September 2015 (protocol code: HP-00065245).

3.8 Informed Consent

Informed consent was obtained from all subjects involved in the study before any procedures were conducted.

3.9 Data Availability

De-identified demographic, clinical, and sensor-derived data have been deposited in the Digital Repository at the University of Maryland (DRUM), `https://doi.org/10.13016/dspace/ddzp-im5e`. The code used in this research paper is also accessible at the same link. The provided notebook contains the code used to implement each component of the signal processing and machine learning pipelines. Detailed instructions on setting up the necessary environment and running the code are provided in the README file.

3.10 Acknowledgments

We thank Aron S. Buchman and Robert J. Dawe for support in the implementation of the mobility testing protocol; Katrina Schrader, Kimberly Bambarger, Veronica Fallon, and Philip Heiser for technical assistance with mobility testing; Vivien Xi and Patrick Curry for assistance with data extraction and processing; and Christina Griffin for regulatory administrative support.

Chapter 4: Machine Learning and Statistical Analyses of Sensor Data Reveal Variability Between Repeated Trials in Parkinson's Disease Mobility Assessments

The work of this chapter has been published in the journal Sensors as: Rana M. Khalil, Lisa M. Shulman, Ann L. Gruber-Baldini, Sunita Shakya, Jeffrey M. Hausdorff, Rainer von Coelln, and Michael P. Cummings. Machine learning and statistical analyses of sensor data reveal variability between repeated trials in Parkinson's disease mobility assessments. *Sensors*, 24(24):8096, 2024 [29].

Abstract

Mobility tasks like the Timed Up and Go test (TUG), cognitive TUG (cogTUG), and walking with turns provide insights into the impact of Parkinson's disease (PD) on motor control, balance, and cognitive function. We assess the test–retest reliability of these tasks in 262 PD participants and 50 controls by evaluating machine learning models based on wearable-sensor-derived measures and statistical metrics. This evaluation examines total duration, subtask duration, and other quantitative measures across two trials. We show that the diagnostic accuracy for distinguishing PD from controls decreases by a mean of 1.8% between the first and the second trial, suggesting that task repetition may not be necessary for accurate diagnosis. Although the

total duration remains relatively consistent between trials (intraclass correlation coefficient (ICC) = 0.62 to 0.95), greater variability is seen in subtask duration and sensor-derived measures, reflected in machine learning performance and statistical differences. Our findings also show that this variability differs not only between controls and PD participants but also among groups with varying levels of PD severity, indicating the need to consider population characteristics. Relying solely on total task duration and conventional statistical metrics to gauge the reliability of mobility tasks may fail to reveal nuanced variations in movement.

4.1 Introduction

Parkinson's disease (PD), a progressive neurodegenerative disorder, poses significant challenges to individuals, affecting their mobility and daily function [235–238]. Mobility tasks, such as the Timed Up and Go test (TUG), cognitive TUG (cogTUG), and walking with turning (e.g., 32-foot walk with 3 turns), offer valuable insights into the impact of PD on dynamic motor control, balance, and cognitive functions and serve as indicators of an individual's ability to navigate daily life [239–241].

Progress in the field of wearable sensors has enabled the collection of quantitative data from accelerometer and gyroscope recordings. Beyond simply measuring the time taken to perform an entire mobility task, these data can be leveraged to calculate the duration of each subtask of the test, along with numerous measures that provide a detailed characterization of movement. Evidence that this type of granular movement analysis of mobility tasks provides added value over simply timing the overall duration with a stopwatch has been accumulating in recent years. For example, a study highlighted significant differences in TUG-derived acceleration measures

between PD participants and age-matched controls, but most measures did not show a correlation with the total duration of TUG [116]. When comparing older adults experiencing falls of unknown origin with healthy controls, the same group of researchers showed significant differences in TUG-derived acceleration measures, but not in overall TUG duration based on a stopwatch [209]. These results suggest that granular sensor-derived TUG measures provide clinically relevant data beyond those captured by overall duration. Similarly, five different sensor-derived TUG subtask measures were associated with clinical parkinsonian gait features in a large community-based cohort study of aging [115], which established the protocol of instrumented mobility tests that we replicated in our study presented here. In a separate study, the use of a single short-duration test and a smartphone-embedded inertial measurement unit showed that individuals with mild to moderate PD exhibit impaired postural control, an altered gait strategy, and slower turn-to-sit performance compared to healthy individuals of the same age [242].

Evaluating the reliability of mobility tasks is essential to ensuring robust and consistent measurements in clinical and research settings [243–246]. In essence, reliability measurements play a crucial role in assessing the degree to which clinical test scores are affected by measurement errors [247]. A high level of reliability suggests consistent results on repeated assessments, often quantified by the intraclass correlation coefficient (ICC) [248]. A higher ICC value indicates better measurement consistency.

Although previous studies have examined the reliability of certain motor tasks, they often had limitations such as small sample sizes [132, 243, 244, 249–253], a focus on a single population [244, 249–251, 253–256], or testing only the reliability of the total time taken to complete the task [243, 255, 256]. Furthermore, all reviewed studies relied solely on statistical measures for the reliability assessment [132, 243, 244, 249–257]. However, machine learning has increasingly been

applied to the diagnosis of PD using sensor data collected during movement tasks. Thus, evaluating the predictive performance of different trials using machine learning provides a purpose-specific (i.e., diagnostic) evaluation that transcends standard statistical evaluation of test–retest reliability. ML is known for its ability to capture complex patterns and relationships that may be challenging to detect using standard observational and statistical methods [12]. Whereas some research has shown differences in time and sensor-derived measures across disease-defined subpopulations [132, 243, 252, 257], they did not investigate how test–retest reliability varies among these disease severity groups.

Here, we present the results of a comprehensive analysis of test–retest reliability for TUG, cogTUG, and walking with turns, employing two trials for each of these three tasks in PD participants and controls. We compare the performance of machine learning models constructed with quantitative sensor-derived features obtained from subtasks and selected independently from each trial. In addition, we compare the total duration, duration of subtasks, and various quantitative measures across both trials and different groups of participants using statistical measures. Our goal is to address the limitations of previous studies by including a large cohort of 262 participants with PD and 50 controls. We aim to go beyond total task duration, investigating the reliability of mobility tasks using more movement-sensitive measures associated with PD. By incorporating both machine learning techniques and statistical methods to assess reliability, we also examine how reliability differs between PD and control groups, as well as within PD subgroups stratified by disease severity.

The study results show that performance differs when features from each trial are analyzed independently by ML models, suggesting non-equivalence of the two trials. Comparison of the accuracy of models based on the first and second trials of each task revealed that repetition of mo-

bility tasks may not be helpful for the diagnostic accuracy, and thus mobility test protocols may be simplified. Statistical analysis also indicates that the duration of the subtasks and the quantitative measures are less consistent between the two trials compared to the total time, which remains more stable. This implies that relying on total duration as the singular metric for assessing reliability may not be sufficient, as it obscures (or rather fails to reveal) differences in movement characteristics detectable with sensor-derived measures. Furthermore, our study demonstrates that the variability between the two trials differs not only between controls and individuals with PD, but also between groups with increasing severity of PD. These findings prompt a reconsideration of how we conceptualize and define a reliable mobility task, emphasizing metrics that better capture changes in movement across different trials.

4.2 Materials and Methods

4.2.1 Participants

The cohort in our study included 50 control participants and 318 individuals diagnosed with parkinsonism. All participants were enrolled in a study on mobility analysis in individuals with parkinsonism at the University of Maryland Movement and Memory Disorders Center (UM-MMDC) during the period from October 2015 to March 2020. All parkinsonism participants were also part of the Health Outcomes Measurement (HOME) study at UM-MMDC, a naturalistic cohort study collecting patient and clinician-reported data on motor symptoms, activities of daily living, and quality of life during routine office visits. Exclusion criteria for the mobility analysis included an inability to stand or walk without assistance (Hoehn and Yahr, H&Y = 5), other conditions unrelated to parkinsonism significantly impacting gait and balance,

or failure to provide informed consent during the evaluation.

Of the 318 parkinsonism participants, 293 were diagnosed with idiopathic PD according to the UK Parkinson's Disease Society Brain Bank Clinical Diagnostic Criteria [188] by a movement disorder specialist at UM-MMDC. The remaining individuals were diagnosed with other forms of parkinsonism and excluded from the data analysis for this study. One participant with PD was unable to complete the study procedures and was subsequently excluded, and thirty others were excluded due to missing data. In this study, 87% of the participating PD patients reported being in a medication ON state. Patients who had previously undergone deep brain stimulation (DBS) surgery were evaluated while in the ON stimulation state. All participants, including controls, provided their informed consent before any study procedures were performed. The study protocols for both the HOME study and the mobility analysis in the parkinsonism study received approval from the University of Maryland Institutional Review Board.

4.2.2 Mobility Assessments

We followed a previously established protocol for instrumented mobility testing [115], commonly used in clinical practice due to the successful application of mobility data in the research of parkinsonism. Participants performed three mobility tasks from that protocol: 1. 32-foot walk, where participants walked a distance of 8 feet (2.4 m), making two round trips without pauses, which involved a total of three turns; 2. two trials of Timed Up and Go (TUG), where participants started by rising from a seated position in a chair, proceeded to walk eight feet to a designated line on the floor, executed a turn, returned to the chair, made a second turn, and finally seated themselves again; and 3. two trials of cognitive Timed Up and Go (cogTUG),

which required participants to perform the TUG while concurrently counting backward from 100 in steps of 3. Participants performed the counting backwards in steps of 3 once in a sitting position as a test run immediately prior to the first cogTUG trial to confirm that they understood the cognitive task and were able to perform it. The second TUG and cogTUG trials for each participant took place in less than one minute after their completion of the first trial. The protocol employed in the second trial was identical to the procedures of the initial trial.

4.2.3 Device and Data Acquisition

Data were acquired using a small (106.6 $\times$ 58 $\times$ 11.5 mm) lightweight (55 g) sensor device (Dynaport MT, McRoberts B.V., The Netherlands, technical specifications: `https://www.mcroberts.nl/products/movetest/`, accessed on 17 June 2024) placed on the participant's lower back using a neoprene waist belt. This device is equipped with a triaxial accelerometer (range: $\pm$8 g, resolution: 1 mg) and a triaxial gyroscope (range: $\pm$2000 dps, resolution: 70 dps), capturing activity data at a sampling frequency of 100 Hz. It records the data in three acceleration axes, including the vertical, mediolateral, and anteroposterior directions, and in three angular velocity axes: yaw (rotation around the vertical axis), pitch (rotation around the mediolateral axis), and roll (rotation around the anteroposterior axis). The sensor recordings were started and stopped remotely via wireless Bluetooth signals from a laptop using a handheld device. Digital markers were strategically placed to mark the initiation and completion of each mobility task, facilitating post hoc segmentation. Sensor readings were then transferred to an encrypted notebook computer and the raw data files were moved to a secure data server.

4.2.4 Data Retrieval

We used a custom Matlab (version 9.1) Runtime-based graphical user interface (GUI) to visualize the complete set of accelerometry data. The identification of the 32-foot walk, TUG, and cogTUG tasks was based on the characteristic accelerometry patterns associated with each task, coupled with digital markers recorded during mobility testing. The raw accelerometry data segmented for each specific mobility task were individually saved as a distinct file for subsequent data processing.

4.2.5 Segmentation

For the TUG and cogTUG tasks, we performed segmentation into subtasks and the extraction of mobility measures as previously described [116, 117, 209]. In brief, specific patterns and peaks in the accelerometry graphs were identified to delineate the stand-to-sit, walk 1, turn, walk 2, and stand-to-sit subtasks for each TUG and cogTUG mobility task, using a second custom-made Matlab (version 9.1) Runtime-based GUI.

The 32-foot walk mobility task was divided into 2 $\times$ 16-foot trials by omitting the second turn. The first trial included the first walk, the first turn, and the second walk, while the second trial consisted of the third walk, the third turn, and the fourth walk. Hence, we refer to each trial as a 16-foot walk. Each trial in the 16-foot walk task was further divided into two walking segments and one turn. To identify the turning segment, we used a previously established algorithm [201], which involved calculating the angular position from the angular velocity along the vertical axis using the trapezoidal integration method [202]. The sections of the signals before and after the turn marked the two walking parts.

4.2.6 Quantitative Measures

For each mobility task, we assessed the total duration calculated from sensor readings as one measure. Additional quantitative measures derived from TUG and cogTUG comprise those derived from walking subtasks, such as total walking duration, step count, step duration, and step and stride regularity [117], and frequency domain measures calculated from the power spectral density (PSD) of the signals, including the amplitude, width, and slope of the PSD peak [258]. For the sit-to-stand and stand-to-sit components, we computed measures including duration, range, jerk, standard deviation, and median measures [209]. The turning measures involved the calculation of the duration, step count, amplitude [117], and dominant frequency [211]. In total, 77 measures were calculated for each TUG and cogTUG trial, providing a comprehensive assessment of various aspects of performance.

For the 16-foot walking task, 36 measures were derived. The measures of the walking segments included walk duration, step count, step duration, stride duration, step regularity, stride regularity, and step symmetry. We used the number of peaks detected in the vertical and anteriorposterior acceleration signals independently to represent the step count, with the step duration and stride duration calculated as the mean time between individual steps and strides, respectively. Step regularity was determined by the amplitude of the first peak in the acceleration autocorrelation signal, and stride regularity corresponded to the amplitude of the second peak in the same signal, as defined in a prior study [259]. Step symmetry was defined as the degree of proximity between the ratio of step regularity to stride regularity and the value 1.0 [259]. For the turning phase of the task, we calculated the time taken to complete the turn, the dominant frequency, and the step count. The dominant frequency was defined as the frequency associated with the highest

amplitude in the discrete Fourier transform (DFT), which was calculated using the fast Fourier transform algorithm (FFT) [260]. All quantitative measures were computed for both the vertical and anteroposterior directions.

A complete list of all measures along with their descriptions is provided in the Supplementary Table B.1.

4.2.7 Feature Selection

To simplify our model while maintaining or improving its performance, we implemented a forward feature selection approach. Our objective was to identify an optimal subset of features that would lead to a more streamlined model. The process began with feature ranking, achieved through a random forest [216] model. The pivotal point in the feature importance plot, known as the elbow point, was determined as the point farthest from a straight line connecting the first and last points on the plot. This elbow point served as the cutoff threshold for selecting a set of candidate features. Multiple random forest models were trained using subsets of the cumulative top-ranked candidate features. The best-performing model was chosen based on its ability to minimize the total Akaike information criterion (AIC) [210], and the corresponding set of features was subsequently used in further analyzes.

4.2.8 Machine Learning Model

We implemented a five-fold cross-validation framework with stratified sampling to generate balanced training and testing splits. This framework was repeated five times, each time using different random seed to produce varied train and test sets, thus improving robustness. Within

each fold, the feature selection method detailed in Section 4.2.7 was applied exclusively to the training data. Selected features from the training set were then applied to the corresponding test set to prevent data leakage and to ensure generalization [213].

Using these selected features, we trained a random forest model with 10K trees on an undersampled version of the training set, matching the minority class size. Each trained model was then used to make predictions on the corresponding test set. This process generated five predictions per participant, one from each cross-validation replicate. The final class for each participant was determined by a majority vote across these independent predictions.

4.2.9 Statistical Analysis

Data analysis was performed with R version 4.2.3 (15 March 2023) and RStudio. Continuous measures were presented as mean $\pm$ standard deviation (SD). To assess test–retest reliability, we used the intraclass correlation coefficient (ICC) and computed 95% confidence intervals (CIs). The analysis used a two-way mixed effects model with absolute agreement using the following formula [261].

$$ICC = \frac{MS_P - MS_E}{MS_P + (k-1)MS_E + \frac{k}{n}(MS_T - MS_E)}$$

where MS_P = mean square for participants; MS_E = mean square error; MS_T = mean square for trials; n = number of participants; and k = number of trials. We interpreted ICC values as follows: values less than 0.5 indicate poor reliability, those between 0.5 and 0.75 denote moderate reliability, values between 0.75 and 0.90 indicate good reliability, and values greater than 0.90 indicate excellent reliability [248]. Additionally, we calculated the standard error of measurement

(SEM) using the formula $SEM = SD \times \sqrt{1 - ICC}$. SEM quantifies the measurement error in the same units as the measurement, with a lower SEM signifying higher reliability [262]. We then calculated the minimal detectable change (MDC) based on the SEM as follows: $MDC = SEM \times 1.96 \times \sqrt{2}$. MDC represents the smallest change required to confidently distinguish the real change between two trials from the changes attributed to the measurement error [263]. Moreover, the level of discrepancy between the two trials was visualized using Bland–Altman plots [264]. These plots feature 95% limits of agreement (LoA), defined as the mean difference $\pm$ 1.96 $\times$ SD. In these plots, the differences between the two trials are plotted against the mean of the trials. Bland–Altman calculations were performed using the blandr R package.

4.3 Results

Wearable sensor data were collected from 262 participants diagnosed with PD and 50 age-matched controls. Within the PD group, participants were categorized based on their H&Y stage into groups with mild (H&Y $\leq$ 2), moderate (H&Y = 2.5 and 3), and severe (H&Y = 4) PD. A summary of the characteristics of the study cohort is presented in Table 4.1. Participants completed the mobility tasks described in Section 4.2.2, with data collection following the procedures outlined in Sections 4.2.3 and 4.2.4. These tasks were then segmented as detailed in Section 4.2.5, producing the quantitative measures described in Section 4.2.6. The variability in the duration of the task and the subtasks, along with quantitative measures, was assessed using statistical (Section 4.2.9) and machine learning (Section 4.2.8) approaches.

Table 4.1: Characteristics of the study cohort.

Feature	**Controls (*n* = 50)**	**PD (*n* = 262)**
Age (years, $\bar{x} \pm SD$)	64.1 $\pm$ 9.8	66.9 $\pm$ 9.3
Gender (%male)	38.0	62.0
Height (cm, $\bar{x} \pm SD$)	168.1 $\pm$ 10.9	172.2$\pm$ 10.4
UPDRS (total, $\bar{x} \pm SD$)	–	35.5 $\pm$ 17.1
UPDRS (motor-part III, $\bar{x} \pm SD$)	–	22.2 $\pm$ 11.8
Disease duration (years, $\bar{x} \pm SD$)	–	7.8 $\pm$ 6.5
H&Y ($\bar{x} \pm SD$)	–	2.2 $\pm$ 0.62
Stage 1 (n)	–	12
Stage 1.5 (n)	–	4
Stage 2 (n)	–	169
Stage 2.5 (n)	–	35
Stage 3 (n)	–	25
Stage 4 (n)	–	17

4.3.1 Reliability of Task/Subtask Duration

We used sensor readings to measure the time it takes participants to complete the TUG, cogTUG, and 16-foot walk tasks. The ICCs for task duration for each mobility task and participant group are provided in Supplementary Table B.2. The reliability of TUG and cogTUG was good to excellent in all groups (ICC = 0.75 to 0.95), and the 16-foot walk task demonstrated moderate to good reliability (ICC = 0.62 to 0.77) across all groups. Participants in the severe PD category showed the highest ICC scores across different tasks (Figure 4.1a,c,e, top row ("total")), even though, in general, gait variability increases with progression of PD.

To further characterize the difference between the first and second trials in our mobility tasks, we computed the relative change (RC) in time between the two trials for each task and participant, as another reliability measure (RC = (duration of the second trial $-$ duration of the first trial)/duration of the first trial). Ideally, for a task to be perfectly reliable, the RC of time

should be zero. Analysis of the median RC for all participants in each group revealed negative RC values for all groups for TUG and cogTUG, indicating that the duration of the second trial was shorter than the first (see Figure 4.1b,d, top row). This observation suggests a learning effect in both the TUG and cogTUG tasks for all groups. However, the 16-foot walk showed positive RC values for all groups (Figure 4.1f, top row), indicating that the second trial took longer than the first, perhaps due to fatigue, as both trials were performed immediately back-to-back (as part of the 32-foot walk mobility task; see Section 4.2.2).

The fact that the severe PD group exhibited the greatest (negative) RC, despite having the highest ICC values, in both TUG and cogTUG, appears contradictory at first glance. However, mathematically, ICC depends both on differences between participants and between trials (see Section 4.2.9). Consequently, there is a positive correlation between variance and ICC, with higher variance leading to higher values of ICC. The severe PD group displayed the highest variance among participants by an order of magnitude compared to other groups (see Supplementary Table B.3). Thus, even though there is a substantial change in performance between the first and second trials (indicated by a high RC value), the ICC remains high because the numerator of the equation (see Section 4.2.9) is large. This analysis suggests that, under certain circumstances, depending only on the ICC to measure test–retest reliability can be misleading, especially when testing patients in more advanced stages of PD. It highlights the need to complement the ICC with the RC measure.

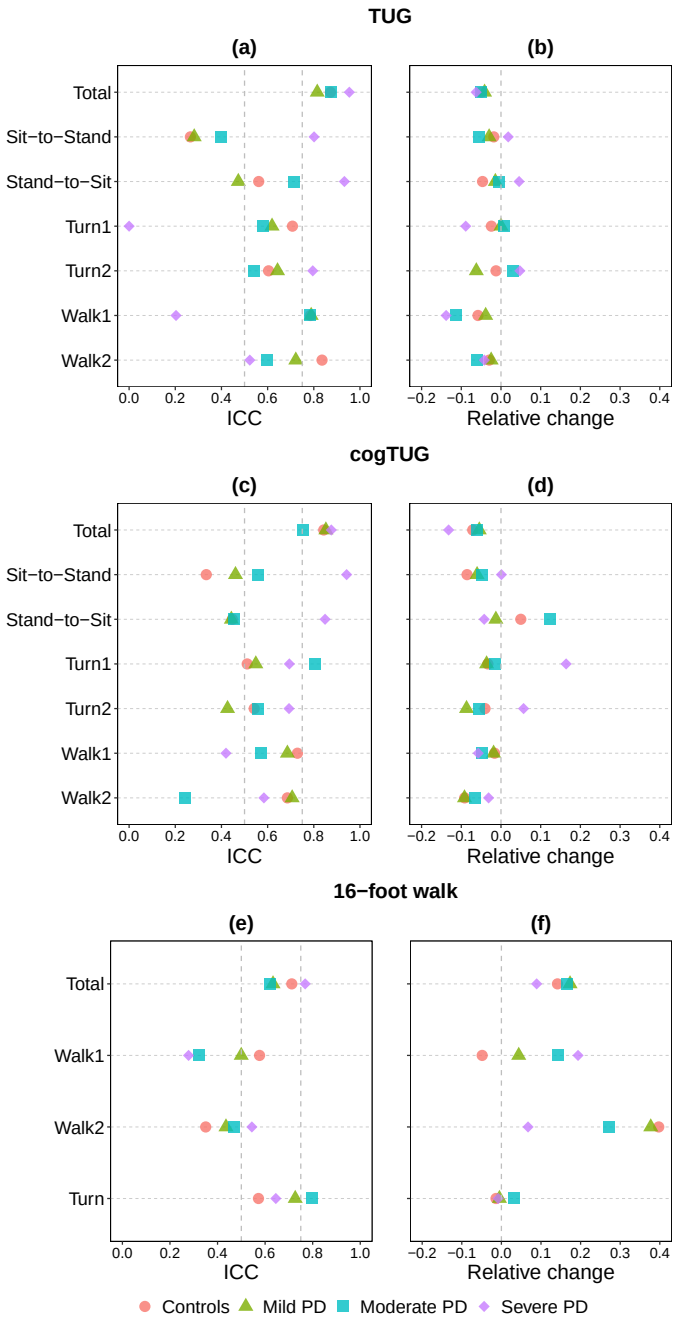


Figure 4.1: Intraclass correlation coefficient (ICC) (left) and median of the relative change (right) in total and subtask duration for TUG (**a**,**b**), cogTUG (**c**,**d**), and 16-foot walk (**e**,**f**). Vertical dashed lines represent thresholds for moderate (ICC = 0.5) and good (ICC = 0.75) reliability.

In addition to the total duration, we investigated the test–retest reliability of each subtask. Figure 4.1a,c,e illustrate the difference in ICC values across various subtasks and different groups of participants. Specifically, in the TUG task, all subtask durations exhibited lower ICC values

compared to the total duration across all participant groups. Similarly, in the cogTUG task, this trend persisted, except for the severe PD group, where the sit-to-stand subtask displayed a higher ICC value than the total duration. Also, the moderate PD group showed the first turn with the highest ICC score. Furthermore, in the 16-foot walk, the turn subtask exhibited a higher ICC value than the total duration for the mild and moderate PD groups.

Comparing ICC values across different subtasks, we observed that for the TUG task, the walking subtasks had the highest ICC values, followed by the turning subtasks and then the chair transition subtasks (sit-to-stand and stand-to-sit) for both the control and mild PD groups. Walking subtasks also had higher ICC values than turning for the moderate group. However, for severe PD stages, the chair transition subtasks had the highest ICC values. In the cogTUG task, the walking subtasks had the highest ICC values, followed by the turning subtasks, and then the chair transition subtasks for controls, similar to the TUG results. Walking subtasks again had the highest ICC values for the mild PD group and no specific ordering was observed for the moderate PD group. For severe PD stages, the chair transition subtasks had the highest ICC values, similar to the TUG. In the 16-foot walk task, the turning subtasks had higher ICC values than the walking subtasks for all groups of participants.

Figure 4.1b,d,f shows that possible learning and fatigue effects are not universally associated with all subtasks. For instance, for cogTUG, the severe PD group exhibited a possible learning effect for the overall duration of the task (negative RC for total duration), but both turns had positive RC, indicating that they were slower in the second trial (Figure 4.1d). Similarly, for the 16-foot walk, controls showed a possible fatigue effect overall (positive RC for total duration), but the first walk component was actually faster in the second trial (negative RC for walk 1; Figure 4.1f). The absolute values of the RC for the duration of some subtasks were also higher

than the absolute RC of the total duration, indicating less consistency between the two trials for those subtasks. On the other hand, the duration of the first turn of TUG, cogTUG, and the 16-foot walk turn showed small RC compared to other subtasks, except for severe PD. In summary, there is a significant degree of variability in test–retest reliability for different subtasks that is not reflected in the high ICC and low RC for the total duration of the task alone.

4.3.2 Reliability of Other Quantitative Measures

Before assessing the test–retest reliability of the sensor-derived quantitative measures, we objectively reduced the number of features, focusing solely on the most important ones. For each task, we constructed three random forest models using all features computed from both trials and aimed to distinguish participants with mild, moderate, and severe PD from the controls. We identified the important features of each model using the feature selection approach described in Section 4.2.7. The combination of the reduced measure sets from these three models yielded 11 measures for each of the TUG, cogTUG, and 16-foot walk tasks. The consistency of these final sets of selected features between both trials was then evaluated. The mean and standard deviation values for each measurement (feature) are presented in Supplementary Table B.4. Supplementary Table B.5 includes the ICC, SEM, and MDC values for each measure. Of the 33 quantitative measures, only 1 variable derived from the cogTUG task demonstrated good reliability, with an ICC > 0.77 across all participant groups.

Taking the median, which is a robust statistic, of the ICC values across all measures, the selected quantitative measures exhibited lower scores compared to the total duration of each of the three mobility tasks (Figure 4.2a,c,e). Although the ICC values for the total duration of both

TUG and cogTUG demonstrated good to excellent reliability for all groups of participants, the median ICC values for other measures ranged from poor to moderate reliability (ICC = 0.36 to 0.61). Similarly, although the total duration of the 16-foot walk displayed moderate to good reliability, the median ICC values for other quantitative measures exhibited poor reliability (ICC = 0.34 to 0.39). ICC scores for duration and other measures were also different among the groups of participants. The median ICC values, along with their respective confidence intervals, across the quantitative measures for each task and group of participants are presented in Supplementary Table B.6.

We also determined the median of the median values of RC for the quantitative measures by calculating the median of the absolute value of RC across all participants in each group and then determining the overall median from all measures. The RC values for the quantitative measures were higher (indicating less consistency) than the RC values for the total duration of the three tasks (Figure 4.2b,d,f). The greater change in other mobility measures between trials compared to total duration is consistent with our observation of a much greater variability of subtask duration than the total duration between trials (demonstrated in Figure 4.1). Among the participant groups, the controls displayed the lowest RC value, and the severe PD group exhibited the highest value (i.e., the greatest amount of change between the first and second trials) in all tasks.

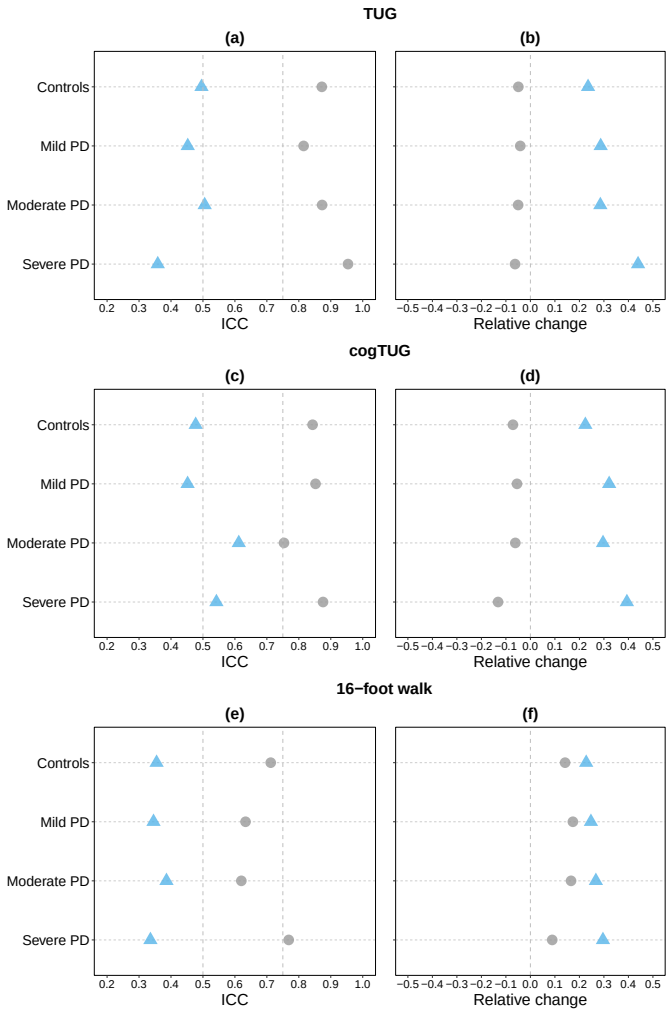


Figure 4.2: Left: Intraclass correlation coefficient (ICC) of total duration and other quantitative measures for TUG (**a**), cogTUG (**c**), and 16-foot walk (**e**). Gray points represent the ICC of the total duration and blue triangles represent the median of the ICC values across the other quantitative measures. Vertical dashed lines represent thresholds for moderate (ICC = 0.5) and good (ICC = 0.75) reliability. Median ICC values, along with their respective confidence intervals, across the quantitative measures for each task and participant group are presented in Supplementary Table B.6. Right: Relative change (RC) in total duration and other quantitative measures for TUG (**b**), cogTUG (**d**), and 16-foot walk (**f**). Gray points represent the median of the RC of the total duration across participants of each group. Blue triangles represent the median of the medians of the absolute RC values across participants and quantitative measures.

For a more granular analysis, for each mobility task, individual measure-level ICC scores and the median of absolute values of RC across all participants in each group were calculated, and these values are visualized in Supplementary Figures B.1–B.3. There are substantial variations

in reliability (ICC and RC scores) among different groups of participants for many measures. A few measures, such as the TUG amplitude yaw of the second turn (Supplementary Figure B.1, first measure) and range of sit-to-stand (Supplementary Figure B.1, fifth measure), as well as amplitude-yaw of first turn from the cogTUG task (Supplementary Figure B.2, tenth measure), showed small differences between participant groups in both ICC and RC values. Only the amplitude yaw of the first turn of the cogTUG task demonstrated good reliability (ICC $>$ 0.77) across all participant groups (Supplementary Figure B.2, tenth measure). The reliability of other measures varied from poor to excellent across the different groups for the three mobility tasks.

To further elucidate the differences between the two trials, Bland–Altman plots were employed, illustrating the relationship between the differences and average values obtained from a measure in both trials. This graphical approach quantifies the agreement and allows for the identification of any potential systematic bias. We plotted the discrepancies in quantitative measures with 95% confidence intervals (CIs) and 95% limits of agreement (LoA) (see Supplementary Figures B.4–B.6). Points outside the CI of the upper and lower LoA for certain measures indicate divergent values between the two trials.

4.3.3 Performance of Machine Learning Models

For each mobility task, trial, and group of PD participants, we applied the machine learning modeling approach described in Section 4.2.8 to distinguish the corresponding group of PD participants from controls (i.e., each analysis is a binary classification problem). The performance of these models is detailed in Table 4.2. Across the three mobility tasks, there were variations in the accuracy of models built using features calculated from each trial separately for classifying

different PD groups and controls. As expected, the accuracy in differentiating PD from control participants was much higher for severe PD compared to mild or moderate disease. When comparing the accuracy of the models based on the first vs. second trial of each task, we observed either a similar or a reduced accuracy from the first to the second trial, with two exceptions (16-foot walk for mild PD and TUG for severe PD). Models using features of the second trial alone showed a decrease in accuracy by a mean of 1.8% compared to models using features from the first trial. Similarly, models incorporating features from both trials showed similar or lower accuracy compared to those using only first-trial features, except in three cases (16-foot walk for mild and severe PD, and TUG for severe PD). These models showed only a slight improvement, with a mean increase of 0.52% over models using first-trial features. This suggests that, for the purpose of differentiating PD from controls, repeating mobility tasks in this study generally provides little to no additional benefit.

To demonstrate how sensor-based measures provide valuable information beyond what task duration reveals, we constructed logistic regression models using only the mean total duration from both trials as a predictor. These models showed a decrease of 7.0% in mean accuracy compared to models that used features from the first trial alone (Table 4.2). Based on these findings, we conclude that sensor-derived measures provide additional value beyond timing the overall task duration.

Table 4.2: Accuracy (percent) of the models. Trial 1 models used features derived only from the first trial; Trial 2 models used features derived only from the second trial; Trials 1&2 models used features derived from the first trial, the second trial, and the mean of corresponding features from both trials; and duration models used the mean durations from both trials. Numbers in parentheses represent the difference in accuracy between each model and the corresponding model that used only features from the first trial.

		Trial 1	**Trial 2**	**Trials 1&2**	**Duration**
Mild PD vs. controls					
	TUG	75.3	73.6 (−1.7)	74.5 (−0.8)	59.3 (−16.0)
	cogTUG	81.7	76.6 (−5.1)	81.7 (0.0)	57.6 (−24.1)
	16-foot walk	66.0	66.8 (+0.8)	69.8 (+3.8)	62.3 (−3.7)
Moderate PD vs. controls					
	TUG	80.9	80.9 (0.0)	80.0 (−0.9)	80.0 (−0.9)
	cogTUG	82.7	79.1 (−3.6)	82.7 (0.0)	80.9 (−1.8)
	16-foot walk	74.5	70.9 (−3.6)	73.6 (−0.9)	71.5 (−3.0)
Severe PD vs. controls					
	TUG	91.0	94.0 (+3.0)	94.0 (+3.0)	85.2 (−5.8)
	cogTUG	95.5	89.6 (−5.9)	89.6 (−5.9)	88.6 (−6.9)
	16-foot walk	86.6	86.6 (0.0)	91.0 (+4.4)	86.2 (−0.4)

4.4 Discussion

Our study presents an evaluation of the test–retest reliability of three commonly employed physical performance tasks for the assessment of PD mobility. Through machine learning and statistical analyses, we demonstrate that, although total duration remains consistent across various trials (ICC = 0.62 to 0.95), there is discernible variation in other sensor-derived measures. These results imply that relying solely on total duration, as commonly practiced, does not adequately capture the true changes in movement between trials. Additionally, we demonstrate that the reliability of sensor-derived metrics varies between different populations and even among subgroups within the same population. Therefore, careful attention should be paid to the population being studied.

We assessed the reliability of mobility tasks by employing machine learning models that go beyond traditional reliance on standard statistical measures [132,243,244,249–257]. Specifically,

we constructed random forest models to distinguish different groups of PD participants from controls using features calculated independently from each trial, as well as features combined from both trials. Across all tasks, we observed a performance difference between models built using the first trial and those built using the second trial (see Table 4.2). The pattern of similar or reduced accuracy from the first to the second trial, or when using combined trials, for most cases indicates that, for the purpose of differentiating PD from controls, repeating mobility tasks offers little to no added benefit. In other words, our results suggest that mobility testing protocols can be simplified to include a single trial of TUG, cogTUG, and walking with turning, if differentiating PD from controls is the purpose of mobility testing. This simplification will save time and effort for both participants and clinicians, thereby improving the feasibility of using wearable sensors as a diagnostic tool in clinical settings. Whereas our previous work demonstrated that reducing the number of sensors and tasks can simplify mobility assessment [28], here we recommend minimizing the number of trials per task to one single trial.

On the other hand, the performance variation of the ML models between trials underscores that the sensor-derived measures of each trial may possess distinct values. These differences between task trials affect the ability of the models to differentiate between PD groups and controls. Additionally, we demonstrate that models built using only total duration performed worse in all cases compared to models that used sensor-derived measures. This suggests that sensor-derived measures offer enhanced informativeness by providing a finer resolution of movement characteristics compared to total duration alone [209].

With respect to test–retest reliability, our procedure, in which participants completed two trials of each task back to back during the same clinical visit, represents a best-case scenario. This design limits potential variability due to factors such as changes in disease progression or

symptoms, day-to-day differences in physical status that can affect movement, and differences in sensor positioning. Thus, this design ensures that participants are in the same disease state for both trials and all other specific characteristics are identical. Despite controlling for these variables, we still observed trial-to-trial variability. We anticipate that altering any of these factors could further increase variability.

Among previous studies that used the ICC of the total task duration to assess the reliability of various mobility tasks, a study evaluating the reliability of the TUG in PD used the time of five TUG trials, which produced high ICC scores ranging from 0.87 to 0.99 [243]. Similar findings were reported in another study that assessed TUG in PD using the elapsed time of two trials (ICC = 0.80) [255]. Several other studies, spanning various populations, have focused on the total duration to test the reliability of different mobility tasks. For instance, in people with chronic stroke, the reliability of dual-task mobility assessments was investigated, demonstrating good reliability (ICC = 0.70 to 0.93) based on walking time across five different walking tasks [257]. Additionally, TUG reliability was assessed in older individuals over three different days, revealing a good reliability (ICC = 0.92) based on the total time required to perform the TUG [256]. Our results align with the literature, revealing good to excellent reliability of the total duration of TUG and cogTUG (ICC = 0.75 to 0.95) and moderate to good reliability for the 16-foot walk (ICC = 0.62 to 0.77). However, the relative change in time between the trials revealed a possible learning effect in TUG and cogTUG and a possible fatigue effect in the 16-foot walk (see Figure 4.1). The fatigue effect observed in the 16-foot walk arises from participants completing four 8-foot walks without any rest, in contrast to the TUG and cogTUG, where participants had a break between trials, leading to the observed learning effect. The learning effect aligns with the findings of other studies [244, 265]. When comparing our results to those of the other studies referenced above, it

is also important to consider the different implications of retesting immediately or within minutes (as in our case), when fatigue and learning effects may play a role, versus retesting over several days, when possible fatigue effects would be expected to be negligible.

In addition to the total duration of the tasks, we conducted an examination of the trial differences by exploring the duration variations in the subtasks that make up the three mobility tasks. As illustrated in Figure 4.1, the consistency among the different components of each task is not uniform, aligning with prior findings [250, 253, 266]. For example, a study of 28 PD participants demonstrated that the walking and turning phases of the TUG were more reliable than the sit-to-stand and stand-to-sit phases [253]. Similarly, the sit-to-stand and stand-to-sit phases showed the lowest reliability in a study of 30 young adults [249], as well as in another study that assessed the reliability of a modified version of the TUG with 28 participants experiencing balance and gait problems [250]. The latter study attributed the low reliability to two factors: the short duration of the subtask (around three seconds), which increases the measurement error, and the difficulty in defining the exact end of the maneuver [267, 268]. Other studies showed that sit-to-stand had the lowest reliability among the components of TUG in healthy adults [266], as well as in patients with mild PD and age-matched controls [252]. The variability in strategies used to perform sit-to-stand, due to its high degrees of freedom, could explain the poor reliability of this metric [269, 270]. Consistent with these findings, we observed that the sit-to-stand subtask was the least reliable in the TUG for both the mild and moderate PD groups, as well as for the controls, and in the cogTUG, it was the least reliable subtask for the controls (see Figure 4.1).

Furthermore, we explored the reliability assessment of a selected subset of quantitative measures derived from sensor data. Figure 4.2 highlights that these measures exhibit lower consistency than the total duration across all participant groups, with reduced ICC and higher RC

scores. These results are in line with previous research that evaluated the reliability of certain gait [244, 251, 252], turn [251, 252], and transition [252] measures. Once again, our results demonstrate that the total duration of the task does not capture the differences in movement revealed by other measures, highlighting the importance of examining individual components for a more comprehensive reliability assessment. Our results presented here offer a much more granular analysis, determining the test–retest reliability of each mobility measure among the groups of participants with different levels of severity of PD. This is in contrast to previous studies referenced above that conducted reliability analyses using data from the entire study population. The observed variation in reliability for each measure among different participant groups, as depicted in Supplementary Figures B.1–B.3, underscores the need for tailored assessments that account for the inherent heterogeneity within the studied population.

Among the measures analyzed for both TUG and cogTUG, we found that features derived from the turning subtask had better test–retest reliability compared to those from the sit-to-stand and stand-to-sit subtasks across all PD groups, with the exception of TUG features in the moderate PD group. This aligns with a study that reported lower reliability for trunk kinematics during sit-to-stand and stand-to-sit compared to turns in a PD population with a median H&Y score of 3 [253], and another study that identified sit-to-stand features as the least reliable among the TUG components in a population of 12 early-to-moderate PD patients (H&Y = 1 to 2.5) [252]. The effect of adding a cognitive task to the TUG has also been explored, and previous work has shown a decrease in reliability when a cognitive task was introduced in a population of 12 community-dwelling older adults [244]. Similarly, in our study, the median ICC values across all measures were lower for cogTUG than for the TUG task in the control group (see Figure 4.2).

The finding that the severe PD group exhibited both the highest ICC and the greatest RC

between trials (see Figure 4.1) suggests that this discrepancy may be an artifact of the ICC calculations, influenced by both within-participant and between-participant variance. A similar observation was made in another study, which noted that variability increases with the time taken to complete the TUG and that ICC is significantly impacted by the range of scores used in its calculation [256]. Specifically, ICC tends to be higher when the differences between measurements are small relative to the range of scores across participants. This underscores a potential source of misinterpretation when relying on ICC, as opposed to RC, which more accurately reflects the temporal differences between trials.

The subsets of sensor-derived quantitative measures analyzed in the test–retest reliability assessment (Section 4.3.2) and the model performance evaluation (Section 4.3.3) differ because these analyses are distinct and designed to address separate objectives. Both analyses aim to explore differences between the two trials, but their focus and criteria for feature inclusion differ. In the reliability assessment, we prioritized features that exhibit high variability between trials and substantial differences across participant groups, as these features best illustrate the inconsistency of sensor-derived measures within and between groups. This approach emphasizes understanding the reproducibility of the measures and their potential sensitivity to the characteristics of the participants. In contrast, in the model performance evaluation, the selection of features was guided by the accuracy of the classification. Here, we focus on identifying the most important features distinguishing participants with PD from the controls. This required selecting features based on their importance in model diagnostic performance rather than their trial-to-trial variability. By addressing these separate goals, the two analyses complement each other, providing insights into both the stability and predictive utility of sensor-derived features.

We acknowledge several limitations in our study. First, we split the 32-foot walk task into

two 16-foot trials due to a lack of separate trials, which could impact the test–retest reliability results. Furthermore, the MDC values reported in this paper may represent underestimates, as our test–retest procedure did not account for additional sources of variability, such as day-to-day fluctuations, disease symptoms, physical status, or sensor repositioning. A more comprehensive evaluation that incorporates these factors would be necessary for a more accurate calculation of MDC. Furthermore, our findings are based on specific mobility tasks (i.e., TUG, cogTUG, and walking with turns), and may not be generalized to other motor or cognitive assessments used in PD. It is also unclear whether our results will apply to other movement assessment contexts, including disorders other than PD, or the evaluation of movement tasks beyond disease diagnostics, such as in sports performance analysis.

4.5 Conclusions

Our findings reveal that repeating mobility tasks may not be helpful for diagnostic accuracy, allowing for the simplification of mobility test protocols; relative change may be superior (or at least offer important complementary information) to ICC when determining test–retest reliability, depending on the disease severity of the population/group in question; and relying exclusively on the overall task duration for reliability assessment is insufficient, as it neglects analytically important differences in aspects of mobility captured by other quantitative measures. Through machine learning and statistical analyses, we demonstrate the nonidentical nature of the different trials when repeating mobility tasks. Specifically, sensor-derived measures exhibit less consistency between the two trials compared to the total duration of the task, which remains more stable. Our findings also reveal that variability between trials differs not only between controls and PD

participants, but also among groups with varying levels of severity of PD. These results prompt a reassessment of how we define a reliable mobility task, urging a more comprehensive and nuanced approach to capturing the complexities of task performance and the inherent heterogeneity within the studied population.

4.6 Funding

Funding was provided by a University of Maryland MPower Seed Grant Award (R.v.C. and M.P.C.), the Rosalyn Newman Foundation (L.M.S.), and the University of Maryland Claude D. Pepper Older Americans Independence Center (P30-AG028747; R.v.C.).

4.7 Institutional Review

All study procedures were approved by the University of Maryland, Baltimore (UMB) Institutional Review Board (IRB) on 10 September 2015 (protocol code: HP-00065245).

4.8 Informed Consent

Informed consent was obtained from all subjects involved in this study before any procedures were conducted.

4.9 Data Availability

De-identified sensor-derived measures have been deposited in the Digital Repository at the University of Maryland (DRUM), `https://doi.org/10.13016/jgw3-6sga` (accessed on 16 December 2024). The code used in this research paper can be accessed publicly at the same

link. The provided notebook contains the code employed for conducting the analysis presented in this study. Comprehensive instructions for configuring the required environment and executing the code are outlined in the `README` file.

4.10 Acknowledgments

We thank: Aron S. Buchman and Robert J. Dawe for their support with implementation of the mobility testing protocol; Katrina Schrader, Kimberly Bambarger, Veronica Fallon, and Philip Heiser for their technical assistance with mobility testing; Rebecca Fenderson, Maxwell Van Hoven, Vivien Xi and Patrick Curry for their assistance with data extraction and processing; and Christina Griffin for her regulatory administrative support.

Chapter 5: Applying Wearable Sensors and Machine Learning to the Diagnostic Challenge of Distinguishing Parkinson's Disease from Other Forms of Parkinsonism

The work of this chapter has been published in the journal Biomedicines as: Rana M. Khalil, Lisa M. Shulman, Ann L. Gruber-Baldini, Stephen G. Reich, Joseph M. Savitt, Jeffrey M. Hausdorff, Rainer von Coelln, and Michael P. Cummings. Applying wearable sensors and machine learning to the diagnostic challenge of distinguishing Parkinson's disease from other forms of parkinsonism. *Biomedicines*, 13(3):572, 2025 [30].

Abstract

Background/Objectives: Parkinson's Disease (PD) and other forms of parkinsonism share motor symptoms, including tremor, bradykinesia, and rigidity. The overlap in their clinical presentation creates a diagnostic challenge, as conventional methods rely heavily on clinical expertise, which can be subjective and inconsistent. This highlights the need for objective, data-driven approaches such as machine learning (ML) in this area. However, applying ML to clinical datasets faces challenges such as imbalanced class distributions, small sample sizes for non-PD parkinsonism, and heterogeneity within the non-PD group. Methods: This study analyzed wearable sensor data from 260 PD participants and 18 individuals with etiologically diverse forms of

non-PD parkinsonism, which were collected during clinical mobility tasks using a single sensor placed on the lower back. We evaluated the performance of ML models in distinguishing these two groups and identified the most informative mobility tasks for classification. Additionally, we examined the clinical characteristics of misclassified participants and presented case studies of common challenges in clinical practice, including diagnostic uncertainty at the patient's initial visit and changes in diagnosis over time. We also suggested potential steps to address the dataset challenges which limited the models' performance. Results: Feature importance analysis revealed the Timed Up and Go (TUG) task as the most informative for classification. When using the TUG test alone, the models' performance exceeded that of combining all tasks, achieving a balanced accuracy of 78.2%, which is within 0.2% of the balanced diagnostic accuracy of movement disorder experts. We also identified differences in some clinical scores between the participants correctly and falsely classified by our models. Conclusions: These findings demonstrate the feasibility of using ML and wearable sensors for differentiating PD from other parkinsonian disorders, addressing key challenges in its diagnosis and streamlining diagnostic workflows.

5.1 Introduction

Distinguishing Parkinson's disease (PD) from other forms of parkinsonism poses a significant clinical challenge due to overlapping clinical characteristics among the different disease entities and the marked heterogeneity in the clinical presentation of each form of parkinsonism, including in idiopathic PD itself [271]. In addition to PD, parkinsonism can be the core manifestation of other neurodegenerative diseases, including multiple system atrophy (MSA), progressive supranuclear palsy (PSP), corticobasal syndrome (CBS), and dementia with Lewy bodies (DLB),

and secondary forms of parkinsonism, such as drug-induced parkinsonism (DIP) and vascular parkinsonism. Furthermore, other movement disorders (especially essential tremor (ET)) can be associated with relatively mild parkinsonism over time [272]. Some ET patients exhibit more complex phenotypes (including parkinsonian signs), while others may eventually progress to PD, adding to the challenge of making a diagnosis [272]. Each of these conditions is characterized by distinct neuropathological and phenotypic features [273]. However, they all share some degree of the motor symptoms typical of PD, including tremor, bradykinesia, and rigidity, although with varying prominence [274]. Additionally, mild parkinsonism without concomitant PD can develop with aging, further complicating an accurate diagnosis [275]. The lack of definitive biomarkers and the reliance on clinical criteria for diagnosis contribute to diagnostic uncertainty, often leading to misclassification, especially early in the disease, when the signs and symptoms of different forms of parkinsonism have greater overlap and distinguishing features can be subtle or absent [223, 276].

Diagnostic accuracy in the clinical diagnosis of PD is highly dependent on the experience and expertise of the treating clinician. A recent comprehensive meta-analysis revealed a balanced diagnostic accuracy ranging from 69.5% for non-experts to 82.4% for movement disorder experts [27]. In other words, up to 26% of patients were initially misdiagnosed, based on the clinical diagnostic criteria of PD [27]. The most common misdiagnoses were due to the significant clinical similarities between PD and other neurodegenerative forms of parkinsonism. The diagnostic accuracy for other parkinsonian disorders is even lower than that of PD [277]. We recently showed that the movement disorder specialists at our own center changed their clinical diagnosis of PD in 6% of patients over a 15-year period [189]. The most frequent diagnostic changes involving PD were observed with DIP, MSA, PSP, and DLB. Treatment and prognoses differ

considerably between these different underlying conditions [273]. Some additional tools commonly used in clinical practice, such as imaging tests (e.g., DaTscan, MIBG scintigraphy), can assist clinicians in differentiating PD from other parkinsonian syndromes by introducing greater objectivity into the diagnostic process [278]. Nonetheless, despite these advancements, the final diagnosis remains largely dependent on clinical examination. Consequently, the development of objective diagnostic tools capable of distinguishing PD from other forms of parkinsonism remains a critical unmet need in both clinical practice and research.

Wearable sensors containing accelerometers and gyroscopes provide continuous, highly granular, and objective measurements that can be used to identify subtle changes in complex movement patterns [119, 279, 280], even in an early stage of the disease when abnormalities may be too subtle to be detected by clinical assessment alone [198]. Although statistical analysis of wearable sensor data has shown promise in the differential diagnosis of parkinsonian disorders [281–285], ML provides a more advanced and adaptable approach for analyzing the large-scale, high-dimensional, and complex datasets generated by these sensors. ML analyses can identify subtle patterns and underlying relationships that may not be detectable through conventional statistical methods or clinical observation [12, 286]. By analyzing sensor-derived movement data from individuals with PD and various parkinsonian syndromes, ML models can extract distinctive kinematic signatures and biomarkers associated with specific parkinsonian syndromes [287]. Previous studies have used ML to distinguish PD from other forms of parkinsonism, demonstrating its potential to support clinical decision-making [288–292]. The ability to classify PD from other types of parkinsonism is clinically relevant in terms of patient/family counseling, the prediction of disease progression, and determining the most appropriate treatment strategies [100, 107].

We conducted a search of the literature using the PubMed, IEEE Xplore, and Nature porto-

folio databases, covering publications from 1995 to 2024. The search keywords included "((artificial intelligence) OR (machine learning) OR (deep learning)) AND (Parkinson) AND ((atypical parkinsonism) OR (differential diagnoses)) AND (wearable OR gait)". We found that previous research applying ML focused predominantly on distinguishing PD from a single type of parkinsonism [226, 290–296]. Only one study applied ML to a smartwatch dataset to differentiate PD from various differential diagnoses, including ET, atypical parkinsonism such as PSP or MSA, secondary causes of parkinsonism, and Multiple Sclerosis [288]. Another study developed an ML system to differentiate between PD, cerebellar ataxia, and PSP-Richardson syndrome (PSP-RS) based on gait and postural instability data [289]. However, this approach involved the construction of two separate binary classifiers: one distinguishing PD from ataxia and another distinguishing PD from PSP-RS.

Using ML to distinguish participants with PD from a diverse group of types of parkinsonism in real-world clinical datasets poses significant challenges. First, the overlapping motor symptoms shared by both groups complicate the ability of ML models to accurately distinguish between them. Second, the non-PD parkinsonism groups generally have a much smaller sample size compared to the PD group, as PD accounts for more than 70% of all cases of parkinsonism [297]. Studies conducted in various countries have shown that PD is the most frequent diagnosis, comprising between 42% and 97% cases of parkinsonism [298–304]. The prevalence and incidence of PD vary widely due to differences in study methodologies, diagnostic criteria, case identification approaches, data collection techniques, and population demographics [298]. The dominance of PD cases limits the number of non-PD samples available for model training, reducing the statistical power and generalizability of these models. Third, the disparity in prevalence results in a highly imbalanced classification problem, where the larger PD group dominates

the dataset. This imbalance often biases the ML models towards prioritizing the majority class, leading to their reduced performance in the minority class [305, 306]. Finally, the heterogeneity of the diseases within the non-PD group introduces additional complexity. Non-PD parkinsonism encompasses a diverse range of syndromes with distinct underlying pathologies but overlapping clinical features, making it difficult for models to identify consistent and discriminatory patterns.

Our study uses ML to differentiate PD patients from a diverse group of participants clinically diagnosed with several non-PD parkinsonian disorders, including "atypical" neurodegenerative diseases and secondary parkinsonism. We analyze data collected from a single sensor worn on the lower back during mobility tests conducted in a hallway in a traditional clinic setting by 260 PD participants and 18 participants with etiologically diverse forms of parkinsonism (non-PD). Our objective is to highlight the potential challenges encountered when ML is applied to datasets that mimic real clinical settings and to demonstrate how these challenges can be addressed. Beyond evaluating model performance, we explored the characteristics of the participants misclassified by the model, identified the most important mobility tasks for classification, and examined whether the mobility testing protocol could be simplified to rely on a single task without compromising accuracy. Furthermore, we deployed the model in four examples of challenging scenarios commonly encountered in clinical practice, where the diagnosis was uncertain at the initial visit or the diagnosis was changed during follow-up visits.

5.2 Materials and Methods

5.2.1 Participants

Our study cohort comprised 318 individuals diagnosed with parkinsonism, all of whom were recruited from the University of Maryland Movement and Memory Disorders Center (UM-MMDC) between October 2015 and March 2020 for a study on mobility analysis in individuals with parkinsonism. Among them, 293 were diagnosed with idiopathic PD, according to the UK Parkinson's Disease Society Brain Bank Clinical Diagnostic [188], by a movement disorder specialist at the UM-MMDC. All participants were also enrolled in the Health Outcomes Measurement (HOME) study at the UM-MMDC, a naturalistic cohort study that collects patient- and clinician-reported data on motor symptoms, activities of daily living, and quality of life during routine office visits and while patients on their regular medications. Each participant underwent a comprehensive clinical evaluation following the guidelines outlined in the HOME study protocol, which have been previously detailed [96, 189–191]. Exclusion criteria for the mobility analysis included the inability to stand or walk without assistance (Hoehn & Yahr stage, H&Y 5), other conditions not related to parkinsonism that significantly affect gait and balance, or the failure to provide informed consent during the evaluation. Additionally, 1 PD participant was unable to complete the study procedures and was subsequently excluded, and 30 other PD participants were excluded due to missing data. All participants provided their informed consent prior to any study procedures. The study protocols for both the HOME study and the mobility analysis in the parkinsonism study were approved by the University of Maryland Institutional Review Board. All clinical neurologists involved in the evaluation and diagnosis of patients are fellowship-trained

and board-certified movement disorder specialists, who use published criteria to diagnose different forms of non-PD parkinsonism (including MSA, PSP, DLB, CBS, ET, and DIP) [307–312]. Participants were enrolled during routine outpatient visits to the UM-MMDC and underwent comprehensive clinical assessments following the HOME study protocol, as detailed in our previous research [28].

5.2.2 Experimental Procedure

Using a wearable sensor, an assessment of mobility was performed following a methodology described previously [115, 198, 199]. Participants performed five specific mobility tasks: a 32-foot walk, standing with their eyes open for 20 seconds, standing with their eyes closed for 20 seconds, two trials of the Timed Up and Go (TUG) test, and two trials of the cognitive TUG (cogTUG). Mobility data were collected using a small (106.6 $\times$ 58 $\times$ 11.5 mm), lightweight (55 g) sensor device (Dynaport MT, McRoberts B.V., The Hague, The Netherlands, technical specifications: `https://www.mcroberts.nl/products/movetest/`, accessed on 7 January 2025) attached to their lower back using a waist belt. This device, equipped with a triaxial accelerometer (range: $\pm$8 g; resolution: 1 mg) and a triaxial gyroscope (range: $\pm$2000 dps; resolution: 70 dps), measured acceleration and rotation along three axes (vertical, mediolateral, anteroposterior) at a rate of 100 Hz. The sensor was remotely controlled via Bluetooth, and digital markers were placed at task start and end points to identify each mobility task later on. After each test, data were transferred to a secure server and analyzed using a custom-made Matlab (version 9.1) GUI. The accelerometry patterns and digital markers were used to segment the data for further analysis. The details of the mobility assessment methodology, devices used,

and data acquisition and retrieval have been described in our previous work [28]. Figure 5.1 presents an overview of the study's workflow.

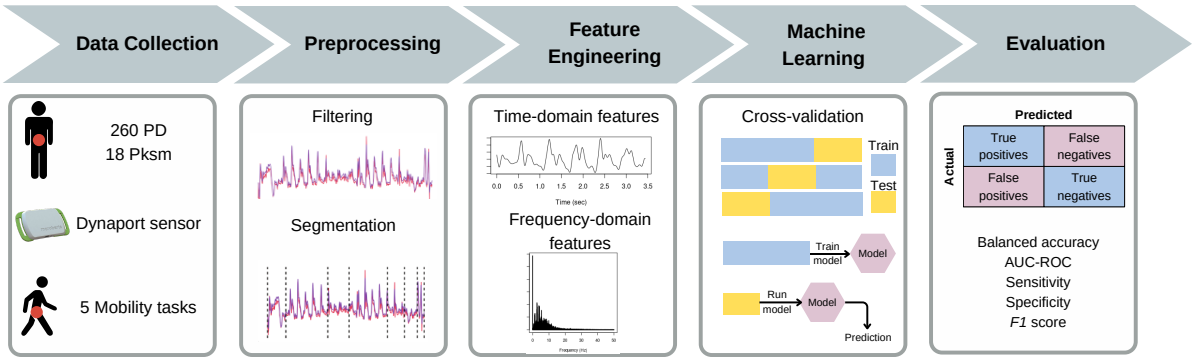


Figure 5.1: Study workflow. Pksm: non-PD parkinsonism; AUC-ROC: area under the receiver operating characteristic curve.

5.2.3 Feature Engineering

To reduce noise in the sensor signals, we implemented a low-pass, 20 Hz, zero-lag, 4th order Butterworth filter [204]. The order and cutoff frequency of the filter were selected based on common practices in the literature [33, 313–315] and further refined through exploratory data analysis. We then segmented the complex motor tasks, such as TUG, cogTUG, and the 32-foot walk, into their respective subtasks. We extended a previous method [201] to identify turns, sit-to-stand, stand-to-sit, and walking segments using the trapezoidal integration [202] of sensor data. Stationary periods were removed to correct timing discrepancies. For a detailed description of the segmentation approach, we refer readers to our previous work [28].

For every participant and task, we computed various sets of features across the six recorded channels, including statistical measures consisting of 22 frequency-domain features (Supplementary Table A.2), 42 time-domain features (Supplementary Table A.3), the top 10 power amplitudes from the Discrete Fourier Transform (DFT) and Lomb-Scargle Periodogram (LSP) [205,

206], and 3 time-domain features computed across the signals from each pair of the six channels (Supplementary Table A.4).

Each participant was characterized by a total of 32,514 sensor-derived features computed from the six channels of the 32-foot walk segmented subtasks, standing with closed eyes task, standing with open eyes task, TUG segmented subtasks (from the first trial, the second trial, and the mean and difference between corresponding features from trials), and cogTUG segmented subtasks (from the first trial, the second trial, and the mean and difference between corresponding features from both trials). Before executing the ML models, preprocessing procedures were applied to the data, including the elimination of variables with constant or infinite values and the imputation of missing values using the median of the observed values.

5.2.4 Mutual Information-Based Feature Selection for EasyEnsemble (MIEE)

We used the mutual information-based feature selection for EasyEnsemble (MIEE) [316] algorithm due to its strong performance on highly imbalanced datasets, such as ours (260 PD and 18 non-PD parkinsonism participants). This method combines EasyEnsemble [317] and feature selection to effectively address class imbalances. EasyEnsemble creates multiple independent balanced subsets by undersampling the majority class, trains classifiers on these subsets, and combines their predictions using ensemble methods. To improve accuracy, MIEE incorporates feature selection by calculating the mutual information (MI) [318] scores between features and the target class, ranking features based on their scores, and selecting the most informative ones. This process helps reduce dimensionality and mitigates the risk of model overfitting. The process begins by generating balanced subsets, applying MI-based feature selection to each sub-

set, training classifiers on these subsets using selected features, and aggregating the predictions of the generated classifiers through ensemble techniques. In our implementation, we created five balanced subsets, identified the top 30 features for each subset using MI scores, and trained a Random Forest (RF) model [216] with 1000 trees for each subset using the R package randomForest. The final predictions for each participant were obtained by averaging the prediction probabilities of the five classifiers.

5.2.5 Machine Learning Model

We used a three-fold cross-validation technique to create training and test splits for training our our models using stratified sampling. This process was repeated five times with different random seeds to improve robustness. Cross-validation ensured that each data point appeared exactly once in the test set, which was necessary given the small sample size of non-PD participants, making a separate test set impractical. Repeating this procedure with different random seeds increases the diversity of the train–test splits and reduces the likelihood of overfitting. The final class for each participant was determined based on the class most commonly predicted by the five replicates. Within each fold, the model described in Section 5.2.4 was trained on the training set and tested on the corresponding test set to prevent data leakage and ensure generalization [213].

Several steps were taken to address the imbalance between the PD and non-PD parkinsonism groups. First, we used stratified sampling when creating training and testing sets to maintain a uniform class representation in the cross-validation folds. Second, the MIEE algorithm was used due to its demonstrated effectiveness in handling highly imbalanced datasets [316]. Third, the performance of the models was evaluated using balanced accuracy, sensitivity, specificity,

F_1 score, and AUC-ROC metrics. These metrics are considered more appropriate for evaluating models in cases of imbalanced datasets compared to overall accuracy, which can be misleading due to its tendency to prioritize common classes over rare ones [319].

5.2.6 Group Feature Importance

We employed the notion of group feature importance to evaluate the predictive contribution of various mobility tasks. Group feature importance involves assessing the combined information of a set of features within a specific category. All features derived from individual mobility tasks were aggregated into five distinct groups, each corresponding to one of the five mobility tasks. To assess the importance of each group, we adapted the permutation importance method for feature groups within the framework of random forests, based on a previously outlined approach [220]. For each feature group, the grouped variables were simultaneously permuted, and the resulting mean decrease in accuracy before and after permutation was calculated.

5.2.7 Statistical Analysis

The data analyses were conducted using R version 4.2.3 (15 March 2023) and RStudio version 2023.6.0.421. We used a two-tailed Student's t-test to compare continuous variables and Pearson's χ^2 test to compare categorical variables between the two groups. Continuous variables were presented as a mean $\pm$ standard deviation (SD), and actual p-values for the statistical tests were reported. For each of the models' performance metrics, we presented both the calculated values and their corresponding 95% two-sided confidence intervals (CIs). These intervals were obtained through balanced bootstrap resampling using 10 K replicates and were adjusted using

bias correction and acceleration to enhance our control over bias and skewness in the bootstrap distribution. We used the boot function in the R package boot version 1.3–30 [221, 222] to generate the bootstrap replicates and the boot.ci function from the same package to generate the CIs.

5.3 Results

Wearable sensor data were collected from a cohort of 260 individuals diagnosed with PD (87% were receiving antiparkinsonian medication) and 18 participants diagnosed with other forms of parkinsonism. In the latter group, there were four cases of parkinsonian-type MSA (MSA-P), three cases of PSP, three cases of DLB, two cases of CBS, two cases of ET, and one case of DIP. Three patients with unspecified parkinsonism were also included. Table 5.1 provides the characteristics of the study cohort.

Table 5.1: Characteristics of the study cohort. UPDRS: Unified Parkinson's Disease Rating Scale. H&Y: modified Hoehn and Yahr.

Feature	**PD** ($n = 260$)	**non-PD Parkinsonism** ($n = 18$)	p
Age (years, $\bar{x} \pm SD$)	66.8 ± 9.3	68.7 ± 8.2	0.377
Gender (%male)	62.3	38.9	0.093
Height (cm, $\bar{x} \pm SD$)	172.2 ± 10.4	165.3 ± 13.8	0.059
UPDRS (total, $\bar{x} \pm SD$)	35.1 ± 17.1	47.9 ± 22.4	0.033
UPDRS (motor-part III, $\bar{x} \pm SD$)	22.1 ± 11.8	28.8 ± 14.9	0.085
Disease duration (years, $\bar{x} \pm SD$)	7.8 ± 6.5	1.7 ± 2.1	4.2×10^{-10}
H&Y ($\bar{x} \pm SD$)	2.2 ± 0.62	2.9 ± 0.77	0.005
stage 1 (n)	12	0	
stage 1.5 (n)	4	0	
stage 2 (n)	168	6	
stage 2.5 (n)	35	3	
stage 3 (n)	24	5	
stage 4 (n)	17	4	

Each participant performed five motor tasks, including a 32-foot walk, standing with eyes open, standing with eyes closed, two trials of the TUG, and two trials of the cogTUG (see Section 5.2.2 for details). For simple tasks such as standing with eyes open and standing with eyes closed, we computed a set of frequency-domain and time-domain features. For more complex tasks such as the 32-foot walk, TUG, and cogTUG, we initially segmented these into their respective subtask components, from which features were then derived for each subtask signal (see Section 5.2.3 for details). Then, we applied the ML model described in Section 5.2.5, using features from all five tasks, to predict the final class for each participant. A summary of the process is provided in Figure 5.1.

5.3.1 Feature Importance Analysis

Across the replicates, folds, and balanced subsets of the model, 82.4% of the selected features were derived from the TUG task, including the first trial, second trial, mean, and difference between corresponding features from both trials. Among these TUG-derived features, 42.5% were calculated specifically from the sit-to-stand phase, while the remaining features were equally distributed across the turn, stand-to-sit, and walking phases. These features captured acceleration and rotational metrics along the vertical, mediolateral, and anteroposterior axes. Additionally, 11% of the selected features were calculated from the cogTUG task, while the remaining features were derived from the 32-foot walk, stand with eyes open, and stand with eyes closed tasks.

The value of the information from the TUG task was also demonstrated by conducting a group feature importance analysis that included the five mobility tasks. The features derived from

the TUG task emerged as the most influential for the classifier, surpassing the importance of other feature groups by a margin of at least 43% (Figure 5.2).

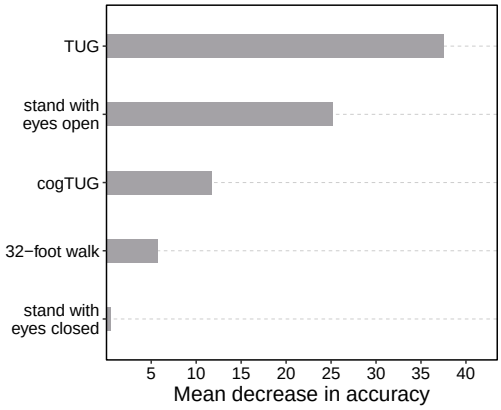


Figure 5.2: Group feature importance for the classification of Parkinson's disease and non-PD parkinsonism. The ranking of groups is determined based on their importance, which is computed by simultaneously permuting features within each group and measuring the average decrease in accuracy between the original and permuted data. Under the null hypothesis, which posits no association between the group of predictor variables and the model's prediction, permutation should result in no impact on predictive performance.

5.3.2 Model Performance

The trained model achieved a balanced accuracy (average accuracy of both classes) of 72.9% (confidence interval [CI] 60.2%, 81.4%), with an area under the receiver operating characteristic curve (AUC-ROC) of 0.73 (CI 0.63, 0.83). Its sensitivity, specificity, and F_1 score were 0.68 (CI 0.62, 0.74), 0.78 (CI 0.50, 0.93), and 0.80 (CI 0.76, 0.84), respectively (Table 5.2, first column). Table 5.3a shows the confusion matrix of the classifier.

Table 5.2: Classification results of all-tasks and TUG-only models using participants with Parkinson's disease (PD) and non-PD parkinsonism. CI: 95% confidence interval.

	All Tasks	**TUG-Only**
Balanced accuracy (CI) [%]	72.9 (60.2, 81.4)	78.2 (65.7, 85.6)
AUC-ROC (CI)	0.73 (0.63, 0.83)	0.78 (0.69, 0.87)
Sensitivity (CI)	0.68 (0.62, 0.74)	0.73 (0.67, 0.78)
Specificity (CI)	0.78 (0.50, 0.93)	0.83 (0.56, 0.95)
F_1 score (CI)	0.80 (0.76, 0.84)	0.84 (0.80, 0.87)

Table 5.3: Confusion matrices of the classifiers using participants with Parkinson's disease (PD) and non-PD parkinsonism (Pksm). Rows represent actual classes and columns represent predictions.

(a) All Tasks

	PD	Pksm
PD	177	83
Pksm	4 *	14

(b) TUG-Only

	PD	Pksm
PD	190	70
Pksm	3 *	15

* One of the misclassified Pksm participants was later diagnosed with PD.

To streamline the complexity of our models, we sought to construct classifiers based solely on features derived from the most informative mobility task. To that end, we focused exclusively on TUG features, given its high ranking in the feature importance analysis. We then implemented the ML model detailed in Section 5.2.5 using these TUG-specific features. The TUG-only model outperformed the initial model, which included all mobility tasks, achieving a 5.3% improvement in balanced accuracy (Table 5.2). Its corresponding confusion matrix is presented in Table 5.3b. This finding suggests that sensor data from the TUG mobility task may be sufficient to achieve a satisfactory performance in differentiating PD from other forms of parkinsonism. Consequently, our subsequent analysis focused on the TUG-only model, given its simplicity and good performance in terms of balanced accuracy compared to the all-tasks model. The features selected for the TUG-only model were distributed across the components of the TUG task as follows: 36.3% from the sit-to-stand phase, 24.7% from the stand-to-sit phase, 23.7% from the turn phase,

and 15.3% from the walking phase.

5.3.3 Misclassification Analysis

To gain insight into the characteristics of participants misclassified by the TUG-only model, we conducted comparisons of various demographic and clinical scores between correctly classified and misclassified participants with PD or non-PD parkinsonism. Two-tailed Student's t-tests were used to compare continuous variables, and Pearson's χ^2 tests were used to compare categorical variables across the groups. The summarized comparison results are presented in Table 5.4. The 70 PD participants falsely classified as having non-PD parkinsonism exhibited significantly ($p < 0.05$) higher Unified Parkinson's Disease Rating Scale [196] part III (UPDRS_PIII) motor scores, lower Montreal Cognitive Assessment (MoCA) [195] scores, higher Cumulative Illness Rating Scale-Geriatrics (CIRS-G) [192] scores, greater age, and longer durations of the disease compared to their 190 correctly classified counterparts. Further examination of the distribution of misclassified PD participants across different H&Y [97, 197] stages revealed a higher-than-expected number of misclassifications at more severe stages (H&Y = 3 and 4). Additionally, the three misclassified non-PD participants had significantly lower UPDRS_PIII motor scores than the 15 correctly classified non-PD participants. These results indicate that the models tend to misclassify PD participants with more severe motor symptoms and non-PD participants with milder motor symptoms. No significant differences in MoCA scores, CIRS-G scores, age, disease duration, or sex were observed between the correctly and incorrectly classified non-PD participants.

Table 5.4: Clinical characteristics of participants with Parkinson's disease (PD) and non-PD parkinsonism correctly and incorrectly classified by the TUG-only model.

	PD			non-PD Parkinsonism		
	Correctly Classified	**Incorrectly Classified**	p	**Correctly Classified**	**Incorrectly Classified**	p
UPDRS_PIII	19.7 ± 10.9	28.7 ± 11.4	3.8×10^{-8}	30.5 ± 14.9	17.7 ± 5.7	0.016
MoCA	27.6 ± 2.8	26.5 ± 3.5	0.02	26.3 ± 4.9	23.3 ± 2.1	0.127
CIRS-G	4.5 ± 3.2	5.7 ± 3.9	0.035	8.0 ± 4.6	7.2 ± 3.6	0.73
Age	66.0 ± 8.9	69.1 ± 9.9	0.023	67.6 ± 8.1	74.0 ± 5.2	0.15
Disease duration	7.1 ± 5.9	8.7 ± 6.9	0.038	2.5 ± 2.3	2.3 ± 2.5	0.47
Sex (% male)	59.5	70.0	0.184	33.3	100.0	0.053
H&Y (n)			1.3×10^{-6}			0.40
1	10	2		0	0	
1.5	3	1		0	0	
2	138	30		4	2	
2.5	27	8		2	1	
3	8	16		5	0	
4	4	13		4	0	

Given the heterogeneous nature of the disorders within the non-PD parkinsonism group, we sought to determine whether there is a specific type of parkinsonism that is more similar to PD and, therefore, more likely to be misclassified by the model. To investigate this, we examined the types of parkinsonism in the three misclassified non-PD participants. We found that one participant had DLB, one had ET, and one had unspecified parkinsonism. This suggests that no particular type of non-PD parkinsonism is more likely to be misclassified as PD. However, the small number of misclassified participants precludes use reaching a definitive conclusion in this regard. Similarly, to explore whether misclassified PD participants share similarities with specific types of non-PD parkinsonism, we used the RF proximity matrix to determine the type of non-PD parkinsonism most similar to each misclassified PD participant (see the Supplementary Results for details). We found that the misclassified PD participants were relatively evenly distributed across the different types of non-PD parkinsonism (Supplementary Table C.1), indicating that they did not show a particular similarity to one type of parkinsonism over others.

5.3.4 Model Predictions for Challenging Cases

To evaluate the performance of the model in particularly challenging scenarios, we analyzed its predictions for four cases initially assessed as a parkinsonian syndrome (no specific diagnosis) at the time of initial gait testing. Three of these participants (between 54 and 69 years old at study enrollment) presented with balance problems early in their disease course, prominent dysautonomia (including urinary urgency and orthostatic hypotension), and mostly symmetrical parkinsonism without a clear response to levodopa. In all three cases, the treating movement disorder neurologists concluded that PD and MSA were equally likely at that point. Therefore, patients were classified as having a parkinsonian syndrome. Over the next one to four years, all three patients showed a robust response to levodopa, slow disease progression, and dysautonomia that was less prominent at follow-up. Therefore, the diagnosis was changed to idiopathic PD for each of these three participants. Due to their indeterminate diagnostic labels (MSA vs. PD) at the time of enrollment, all three participants were excluded from the training dataset and were instead classified by the models trained on the non-excluded participants. We used the TUG-only model to determine the mean prediction score for the three participants. One participant (participant a, H&Y = 2) was classified as having PD by the model, with a score in the 18.6 percentile of all PD-classified participants. The other two participants were classified as having non-PD parkinsonism by the model. One of them (participant b, H&Y = 3) was assigned to the non-PD group with a model score in the 35.2 percentile of all participants classified as having non-PD parkinsonism. The other participant (participant c, H&Y = 2) was also classified as having non-PD parkinsonism, but the model score was only in the 12.1 percentile among the participants classified as having non-PD parkinsonism.

In the fourth challenging case, with a change in diagnosis over time, the participant was a 68-year-old patient (participant d, H&Y = 2), with prominent hypophonia, stuttering, and symmetrical parkinsonism without a clear response to levodopa. In the absence of clear features suggestive of a specific form of parkinsonism, the treating neurologist diagnosed that person with an unspecified parkinsonian syndrome. Over the next four years, the patient reported significant worsening of his motor symptoms after stopping levodopa and developed typical levodopa-induced motor fluctuations and choreiform dyskinesias after re-starting levodopa. At that point, the movement disorder specialist made the diagnosis of idiopathic PD. Since this participant was diagnosed with an unspecified parkinsonian syndrome at the time of enrollment and mobility testing, they were included in our models as a non-PD parkinsonism participant. However, both the all-tasks and TUG-only models misclassified this participant as having PD, agreeing with the diagnosis made by the movement disorder specialist during follow-up evaluations.

5.3.5 The Performance of Alternative Models

In addition, we evaluated the performance of various ML models and sampling techniques using TUG-only features. Each of them was tested within the three-fold cross-validation and five-repeat framework: features were selected from the training set, the model was trained on the training data, and its performance was evaluated on the corresponding test set using the same features selected from the training data. First, we applied an unsupervised RF to rank the features by importance, selected a subset of the top 30 features that maximized the balanced accuracy of the training set, and trained a final RF model using the selected features (Table C.2a; Table 5.5, first column). Second, we calculated the mutual information between all features and the target

variable, ranked them by mutual information scores, and selected the top 30 features to train an RF model (Table C.2b; Table 5.5, second column). Third, we implemented the feature selection method described in a previous work [320], which calculates the F_1 scores for the features using a decision tree classifier. The features were ranked according to their scores, and the top 30 features were used to train an RF model (Table C.2c; Table 5.5, third column). Fourth, we employed a supervised RF approach to rank features by importance, selecting a subset that minimized the Akaike information criterion (AIC) [210] for the training set. The final RF model was built using the selected features (Table C.2d; Table 5.5, fourth column).

To address the class imbalance between PD and non-PD parkinsonism, we explored the impact of sampling the training set by undersampling the PD class and oversampling the non-PD class, as proposed in previous research [321]. We then employed three sampling placement strategies described in another work [322]. For each approach, the features were ranked using RF importance scores, and a subset was selected that minimized the AIC of the model. The approaches were as follows: 1. sample the training set first, followed by feature selection and RF model training using the original training data (Table C.2e; Table 5.6, first column); 2. sample the training set first, performing feature selection, and train the RF model on the sampled data (Table C.2f; Table 5.6, second column); and 3. perform feature selection on the original training set, followed by sampling and training of the RF model on the sampled data (Table C.2g; Table 5.6, third column).

Additionally, we implemented random undersampling of the PD group to match the size of the non-PD parkinsonism group each time an RF model was constructed. However, none of these models surpassed the performance of our suggested model (Section 5.2.5).

Table 5.5: Classification results of alternative machine learning models using TUG-only features. RF: random forest; FS: feature selection; bacc: balanced accuracy; MI: mutual information; DT: decision tree; AIC: Akaike information criterion.

	Unsupervised RF FS with Max Bacc	**MI-Based Ranking for Top FS**	**F_1 Ranking Using DT for Feature Scoring**	**Supervised RF FS with Min AIC**
Balanced accuracy (%)	67.8	68.2	68.3	52.1
AUC-ROC	0.68	0.68	0.68	0.52
Sensitivity	0.80	0.92	0.87	0.93
Specificity	0.56	0.44	0.50	0.11
F_1 score	0.87	0.94	0.91	0.93

Table 5.6: Classification results of sampling techniques using TUG-only features. FS: feature selection.

	Sampling Before FS, Model Trained on Original Data	**Sampling Before FS, Model Trained on Sampled Data**	**FS Before Sampling, Model Trained on Sampled Data**
Balanced accuracy (%)	56.4	49.6	59.4
AUC-ROC	0.56	0.50	0.59
Sensitivity	0.96	0.99	0.97
Specificity	0.17	0.00	0.22
F_1 score	0.95	0.96	0.96

5.4 Discussion

We demonstrate the potential of the machine learning analysis of sensor data collected during mobility testing from a single wearable accelerometer and gyroscope to distinguish PD from other forms of parkinsonism. The inherent similarities between these movement disorders make distinguishing idiopathic PD from other parkinsonian conditions a well-recognized diagnostic challenge for clinicians [271]. Based on findings from a previous meta-analysis, the balanced accuracy of the clinical diagnosis of PD was reported to be 69.5% when conducted primarily by non-experts, 78.4% when conducted by movement disorders experts during their initial assessment, and 82.4% for their diagnosis after follow-up [27]. Despite the overlap in motor symptoms between the groups, our TUG-only model achieved a balanced accuracy of 78.2%, approaching

the diagnostic accuracy of experts during their initial assessments. Expert clinicians reported a sensitivity of 81.8% and specificity of 74.9% for their initial diagnoses [27], while our TUG-only model achieved a sensitivity of 73.1% and specificity of 83.3%. This suggests that such a model could serve as an additional diagnostic "expert", which is particularly valuable for challenging cases, by providing an objective assessment based on a single mobility task and without requiring the full clinical evaluation used by clinicians. Additionally, wearable sensors offer additional benefits for remote monitoring and early-stage diagnosis, presenting a data-driven alternative to traditional clinical assessments [198].

One substantial challenge limiting model performance is the small sample size of the non-PD group. This imbalance is a common issue in clinical settings, given the higher prevalence of PD compared to other parkinsonian disorders [297, 298]. To address this, we applied a sampling approach recommended in a prior study [321], which combined the undersampling of the majority (PD) class, using the Neighborhood Cleaning Rule (NCL) [323], and the oversampling of the minority (non-PD) class using the synthetic minority oversampling technique (SMOTE) [324]. We integrated these sampling techniques with three strategies proposed in another study, comparing model performance when sampling occurred before or after feature selection and when classifiers were built using original (unsampled) or sampled data [322]. However, none of these approaches improved upon the performance of our suggested model (Supplementary Table C.2a–C.2g).

Other challenges included significant class imbalance and heterogeneity within the non-PD group. The steps taken to address this class imbalance, described in Section 5.2.5, included stratified sampling for the training and testing sets in cross-validation folds, using the MIEE algorithm for its effectiveness in imbalanced datasets [316] and evaluating models with appropriate metrics

for imbalanced data, such as balanced accuracy [319]. Although the heterogeneity of the non-PD group negatively affected model performance, it can also be seen as a strength of this study, because it represents an actual clinical population with a mixed group of parkinsonism patients, where all different forms of non-PD parkinsonism must be considered as differential diagnoses. Studies that only compare PD with one form of non-PD parkinsonism (e.g., PSP [325]) have a clear advantage in terms of statistics and the homogeneity of their two groups, but have the disadvantage of analyzing a somewhat artificially simplified situation. The same principles apply to the heterogeneity within the PD group, as our dataset includes a large PD cohort with varying stages of the disease, further reflecting the diversity encountered in clinical practice.

Analysis of group and individual feature importance identified the TUG task as the most important mobility task for distinguishing PD participants from those with other forms of parkinsonism (Figure 5.2). The model using only TUG task features achieved a 5.3% higher balanced accuracy compared to the model incorporating features from all five mobility tasks. Previous studies have emphasized the significance of gait and sway tasks [281, 284, 326, 327]. However, in the context of our study and the included mobility tasks, we found that gait and sway tasks were less important compared to the TUG. The TUG encompasses a greater diversity of movements, including movements that appear to be especially sensitive to parkinsonism, such as rising from a chair, sitting down, and turning 180°. This observation is reflected in the features of the TUG selected by our model, which were primarily derived from the sit-to-stand, stand-to-sit, and turn components. Those specific parts of the TUG might be particularly sensitive to atypical forms of parkinsonism, which tend to affect axial motor function earlier and more severely than idiopathic PD [283]. Although sensor-derived features from the TUG task identified it as the most informative, statistical tests (Student's t-test) did not reveal significant differences between the

UPDRS_III axial items (e.g., 3.9, 3.10, or 3.12) in the PD and non-PD groups.

A study highlighting TUG as one of the mobility tests with superior discriminative abilities between PD and atypical parkinsonism, with an AUC of 0.77 [283], aligns well with the results from our TUG-only model (AUC = 0.78). Our results show a better performance compared to a previous study that reported a balanced accuracy of 72.42% for distinguishing PD from a diverse group of non-PD parkinsonism by combining smartwatch data from movement-based assessments with an electronic questionnaire [288]. When questionnaire data were excluded, as in our models, the balanced accuracy in that study decreased to 69.18%.

Using a single sensor and a single task (TUG in our study) significantly simplifies both the mobility testing process in clinical settings and the data processing and analysis required for the model. This streamlined approach enhances the feasibility of adopting sensor-based mobility testing as a diagnostic tool in routine evaluations of movement disorders. These findings complement our previous work, which demonstrated that the clinical protocol for distinguishing PD from a control group could be simplified using a single sensor and task [28], with further simplification achieved by reducing the number of trials per task to just one [29].

Our analysis of the participants misclassified by the TUG model revealed a tendency to misclassify PD participants with more severe disease symptoms and non-PD participants with milder symptoms (Table 5.4). Although the non-PD group in this study had a shorter disease duration than the PD group, they displayed more severe motor symptoms, as reflected by their higher UPDRS_PIII scores (Table 5.1). Matching both groups by disease duration would result in differing severity levels, as other forms of parkinsonism are known to progress more rapidly than PD [328, 329].

For a small group of study participants, the diagnosis changed from the time of the initial

mobility testing to their clinical follow-up visit(s). We explored the characteristics of those participants more closely. Among them, three participants (excluded from the models) initially had indeterminate diagnostic labels (MSA vs. PD) but were later diagnosed with idiopathic PD during follow-up visits. Of these, our TUG-only model correctly classified one participant as having PD, while the other two were predicted to have non-PD parkinsonism, although their model scores were relatively low among those classified in the non-PD group. Another participant, initially diagnosed with an unspecified parkinsonian syndrome, was later confirmed to have idiopathic PD at follow-up. This participant, included in our models as a non-PD case, was misclassified as having PD by the model, aligning with the revised clinical diagnosis made at follow-up. These cases serve as examples of challenging scenarios encountered in real clinical settings, where diagnostic decisions can be as difficult for ML models as they are for clinicians.

In that context, it is also important to remember that there is no true diagnostic "gold standard" for these conditions, apart from a postmortem pathological examination of brain tissue. It is therefore quite possible that in a cohort of close to 300 participants across the two groups, a few participants may have been clinically misdiagnosed and, in fact, categorized correctly by our ML approach. This is an inherent problem for this type of analysis that will only be overcome once reliable and specific diagnostic biomarkers for the different forms of parkinsonism have been established.

A limitation of our work is the small sample size of our non-PD parkinsonism group, which may limit the generalizability of our findings. Subsequent research efforts should prioritize validating our results within larger cohorts to enhance the robustness of our conclusions. Additionally, the small sample size prevented us from addressing the heterogeneity issue by developing a multiclass classifier, where each form of parkinsonism could be represented by a separate class.

Furthermore, given the differences in UPDRS_PIII, H&Y scores, age, and other clinical measures between the correctly classified and misclassified participants, stratifying the model by one or more of these characteristics could potentially improve its performance. However, due to the limited sample size, this approach was not feasible in our study. Future studies should also explore the impact of including clinical measures (e.g., UPDRS, H&Y, disease duration) and other key diagnostic factors (e.g., severity of dysautonomia, early cognitive impairment, early falls, and vertical eye movement abnormalities) in ML models to determine whether incorporating these features improves their diagnostic accuracy.

5.5 Conclusions

Our study demonstrates the potential of machine learning applied to wearable sensor data to distinguish PD from other parkinsonian disorders while using a single sensor. The TUG task was the most valuable mobility task for our classifier, and using only this task would simplify mobility testing protocols. The TUG-only model achieved a balanced accuracy comparable to that of expert clinicians. We also discussed the challenges encountered when applying ML to datasets that reflect real clinical settings and presented strategies to address these issues.

5.6 Funding

Funding was provided by a University of Maryland MPower Seed Grant Award (R.v.C. and M.P.C.), the Rosalyn Newman Foundation (L.M.S.), and the University of Maryland Claude D. Pepper Older Americans Independence Center (P30-AG028747; R.v.C.).

5.7 Institutional Review

All study procedures were approved by the University of Maryland, Baltimore (UMB) Institutional Review Board (IRB) on 10 September 2015 (protocol code: HP-00065245).

5.8 Informed Consent

Informed consent was obtained from all subjects involved in the study before any procedures were conducted.

5.9 Data Availability

De-identified demographic, clinical, and sensor-derived data have been archived in the Digital Repository at the University of Maryland (DRUM), `https://doi.org/10.13016/2qhy-961s`. The code used in this research paper is also accessible via the same link. The provided notebook contains the code necessary to execute each aspect of the signal processing and machine learning models. Comprehensive instructions for configuring the required environment and executing the code are outlined in the `README` file.

5.10 Acknowledgments

We thank Aron S. Buchman and Robert J. Dawe for their support with the implementation of the mobility testing protocol; Katrina Schrader, Kimberly Bambarger, Veronica Fallon, and Philip Heiser for their excellent technical assistance with participant enrollment, clinical data collection, and mobility testing; and Christina Griffin for her regulatory administrative support.

Part III

Human Hearing

Chapter 6: Using Machine Learning to Understand the Relationship Between Audiometric Data, Speech Perception, Temporal Processing, and Cognition

The work of this chapter has been published in the IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP) as: Rana M. Khalil, Alexandra Papanicolaou, Renee Ti Chou, Bobby E. Gibbs, Samira Anderson, Sandra Gordon-Salant, Michael P. Cummings, and Matthew J. Goupell. Using machine learning to understand the relationships between audiometric data, speech perception, temporal processing, and cognition. In ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5, 2023 [31].

Abstract

Aging and hearing loss cause communication difficulties, particularly for speech perception in demanding situations, which have been associated with factors including cognitive processing and extended high-frequency ($>$ 8 kHz) hearing. Quantifying such associations and finding other (possibly unintuitive) associations is well suited to machine learning. We constructed ensemble models for 443 participants who varied in age and hearing loss. Audiometric, perceptual, electrophysiological, and cognitive data were used to predict speech perception in noise, reverberation,

and with time compression. Speech perception was best predicted by variables associated with audiometric thresholds (including new across-frequency composite variables) between 1–4 kHz, followed by basic temporal processing ability. Cognitive factors and extended high-frequency thresholds had little to no predictive ability of speech perception. Future associations or lack thereof will inform the field as we attempt to better understand the intertwined effects of speech perception, aging, hearing loss, and cognition.

6.1 Introduction

Over 50% of adults in the US >65 years suffer from significant hearing loss [330]. As people age, their ability to understand speech decreases and places a great burden on their ability to effectively communicate, particularly in demanding situations that include background noise, reverberation, multiple talkers, rapid talkers, or accented talkers [331]. These declines are a result of peripheral hearing loss (i.e., some speech sounds are less audible), central age-related temporal-processing deficits (i.e., neural encoding of speech sounds is distorted), and a decrease in higher-level cognitive abilities (e.g., memory, non-auditory processing speed) [63].

The complexity of understanding how age, hearing loss, and cognition intertwine as factors to predict speech-perception performance argues for modern data-driven analysis approaches on human measurements [332]. This includes machine learning, if the data set is large enough. For the present data study, we collected a large battery of audiometric, perceptual, electrophysiological, and cognitive data in 443 participants [333]. Novel to this data set is the number of subjects and suprathreshold perceptual measures of auditory temporal processing (e.g., pulse-rate discrimination).

Numerous studies have shown that among older listeners, speech perception is related to working memory [334, 335] and speed of processing [335, 336]. There is a possible relationship between hearing loss and cognitive decline as a possible precursor to dementia and Alzheimer's disease [64, 65].

The large multidimensional data set contained highly correlated variables, a challenge requiring variable reduction for improved predictions. Variable reduction for hearing thresholds has involved cluster analyses [337, 338]. Our variable reduction approach involved multiple hearing threshold composite measures from 0.25–14 kHz. Evaluating extended high-frequency thresholds between 8–14 kHz has become increasingly common because they appear to predict speech perception in noise [80, 339].

Here we determine variables predictive of speech perception to verify previously identified associations in the literature and identify possible novel associations. The goal is to produce hypotheses for future studies.

6.2 Methods

6.2.1 Participants

We recruited 443 adult participants (293 female, 125 male, 25 no response) for a separate study [333]. They had a range of ages (18–85 y, $\bar{x}$ = 59.7, SD = 22.5) and hearing loss. All participants had a rigorous audiometric evaluation and cognitive assessment done. A subset of 130 participants performed additional speech perception and non-speech auditory temporal processing tasks, and participated in an electrophysiological assessment.

6.2.2 Audiometric Tests

Hearing thresholds were assessed at octave intervals between 0.125 and 8 kHz using standard audiometric techniques and a calibrated audiometer in a sound-attenuating booth. Audiometric thresholds were also collected at 3, 6, 10, 12.5, and 14 kHz, and tympanometry was performed in each ear. Many participants had audiometrically normal hearing, defined as pure-tone thresholds $\leq$ 25 dB hearing level (HL) [340] at 0.25, 0.5, 1, 2, and 4 kHz in the right ear. There were 100 younger normal-hearing listeners with a high-frequency pure-tone average (HF-PTA: 1, 2, 4 kHz) = 7.3 $\pm$ 12.3 dB HL and 187 older normal-hearing listeners with a HF-PTA = 24.4 $\pm$ 17.2 dB HL. Some older listeners (n = 120) had a mild-to-moderate hearing loss [defined as a HF-PTA $>$ 30 dB HL and thresholds at 2 kHz and 4 kHz $<$ 70 dB HL], with an average HF-PTA = 40.9 $\pm$ 8.0 dB HL.

6.2.3 Perception, Electrophysiological, and Cognitive Tests

Temporal processing was assessed for several psychoacoustical tasks, including pulse-rate discrimination, gap detection, gap duration discrimination, and tempo discrimination [333]. Auditory brainstem responses were recorded to 100-μs broad-band clicks. Auditory steady-state responses (ASSR) were recorded to band-limited pulse trains presented at rates of 100, 200, 300, and 400 Hz [333]. Sentence recognition in quiet was measured using the IEEE corpus [341] in five conditions: a normal rate with no reverberation (i.e., clean speech), with 40% and 60% time compression, and with reverberation (0.6-s and 1.2-s RT_{60}). There were 10 sentences presented in each condition. Sentence recognition in noise was measured using the QuickSIN [342]. Word recognition scores were collected for single 25-word lists of the NU-6 test [343] presented bi-

laterally at 75 dB HL in quiet. The Montreal Cognitive Assessment (MoCA) [195], assessments from the National Institutes of Health Cognition Toolbox (List Sorting Working Memory Test, the Flanker Inhibitory Control and Attention Test, the Pattern Comparison Processing Speed Test, and the Dimensional Card Sort Test) [344], and a subset of the Speech, Spatial, and Qualities of Hearing Scale [345] were administered.

6.2.4 Machine Learning Analysis

We derived a total of 147 calculated features from the audiometric measurements. We recorded HL of each audiogram and computed new features based on the mean and difference between the two ears for each frequency. Using these features, we computed profile variables by applying a simple linear model to each participant: slope, intercept, coefficient of determination of the line of best fit, and sum of hearing thresholds across frequencies. We calculated these profile features for three different frequency ranges: standard frequencies (SF; 2.5–8 kHz), extended high frequencies (EHF; 10–14 kHz), and all frequencies.

We extracted additional features from the audiograms of both ears including total length of segments joining consecutive thresholds, highest frequency each subject can hear, inflection point, notch index [346], and a notch presence using to three definitions. One type of notch definition was a point at 3, 4, or 6 kHz where there are differences of at least 15 dB and 10 dB compared to the previous and next frequencies, respectively. The two other definitions were described by others [347, 348]. Finally, we added features that describe the symmetry, configuration, and severity of an audiogram [349].

Some subjects did not respond at the highest-intensity limits of the equipment, indicat-

ing a profound hearing loss at that frequency (typically at extended high frequencies > 8 kHz). To reflect this hearing loss, we replaced the missing hearing thresholds with +5 dB above the maximum intensity level that our equipment allows at each frequency, as is typical in other publications [350]. This way, the hearing thresholds reflect the profound degree of the true thresholds but remain distinguishable from those who actually heard the signals at the equipment limits. For the other variables, missing categorical values were imputed by using the most frequent category, and the numeric values were replaced with variable medians. All features are grouped into the following seven categories: subject information, audiometric measures, behavioral-speech tests, behavioral-non-speech tests, electrophysiological measures, cognitive tests, and subjective questionnaires. We reduced the number of features by selecting those that minimized the total Akaike information criterion based on the results of random forest analyses.

We built super learner [214] stacked generalization models with an ensemble of methods using SuperLearner package [351] with ten-fold cross-validation and non-negative least square error as a loss function. For base models, we included random forests, lasso and elastic-net regularized generalized linear models, extreme gradient boosting, and feed-forward neural networks. To evaluate the performance of the super learner model, an additional layer of 5-fold cross-validation was performed and the whole process was repeated five times. The performance measure (percent variance explained) was then averaged across all folds and repeats. All analyses for this study were carried out in the Statistical Computing Programming Language R (Version: 4.1.3) [352].

Model performance was measured as the percentage of variance explained, which is defined as the fraction of the variance of the response variable that can be explained using the predictors. Permutation-based feature importance was calculated from 25 replications and reported

as relative values such that the maximum score has a value of 100.

6.3 Results

We used super learner to model the scores of three different speech-perception tests: QuickSIN speech-in-noise, sentence recognition with 60% time compression, and sentence recognition with 1.2-s reverberation time. To eliminate the dependence between different behavioral-speech features, we excluded all other features that fall under the same category before performing feature elimination and building the super learner model. For simplicity, the plots are restricted to features with average relative importance > 15 (an arbitrarily chosen value). Group variable importance analysis selected the audiometric category as the most important predictor for the scores of the three speech-perception tests.

The proportion of variance in QuickSIN test scores explained was 66.91%, and 10 of 12 important features were from the audiometric category including (ordered by importance) HL of the right ear at 1 kHz, HL of the left ear at 3 kHz, minimum threshold of the right ear at the mid-frequency range (1–3 kHz), mean and maximum of thresholds of the left ear at the mid-frequency range, HL of the right ear at 0.5 and 2 kHz, mean of thresholds of the right ear at the mid-frequency range, mean of HL of two ears at 2 kHz, and total length of segments joining consecutive left ear thresholds (Figure 6.1). The other two were from the behavioral-non-speech category (i.e., basic temporal processing), which includes the average of pulse-rate discrimination values across frequencies (100–400 Hz) and at the single pulse rate of 300 Hz.

The proportion of variance in 60% time compression sentence recognition scores explained was 74.47%, and 5 of 6 important features were from the audiometric category including (ordered

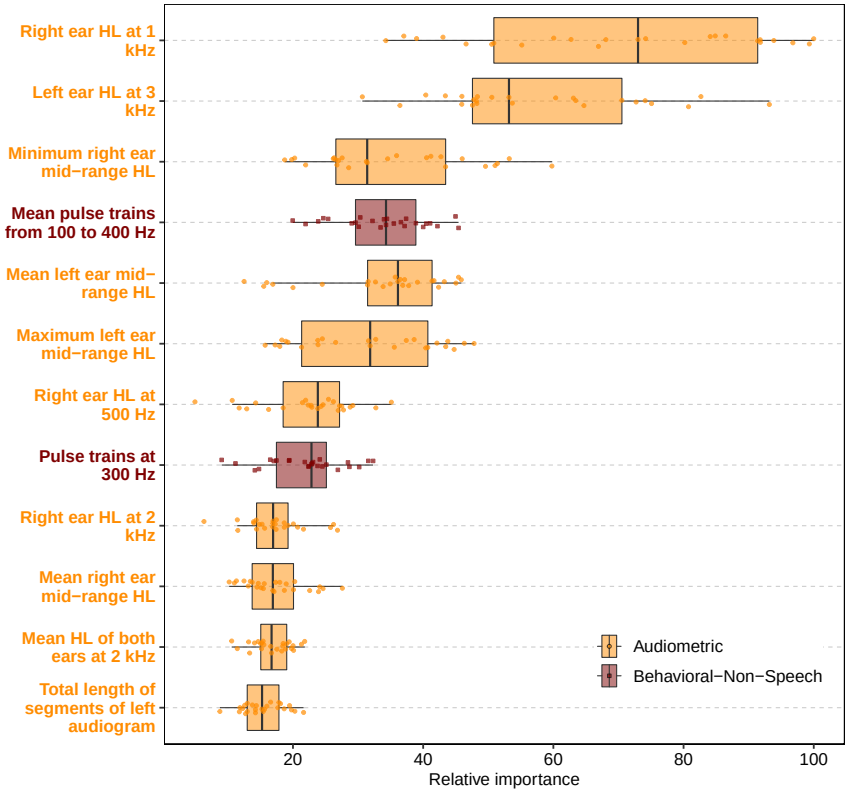


Figure 6.1: Variable importance for predicting QuickSIN Speech-in-Noise. Points represent importance scores from 25 replications. Box plots indicate median (middle line), 25^{th}, 75^{th} percentile (box). Whiskers extend 1.5 $\times$ interquartile range.

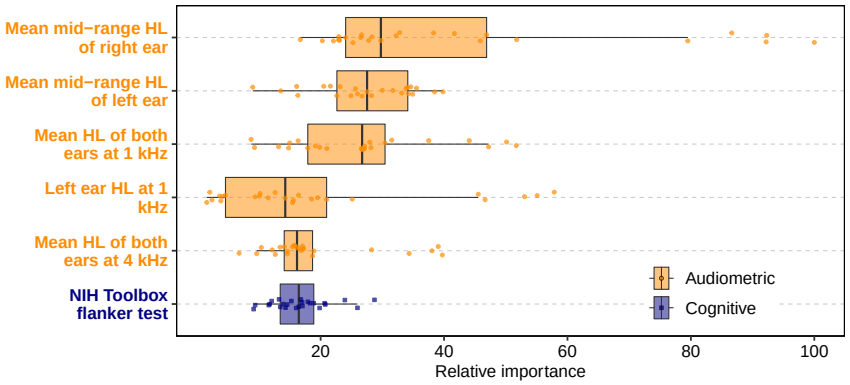


Figure 6.2: Variable importance for predicting sentence recognition with 60% time compression. Points represent importance scores from 25 replications. Box plots indicate median (middle line), 25^{th}, 75^{th} percentile (box). Whiskers extend 1.5 $\times$ interquartile range.

by importance) mean of thresholds of the right ear at the mid-frequency range, mean of thresholds of the left ear at the mid-frequency range, mean of HL of two ears at 1 kHz, HL of the left ear at

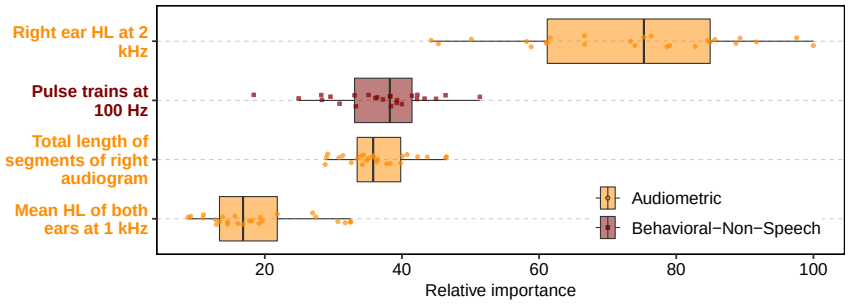


Figure 6.3: Variable importance for predicting sentence recognition with 1.2-s reverberation time. Points represent importance scores from 25 replications. Box plots indicate median (middle line), 25^{th}, 75^{th} percentile (box). Whiskers extend 1.5 $\times$ interquartile range.

1 kHz, and mean of HL of two ears at 4 kHz (Figure 6.2). The last feature was the Flanker test (inhibitory control and attention cognitive test).

The proportion of variance in 1.2-s reverberation time sentence recognition scores explained was 71.77%, and 3 of 4 important features were from the audiometric category including (ordered by importance) HL of the right ear at 2 kHz, total length of segments joining consecutive right ear thresholds, and mean of HL of two ears at 1 kHz (Figure 6.3). The last feature was from the behavioral-non-speech category, pulse-rate discrimination for 100-Hz pulse trains.

Information about model parameters, variance explained, model weight, R code and models are available in the supplementary material at `https://osf.io/ta7kf`.

6.4 Discussion

We used machine learning to determine variables predictive of demanding speech-perception conditions to verify previously identified associations and identify possible novel associations, producing hypotheses for future studies. We also developed new variables that had significant contributions to the variance explained by our models, such as the total length of segments join-

ing consecutive thresholds (Figures 6.1 and 6.3). Such variables capture properties of the entire hearing range, rather than single frequencies, showing the importance of using composite variables to reduce the audiometric variable space.

Although others have used clustering and ensemble-based approaches to examine communication difficulties in relation to audiometric profiles [83, 353], our results suggest that the largest amount of speech-perception variance is explained by audibility from 1–4 kHz, followed by temporal processing ability, and attention (Figures 6.1, 6.2, and 6.3). Audiometric variables were the most predictive of demanding speech perception, particularly for variables that were associated with thresholds between 1–4 kHz (e.g., the mid-range composite variables; Figures 6.1 and 6.2), consistent with previous studies [354]. Conspicuous in their absence were extended high-frequency (>8 kHz) threshold variables. It could be that the importance of extended high-frequency thresholds is diminished by hearing loss around 1–4 kHz to a negligible level. Similarly, hearing loss can obscure effects of aging, and one needs to carefully design their study to include groups that vary both by hearing loss and age [333] to see effects of the relatively weaker factor of age.

We found a limited role of cognitive processing abilities in explaining the variance in speech-perception performance. The only cognitive factor to appear in our results was the Flanker score, which is a measure of inhibition of irrelevant stimuli and attention; this factor was the second most important factor in a meta-analysis [335] and was negatively correlated with cortical envelope in older listeners in [355]. Absent was working memory and processing speed, which was unexpected given past studies [333, 335, 336]. Noteworthy is that the MoCA, which is prevalent in its use as a cognitive screener in the field of hearing and aging, had no predictive power for our outcome variables. In fact, in a separate analysis (not shown) the combined MoCA score

or any individual question was not predictive of any variable in this dataset.

These findings may suggest that researchers need to be cautious in future study design when investigating extended high-frequency threshold and cognitive effects on auditory processing like speech perception. Controlling for more prominent variables like hearing thresholds appears critical.

6.5 Acknowledgments

Funding was provided by a University of Maryland Brain and Behavior Institute Seed Grant (MPC, MJG), award T32-DC000046 from the National Institute on Deafness and Other Communication Disorders (trainee BEGII), and award P01AG055365 from the National Institute of Aging of the National Institutes of Health (SGS).

Part IV

Future Work

Chapter 7: Future Work

7.1 Movement Disorders

7.1.1 Larger Sample Size

This dissertation analyzed data from 260 participants with PD, 50 age-matched controls, and 18 participants with non-PD parkinsonian disorders.

Differentiating mildly affected PD patients from older adults without PD remains a clinical challenge, making wearable sensors valuable as supportive biomarkers. Expanding the sample size, particularly for these groups, could enable the development of more robust models that better capture movement characteristics, thereby improving classification accuracy and reliability.

The initial study design aimed for approximately 50 participants per H&Y stage, with healthy controls categorized as "H&Y 0". However, actual enrollment varied due to the prevalence of PD patients across different stages at the UM-MMDC clinic. Expanding the control group size would create a more balanced dataset, improve model generalization, and enhance PD vs. control classification performance beyond the results presented in Chapter 3.

Similarly, the small sample size of the non-PD parkinsonism group in Chapter 5 limits the generalizability of the findings. Future research should validate these results in larger cohorts to strengthen their robustness. Additionally, the limited sample size restricted the ability to address

heterogeneity by developing a multiclass classifier that could distinguish between different forms of parkinsonism.

7.1.2 Non-motor Features for PD Identification

Beyond the typical motor symptoms, PD also presents with non-motor features such as cognitive impairment and autonomic dysfunction, including orthostatic hypotension. These non-motor symptoms were not assessed in control participants in this study and, therefore, were not incorporated into the machine learning models. Including these features in future analyses could enhance model accuracy in predicting and identifying PD, potentially improving diagnostic capabilities and clinical applications.

7.1.3 Longitudinal Analysis

Objective and detailed assessments of disease manifestation through wearable sensors hold promise as superior markers of disease severity and progression. However, validating this hypothesis requires longitudinal studies. Currently, longitudinal analysis of a subgroup of the study cohort is ongoing.

7.1.4 Impact of Confounding Factors on Classification Performance

In Chapter 3, the analysis of misclassified participants revealed that certain variables, likely related to the level of medical comorbidity (e.g., age), may influence sensor data and classification outcomes. Specifically, older controls were incorrectly classified as PD patients, and PD participants with relatively mild motor symptoms (low UPDRS_PIII motor score) were misclas-

sified as controls. To improve the accuracy and reliability of PD classification, it is essential to account for these factors and develop more comprehensive models that stratify participants based on such variables.

Similarly, in Chapter 5, the differences in clinical measures such as UPDRS_PIII, H&Y scores, and age between correctly classified and misclassified PD and non-PD participants suggest that stratifying the model by one or more of these characteristics could enhance performance. However, due to the small sample size in this study, this stratification approach was not feasible.

7.1.5 Unsupervised Analysis of Sensor-based Movement Data

Applying unsupervised analysis techniques to wearable sensor data from movement disorder motor tasks offers a valuable approach to exploring and understanding the underlying patterns and characteristics of movement in individuals with the disease. Unsupervised analysis methods, such as clustering, allow for the identification of hidden structures and similarities within the sensor data without relying on predefined labels or annotations. Especially, each participant can be characterized by multiple measures, encompassing time-domain and frequency-domain features derived from sensor data, with each measure representing a distinct dimension. Clustering algorithms like k-means, hierarchical clustering, density-based clustering, and fuzzy clustering can be applied with varying parameters to find the optimal configurations for the data. Once the clusters are identified, dimensionality reduction techniques such as PCA (principal component analysis) [356], t-SNE (t-distributed stochastic neighbor embedding) [357], MDS (multidimensional scaling) [358], and UMAP (uniform manifold approximation and projection) [359] can be used to project the high-dimensional data into a lower-dimensional space for visualization. This

visualization might provide insights into the underlying patterns and structures within the dataset.

The goal of this analysis is to investigate the presence of clusters within the data and ascertain the defining characteristics of these clusters if any exist. The clusters may correspond to varying levels of movement impairment, which could be reflected in clinical scores like H&Y or UPDRS (motor-part III) scores. Furthermore, the clusters can be analyzed as distinct classes to investigate which features and mobility tasks most influence the differentiation between them and to reveal the key factors that contribute to the classification of the participants.

7.1.6 WearGait-PD Dataset

The FDA, in collaboration with the U.S. Department of Veterans Affairs (VA) and Johns Hopkins School of Medicine, has established an open-access dataset containing raw inertial measurement unit (IMU) and sensorized insole data synchronized to a gait walkway reference system. The dataset, which can be accessed here `https://cdrh-rst.fda.gov/weargait-pd-wearables-dataset-gait-parkinsons-disease-and-age-matched-controls`, includes data from 61 individuals with PD and 65 age-matched controls. The data includes readings from 13 IMU sensors placed at various body locations, sensorized insole data capturing absolute pressure from 16 sensors in each insole, and walkway data with over 18,000 sensors that provide 2D position and relative pressure during each footfall. Participants performed six walking/gait tasks, three trials of the TUG, and six balance tasks.

This dataset offers a rich resource with three different devices, varied sensor positions, and multiple tasks. The analyses conducted in this dissertation could also be applied to this dataset for comparison. Possible analyses include building a classifier to distinguish PD from controls,

predicting PD severity, analyzing misclassified participants, and exploring simplification techniques through group feature importance analysis to identify the most important sensor devices, tasks, and sensor placements.

7.1.7 Genetic Markers

Identifying candidate genetic markers associated with movement disorders, including PD and other parkinsonian disorders, is crucial for advancing our understanding of their genetic foundations and improving diagnostic and therapeutic approaches. Genome-wide association studies (GWAS) can be used to uncover genetic variants linked to PD susceptibility, progression, and severity. A deeper understanding of these genetic markers holds promise for enhancing early diagnosis, enabling personalized treatment strategies, and developing targeted therapies. Currently, genetic data from a subgroup of our study cohort is available, and efforts are underway to extend this data to all participants.

7.2 Human Hearing

7.2.1 Inherent Biases in Audiological Thresholds and Standards

While working with the hearing dataset used in the analysis of Chapter 6, and collaborating with the hearing team, we noticed some inherent biases in audiological thresholds and standards. After conducting a literature review, we identified several issues with current audiological practices and data representation. One of the key problems is the conceptualization of hearing capability as "hearing loss", which leads to a narrow and often exclusionary view of what constitutes normal hearing. This model treats hearing as an idealized concept, measuring individ-

uals' abilities against a "perfect" standard, with deviations from this ideal defined as hearing loss. This approach does not adequately account for the natural variation in hearing across individuals. For instance, similar to how height or weight is measured, hearing could be better understood as a range of values, but audiology often fails to accommodate these variations, leading to inaccurate and limited representations of hearing abilities in the data.

Moreover, many audiological studies exclude individuals based on specific criteria, contributing to survivorship bias. For example, individuals with otological diseases or those outside the "normal" hearing range are often excluded, distorting the representation of hearing capabilities in the general population. The current reliance on the audiogram as the standard tool also fails to capture the full spectrum of hearing, especially as it often involves unilateral assessments. This focus neglects the crucial aspect of binaural hearing, leaving out significant insights into hearing asymmetry. Additionally, traditional audiometric standards do not sufficiently reflect the differences in hearing abilities related to demographic factors such as age, sex, and race, which are known to influence hearing outcomes. By excluding certain groups or not accounting for these factors, audiology fails to generate accurate population standards or provide a reliable understanding of hearing.

To address these shortcomings, we propose alternative methods to the traditional audiogram, such as using contour plots and percentiles to represent hearing data. These methods would allow for a more inclusive and detailed representation of hearing across a wider range of abilities. Moreover, it is critical to include demographic factors like sex, gender, age, and race in the interpretation of audiological data, as these factors significantly affect hearing outcomes. This shift towards more inclusive practices would not only enhance the accuracy of audiological assessments but also provide a more equitable framework for understanding hearing in diverse

populations. By adopting robust statistical methods, such as percentiles, audiology could present hearing data in a way that accurately reflects the full range of human hearing abilities.

Currently, progress is ongoing toward a manuscript that discusses these current pitfalls and suggests possible solutions for a more inclusive and representative approach to audiology.

7.2.2 Left-Right Ear Asymmetry

In Chapter 6, the feature importance analysis revealed interesting findings regarding the prediction of speech perception tests. Typically, one would expect the important audiometric features to be predominantly from the same ear as the tests were performed (right ear). However, the results indicate that there is no significant distinction in the ranking of the right and left ear features. The Wilcoxon Rank-Sum tests [360] comparing the rankings of various speech perception tests, such as QuickSIN speech-in-noise, sentence recognition with 60% time compression, and sentence recognition with 1.2-s reverberation time, yielded p-values of 0.26, 0.91, and 0.13 respectively. These findings suggest the need to delve deeper into understanding the observed differences between the two ears leading to binaural hearing deficits and posing difficulties for people communicating in everyday environments with competing background sounds and noise.

Left-right ear asymmetry can arise from various sources, but one of the most prevalent causes is exposure to noise, such as during shooting activities [361–363]. In the case of right-handed individuals firing a rifle or shotgun, their left ear is in closer proximity to the muzzle of the firearm barrel, while the right ear is naturally protected by the head. Furthermore, studies have shown that individuals exposed to occupational noise also exhibit a similar hearing asymmetry, with the left ear being more affected than the right ear [364–366]. In light of these observations,

possible steps can be taken to study the phenomenon of ear asymmetry and its implications for hearing.

Change from the right and left audiometric variables to better and worse. An investigation should be done here to find the best definition of the better ear.

Build supervised models with only subsets of hearing level values including only left ear, right ear, better ear, worse ear, handed ear (i.e., right-handed people use right ear), opposite-handed ear (i.e., right-handed people use left ear), mean of both ears, right − left ear difference, better − worse ear difference. Feature importance analysis can then be done to examine the most important features of each model.

Study if the asymmetry between the two ears is related to age and hearing loss.

Design a new study where speech perception tests are performed in both ears.

Analyze more accessible/open-access datasets that measure speech perception tests in both ears.

So, though left-right hearing asymmetry can be observed, which has been shown in previous research [367], what will be new is the consideration of better and worse ears, implications for suprathreshold measures, and the use of machine learning.

Part V

Appendix

Appendix A: Supplementary Material for Simplification of Mobility Tests and Data Processing to Increase Applicability of Wearable Sensors as Diagnostic Tool for Parkinson's Disease

Table A.1: Number of sensors and mobility tasks used in a survey of literature from 2017 to 2024.

Objective	Authors	No. of sensors	No. of mobility tasks
PD diagnosis	Schlachetzki et al. [119]	2	1
	Krupicka et al. [120]	2	3
	Abtahi et al. [121]	17	8
	Castaño-Pino et al. [122]	18	1
	Phan et al. [123]	4	4
	Del et al.Din et al. [124]	1	4
	Rovini et al. [125]	1	4
	Rehman et al. [126]	2	2
	Buckley et al. [127]	3	1
	Cai et al. [128]	10	1
	Jovanovic et al. [129]	16	1
	Battista et al. [130]	1	4
	Yue et al. [131]	5	12
	Bailo et al. [132]	1	1
	Nair et al. [133]	16	1
	Keloth et al. [134]	2	3
	Trabassi et al. [135]	1	1
	Marin et al. [136]	1	5
	Gourrame et al. [137]	6	1
	Meng et al. [138]	11	1
	Caramia et al. [139]	8	1
Quantifying/monitoring the	Mahadevan et al. [140]	1	14
severity of PD symptoms	Lin et al. [141]	1	8
	Borzì et al. [142]	1	1

	Ymeri et al. [143]	1	5
	Bobic et al. [144]	2	1
	Han [145]	2	1
	Heijmans et al. [146]	6	1
	Wu et al. [147]	6	1
	Wu et al. [148]	6	1
	Pugh [149]	6	1
	Sotirakis et al. [110]	6	2
	Delrobaei et al. [150]	17	2
	Memar et al. [151]	17	2
	di Biase et al. [152]	5	3
	Delrobaei et al. [153]	17	3
	Singh et al. [154]	4	5
	Prakash et al. [155]	4	5
	Bremm et al. [156]	2	6
	Huo et al. [157]	8	6
	Yan et al. [158]	8	6
	Hssayeni et al. [56]	2	8
	Zajki-Zechmeister et al. [159]	2	8
	Lonini et al. [44]	6	13
	Tsakanikas et al. [160]	5	5
Detecting freezing of gait in PD	Chomiak et al. [161]	1	4
	Punin et al. [162]	1	3
	Li et al. [163]	1	3
	Camps et al. [164]	1	3
	Samà et al. [165]	1	4
	Capecci et al. [166]	1	3
	Reches et al. [167]	3	1
	Pham et al. [168]	3	1
	Masiala et al. [169]	3	3
	Mancini et al. [170]	8	2
	Marcante et al. [171]	13	5
	Sigcha et al. [172]	1	7
	Bikias et al. [173]	1	3
	Pardoel et al. [174]	4	2
	Shi et al. [175]	3	2
Quantitative characterization of gait/	Rennie et al. [176]	1	3
identifying imbalance in PD	Myers et al. [177]	1	3
	Zadka et al. [178]	1	3
	Haji et al.Ghassemi et al. [179]	2	1
	Nguyen et al. [180]	2	1
	Castelli Gattinara Di Zubiena et al. [111]	3	1
	Liuzzi et al. [181]	3	1

Stack et al. [182]	5	7
Jehu et al. [183]	7	2
Zhang et al. [184]	8	2
Romijnders et al. [185]	2	3
Lukšys et al. [186]	6	1

Table A.2: Frequency domain features. The power spectrum is obtained using the Discrete Fourier Transform (DFT) of the signal with a fast algorithm, the Fast Fourier Transform (FFT). Variables: f_j is the frequency of the spectrum at frequency bin j; P_j is the power spectrum at frequency bin j; and M is the length of the frequency bin.

Name	**Description**	**Mathematical equation**	**References**
MNF	Mean frequency	$\sum_{j=1}^{M} f_j P_j \ / \ \sum_{j=1}^{M} P_j$	[37, 368, 369]
MDF	Median frequency	$\frac{1}{2} \sum_{j=1}^{M} P_j$	[37, 368, 369]
PSD	Power spectrum deformation	$\frac{\sqrt{\frac{M_2}{M_0}}}{\frac{M_1}{M_0}}$; $M_n = \sum_{j=1}^{M} f_j P_j$	[369, 370]
FI	Freeze index	$\sum_{j=3}^{8} P_j \ / \ \sum_{j=0.5}^{3} P_j$	[37, 371]
ENT	Entropy	$-\sum_{j=1}^{M} Pr_j log(Pr_j)$; $Pr_j = \frac{P_j}{\sum_{j=1}^{M} P_j}$	–
TTP	Total power	$\sum_{j=1}^{M} P_j$	[37, 368]
MNP	Mean power	$\sum_{j=1}^{M} P_j \ / \ M$	[37, 368]
PKF	Peak frequency	$f_{argmax(P_j)}$, $j=1...M$	[37, 368]
PKF	Peak frequency	$max(P_j)$, $j = 1...M$	[372]
FR	Frequency ratio	$\frac{max(f_j)}{min(f_j)}$, $j = 1...M$	[37, 368]
PSR	Power spectrum ratio	$\frac{P_0}{P} = \sum_{f_0-n}^{f_0+n} P_j \ / \ \sum_{j=-\infty}^{\infty} P_j$ f_0: value of PKF, n: integral limit.	[368]
SM	Spectral moments	$SM_i = \sum_{j=1}^{M} f_j^i P_j$, $i = 1, 2, 3$	[37, 368, 370]
VR	Variance	$\frac{1}{M-1} \sum_{j=1}^{M} (P_j - MNP)^2$	–
SD	Standard deviation	$\sqrt{\frac{1}{M-1} \sum_{j=1}^{M} (P_j - MNP)^2}$	–

SS	Skewness	$\frac{1}{M-1}\sum_{j=1}^{M}(P_j - MNP)^3 \,/\, SD^3$	–
SK	Kurtosis	$\frac{1}{M-1}\sum_{j=1}^{M}(P_j - MNP)^4 \,/\, SD^4$	–
SBW	Spectral bandwidth	$\frac{\sum_{j=1}^{M}(f_j - MNF)^2 P_j}{TTP}$	–
SR	Spectral roll-off	$c\sum_{j=1}^{M} P_j$	–
VCF	Variance of central frequency	$\frac{SM_2}{SM_0} - \left(\frac{SM_1}{SM_0}\right)^2$	[368]
MSE	Mean spectral energy	$\sum_{j=1}^{M} \lvert P_j \rvert^2 \,/\, M$	[208, 373]
DPR	Maximum to minimum drop in power density	$\frac{\text{highest mean power density value}}{\text{lowest mean power density value}}$	[369, 370]
SN	Signal to noise ratio	$\frac{\sum P_j \text{ in the upper 20\% frequency range}}{TTP}$	[369, 370]
HR	Harmonic ratio	$\frac{\sum \text{amplitudes of even harmonics}}{\sum \text{amplitudes of odd harmonics}}$	[127]

Table A.3: Time domain features. Variables: x_n represents n^{th} sample of the signal; and N is the length of the signal.

Name	**Description**	**Mathematical equation**	**References**
MN	Mean	$\frac{1}{N}\sum_{n=1}^{N} x_n$	[59, 118, 369]
VR	Variance	$\frac{1}{N-1}\sum_{n=1}^{N}(x_n - MN)^2$	[369]
SD	Standard deviation	$\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_n - MN)^2}$	[59, 118, 369]
SS	Skewness	$\frac{1}{N-1}\sum_{n=1}^{N}(x_n - MN)^3 \,/\, SD^3$	[59, 118, 369]
SK	Kurtosis	$\frac{1}{N-1}\sum_{n=1}^{N}(x_n - MN)^4 \,/\, SD^4$	[59, 118, 369]
IAV	Integrated absolute value	$\sum_{n=1}^{N} \lvert x_n \rvert$	[37, 368]
MAV	Mean absolute value	$\frac{1}{N}\sum_{n=1}^{N} \lvert x_n \rvert$	[37, 368]
SSI	Simple square interval	$\sum_{n=1}^{N} \lvert x_n \rvert^2$	[37, 368]

RMS	Root mean square	$\sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2}$	[37, 368]
V3	V-order 3	$\sqrt[3]{\frac{1}{N}\sum_{n=1}^{N} \lvert x_n\rvert^3}$	[37, 368]
WL	Waveform length	$\sum_{n=1}^{N-1} \lvert x_{n+1} - x_n\rvert$	[37, 368, 369]
AAC	Average amplitude change	$\frac{1}{N}\sum_{n=1}^{N-1} \lvert x_{n+1} - x_n\rvert$	[37, 368]
DASDV	Difference absolute standard deviation value	$\sqrt{\frac{1}{N-1}\sum_{n=1}^{N-1}(x_{n+1} - x_n)^2}$	[37, 368]
MFL	Maximum fractal length	$log_{10}\left(\sqrt{\sum_{n=1}^{N-1}(x_{n+1} - x_n)^2}\right)$	[37, 374]
ZC	Zero crossing	$\sum_{n=1}^{N-1} sgn(x_n * x_{n+1}) \cap \lvert x_n - x_{n+1}\rvert \geq threshold$ $sgn(x) = 1 \quad if\ x \geq threshold;\ \ 0\ \ o.w$	[368, 369]
RC	Rate of change	$\sum_{n=1}^{N-1} f\lvert x_n - x_{n+1}\rvert$ $f(x) = 1 \quad if\ x \geq threshold;\ \ 0\ \ o.w$	[368, 369]
SSC	Slope sign change	$\sum_{n=2}^{N-1} f[(x_n - x_{n-1}) * (x_n - x_{n+1})]$ $f(x) = 1 \quad if\ x \geq threshold;\ \ 0\ \ o.w$	[368, 369]
DR	Data range	$max(x_n) - min(x_n)\ ;\ n = 1...N$	[118]
ENT	Entropy	$-\sum_{b=1}^{numBins} Pr_b log(Pr_b)\ ;\ numBins = 10$	[118]
LOG	Log detector	$e^{\frac{1}{N}\sum_{n=1}^{N} log\lvert x_n\rvert}$	[368]
MAD	Mean absolute deviation	$\frac{1}{N}\sum_{n=1}^{N} \lvert x_n - MN\rvert$	[369]
Q1, Q2, Q3	1st, 2nd, 3rd quartiles		[118]
IQR	Interquartile range	$Q_3 - Q_1$	[59, 118]
CV	Coefficient of variation	SD/MN	[375, 376]
MD	Median		[118]
MOD	Mode		[118]

TKEO	Teager-Kaiser energy operator	$mean(x_n^2 - (x_{n-1} * x_{n+1}))$; $n = 1...N$	[118, 377]
AR	Auto-regressive coefficients	$x_i = \sum_{p=1}^{P} a_p x_{i-p} + w_i$ $P = 4$, a_i : AR coefficient , w_i : noise term	[368, 369]
DFA	Detrended fluctuation analysis	$y_k = \sum_{t=1}^{k} x_n - MN$; $k = 1...N$ $F_{s,t} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (y_k - y_{k,s})^2}$; $s = 1...N_t$, $N_t = N/t$, t : num of nonoverlapping segments , $y_{k,s}$: fitting polynomial in segment s $DFA = F(t) = \left[\frac{1}{2N_t} \sum_{s=1}^{2N_t} F_{s,t}^2\right]^{1/2}$ $X_k^n = X(n + ik)$; $i \in [0, [(N - n)/k]]$, $n \in [1, k]$, n : initial time , k : time interval $L_n(k) = \frac{\left(\sum_{i=1}^{\left[\frac{N-n}{k}\right]} \left\lvert X(n+ik) - X(n+(i-1)k)\right\rvert\right) \frac{N-1}{\left[\frac{N-n}{k}\right]k}}{k}$	[118, 374, 378]
HFD	Higuchi's fractal dimension	$L_n(k)$: length of the curve X_k^n $\langle L(k) \rangle \propto k^{-D}$ $\langle L(k) \rangle$: length of the curve for time interval k $= mean(L_n(k))$ over k sets , D: fractal dimension , [ ]: Gauss' notation $\frac{log(L/a)}{log(d/a)}$	[374, 379]
KATZ	Katz's fractal dimension	L : $\sum$(distances b/w successive points) a : avg(distances b/w successive points) d : max(distances b/w first and other points)	[380]

MAV1	Modified mean absolute value type 1	$\frac{1}{N}\sum_{n=1}^{N} w_i \lvert x_n \rvert$ $w_i = 1 \quad if\ 0.25N \leq i \leq 0.75N;\ \ 0.5 \ \ o.w$ $\frac{1}{N}\sum_{n=1}^{N} w_i \lvert x_n \rvert$	[368]
MAV2	Modified mean absolute value type 2	$w_i = 1 \quad if\ 0.25N \leq i \leq 0.75N;$ $4i/N \quad if\ i < 0.25N;\ \ 4(i-N)/N \ \ o.w$	[368]
MAVS	Mean absolute value slope	$MAV_{k+1} - MAV_k\ ;\ k = 1...K-1\ ,\ K = 2$	[368]
MBV	Mean binarized values	$\frac{1}{N}\sum_{n=1}^{N} f(x)$ $f(x) = 1 \quad if\ x \geq threshold;\ \ 0 \ \ o.w$	[368]
TM4	Absolute temporal moment	$\frac{1}{N}\sum_{n=1}^{N} x^4$	[368]
VFD	Variation fractal dimension	$\widehat{V}_p(l/N) = \frac{1}{2(N-l)}\sum_{i=l}^{N} \lvert X_{i/N} - X_{(i-l)/N} \rvert^p\ ,$ $\widehat{V}_p =$ moment estimator of order $p\ ;\ \ l = 1, 2, ..$ $\widehat{D}_{V;p} = 2 - \frac{1}{p}\frac{log(\widehat{V}_p(2/N)) - log(\widehat{V}_p(1/N))}{log 2}$ $\widehat{D}_{V;p}$: estimator for the fractal dimension	[381]
MAX	Maximum	$max(x_n)\ ;\ n = 1..N$	[382]
MIN	Minimum	$min(x_n)\ ;\ n = 1..N$	[382]
GMN	Geometric mean	$\left(\prod_{n=1}^{N} x_n\right)^{1/N}$	–
HMN	Harmonic mean	$N/\sum_{n=1}^{N}(1/x_n)$	[383, 384]
MDAD	Median absolute deviation	$median(\lvert x_n - MD \rvert)$	[382]

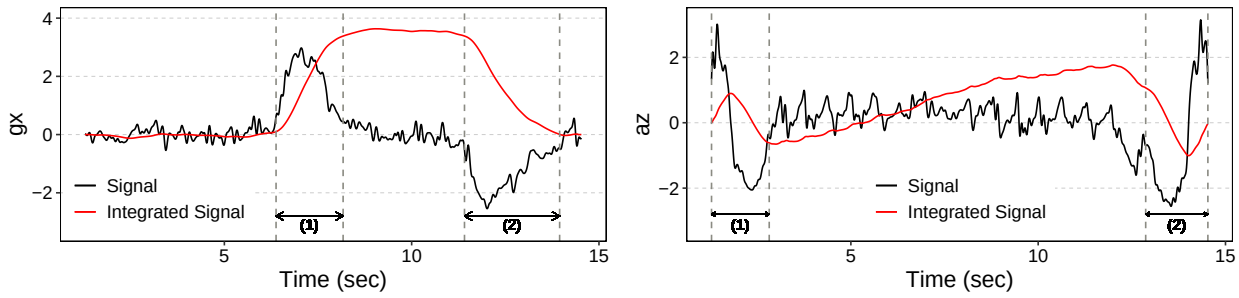
		$MSE(x, \tau, m, r) = SampEn(y_1^{(\tau)}, m, r)$	
		$y_{k,j}^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+k}^{j\tau+k-1} x_i \ , \ 1 \leq j \leq \frac{N}{\tau} \ , \ 1 \leq k \leq \tau$	
MSENT	Multiscale entropy, scale factor = 4, 5	$SampEn(x, m, r) = -\ln \frac{n^{m+1}}{n^m}$ τ: scale factor m: embedding dimension, r: tolerance n^{m+1}: # of $(m+1)$ dimensional vectors within r n^m: # of m dimensional vectors within r	[385]

Table A.4: Cross-time domain features.

Name	Description	Mathematical equation	References
ENT	Cross-entropy	$H(f, g) = -\int_t f(t) \ \ log(g(t)) \ dt$	–
CCP	Cross-correlation	$(f \star g)(t) = \int_{-\infty}^{\infty} \overline{f}(\tau) \ \ g(t+\tau) \ d\tau$ $\overline{f}(\tau)$ is the complex conjugate of $f(\tau)$	[386]
MI	Mutual information	$MI(f; g) = \int_t \int_t P_{(f,g)} log\left(\frac{P_{(f,g)}}{P_f P_g}\right) dt \ dt$ $P_{(f,g)}$: joint probability mass function of f and g P_f, P_g : marginal probability mass functions of f and g	[387]

Table A.5: Features selected by the feature selection approach for TUG-only and cogTUG-only models using features from the first trial, the second trial, and the mean of the corresponding features from both trials.

	TUG-only	cogTUG-only
PD vs controls	kurtosis of DFT of sit-to-stand (TUG1, ay), detrended fluctuation analysis of sit-to-stand (TUG1, gz), mean binarized values of stand-to-sit (TUG1, ax), maximum to minimum drop in power density of second turn (TUG1, gy), mean absolute value slope of first turn (TUG2, ax), minimum of second turn (TUG2, gx), frequency at 3rd peak of LSP of stand-to-sit (mean TUG, gy), median frequency of DFT of first walk (mean TUG, az), freeze index of DFT of second walk (mean TUG, az), standard deviation of DFT of second walk (mean TUG, gz), OARS total IADLs score	mean absolute value slope of stand-to-sit (cogTUG2, gx), variation fractal dimension of first turn (cogTUG2, ax), modified mean absolute value type 2 of second turn (cogTUG2, ay), frequency at fifth peak of LSP of sit-to-stand (mean cogTUG, az), frequency at first peak of DFT of stand-to-sit (mean cogTUG, az), peak frequency of DFT of stand-to-sit(mean cogTUG, az), maximum to minimum drop in power density of DFT of stand-to-sit(mean cogTUG, gx), slope sign change of stand-to-sit(mean cogTUG, gz), variation fractal dimension of first turn(mean cogTUG, ax), detrended fluctuation analysis of DFT of first turn(mean cogTUG, az)
Mild PD vs controls	spectral bandwidth of DFT of stand-to-sit (TUG1, gx), signal to noise ratio of DFT stand-to-sit (TUG1, gz), frequency of 3rd peak of LSP of stand-to-sit (TUG2, gx), mode of second turn (TUG2, gx), minimum of second turn (TUG2, gx), frequency of 6th peak of LSP of stand-to-sit (mean TUG, gx), frequency of 7th peak of LSP of second turn (mean TUG, gx)	variation fractal dimension of first turn (cogTUG2, ax), root mean square of second turn (cogTUG2, ay), modified mean absolute value type 2 of second turn (cogTUG2, ay), mode of first walk (cogtUG2, ay), slope sign change of stand-to-sit (mean cogTUG, ay), peak frequency of DFT of stand-to-sit (mean cogTUG, az), slope sign change of stand-to-sit (mean cogTUG, az), slope sign change of stand-to-sit (mean cogTUG, gx), slope sign change of stand-to-sit (mean cogTUG, gz)
Moderate PD vs controls	waveform length of second walk (TUG1, gx), integrated absolute value of second walk (TUG2, gy), integrated absolute value of second walk (mean TUG, gy), OARS total ADLs score, OARS total IADLs score, OARS total score	geometric mean of first walk (cogTUG1, gx), frequency at 5th peak of LSP of sit-to-stand (cogTUG2, az), mean power of DFT of stand-to-sit (cogTUG2, gy), integrated absolute value of second walk (cogTUG2, gy), simple square interval of second walk (cogTUG2, gy), slope sign change of stand-to-sit (mean cogTUG, az), OARS total ADLs score, OARS total score
Severe PD vs controls	slope sign change of first walk (TUG1, ay), frequency at eight peak of LSP of first walk (TUG1, gy), OARS total ADLs score, OARS total IADLs score, OARS total score	integrated absolute value of first walk (cogTUG2, gx), OARS total score



(a) (b)

Figure A.1: **(a)** An example signal for the angular velocity around the roll axis (gx, black line) and the trapezoidal integration of the signal (red line) for the TUG task. Segments (1) and (2) represent the two turns. **(b)** The signal for the vertical acceleration (az, black line) and the trapezoidal integration of the signal (red line). Segments (1) and (2) represent the sit-to-stand and stand-to-sit parts.

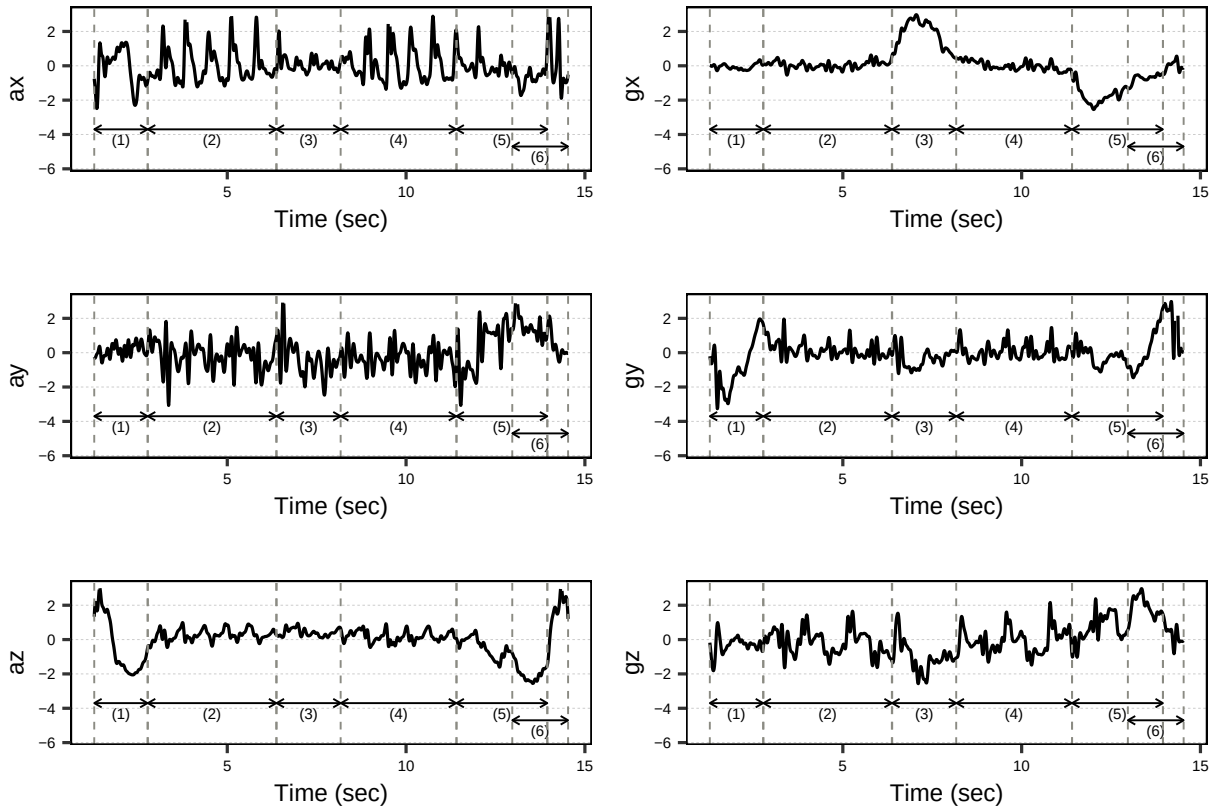


Figure A.2: Example results of our segmentation approach for the TUG task applied to all six channels. The components are (1) sit-to-stand, (2) first walk, (3) first turn, (4) second walk, (5) second turn, and (6) stand-to-sit.

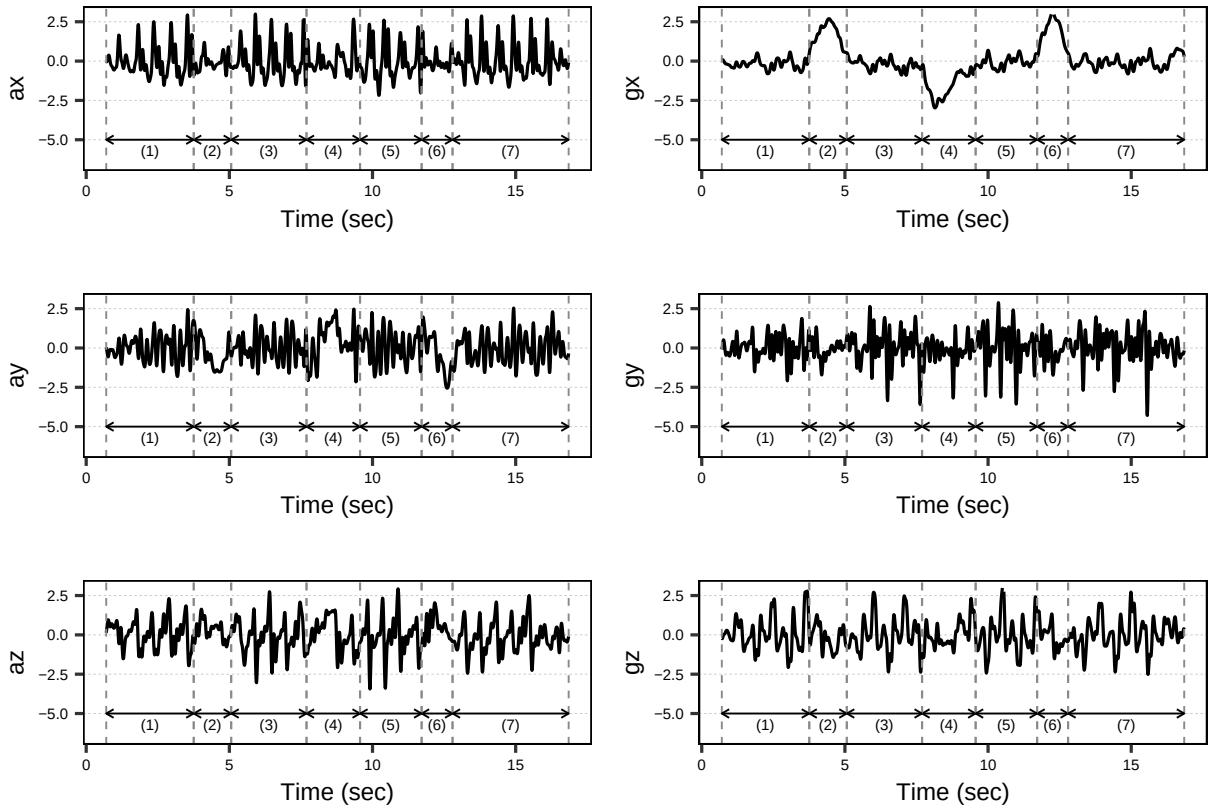


Figure A.3: Example results of our segmentation approach for the 32-foot walk task applied to all six channels. The components are (1) first walk, (2) first turn, (3) second walk, (4) second turn, (5) third walk, (6) third turn, and (7) fourth walk.

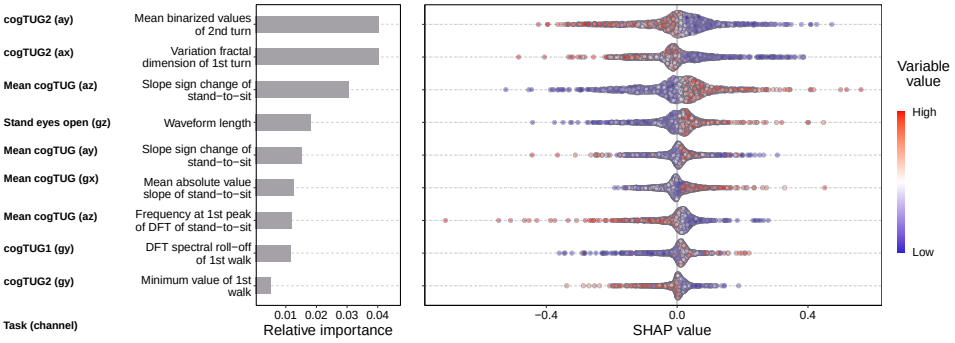


Figure A.4: Feature importance scores (left) and SHAP values (right) of the features included in the all PD vs controls classification model. The first column conveys the task and signal channel. Features are ranked by their importance defined as the mean decrease in area under the curve (AUC) between the original and permuted models. SHAP values were calculated using the kernel SHAP method (see Supplementary Methods). Points in the SHAP plot represent participants from the test sets and SHAP values indicate the impact on model output. The color gradient represents the variable values normalized based on percentile ranks. DFT: Discrete Fourier Transform.

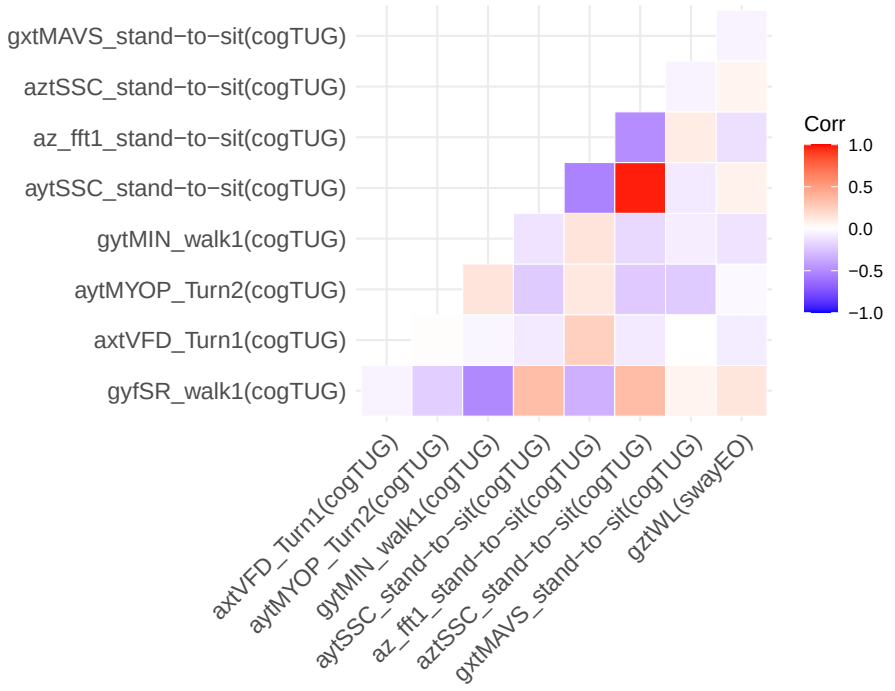


Figure A.5: Correlation matrix of the features included in the all PD vs controls classification model.

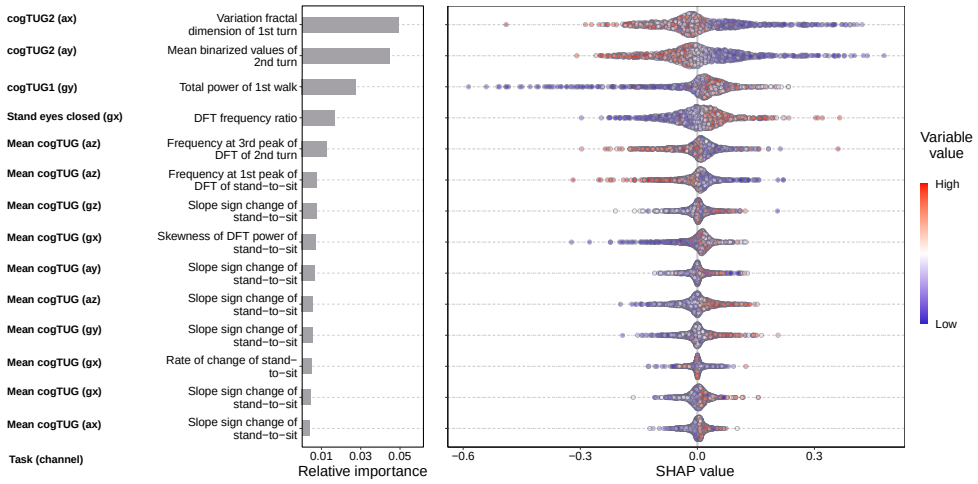


Figure A.6: Feature importance scores (left) and SHAP values (right) of the features included in the mild PD (H&Y values ≤ 2) vs controls classification model. The first column conveys the task and signal channel. Features are ranked by their importance defined as the mean decrease in area under the curve (AUC) between the original and permuted models. SHAP values were calculated using the kernel SHAP method (see Supplementary Methods). Points in the SHAP plot represent participants from the test sets and SHAP values indicate the impact on model output. The color gradient represents the variable values normalized based on percentile ranks. DFT: Discrete Fourier Transform.

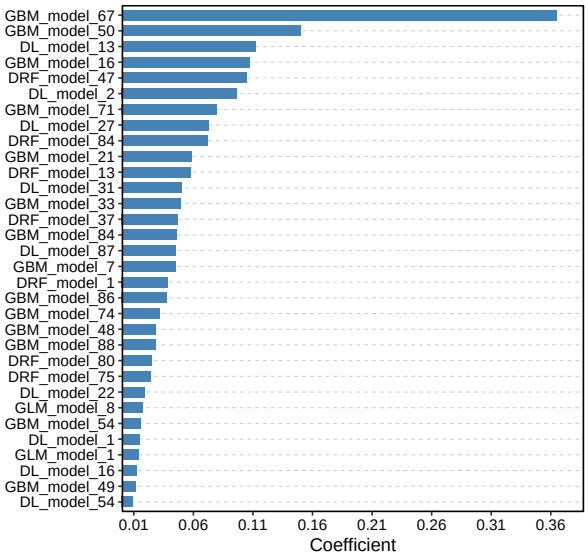


Figure A.7: Base model coefficients in final super learner classifier using all PD participants and controls. Models with zero coefficients are not shown. GBM: gradient boosting machines, DL: deep learning, DRF: distributed random forests, GLM: generalized linear model.

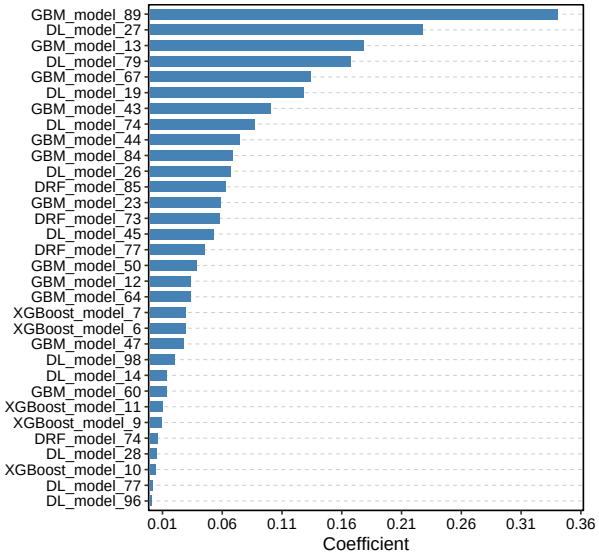


Figure A.8: Base model coefficients in final super learner classifier using mild PD (H&Y $\leq$ 2) participants and controls. Models with zero coefficients are not shown. GBM: gradient boosting machines, DL: deep learning, DRF: distributed random forests, XGBoost: extreme gradient boosting.

A.1 Supplementary Methods

A.1.1 SHAP Values and Individual Feature Importance

SHAP (SHapley Additive exPlanations) is a comprehensive approach that applies concepts from game theory to machine learning models for interpretation of results [388]. By assigning a value to each input feature, SHAP illustrates how each feature contributed to the final prediction result. Unlike traditional feature importance metrics, SHAP considers the impact of features on individual samples, revealing both positive and negative associations with the target variable. To calculate feature importance scores and SHAP values of our classification models, we used the model with the highest contribution weight to the final super learners. By applying the same nested loop framework applied to split the data into training and testing sets, we calculated feature importance scores and SHAP values using the corresponding top model, training, and testing sets. We used the KernelExplainer function in the Python package SHAP to calculate SHAP values. Kernel SHAP is a model-agnostic method that uses a special weighted linear regression to compute the importance of each feature. Feature importance was defined using a permutation-based scoring scheme where each feature was permuted ten times and the mean decrease in area under the curve (AUC) between the permuted and original models represented the importance score.

A.2 Supplementary Results

A.2.1 Classification of Moderate and Severe PD versus Control

Employing the forward feature selection approach (see Section 3.2.8), the number of features was reduced from 25,705 to 8 and 4 features for moderate and severe PD classification, respectively. Among the 8 selected features for the moderate classification, 4 were derived from the cogTUG task, one from the TUG task, and 3 were non-sensor features (Supplementary Figure A.9). Similarly, for the severe classification, 3 out the 4 selected features were derived from the cogTUG task, and one was a non-sensor feature (Supplementary Figure A.10). The selected cogTUG features for both classifiers were derived from the walk, sit-to-stand, and stand-to-sit segments of the task. These features represented acceleration and rotational measures along the vertical, mediolateral, or anteroposterior axes. An extreme gradient boosting machine had the highest contribution to the moderate PD ensemble (Supplementary Figure A.11) and a gradient boosting machine had the highest weighted contribution to the final ensemble of severe PD (Supplementary Figure A.12).

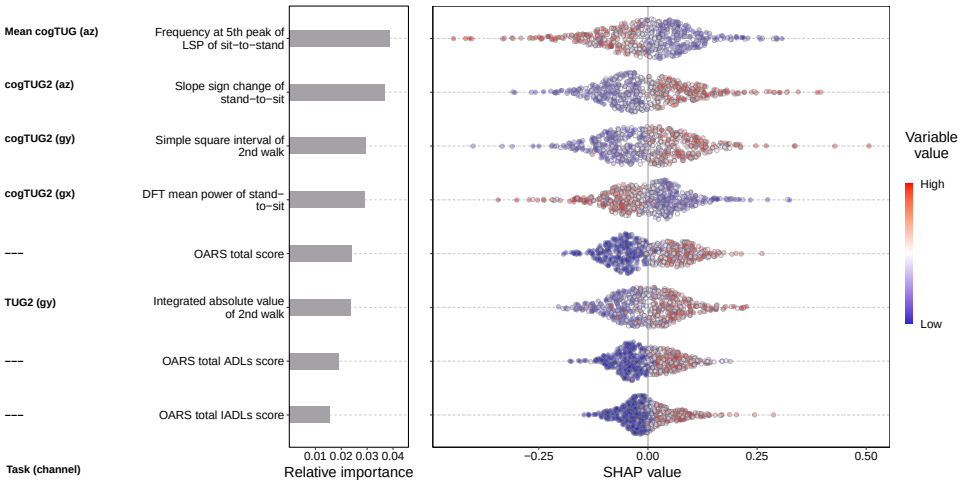


Figure A.9: Feature importance scores (left) and SHAP values (right) of the features included in the moderate PD (H&Y = 2.5, 3) vs controls classification model. The first column conveys the task and signal channel. Features are ranked by their importance defined as the mean decrease in AUC between the original and permuted models. SHAP values were calculated using the kernel SHAP method (see Supplementary Methods). Points in the SHAP plot represent participants from the test sets and SHAP values indicate the impact on model output. The color gradient represents the variable values normalized based on percentile ranks. DFT: Discrete Fourier Transform, LSP: Lomb-Scargle Periodogram, OARS: Older Americans Resource Survey-Instrumental Activities of Daily Living, ADL: activities of daily living, IADL: instrumental activities of daily living.

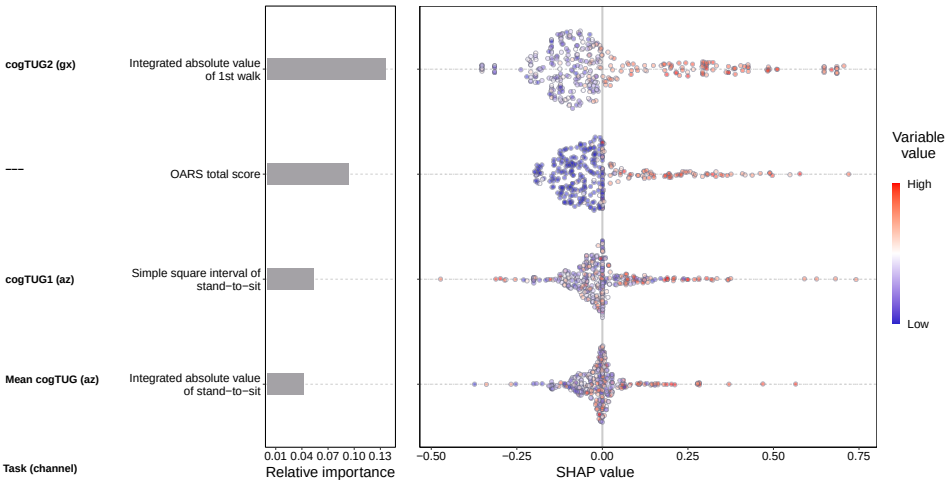


Figure A.10: Feature importance scores (left) and SHAP values (right) of the features included in the severe PD (H&Y = 4) vs controls classification model. The first column conveys the task and signal channel. Features are ranked by their importance defined as the mean decrease in AUC between the original and permuted models. SHAP values were calculated using the kernel SHAP method (see Supplementary Methods). Points in the SHAP plot represent participants from the test sets and SHAP values indicate the impact on model output. The color gradient represents the variable values normalized based on percentile ranks. OARS: Older Americans Resource Survey-Instrumental Activities of Daily Living.

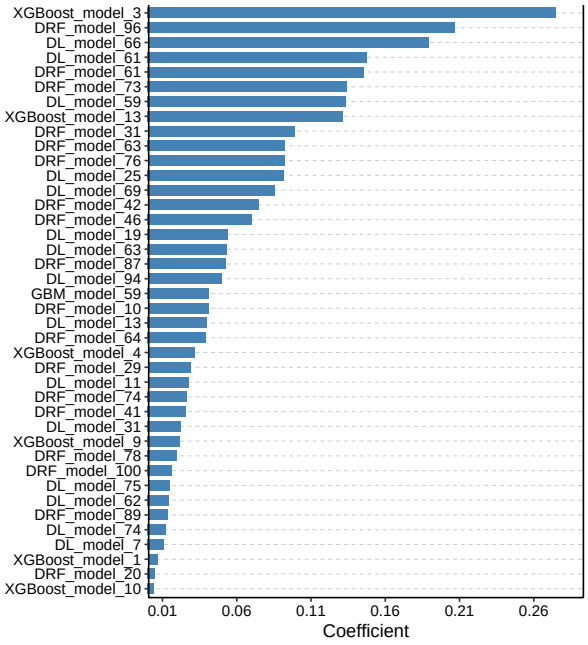


Figure A.11: Base model coefficients in final super learner classifier using moderate PD (H&Y 2.5 and 3) and controls. Models with zero coefficients are not shown. DL: deep learning, DRF: distributed random forests, XGBoost: extreme gradient boosting.

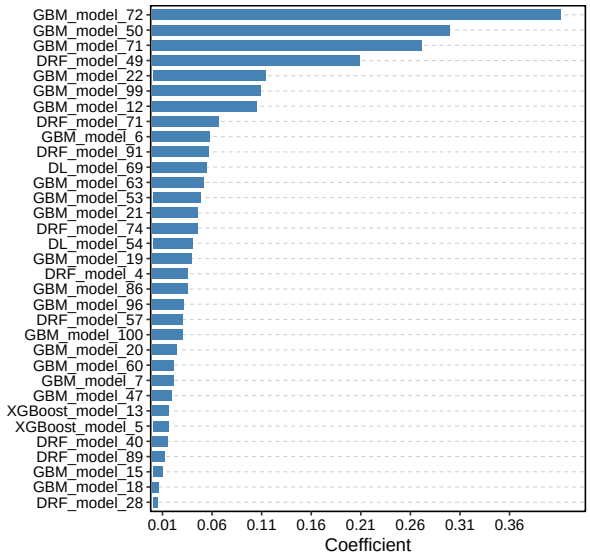


Figure A.12: Base model coefficients in final super learner classifier using severe PD (H&Y 4) and controls. Models with zero coefficients are not shown. GBM: gradient boosting machines, DL: deep learning, DRF: distributed random forests, XGBoost: extreme gradient boosting.

The accuracy of the classifiers was 90.0% (confidence interval [CI] 81.8%, 93.6%) in distinguishing moderate PD from controls and 97.0% (CI 86.6%, 98.5%) in distinguishing those with severe PD from controls. Supplementary Tables A.6 and A.7 display confusion matrices and performance metrics for the classifiers.

Table A.6: Confusion matrix of the classifier using controls and participants with moderate (H&Y 2.5, 3) and severe (H&Y 4) Parkinson's disease (PD). Rows represent actual class and columns represent predictions.

(a) Controls vs Moderate PD

	Controls	PD
Controls	44	6
PD	5	55

(b) Controls vs Severe PD

	Controls	PD
Controls	50	0
PD	2	15

Table A.7: Classification results of models distinguishing moderate (H&Y = 2.5, 3) and severe (H&Y = 4) Parkinson's disease (PD) participants from controls. CI: 95% confidence interval.

PD vs controls	Moderate PD	Severe PD
Number of PD	60	17
Accuracy (CI) [%]	90.0 (81.8, 93.6)	97.0 (86.6, 98.5)
AUC-ROC (CI)	0.90 (0.84, 0.96)	0.94 (0.86, 1.0)
Sensitivity (CI)	0.92 (0.82, 0.97)	0.88 (0.63, 1.0)
Specificity (CI)	0.88 (0.76, 0.95)	1.0
F_1 score (CI)	0.91 (0.84, 0.95)	0.94 (0.75, 1.0)

Analogous to the findings presented earlier for the all PD and mild PD classifiers, an analysis of groupwise feature importance across the five mobility tasks and non-sensor features revealed that features derived from the cogTUG task were by far the most important in distinguishing individuals with PD at different severity levels from control participants. There was a substantial decline of at least 38% in importance when comparing the cogTUG task to other mobility tasks (Supplementary Figure A.13). Additionally, we conducted a detailed analysis of individual feature-level importance scores and SHAP values for the moderate and severe PD classifiers (Supplementary Figures A.9 and A.10, Supplementary Results (Section A.2.2)).

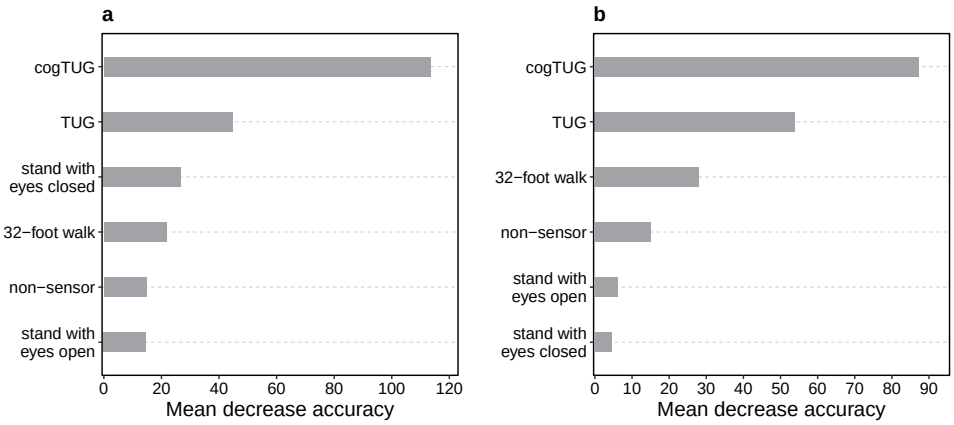


Figure A.13: Group feature importance of models distinguishing **(a)** moderate (H&Y = 2.5, 3) Parkinson's disease (PD) and **(b)** severe PD (H&Y = 4) participants from controls. Groups are ranked by their importance calculated by permuting features within each group simultaneously and reporting the mean decrease in accuracy between the original and permuted models. Under the null hypothesis that there is no association of the group of predictor variables and the model prediction, permutation should have no or little impact on predictive performance. More detailed individual feature-level importance scores and SHapley Additive exPlanations (SHAP) values are illustrated in Supplementary Figures A.9 and A.10.

Comparing demographic and clinical characteristics between correctly classified and misclassified participants (see Supplementary Table A.8), we found that there was no significant age difference between the correctly classified controls and those falsely classified as moderate PD. No control participants were falsely classified as severe PD. Likewise, there was no significant age difference between correctly classified PD participants (TP) and those falsely classified as controls (FN), regardless of PD severity. UPDRS_PIII motor scores of PD participants falsely classified as controls were lower than in correctly classified participants at moderate and severe H&Y stages, although these differences were not statistically significant. In contrast, MoCA scores were higher for the PD participants falsely classified as controls compared to the correctly classified participants with significant differences for severe PD. None of the other demographic and clinical variables (H&Y stage, medication state, sex, CIRS-G) were significantly different

between the correctly and falsely classified PD participants, regardless of PD severity.

Table A.8: Clinical characteristics of correctly classified and misclassified controls and moderate and severe Parkinson's disease (PD) participants.

		Controls			PD		
		False positive	True negative	p	False negative	True positive	p
Moderate PD							
	UPDRS_PIII	–	–	–	23.2 ± 13.2	28.8 ± 9.8	0.20
	MoCA	–	–	–	27.4 ± 2.3	27.2 ± 3.1	0.85
	CIRS-G	–	–	–	5.3 ± 3.7	5.7 ± 3.9	0.84
	Age	63.8 ± 12.0	64.2 ± 12.0	0.52	65.4 ± 7.4	69.8 ± 8.7	0.13
	Sex (% male)	16.7	40.1	0.28	80.0	69.1	0.63
	Medication state (% ON)	–	–	–	100	67.3	0.34
	H&Y (n)			–			0.72
	2.5				4	31	
	3				1	24	
Severe PD							
	UPDRS_PIII	–	–	–	19.5 ± 9.2	42.9 ± 9.5	0.07
	MoCA	–	–	–	29.5 ± 0.71	24.9 ± 5.0	0.005
	CIRS-G	–	–	–	5.0 ± 2.8	7.5 ± 3.8	0.40
	Age	–	–	–	65.0 ± 15.6	74.4 ± 7.8	0.27
	Sex (% male)	–	–	–	50.0	66.7	0.66
	Medication state (% ON)	–	–	–	100	100	–
	H&Y (n)			–			–
	4				2	15	

Following the same simplification strategies applied to all PD and mild PD classification, the effectiveness of the automated segmentation and the redundancy of complex calculations of kinesiological features were once more demonstrated through a comparison between our standard models, which used data from segmented tasks and excluded kinesiological features to alternative models. These alternative models either used features derived from data of the unsegmented tasks or included the kinesiological features in addition to the time- and frequency-domain features from the algorithmic segmented tasks. Models employing data from unsegmented tasks reduced the accuracy by 4.5% for moderate classification and had similar performance for severe PD classification (Supplementary Table A.9, 2nd column vs 1st column). Similarly, models incorporating

kinesiological features dropped the accuracy by 2.7% and 1.5% compared to our standard models for moderate and severe PD classification, respectively (Supplementary Table A.9, 3rd column vs 1st column).

Table A.9: Accuracy (%) of super learner models using the unsegmented tasks and combination of segmented tasks and calculated kinesiological features to distinguish controls from PD participants. The numbers in parentheses represent the difference in accuracy between each model and the corresponding standard model, which used data from segmented tasks without incorporating kinesiological features (same row).

	With segmented tasks and no kinesiological features	With unsegmented tasks and no kinesiological features	With segmented tasks and kinesiological features
Moderate PD (H&Y 2.5, 3) vs controls	90.0	85.5 (−4.5)	87.3 (−2.7)
Severe PD (H&Y 4) vs controls	97.0	97.0 (0.0)	95.5 (−1.5)

Additionally, we compared the performance of models based solely on TUG or cogTUG features to our standard models incorporating features from all segmented mobility tasks (see Supplementary Table A.5 for the selected features). We observed that cogTUG-only models yielded better results than TUG-only models and increased the accuracy by 2.7% compared to the standard models for the moderate stages of the disease (Supplementary Table A.10, columns 2, 3). In severe stages, both the TUG-only and cogTUG-only models showed the same performance, with a slight decrease in accuracy (1.5%) compared to the standard model (Supplementary Table A.10, columns 2, 3). In summary, these findings demonstrate again that sensor data from a single, more complex mobility task incorporating multiple movements may achieve satisfactory performance in distinguishing PD and control participants.

Furthermore, the possibility of achieving further simplification without significantly compromising model performance was examined by assessing whether total duration of TUG and cogTUG could serve as a single feature sufficient to construct a classifier with comparable predictive performance. Consistent with the results for the all PD and mild PD classification, par-

Table A.10: Accuracy (%) of alternative models. TUG-only and cogTUG-only models used features derived from the first trial, the second trial, and the mean of corresponding features from both trials. TUG-duration and cogTUG-duration models used the duration of TUG and cogTUG tasks. The numbers in parentheses represent the difference in accuracy between the respective model and the corresponding all-tasks model (same row).

	All tasks	TUG-only	cogTUG-only	TUG-duration	cogTUG-duration
Moderate PD (H&Y 2.5, 3) vs controls	90.0	81.8 (-8.2)	92.7 (+2.7)	79.1 (-10.9)	82.7 (-7.3)
Severe PD (H&Y 4) vs controls	97.0	95.5 (-1.5)	95.5 (-1.5)	94.0 (-3.0)	92.5 (-4.5)

ticipants at moderate and severe PD stages exhibited significantly longer durations than control participants. However, the constructed logistic regression models using total duration as the only predictor decreased the accuracy in distinguishing PD from control participants by at least 3.0% compared to other more complex and comprehensive models (Supplementary Table A.10, columns 4, 5). Once more, TUG/cogTUG duration alone did not reflect the discriminatory potential of these complex tasks.

A.2.2 Feature-level Importance

SHAP values in Supplementary Figure A.4 show that low values of mean binarized values of the second turn, variation fractal dimension of the first turn, frequency at the first peak of Discrete Fourier Transform (DFT) of stand-to-sit transition, and minimum value of first walk features derived from the cogTUG task increased the probability of belonging to the PD class. Inversely, high values of slope sign change of stand-to-sit and mean absolute value slope of stand-to-sit, which were calculated from the cogTUG task, and waveform length derived from the standing with eyes closed increased the probability of belonging to the PD class.

In Supplementary Figure A.6, using SHAP values of the cogTUG features we found that low values of variation fractal dimension of the first turn, mean binarized values of the second

turn, the frequency at the third peak of DFT of the second turn, and the frequency at the first peak of DFT of stand-to-sit increased the probability of belonging to the PD class. Whereas, high values of the total power of the first walk and skewness of DFT power of stand-to-sit increased the probability of belonging to the PD class. Similarly, high values of the DFT frequency ratio for the standing with eyes closed task increased the probability of being classified as PD.

Supplementary Figures A.9 and A.10 show that scores from the Older Americans Resource Survey-Instrumental Activities of Daily Living (OARS) questionnaire were from among the selected by our models for classifying later stages of the disease. OARS comprise 14 questions: 7 activities of daily living (ADLs); and 7 instrumental activities of daily living (IADLs). ADLs are essential for self-care (walking, eating, bathing, dressing, grooming, getting in/out of bed, and using the toilet) and IADLs are necessary to function independently in the environment (housework, shopping, traveling, preparing meals, using the telephone, managing medication, and handling money). Each question has five possible responses: no difficulty, slower/with greater difficulty, needs some help, needs moderate help, or is completely unable. Higher scores indicate increased disability. Total OARS, ADL, and IADL scores were important in identifying moderate PD participants, and the total OARS score was important in identifying severe PD participants. SHAP values show that low values of the frequency at the fifth peak of Lomb-Scargle Periodogram (LSP) of sit-to-stand, and DFT mean power of stand-to-sit features increased the probability of belonging to the PD class. Whereas, high values of slope sign change of stand-to-sit, the simple square interval of the second walk, integrated absolute values of first walk and stand-to-sit cog-TUG features, integrated absolute values of the second walk from the second TUG trial, and total OARS score features increased the probability of being classified as PD.

A.2.3 Single Feature-based Models: Total TUG/cogTUG Duration

We compared the total duration of TUG and cogTUG using two-tailed Student's t-test. The analysis revealed a significant time difference between PD and control (ctrl) participants (TUG $ctrl = 11.7 \pm 2.3\ s.d.,\ PD = 15.6 \pm 6.4\ s.d.,\ p = 1.3 \times 10^{-12}$; cogTUG $ctrl = 14.4 \pm 3.4\ s.d.,\ PD = 21.3 \pm 9.3\ s.d.,\ p = 2.2 \times 10^{-16}$, see Supplementary Figure A.14, rows 1, 2). Comparison of subtask durations within TUG and cogTUG shows significant differences between PD and control groups, and that these differences are distributed among all subtasks (Supplementary Tables A.11 and A.12, top row).

Total duration of the task (in seconds) increased with PD severity (TUG mild PD $= 14.1 \pm 3.1\ s.d.$, moderate PD $= 16.6 \pm 4.6\ s.d.$, severe PD $= 27.8 \pm 17.2\ s.d.$), and with adding the cognitive test element to the movement task (cogTUG mild PD $= 19.1 \pm 6.3\ s.d.$, moderate PD $= 23.6 \pm 7.0\ s.d.$, severe PD = $38.0 \pm 19.5\ s.d.$) (Supplementary Figure A.14, rows 3-5). Similar patterns were observed for the duration of each subtask (Supplementary Tables A.11 and A.12, rows 2-4). In comparison to the control participants, participants at different PD stages exhibited significantly longer durations when compared using two-tailed Student's t-test ($p < 0.0014$).

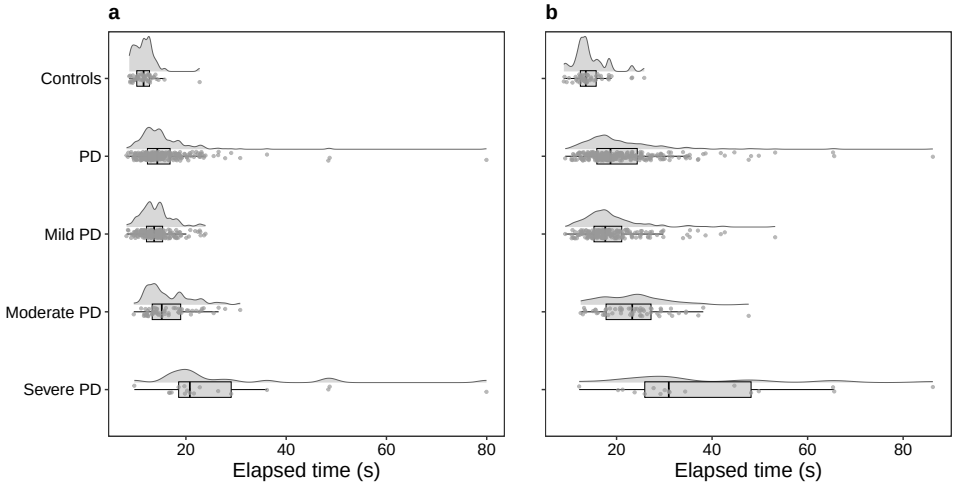


Figure A.14: Comparison of the total duration of **(a)** TUG and **(b)** cogTUG tasks between different groups of PD participants and controls. Different PD groups shows significantly longer durations when compared to the controls using two-tailed Student's t-test ($p < 0.0014$).

Table A.11: Comparing the duration of TUG sub-tasks between different groups of PD participants and controls (ctrl).

	sit-to-stand	stand-to-sit	first turn	second turn	first walk	second walk
PD vs controls	PD: 1.9 ± 0.93 ctrl: 1.5 ± 0.34 $p = 8.4 \times 10^{-8}$	PD: 3.4 ± 2.2 ctrl: 2.4 ± 0.68 $p = 2.2 \times 10^{-9}$	PD: 1.8 ± 0.61 ctrl: 1.5 ± 0.35 $p = 7.0 \times 10^{-8}$	PD: 1.8 ± 0.72 ctrl: 1.3 ± 0.31 $p = 2.7 \times 10^{-10}$	PD: 3.1 ± 1.5 ctrl: 2.5 ± 0.63 $p = 5.1 \times 10^{-6}$	PD: 3.2 ± 2.0 ctrl: 2.5 ± 0.82 $p = 3.2 \times 10^{-5}$
Mild PD vs controls	PD: 1.8 ± 0.61 ctrl: 1.5 ± 0.34 $p = 0.0002$	PD: 3.0 ± 1.1 ctrl: 2.4 ± 0.68 $p = 3.3 \times 10^{-7}$	PD: 1.7 ± 0.46 ctrl: 1.5 ± 0.35 $p = 6.5 \times 10^{-4}$	PD: 1.6 ± 0.48 ctrl: 1.3 ± 0.31 $p = 7.7 \times 10^{-7}$	PD: 2.8 ± 0.83 ctrl: 2.5 ± 0.63 $p = 0.04$	PD: 2.8 ± 0.89 ctrl: 2.5 ± 0.82 $p = 0.013$
Moderate PD vs controls	PD: 2.1 ± 1.1 ctrl: 1.5 ± 0.34 $p = 4.0 \times 10^{-5}$	PD: 3.6 ± 1.6 ctrl: 2.4 ± 0.68 $p = 3.9 \times 10^{-7}$	PD: 2.0 ± 0.63 ctrl: 1.5 ± 0.35 $p = 6.3 \times 10^{-8}$	PD: 1.8 ± 0.60 ctrl: 1.3 ± 0.31 $p = 1.2 \times 10^{-7}$	PD: 3.6 ± 1.5 ctrl: 2.5 ± 0.63 $p = 1.6 \times 10^{-6}$	PD: 3.5 ± 1.1 ctrl: 2.5 ± 0.82 $p = 2.9 \times 10^{-7}$
Severe PD vs controls	PD: 2.9 ± 2.0 ctrl: 1.5 ± 0.34 $p = 0.011$	PD: 6.7 ± 6.7 ctrl: 2.4 ± 0.68 $p = 0.017$	PD: 2.7 ± 0.97 ctrl: 1.5 ± 0.35 $p = 8.0 \times 10^{-5}$	PD: 2.8 ± 1.8 ctrl: 1.3 ± 0.31 $p = 0.004$	PD: 5.6 ± 3.5 ctrl: 2.5 ± 0.63 $p = 0.002$	PD: 6.6 ± 6.5 ctrl: 2.5 ± 0.82 $p = 0.019$

Table A.12: Comparing the duration of cogTUG sub-tasks between different groups of PD participants and controls (ctrl).

	sit-to-stand	stand-to-sit	first turn	second turn	first walk	second walk
PD vs controls	PD: 2.8 ± 1.6 ctrl: 1.9 ± 0.50 $p = 1.2 \times 10^{-12}$	PD: 4.0 ± 2.0 ctrl: 2.5 ± 0.73 $p = 3.4 \times 10^{-16}$	PD: 2.2 ± 0.90 ctrl: 1.6 ± 0.37 $p = 2.1 \times 10^{-13}$	PD: 2.2 ± 1.0 ctrl: 1.5 ± 0.44 $p = 9.2 \times 10^{-12}$	PD: 4.6 ± 3.8 ctrl: 3.2 ± 1.2 $p = 7.0 \times 10^{-7}$	PD: 5.0 ± 2.8 ctrl: 3.6 ± 1.3 $p = 1.3 \times 10^{-7}$
Mild PD vs controls	PD: 2.5 ± 1.1 ctrl: 1.9 ± 0.50 $p = 1.5 \times 10^{-8}$	PD: 3.6 ± 1.3 ctrl: 2.5 ± 0.73 $p = 2.5 \times 10^{-12}$	PD: 2.0 ± 0.62 ctrl: 1.6 ± 0.37 $p = 3.0 \times 10^{-9}$	PD: 2.0 ± 0.73 ctrl: 1.5 ± 0.44 $p = 4.3 \times 10^{-8}$	PD: 4.1 ± 3.3 ctrl: 3.2 ± 1.2 $p = 0.003$	PD: 4.3 ± 2.1 ctrl: 3.6 ± 1.3 $p = 0.0018$
Moderate PD vs controls	PD: 3.2 ± 1.3 ctrl: 1.9 ± 0.50 $p = 6.2 \times 10^{-10}$	PD: 4.2 ± 1.4 ctrl: 2.5 ± 0.73 $p = 1.6 \times 10^{-11}$	PD: 2.4 ± 1.0 ctrl: 1.6 ± 0.37 $p = 2.9 \times 10^{-7}$	PD: 2.4 ± 1.0 ctrl: 1.5 ± 0.44 $p = 7.1 \times 10^{-8}$	PD: 5.1 ± 3.4 ctrl: 3.2 ± 1.2 $p = 1.2 \times 10^{-4}$	PD: 5.9 ± 2.9 ctrl: 3.6 ± 1.3 $p = 2.1 \times 10^{-7}$
Severe PD vs controls	PD: 4.3 ± 4.1 ctrl: 1.9 ± 0.50 $p = 0.025$	PD: 6.9 ± 5.2 ctrl: 2.5 ± 0.73 $p = 0.003$	PD: 3.2 ± 1.8 ctrl: 1.6 ± 0.37 $p = 0.0013$	PD: 3.2 ± 2.2 ctrl: 1.5 ± 0.44 $p = 0.0049$	PD: 9.4 ± 6.4 ctrl: 3.2 ± 1.2 $p = 0.001$	PD: 8.4 ± 5.0 ctrl: 3.6 ± 1.3 $p = 1.0 \times 10^{-3}$

Appendix B: Supplementary Material for Machine Learning and Statistical Analyses of Sensor Data Reveal Variability Between Repeated Trials in Parkinson's Disease Mobility Assessments

Table B.1: Sensor-derived quantitative parameters. All quantitative measures were computed for both the vertical and anteroposterior directions.

Task	**Subtask**	**Measure**
TUG, cogTUG	overall TUG	Duration [s] -acc
TUG, cogTUG	overall TUG	Duration [s]- marks
TUG, cogTUG	overall TUG	Standard deviation [g]
TUG, cogTUG	overall TUG	Median [g]
TUG, cogTUG	Walking	Duration (walk 1+ walk 2) [s]
TUG, cogTUG	Walking 1,2	Duration [s]
TUG, cogTUG	Walking 1,2	Number of steps
TUG, cogTUG	Walking 1,2	Step regularity [g^2]
TUG, cogTUG	Walking 1,2	Stride regularity [g^2]
TUG, cogTUG	Walking 1,2	Step symmetry
TUG, cogTUG	Walking 1,2	Step duration [s]
TUG, cogTUG	Walking 1,2	Stride duration [s]
TUG, cogTUG	Walking 1,2	Step duration/stride duration
TUG, cogTUG	Walking 1,2	Frequency (7 s) [hz]
TUG, cogTUG	Walking 1,2	Amplitude (7 s) [psd]
TUG, cogTUG	Walking 1,2	Width (7 s) [hz]
TUG, cogTUG	Walking 1,2	Slope (7 s) [psd/hz]

TUG, cogTUG	Walking 1,2	Number of steps
TUG, cogTUG	Walking 1,2	Step regularity [g^2]
TUG, cogTUG	Walking 1,2	Stride regularity [g^2]
TUG, cogTUG	Walking 1,2	Step symmetry
TUG, cogTUG	Walking 1,2	Step duration [s]
TUG, cogTUG	Walking 1,2	Stride duration [s]
TUG, cogTUG	Walking 1,2	Step duration/stride duration
TUG, cogTUG	Walking 1,2	Frequency (7 s) [hz]
TUG, cogTUG	Walking 1,2	Amplitude (7 s) [psd]
TUG, cogTUG	Walking 1,2	Width (7 s) [hz]
TUG, cogTUG	Walking 1,2	Slope (7 s) [psd/hz]
TUG, cogTUG	Sit-to-Stand	Range [g]
TUG, cogTUG	Stand-to-Sit	Range [g]
TUG, cogTUG	Sit-to-Stand	Duration [s]
TUG, cogTUG	Stand-to-Sit	Duration [s]
TUG, cogTUG	Sit-to-Stand	Jerk [g/s]
TUG, cogTUG	Stand-to-Sit	Jerk [g/s]
TUG, cogTUG	Sit-to-Stand	Jerk sit-to-stand A [g/s]
TUG, cogTUG	Sit-to-Stand	Jerk sit-to-stand B [g/s]
TUG, cogTUG	Stand-to-Sit	Jerk stand-to-sit A [g/s]
TUG, cogTUG	Stand-to-Sit	Jerk stand-to-sit B [g/s]
TUG, cogTUG	Sit-to-Stand	Range sit-to-stand A [g]
TUG, cogTUG	Sit-to-Stand	Range sit-to-stand B [g]
TUG, cogTUG	Stand-to-Sit	Range stand-to-sit A [g]
TUG, cogTUG	Stand-to-Sit	Range stand-to-sit B [g]
TUG, cogTUG	Sit-to-Stand	Median [g]
TUG, cogTUG	Stand-to-Sit	Median [g]
TUG, cogTUG	Sit-to-Stand	Standard deviation [g]
TUG, cogTUG	Stand-to-Sit	Standard deviation [g]
TUG, cogTUG	Sit-to-Stand	Minimum amplitude [deg/sec]
TUG, cogTUG	Stand-to-Sit	Minimum amplitude [deg/sec]

TUG, cogTUG	Sit-to-Stand	Duration pitch [s]
TUG, cogTUG	Sit-to-Stand	Range pitch [deg/s]
TUG, cogTUG	Sit-to-Stand	Jerk pitch [deg/s^2]
TUG, cogTUG	Stand-to-Sit	Duration pitch [s]
TUG, cogTUG	Stand-to-Sit	Range pitch [deg/s]
TUG, cogTUG	Stand-to-Sit	Jerk pitch [deg/s^2]
TUG, cogTUG	Sit-to-Stand	Duration integrated pitch [s]
TUG, cogTUG	Sit-to-Stand	Range of integrated pitch [deg]
TUG, cogTUG	Sit-to-Stand	Minimum amplitude of integrated pitch [deg]
TUG, cogTUG	Stand-to-Sit	Duration integrated pitch [s]
TUG, cogTUG	Stand-to-Sit	Range of integrated pitch [deg]
TUG, cogTUG	Stand-to-Sit	Minimum amplitude of integrated pitch [deg]
TUG, cogTUG	Sit-to-Stand	Slope of the pitch [deg/s^2]
TUG, cogTUG	Stand-to-Sit	Slope of the pitch [deg/s^2]
TUG, cogTUG	Sit-to-Stand	Mean error between the pitch and its linear fit [deg/s^2]
TUG, cogTUG	Stand-to-Sit	Mean error between the pitch and its linear fit [deg/s^2]
TUG, cogTUG	Turn 1,2	Amplitude-yaw [deg/s]
TUG, cogTUG	Turn 1,2	Duration [s]
TUG, cogTUG	Turn 1,2	Dominant frequency [hz]
TUG, cogTUG	Turn 1,2	Number of steps
16-foot walk	Turn	Duration [s]
16-foot walk	Turn	Dominant frequency [hz]
16-foot walk	Turn	Number of steps
16-foot walk	Walking 1,2	Duration [s]
16-foot walk	Walking 1,2	Number of steps
16-foot walk	Walking 1,2	Step duration [s]
16-foot walk	Walking 1,2	Stride duration [s]
16-foot walk	Walking 1,2	Step duration/stride duration
16-foot walk	Walking 1,2	Frequency (7 s) [hz]
16-foot walk	Walking 1,2	Amplitude (7 s) [psd]
16-foot walk	Walking 1,2	Width (7 s) [hz]

16-foot walk	Walking 1,2	Slope (7 s) [psd/hz]
16-foot walk	Walking 1,2	Number of steps
16-foot walk	Walking 1,2	Step regularity [g^2]
16-foot walk	Walking 1,2	Stride regularity [g^2]
16-foot walk	Walking 1,2	Step symmetry

Table B.2: Intraclass correlation coefficient (ICC, 95% CI) of sensor-based total duration of each mobility task. CI: confidence interval.

Task	Controls	Mild PD	Moderate PD	Severe PD
TUG	0.87 (0.63-0.94)	0.82 (0.72-0.87)	0.87 (0.76-0.93)	0.95 (0.87-0.98)
cogTUG	0.84 (0.50-0.94)	0.85 (0.76-0.90)	0.75 (0.61-0.85)	0.88 (0.69-0.95)
16-foot walk	0.71 (-0.03-0.90)	0.63 (0.06-0.83)	0.62 (0.15-0.82)	0.77 (0.45-0.91)

Table B.3: Variance of sensor-based task duration (s^2).

Task	Controls		Mild PD		Moderate PD		Severe PD	
	Trial 1	Trial2	Trial 1	Trial2	Trial 1	Trial2	Trial 1	Trial2
TUG	6.46	4.64	11.34	9.32	26.44	18.71	327.84	277.06
cogTUG	13.05	11.96	46.28	39.65	56.97	54.95	434.26	371.90
16-foot walk	1.82	2.62	3.00	5.52	5.90	9.34	36.24	109.10

Table B.4: Values of assessments measures.

Assessment	Measure	Trial 1 mean ± SD				Trial 2 mean ± SD			
		Controls	mild PD	moderate PD	severe PD	Controls	mild PD	moderate PD	severe PD
TUG	Total Duration [s]	12.05 ± 2.54	14.48 ± 3.37	17.1 ± 5.14	29.46 ± 18.09	11.34 ± 2.15	13.77 ± 3.01	16.08 ± 4.33	28.38 ± 16.48
	[Turn 2] Amplitude-Yaw [deg/s]	199.86 ± 42.16	167.23 ± 46.61	138.74 ± 53.79	93.36 ± 35.49	208.19 ± 40.37	170.7 ± 42.7	146.4 ± 47.7	89.46 ± 34.86
	[Sit-to-Stand] Jerk [g/s]	-2.54 ± 0.94	-3.91 ± 11.2	-2.92 ± 9.25	-0.56 ± 0.76	-3.44 ± 3.05	-2.76 ± 5.98	-1.52 ± 1.37	-2.53 ± 6.99
	[Sit-to-Stand] Range pitch [deg/s]	361.84 ± 128.48	280.14 ± 140.56	210.5 ± 88.78	141.85 ± 110.9	360.8 ± 118.4	262.3 ± 130.3	214.3 ± 113.8	182.9 ± 130.3
	[Sit-to-Stand] Jerk Sit-to-Stand A [g/s]	-2.6 ± 2.36	-2.99 ± 10.11	-2.85 ± 8.86	-0.71 ± 0.91	-3.7 ± 3.63	-3.01 ± 6.74	-1.6 ± 1.43	-2.16 ± 5.52
	[Sit-to-Stand] Range [g]	1.48 ± 0.35	1.28 ± 0.48	1.02 ± 0.31	0.82 ± 0.32	1.52 ± 0.35	1.27 ± 0.46	1.07 ± 0.35	0.98 ± 0.5
	[Sit-to-Stand] Mean error between the pitch and it's linear fit [deg/s^2]	37.57 ± 21.99	24.07 ± 19.51	17.4 ± 15.16	13.2 ± 13.34	36.66 ± 20.88	20.59 ± 18.9	16.67 ± 18.63	15.14 ± 11.29
	[Stand-to-Sit] Mean error between the pitch and it's linear fit [deg/s^2]	14.71 ± 15.48	18.44 ± 11.75	15.51 ± 8.82	10.72 ± 6.16	13.48 ± 10.61	14.85 ± 11.58	12.82 ± 9.62	10.87 ± 5.86
	[Stand-to-Sit] Jerk pitch [deg/s^2]	205.33 ± 145.88	368.49 ± 2607.44	314.36 ± 1128.7	58.67 ± 46.01	201.83 ± 126.64	109.4 ± 609.3	374.1 ± 1921.4	55.82 ± 43.87
	[Sit-to-Stand] Maximum amplitude [deg/s]	65.71 ± 35.45	65.11 ± 51.54	58.3 ± 31.99	27.47 ± 20.21	66.82 ± 38.41	62.69 ± 41.81	53.6 ± 30.73	50.54 ± 36.72
	[Stand-to-Sit] Range [g]	1.83 ± 0.73	1.44 ± 0.51	1.3 ± 0.44	1.36 ± 0.56	1.75 ± 0.58	1.42 ± 0.53	1.36 ± 0.52	1.01 ± 0.36
	[Stand-to-Sit] Range Stand-to-Sit B [g]	1.41 ± 0.74	1.02 ± 0.57	0.94 ± 0.51	1.08 ± 0.63	1.25 ± 0.65	1 ± 0.58	1 ± 0.59	0.67 ± 0.4
cogTUG	Total Duration [s]	15 ± 3.61	19.8 ± 6.8	24.37 ± 7.55	39.98 ± 20.84	13.71 ± 3.46	18.33 ± 6.3	22.75 ± 7.41	36.1 ± 19.28
	[Sit-to-Stand] Jerk [g/s]	-2.39 ± 1.14	-1.45 ± 1.68	-0.95 ± 0.74	-0.37 ± 0.74	-2.59 ± 1.65	-1.4 ± 0.86	-1.04 ± 1.04	-0.41 ± 0.75
	[Sit-to-Stand] Slope of the pitch [deg/s^2]	410.67 ± 167.04	317.21 ± 279.54	238.99 ± 207.02	152.75 ± 134.4	424.02 ± 217.61	302.34 ± 207.53	221.39 ± 181.47	155.21 ± 94.84
	[Stand-to-Sit] Jerk Pitch [deg/s^2]	240.05 ± 160.25	156.17 ± 350.04	107.88 ± 78.21	69.71 ± 172.68	246.95 ± 168.12	158.7 ± 129.8	103.48 ± 92.57	290.1 ± 989.16
	[Sit-to-Stand] Jerk Sit-to-Stand A [g/s]	-2.03 ± 1.54	-0.9 ± 1.6	-0.8 ± 0.84	-0.36 ± 0.49	-2.3 ± 1.73	-0.88 ± 1.07	-0.94 ± 1.21	-0.06 ± 1.54
	[Turn 2] Amplitude-Yaw [deg/s]	183.92 ± 48.01	143.98 ± 43.77	116.08 ± 39.71	94.58 ± 75.08	198.91 ± 58.91	149.17 ± 48.93	126.97 ± 49.77	95.28 ± 60.64
	[Sit-to-Stand] Range [g]	1.43 ± 0.4	1.17 ± 0.45	0.98 ± 0.24	0.93 ± 0.52	1.48 ± 0.36	1.12 ± 0.39	0.97 ± 0.25	0.88 ± 0.38
	[Stand-to-Sit] Range [g]	1.94 ± 0.72	1.33 ± 0.49	1.48 ± 0.62	1.25 ± 1.05	1.84 ± 0.69	1.4 ± 0.6	1.41 ± 0.55	1.07 ± 0.38
	[Sit-to-Stand] Standard deviation [g]	0.39 ± 0.09	0.31 ± 0.11	0.28 ± 0.07	0.22 ± 0.09	0.4 ± 0.09	0.3 ± 0.09	0.28 ± 0.08	0.24 ± 0.09
	[Stand-to-Sit] Jerk Stand-to-Sit B [g/s]	1.31 ± 0.84	1 ± 2.7	1.02 ± 1.23	0.47 ± 0.93	1.48 ± 0.86	1.27 ± 1.88	1.16 ± 0.96	0.54 ± 0.69
	[Turn 1] Amplitude-Yaw [deg/s]	181.3 ± 46.29	148.01 ± 41.12	130.09 ± 38.96	106.79 ± 75.12	180.36 ± 44.81	152.87 ± 39.82	128.77 ± 36.15	103.44 ± 53.24
	[Sit-to-Stand] Range Sit-to-Stand A [g]	0.78 ± 0.5	0.57 ± 0.37	0.54 ± 0.29	0.51 ± 0.44	0.88 ± 0.41	0.56 ± 0.41	0.54 ± 0.31	0.5 ± 0.46
16-foot walk	Total Duration [s]	7.35 ± 1.35	7.79 ± 1.73	9.51 ± 2.43	14.37 ± 6.02	8.36 ± 1.62	9.2 ± 2.35	11.23 ± 3.06	17.06 ± 10.45
	[Turn] Duration [s]	1.56 ± 0.36	1.7 ± 0.48	2.06 ± 0.68	2.5 ± 0.82	1.55 ± 0.38	1.72 ± 0.53	2.18 ± 0.75	2.44 ± 0.8
	[Walking 1] Stride Regularity [g^2]	0.65 ± 0.17	0.64 ± 0.19	0.66 ± 0.32	0.58 ± 0.22	0.77 ± 0.18	0.66 ± 0.26	0.63 ± 0.22	0.61 ± 0.19
	[Walking 1] Duration- Walk 1 [s]	2.96 ± 0.62	3.07 ± 0.71	3.7 ± 1.04	4.98 ± 2.61	2.89 ± 0.92	3.29 ± 1.12	4.27 ± 1.51	7.74 ± 8.7
	[Walking 1] Step Regularity [g^2]	0.73 ± 0.18	0.65 ± 0.23	0.52 ± 0.29	0.39 ± 0.31	0.69 ± 0.27	0.6 ± 0.26	0.5 ± 0.24	0.34 ± 0.24
	[Turn] Number of steps to complete turn	1.8 ± 1.32	1.53 ± 1.75	2.45 ± 2.14	0.83 ± 1.19	1.41 ± 1.26	1.92 ± 2.15	1.95 ± 2.22	1.07 ± 0.97
	[Walking 2] Step duration [s]	0.62 ± 0.1	0.59 ± 0.11	0.61 ± 0.09	0.71 ± 0.28	0.62 ± 0.17	0.6 ± 0.14	0.62 ± 0.12	0.68 ± 0.24
	[Walking 1] Number of steps to complete walk	5.1 ± 1.2	5.38 ± 1.36	6.25 ± 1.93	7.41 ± 2.32	4.74 ± 1.14	5.48 ± 1.88	6.52 ± 2.38	9.35 ± 6.44
	[Walking 2] Step Regularity [g^2]	0.69 ± 0.25	0.65 ± 0.26	0.49 ± 0.29	0.32 ± 0.22	0.64 ± 0.15	0.59 ± 0.2	0.46 ± 0.24	0.39 ± 0.25
	[Walking 2] Stride duration [s]	1.24 ± 0.23	1.17 ± 0.18	1.23 ± 0.16	1.47 ± 0.68	1.24 ± 0.34	1.18 ± 0.22	1.23 ± 0.22	1.34 ± 0.44
	[Turn] Dominant frequency [Hz]	1.12 ± 0.8	0.91 ± 1.13	1.31 ± 1.28	0.34 ± 0.5	0.87 ± 0.69	1.13 ± 1.24	0.97 ± 1.12	0.49 ± 0.47
	[Walking 2] Stride Regularity [g^2]	0.72 ± 0.2	0.73 ± 0.31	0.58 ± 0.27	0.4 ± 0.28	0.59 ± 0.16	0.59 ± 0.19	0.5 ± 0.17	0.52 ± 0.23

Table B.5: Reliability of each of TUG, cogTUG, and 16-foot walk assessments. ICC intraclass correlation coefficient, CI confidence interval, SEM standard error of measurement agreement, MDC minimal detectable change.

Assessment	Measure	ICC (95%CI)				SEM				MDC			
		Controls	mild PD	moderate PD	severe PD	Controls	mild PD	moderate PD	severe PD	Controls	mild PD	moderate PD	severe PD
TUG	Total Duration [s]	0.87 (0.63-0.94)	0.82 (0.72-0.87)	0.87 (0.76-0.93)	0.95 (0.87-0.98)	0.85	1.38	1.7	3.78	2.35	3.82	4.7	10.47
	[Turn 2] Amplitude-Yaw [deg/s]	0.78 (0.64-0.87)	0.71 (0.63-0.77)	0.73 (0.59-0.83)	0.74 (0.41-0.90)	19.24	24.11	26.35	17.73	53.34	66.83	73.05	49.14
	[Sit-to-Stand] Jerk [g/s]	0.00 (-0.26-0.26)	0.38 (0.24-0.49)	0.02 (-0.23-0.27)	0.03 (-0.44-0.50)	2.29	7.1	6.54	4.91	6.35	19.68	18.12	13.61
	[Sit-to-Stand] Range pitch [deg/s]	0.54 (0.32-0.71)	0.59 (0.49-0.68)	0.75 (0.62-0.84)	0.66 (0.27-0.86)	83.13	86.92	50.48	70.64	230.41	240.92	139.92	195.8
	[Sit-to-Stand] Jerk Sit-to-Stand A [g/s]	0.27 (0.01-0.50)	0.33 (0.19-0.45)	0.01 (-0.24-0.26)	0.16 (-0.34-0.59)	2.65	7.02	6.31	3.64	7.34	19.47	17.48	10.09
	[Sit-to-Stand] Range [g]	0.56 (0.34-0.72)	0.56 (0.45-0.65)	0.58 (0.39-0.73)	0.60 (0.19-0.84)	0.23	0.31	0.21	0.27	0.64	0.86	0.59	0.74
	[Sit-to-Stand] Mean error between the pitch and it's linear fit [deg/s^2]	0.68 (0.50-0.80)	0.49 (0.37-0.59)	0.73 (0.59-0.83)	0.40 (-0.09-0.74)	12.06	13.76	8.8	9.47	33.43	38.14	24.4	26.26
	[Stand-to-Sit] Mean error between the pitch and it's linear fit [deg/s^2]	0.38 (0.12-0.59)	0.44 (0.30-0.55)	0.47 (0.25-0.64)	0.09 (-0.41-0.55)	10.42	8.85	6.77	5.64	28.88	24.54	18.76	15.63
	[Stand-to-Sit] Jerk pitch [deg/s^2]	0.70 (0.53-0.82)	0.01 (-0.13-0.16)	0.84 (0.74-0.90)	0.49 (0.02-0.78)	74.1	1881.68	634.86	31.64	205.4	5215.75	1759.74	87.71
	[Sit-to-Stand] Maximum amplitude [deg/s]	0.90 (0.82-0.94)	0.30 (0.16-0.43)	0.00 (-0.25-0.25)	0.36 (-0.07-0.71)	11.86	39.25	31.32	25.06	32.88	108.78	86.83	69.47
	[Stand-to-Sit] Range [g]	0.33 (0.06-0.55)	0.61 (0.51-0.69)	0.70 (0.54-0.81)	0.26 (-0.14-0.64)	0.54	0.33	0.26	0.42	1.49	0.9	0.73	1.18
	[Stand-to-Sit] Range Stand-to-Sit B [g]	0.31 (0.04-0.54)	0.57 (0.46-0.66)	0.72 (0.58-0.83)	0.15 (-0.22-0.55)	0.58	0.38	0.29	0.52	1.6	1.05	0.8	1.43
cogTUG	Total Duration [s]	0.84 (0.49-0.94)	0.85 (0.76-0.90)	0.75 (0.61-0.85)	0.88 (0.69-0.95)	1.42	2.53	3.72	7	3.93	7.02	10.31	19.41
	[Sit-to-Stand] Jerk [g/s]	0.31 (0.04-0.54)	0.18 (0.03-0.31)	0.53 (0.32-0.69)	0.86 (0.65-0.94)	1.18	1.21	0.62	0.28	3.26	3.36	1.72	0.77
	[Sit-to-Stand] Slope of the pitch [deg/s^2]	0.57 (0.35-0.73)	0.51 (0.40-0.61)	0.67 (0.50-0.79)	0.74 (0.42-0.90)	126.65	171.32	112.17	58.65	351.05	474.87	310.91	162.56
	[Stand-to-Sit] Jerk pitch [deg/s^2]	0.81 (0.69-0.89)	0.00 (-0.14-0.14)	0.66 (0.49-0.78)	0.00 (-0.47-0.47)	71.28	263.63	50.05	708.07	197.56	730.74	138.74	1962.66
	[Sit-to-Stand] Jerk Sit-to-Stand A [g/s]	0.07 (-0.21-0.34)	0.15 (0.00-0.29)	0.66 (0.50-0.78)	0.34 (-0.14-0.70)	1.58	1.26	0.6	0.92	4.38	3.48	1.67	2.55
	[Turn 2] Amplitude-Yaw [deg/s]	0.67 (0.47-0.80)	0.69 (0.61-0.76)	0.72 (0.56-0.83)	0.86 (0.66-0.95)	31.14	25.73	23.83	25.18	86.32	71.31	66.05	69.8
	[Sit-to-Stand] Range [g]	0.39 (0.13-0.60)	0.67 (0.58-0.74)	0.44 (0.21-0.62)	0.79 (0.52-0.92)	0.3	0.24	0.18	0.21	0.83	0.67	0.5	0.57
	[Stand-to-Sit] Range [g]	0.68 (0.50-0.81)	0.56 (0.46-0.66)	0.71 (0.56-0.82)	0.35 (-0.13-0.70)	0.4	0.36	0.31	0.63	1.1	1	0.87	1.74
	[Sit-to-Stand] Standard deviation [g]	0.46 (0.22-0.66)	0.57 (0.47-0.66)	0.56 (0.36-0.71)	0.79 (0.52-0.92)	0.06	0.06	0.05	0.04	0.17	0.18	0.14	0.11
	[Stand-to-Sit] Jerk Stand-to-Sit B [g/s]	0.18 (-0.10-0.43)	0.56 (0.45-0.65)	0.63 (0.45-0.76)	0.34 (-0.14-0.70)	0.77	1.55	0.67	0.65	2.14	4.29	1.85	1.81
	[Turn 1] Amplitude-Yaw [deg/s]	0.90 (0.83-0.94)	0.77 (0.70-0.82)	0.81 (0.70-0.88)	0.89 (0.74-0.96)	14.55	19.56	16.4	20.96	40.32	54.22	45.46	58.09
	[Sit-to-Stand] Range Sit-to-Stand A [g]	0.21 (-0.06-0.46)	0.31 (0.17-0.43)	0.34 (0.10-0.55)	0.00 (-0.47-0.47)	0.4	0.33	0.24	0.44	1.12	0.9	0.67	1.23
16-foot walk	Total Duration [s]	0.71 (-0.03-0.90)	0.63 (0.06-0.83)	0.62 (0.15-0.82)	0.77 (0.45-0.91)	0.84	1.32	1.78	4.09	2.33	3.66	4.92	11.34
	[Turn] Duration [s]	0.57 (0.35-0.73)	0.73 (0.65-0.79)	0.8 (0.68-0.88)	0.65 (0.26-0.85)	0.24	0.26	0.32	0.48	0.67	0.73	0.89	1.32
	[Walking 1] Stride Regularity [g^2]	0.00 (-0.21-0.24)	0.22 (0.08-0.35)	0.00 (-0.25-0.25)	0.12 (-0.37-0.55)	0.18	0.2	0.27	0.19	0.5	0.55	0.76	0.53
	[Walking 1] Duration- Walk 1 [s]	0.58 (0.36-0.74)	0.5 (0.38-0.60)	0.32 (0.09-0.53)	0.28 (-0.18-0.65)	0.51	0.67	1.09	5.5	1.41	1.86	3.01	15.25
	[Walking 1] Step Regularity [g^2]	0.08 (-0.20-0.34)	0.34 (0.21-0.46)	0.38 (0.14-0.57)	0.45 (-0.01-0.76)	0.22	0.2	0.21	0.2	0.61	0.56	0.59	0.56
	[Turn] Number of steps to complete turn	0.40 (0.15-0.61)	0.06 (-0.08-0.20)	0.39 (0.16-0.59)	0.34 (-0.14-0.70)	1	1.91	1.7	0.88	2.78	5.28	4.72	2.43
	[Walking 2] Step duration [s]	0.52 (0.29-0.70)	0.36 (0.22-0.48)	0.54 (0.33-0.70)	0.37 (-0.11-0.71)	0.1	0.1	0.07	0.2	0.27	0.28	0.19	0.56
	[Walking 1] Number of steps to complete walk	0.26 (0.00-0.05)	0.39 (0.26-0.51)	0.41 (0.18-0.60)	0.03 (-0.43-0.48)	1.01	1.28	1.66	4.8	2.8	3.55	4.6	13.31
	[Walking 2] Step Regularity [g^2]	0.31 (0.05-0.54)	0.42 (0.29-0.53)	0.42 (0.19-0.61)	0.50 (0.06-0.78)	0.17	0.18	0.2	0.16	0.48	0.5	0.56	0.45
	[Walking 2] Stride duration [s]	0.54 (0.31-0.71)	0.5 (0.38-0.60)	0.47 (0.25-0.65)	0.22 (-0.27-0.62)	0.19	0.14	0.14	0.5	0.54	0.4	0.39	1.39
	[Turn] Dominant frequency [Hz]	0.35 (0.10-0.57)	0.02 (-0.13-0.16)	0.23 (-0.02-0.45)	0.17 (-0.32-0.59)	0.61	1.18	1.06	0.44	1.68	3.27	2.95	1.22
	[Walking 2] Stride Regularity [g^2]	0.10 (-0.13-0.33)	0.19 (0.05-0.33)	0.30 (0.06-0.51)	0.53 (0.10-0.80)	0.18	0.24	0.19	0.18	0.5	0.67	0.52	0.5

Table B.6: Median of intraclass correlation coefficient (median, 95% confidence interval (CI)) across the quantitative measures.

Task	Controls	Mild PD	Moderate PD	Severe PD
TUG	0.50 (0.32-0.67)	0.45 (0.32-0.58)	0.51 (0.28-0.73)	0.36 (0.29-0.52)
cogTUG	0.48 (0.29-0.66)	0.45 (0.28-0.62)	0.61 (0.52-0.70)	0.54 (0.31-0.75)
16-foot walk	0.35 (0.23-0.48)	0.35 (0.23-0.46)	0.39 (0.28-0.49)	0.34 (0.23-0.44)

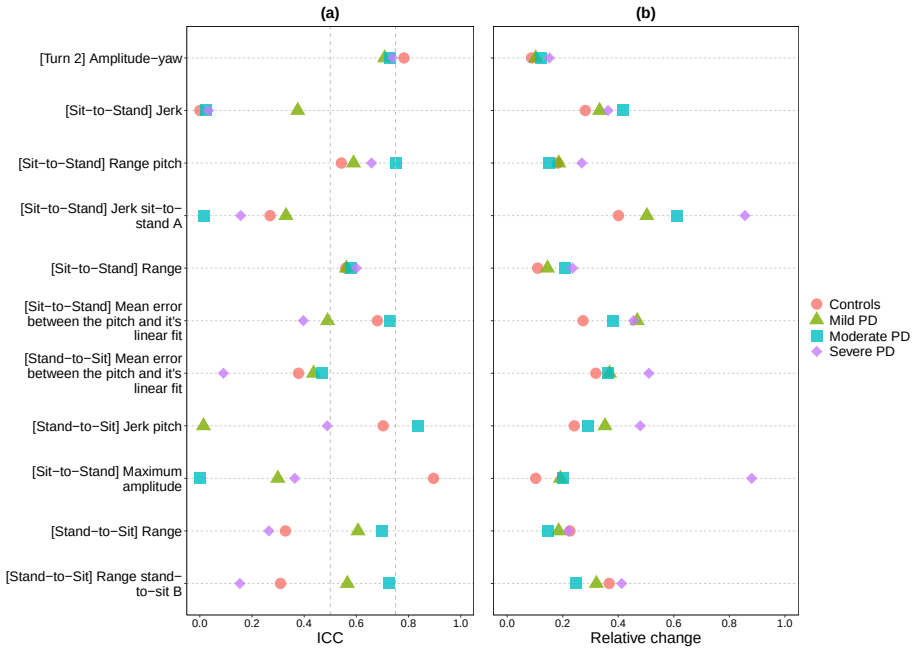


Figure B.1: TUG intraclass correlation coefficient (ICC) **(a)** and median of the relative change **(b)** of each measure. Vertical dashed lines represent thresholds for moderate and good reliability.

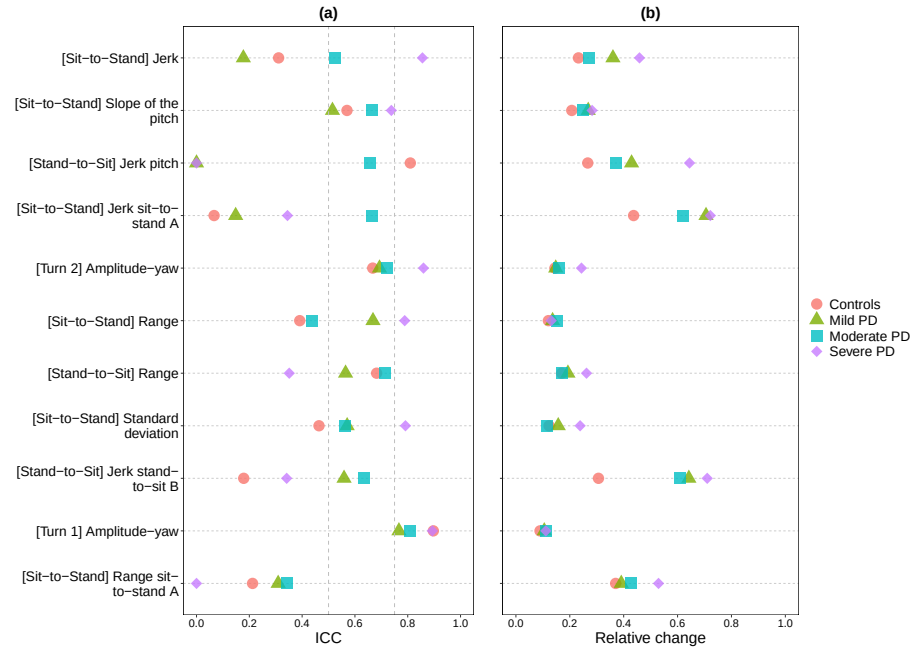


Figure B.2: cogTUG intraclass correlation coefficient (ICC) **(a)** and median of the relative change **(b)** of each measure. Vertical dashed lines represent thresholds for moderate and good reliability.

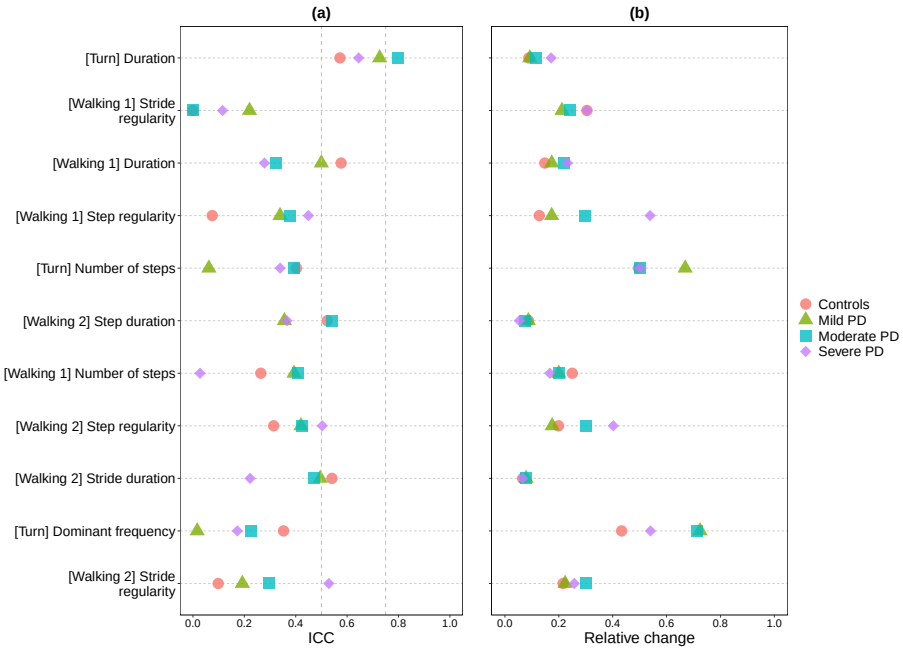


Figure B.3: 16-foot walk intraclass correlation coefficient (ICC) **(a)** and median of the relative change **(b)** of each measure. Vertical dashed lines represent thresholds for moderate and good reliability.

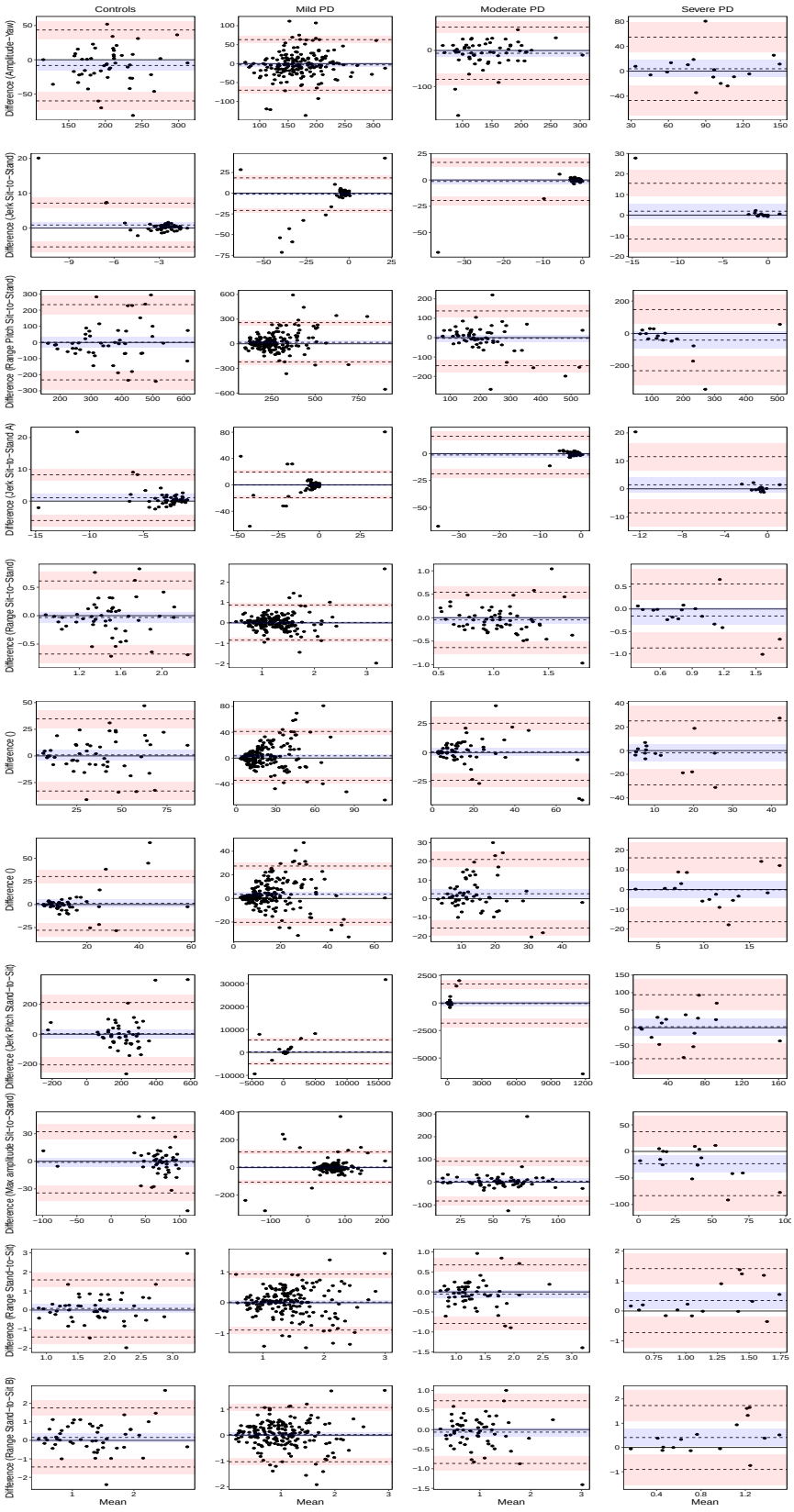


Figure B.4: Bland-Altman plots of each measure for the TUG task. The dashed lines represent the mean of the differences and the 95% limits of agreement (LoA), defined as mean $\pm$ 1.96 SD. The blue and red shaded areas are the confidence intervals of the mean and LoA respectively.

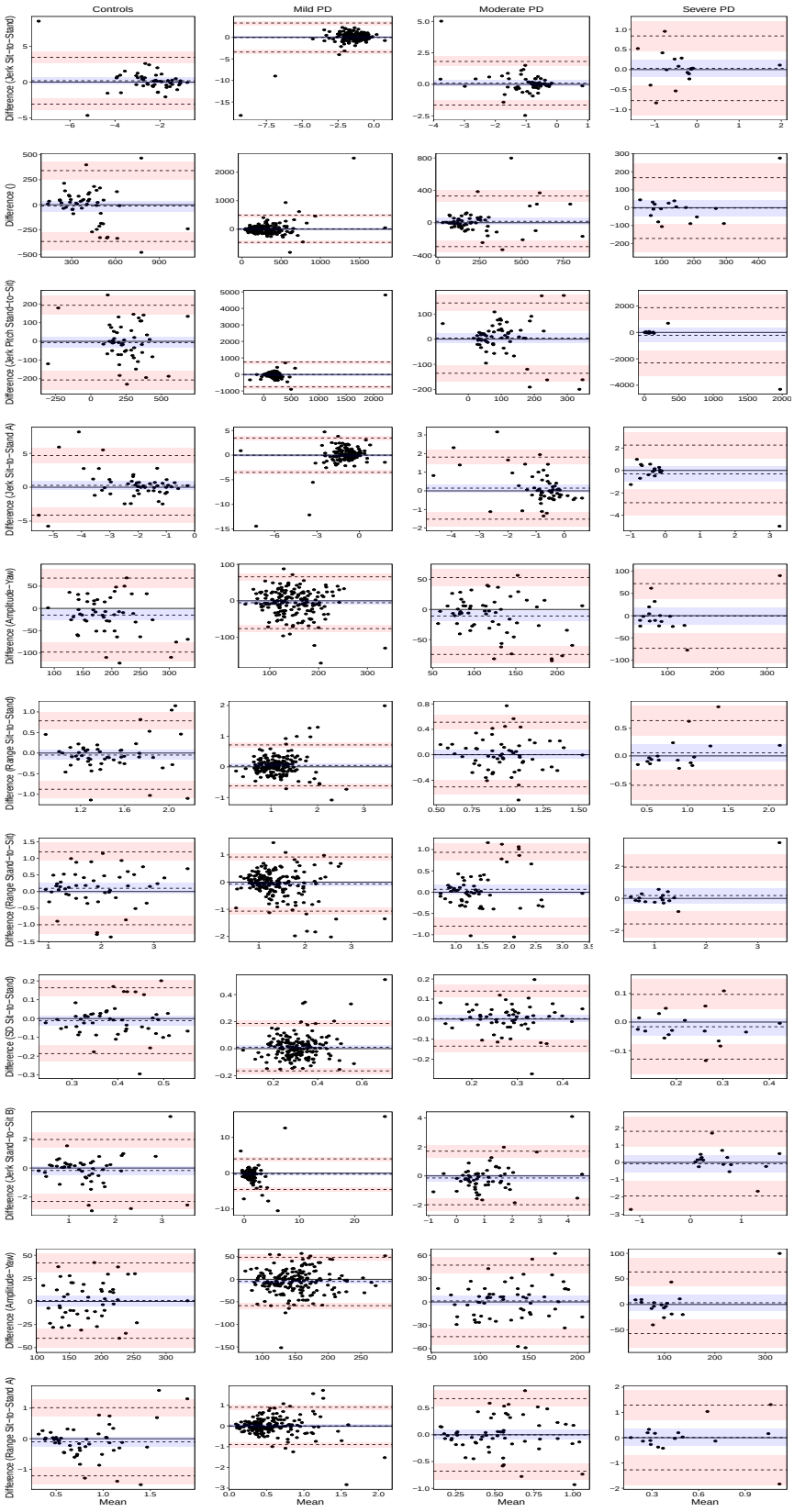


Figure B.5: Bland-Altman plots of each measure for the cogTUG task. The dashed lines represent the mean of the differences and the 95% limits of agreement (LoA), defined as mean $\pm$ 1.96 SD. The blue and red shaded areas are the confidence intervals of the mean and LoA respectively.

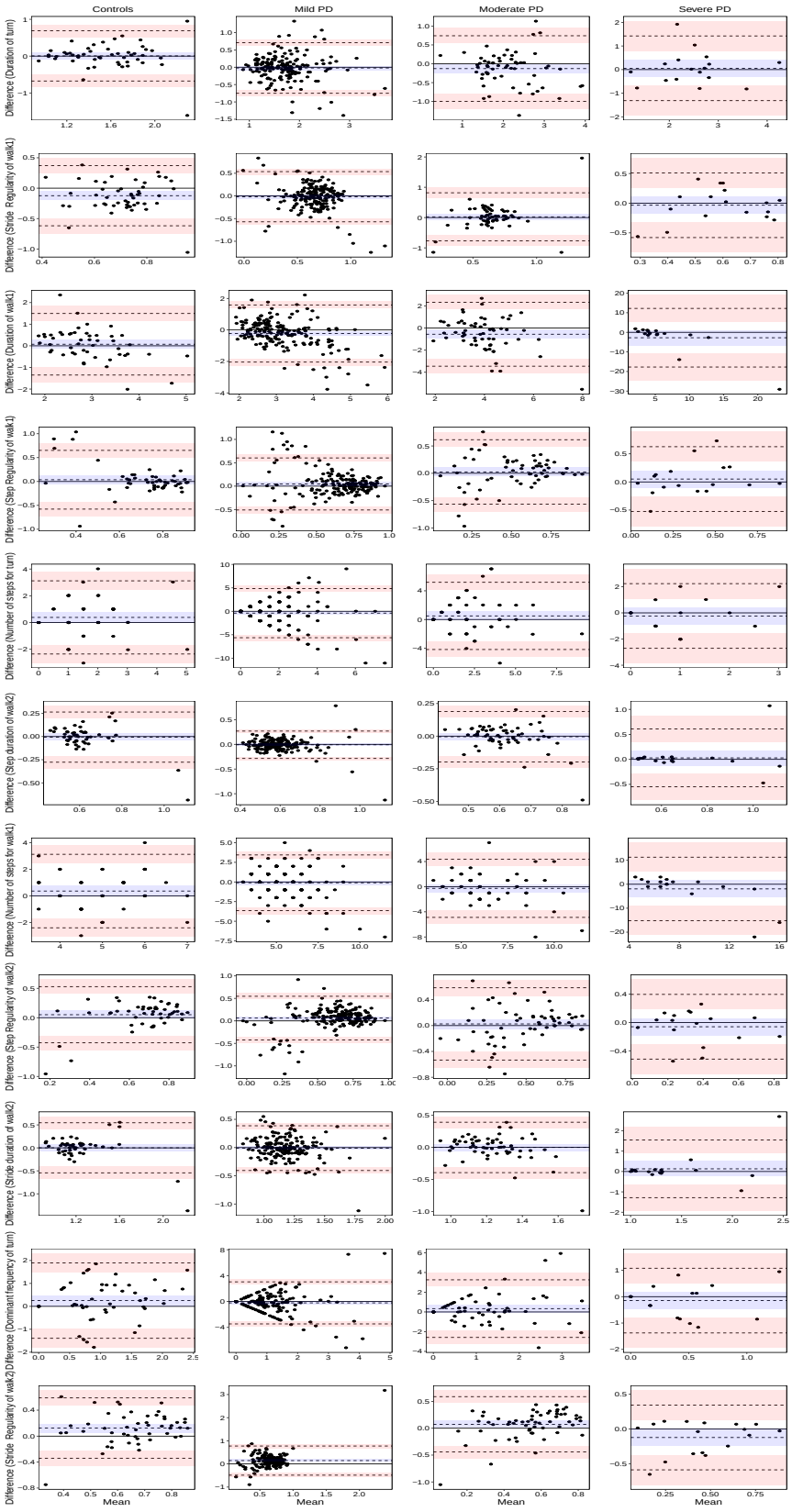


Figure B.6: Bland-Altman plots of each measure for the 16-foot walk task. The dashed lines represent the mean of the differences and the 95% limits of agreement (LoA), defined as mean $\pm$ 1.96 SD. The blue and red shaded areas are the confidence intervals of the mean and LoA respectively.

Appendix C: Supplementary Material for Applying Wearable Sensors and Machine Learning to the Diagnostic Challenge of Distinguishing Parkinson's Disease from Other Forms of Parkinsonism

Table C.1: The count of misclassified participants with Parkinson's disease (PD) that closely resemble each type of non-PD parkinsonism. UnspecPksm: unspecified parkinsonism.

MSA	Unspec Pksm	PSP	DLB	CBS	ET	DIP
13	6	21	8	11	9	2

Table C.2: Confusion matrices of the classifiers using Parkinson's disease (PD) and non-PD parkinsonism (Pksm) participants using features from TUG mobility task. Rows represent actual class and columns represent predictions.

(a) Unsupervised RF feature selection with balanced accuracy maximization

	PD	Pksm
PD	208	52
Pksm	8	10

(b) Mutual information-based ranking for top feature selection

	PD	Pksm
PD	239	21
Pksm	10	8

(c) F_1 score ranking using decision tree classifier for feature scoring

	PD	Pksm
PD	225	35
Pksm	9	9

(d) Supervised RF feature selection with AIC minimization

	PD	Pksm
PD	242	18
Pksm	16	2

(e) Sampling before feature selection with final model trained on original data

	PD	Pksm
PD	250	10
Pksm	15	3

(f) Sampling before feature selection with final model trained on sampled data

	PD	Pksm
PD	258	2
Pksm	18	0

(g) Feature selection before sampling with final model trained on sampled data

	PD	Pksm
PD	251	9
Pksm	14	4

C.1 Supplementary Results

We used the RF proximity matrix to determine, for each misclassified PD participant, the type of non-PD parkinsonism they are most like. This matrix gauges similarity between two observations (participants) points by tracking their frequency of ending up at the same terminal node across all trees.

When participants exhibit similar features, they frequently traverse the same decision paths. Each time a misclassified PD participant and a participant with non-PD parkinsonism appeared together in the balanced training or testing sets, we used the corresponding proximity matrix of the RF model to calculate their similarity. We then averaged these similarity scores to represent the final similarity between the two participants. This allowed us to calculate the overall similarity between each misclassified PD participant and all participants with non-PD parkinsonism. Then, we assigned the type of non-PD parkinsonism participant with the highest similarity value to the misclassified PD participant (Table C.1).

Appendix D: Supplementary Material for Using Machine Learning to Understand the Relationship Between Audiometric Data, Speech Perception, Temporal Processing, and Cognition

D.1 More Detailed Results

Table D.1 describes the base learners included in the SuperLearner models. For randomForest and ranger, we used the default R implementation but increased the number of trees to $10k$. None of the parameters of glmnet or nnet were changed. We included six different setups for the xgboost models where the number of trees was set to $10k$, the maximum depth of trees changed from 1 to 3, and the shrinkage of predictions was either 0.01 or 0.1.

Table D.1: Base models used to build the SuperLearner model and the parameters differing from the defaults.

model	**parameters**
randomForest	$num_trees = 10k$
ranger	$num_trees = 10k$
glmnet	–
nnet	–
xgboost-1	$num_trees = 10k$, $max_depth = 1$, $shrinkage = 0.01$
xgboost-2	$num_trees = 10k$, $max_depth = 2$, $shrinkage = 0.01$
xgboost-3	$num_trees = 10k$, $max_depth = 3$, $shrinkage = 0.01$
xgboost-4	$num_trees = 10k$, $max_depth = 1$, $shrinkage = 0.1$
xgboost-5	$num_trees = 10k$, $max_depth = 2$, $shrinkage = 0.1$
xgboost-6	$num_trees = 10k$, $max_depth = 3$, $shrinkage = 0.1$

Tables D.2, D.3, and D.4 show the average percent of variance explained by each base model and the average weighted contribution in the overall ensemble for the three speech-perception tests: QuickSIN speech-in-noise, sentence recognition with 60% time compression, and sentence recognition with 1.2-s reverberation time respectively. The highest weight was assigned to randomForest when predicting QuickSIN speech-in-noise and to glmnet when predicting the others. nnet had the lowest percentage of variance explained and weighted contribution.

Table D.2: Base models used to predict QuickSIN Speech-in-Noise, the average percent of variance explained by each model, and the average weighted contribution in the overall ensemble.

	variance explained	
model	**(percent)**	**weight**
randomForest	66.69	0.3604
ranger	66.54	0.2152
glmnet	47.65	0.2334
nnet	1.08	0.0001
xgboost-1	61.73	0.0197
xgboost-2	64.56	0.0473
xgboost-3	63.46	0.0127
xgboost-4	60.32	0.0451
xgboost-5	64.56	0.0507
xgboost-6	63.95	0.0154

D.2 R Code

All analyses for this study were carried out in the Statistical Computing Programming Language R (Version: 4.1.3) [352]. The SuperLearner [351] R models included random forests [216,389], ranger [390], lasso and elastic-net regularized generalized linear models (glmnet) [391], extreme gradient boosting (XGBoost) [218], and feed-forward neural networks (nnet) [392] as base models. For feature importance, we used the R package vip [393], to calculate the importance scores for the predictors using the permutation method and the difference in mean squared error (MSE) as an estimator.

Table D.3: Base models used to predict sentence recognition with 60% time compression, the average percent of variance explained by each model, and the average weighted contribution in the overall ensemble.

	variance explained	
model	**(percent)**	**weight**
randomForest	73.10	0.1128
ranger	73.38	0.2499
glmnet	72.13	0.3829
nnet	8.99	0.0026
xgboost-1	74.45	0.1112
xgboost-2	71.83	0.0136
xgboost-3	72.37	0.0104
xgboost-4	74.04	0.0640
xgboost-5	71.62	0.0252
xgboost-6	72.16	0.0274

Table D.4: Base models used to predict sentence recognition with 1.2-s reverberation time, the average percent of variance explained by each model, and the average weighted contribution in the overall ensemble.

	variance explained	
model	**(percent)**	**weight**
randomForest	69.64	0.1475
ranger	69.22	0.2094
glmnet	68.53	0.3143
nnet	4.80	0.0000
xgboost-1	67.53	0.0432
xgboost-2	69.84	0.0921
xgboost-3	69.14	0.0156
xgboost-4	66.31	0.1127
xgboost-5	69.96	0.0351
xgboost-6	69.39	0.0301

Bibliography

[1] Chiranjib Chakraborty, Manojit Bhattacharya, Soumen Pal, and Sang-Soo Lee. From machine learning to deep learning: advances of the recent data-driven paradigm shift in medicine and healthcare. *Current Research in Biotechnology*, 7:100164, 2024.

[2] Arunim Garg and Vijay Mago. Role of machine learning in medical research: a survey. *Computer Science Review*, 40:100370, 2021.

[3] Rahul C. Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

[4] Sri Venkat Gunturi Subrahmanya, Dasharathraj K. Shetty, Vathsala Patil, BM. Zeeshan Hameed, Rahul Paul, Komal Smriti, Nithesh Naik, and Bhaskar K. Somani. The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal of Medical Science (1971-)*, 191(4):1473–1483, 2021.

[5] Pierre Baldi. Deep learning in biomedical data science. *Annual Review of Biomedical Data Science*, 1:181–205, 2018.

[6] David Soriano-Valdez, Ingris Pelaez-Ballestas, Amaranta Manrique de Lara, and Alfonso Gastelum-Strozzi. The basics of data, big data, and machine learning in clinical practice. *Clinical Rheumatology*, 40(1):11–23, 2021.

[7] William S. Weintraub, Akl C. Fahed, and John S. Rumsfeld. Translational medicine in the era of big data and machine learning. *Circulation Research*, 123(11):1202–1204, 2018.

[8] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001.

[9] Douglas G. Altman. *Practical statistics for medical research*. Chapman and Hall/CRC, 1990.

[10] Martin Bland. *An introduction to medical statistics*. Oxford university press, 2015.

[11] Robert H. Riffenburgh. *Statistics in medicine*. Academic press, 2012.

[12] Danilo Bzdok, Naomi Altman, and Martin Krzywinski. Statistics versus machine learning. *Nature Methods*, 15:233–234, 2018.

[13] Qi An, Saifur Rahman, Jingwen Zhou, and James Jin Kang. A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges. *Sensors*, 23(9):4178, 2023.

[14] Mohammad Shehab, Laith Abualigah, Qusai Shambour, Muhannad A. Abu-Hashem, Mohd Khaled Yousef Shambour, Ahmed Izzat Alsalibi, and Amir H. Gandomi. Machine learning in medical applications: a review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458, 2022.

[15] Arwinder Dhillon and Ashima Singh. Machine learning in healthcare data analysis: a survey. *Journal of Biology and Today's World*, 8:1–10, 2019.

[16] Andrew L. Beam and Isaac S. Kohane. Big data and machine learning in health care. *JAMA*, 319(13):1317–1318, 2018.

[17] Abdullah Alanazi. Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, 30:100924, 2022.

[18] Sughasiny Madhavakanna and J. Rajeshwari. Application of machine learning techniques, big data analytics in health care sector – a literature survey. In *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on*, pages 741–749, 2018.

[19] Amir Masoud Rahmani, Efat Yousefpoor, Mohammad Sadegh Yousefpoor, Zahid Mehmood, Amir Haider, Mehdi Hosseinzadeh, and Rizwan Ali Naqvi. Machine learning (ML) in medicine: review, applications, and challenges. *Mathematics*, 9(22):2970, 2021.

[20] Hasmat Malik, Nuzhat Fatema, and Jafar A. Alzubi. *AI and machine learning paradigms for health monitoring system : intelligent data analytics*. Springer, 2021.

[21] Federico Del Giorgio Solfa and Fernando Rogelio Simonato. Big data analytics in healthcare: exploring the role of machine learning in predicting patient outcomes and improving healthcare delivery. *International Journal of Computations Information and Manufacturing (IJCIM)*, 3(1):1–9, 2023.

[22] Jeffrey P. Anderson, Jignesh R. Parikh, Daniel K. Shenfeld, Vladimir Ivanov, Casey Marks, Bruce W. Church, Jason M. Laramie, Jack Mardekian, Beth Anne Piper, Richard J. Willke, and Dale A. Rublee. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *Journal of Diabetes Science and Technology*, 10(1):6–18, 2016.

[23] Akhilesh Kumar Sharma, Shamik Tiwari, Gaurav Aggarwal, Nitika Goenka, Anil Kumar, Prasun Chakrabarti, Tulika Chakrabarti, Radomir Gono, Zbigniew Leonowicz, and Michał Jasiński. Dermatologist-level classification of skin cancer using cascaded ensembling of convolutional neural network and handcrafted features based deep neural network. *IEEE Access*, 10:17920–17932, 2022.

[24] Maha Gharaibeh, Dalia Alzu'bi, Malak Abdullah, Ismail Hmeidi, Mohammad Rustom Al Nasar, Laith Abualigah, and Amir H. Gandomi. Radiology imaging scans for early diagnosis of kidney tumors: a review of data analytics-based machine learning and deep learning approaches. *Big Data and Cognitive Computing*, 6(1):29, 2022.

[25] Tamara Pringsheim, Nathalie Jette, Alexandra Frolkis, and Thomas D.L. Steeves. The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Movement Disorders*, 29(13):1583–1590, 2014.

[26] Stephen K. Van Den Eeden, Caroline M. Tanner, Allan L. Bernstein, Robin D. Fross, Amethyst Leimpeter, Daniel A. Bloch, and Lorene M. Nelson. Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. *American Journal of Epidemiology*, 157(11):1015–1022, 2003.

[27] Giovanni Rizzo, Massimiliano Copetti, Simona Arcuti, Davide Martino, Andrea Fontana, and Giancarlo Logroscino. Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis. *Neurology*, 86:566–576, 2016.

[28] Rana M. Khalil, Lisa M. Shulman, Ann L. Gruber-Baldini, Sunita Shakya, Rebecca Fenderson, Maxwell Van Hoven, Jeffrey M. Hausdorff, Rainer von Coelln, and Michael P. Cummings. Simplification of mobility tests and data processing to increase applicability of wearable sensors as diagnostic tools for Parkinson's disease. *Sensors*, 24(15):4983, 2024.

[29] Rana M. Khalil, Lisa M. Shulman, Ann L. Gruber-Baldini, Sunita Shakya, Jeffrey M. Hausdorff, Rainer von Coelln, and Michael P. Cummings. Machine learning and statistical analyses of sensor data reveal variability between repeated trials in Parkinson's disease mobility assessments. *Sensors*, 24(24):8096, 2024.

[30] Rana M. Khalil, Lisa M. Shulman, Ann L. Gruber-Baldini, Stephen G. Reich, Joseph M. Savitt, Jeffrey M. Hausdorff, Rainer von Coelln, and Michael P. Cummings. Applying wearable sensors and machine learning to the diagnostic challenge of distinguishing parkinson's disease from other forms of parkinsonism. *Biomedicines*, 13(3):572, 2025.

[31] Rana M. Khalil, Alexandra Papanicolaou, Renee Ti Chou, Bobby E. Gibbs, Samira Anderson, Sandra Gordon-Salant, Michael P. Cummings, and Matthew J. Goupell. Using machine learning to understand the relationships between audiometric data, speech perception, temporal processing, and cognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[32] Hanrui Zhang, Kaiwen Deng, Hongyang Li, Roger L Albin, and Yuanfang Guan. Deep learning identifies digital biomarkers for self-reported Parkinson's disease. *Patterns*, 1:100042, 2020.

[33] Luca Palmerini, Sabato Mellone, Guido Avanzolini, Franco Valzania, and Lorenzo Chiari. Quantification of motor impairment in Parkinson's disease using an instrumented timed up and go test. *IEEE Transactions Neural Systems and Rehabilitation Engineering*, 21:664–673, 2013.

[34] Siddharth Arora, Vinayak Venkataraman, Andong Zhan, Sean Donohue, Kevin M. Biglan, Earl Ray Dorsey, and Max A. Little. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study. *Parkinsonism & Related Disorders*, 21(6):650–653, 2015.

[35] Andreas Kuhner, Tobias Schubert, Massimo Cenciarini, Isabella Katharina Wiesmeier, Volker Arnd Coenen, Wolfram Burgard, Cornelius Weiller, and Christoph Maurer. Correlations between motor symptoms across different motor tasks, quantified via random forest feature classification in Parkinson's disease. *Frontiers in Neurology*, 8:607, 2017.

[36] Jochen Klucken, Jens Barth, Patrick Kugler, Johannes Schlachetzki, Thore Henze, Franz Marxreiter, Zacharias Kohl, Ralph Steidl, Joachim Hornegger, Bjoern Eskofier, and Juergen Winkler. Unbiased and mobile gait analysis detects motor impairment in Parkinson's disease. *PLOS ONE*, 8(2):1–9, 2013.

[37] Oung Qi Wei, Hariharan M, Hoileong Lee, Shafriza Basah, Mohamed Sarillee, and C. Lee. Wearable multimodal sensors for evaluation of patients with Parkinson disease. In *2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, pages 269–274, 2015.

[38] Iris Tien, Steven D. Glaser, and Michael J. Aminoff. Characterization of gait abnormalities in Parkinson's disease using a wireless inertial sensor system. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2010:3353—3356, 2010.

[39] Hamza Abujrida, Emmanuel Agu, and Kaveh Pahlavan. Smartphone-based gait assessment to infer Parkinson's disease severity using crowdsourced data. In *2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, pages 208–211, 2017.

[40] Luca Palmerini, Laura Rocchi, Sabato Mellone, Franco Valzania, and Lorenzo Chiari. Feature selection for accelerometer-based posture analysis in Parkinson's disease. *IEEE Transactions on Information Technology in Biomedicine*, 15(3):481–490, 2011.

[41] Adriane Muniz, H. Liu, K. Lyons, Rajesh Pahwa, W. Liu, Flavio Nobre, and Jurandir Nadal. Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait. *Journal of Biomechanics*, 43(4):720–726, 2010.

[42] Resul Das. A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Systems with Applications*, 37:1568, 2010.

[43] Shyamal Patel, Konrad Lorincz, Richard Hughes, Nancy Huggins, John Growdon, David Standaert, Metin Akay, Jennifer Dy, Matt Welsh, and Paolo Bonato. Monitoring motor fluctuations in patients with parkinson's disease using wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 13(6):864–873, 2009.

[44] Luca Lonini, Andrew Dai, Nicholas Shawen, Tanya Simuni, Cynthia Poon, Leo Shimanovich, Margaret Daeschler, Roozbeh Ghaffari, John Rogers, and Arun Jayaraman.

Wearable sensors for Parkinson's disease: which data are worth collecting for training symptom detection models. *npj Digital Medicine*, 1:64, 2018.

[45] Bryan Cole, Serge Roy, Carlo De Luca, and S. Nawab. Dynamical learning and tracking of tremor and dyskinesia from wearable sensors. *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, 22(5):982–991, 2014.

[46] Bjoern M. Eskofier, Sunghoon I. Lee, Jean-Francois Daneault, Fatemeh N. Golabchi, Gabriela Ferreira-Carvalho, Gloria Vergara-Diaz, Stefano Sapienza, Gianluca Costante, Jochen Klucken, Thomas Kautz, and Paolo Bonato. Recent machine learning advancements in sensor-based mobility analysis: deep learning for Parkinson's disease assessment. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 655–658, 2016.

[47] Albert Samà Monsonís, Carlos Pérez, Daniel Rodríguez-Martín, Andreu Català, J.M. Moreno-Aróstegui, Joan Cabestany, Eva de Mingo, and Alejandro Rodríguez-Molinero. Estimating bradykinesia severity in Parkinson's disease by analysing gait through a waist-worn sensor. *Computers in Biology and Medicine*, 84:114–123, 2017.

[48] Jorge Cancela, Mario Pansera, María Teresa Arredondo, Juan Estrada, Matteo Pastorino, Laura Pastor-Sanz, and Juan Luis Villalar. A comprehensive motor symptom monitoring and management system: the bradykinesia case. *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 1008–1011, 2010.

[49] Noël L.W. Keijsers, Martin W.I.M. Horstink, and Stan C.A.M. Gielen. Automatic assessment of levodopa-induced dyskinesias in daily life by neural networks. *Movement disorders*, 18(1):70—80, 2003.

[50] Giulia Belgiovine, Marianna Capecci, Lucio Ciabattoni, Maria Chiara Fiorentino, Gabriele Foresi, Andrea Monteriù, and Lucia Pepa. Upper and lower limbs dyskinesia detection for patients with Parkinson's disease. In *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pages 704–705, 2018.

[51] Evanthia E. Tripoliti, Alexandros T. Tzallas, Markos G. Tsipouras, George Rigas, Panagiota Bougia, Michael Leontiou, Spiros Konitsiotis, Maria Chondrogiorgi, Sofia Tsouli, and Dimitrios I. Fotiadis. Automatic detection of freezing of gait events in patients with Parkinson's disease. *Computer methods and programs in biomedicine*, 110(1):12–26, 2013.

[52] Satyabrata Aich, Pyari Mohan Pradhan, Jinse Park, Nitin Sethi, Vemula Sai Sri Vathsa, and Hee-Cheol Kim. A validation study of freezing of gait (FoG) detection and machine-learning-based FoG prediction using estimated gait characteristics with a wearable accelerometer. *Sensors*, 18(10):3287, 2018.

[53] Nader Naghavi, Aaron Miller, and Eric Wade. Towards real-time prediction of freezing of gait in patients with Parkinson's disease: addressing the class imbalance problem. *Sensors*, 19:3898, 2019.

[54] Amira S. Ashour, Amira El-Attar, Nilanjan Dey, Hatem Abd El-Kader, and Mostafa M. Abd El-Naby. Long short term memory based patient-dependent model for FOG detection in Parkinson's disease. *Pattern Recognition Letters*, 131:23–29, 2020.

[55] Sinziana Mazilu, Michael Hardegger, Zack Zhu, Daniel Roggen, Gerhard Tröster, Meir Plotnik, and Jeffrey M. Hausdorff. Online detection of freezing of gait with smartphones and machine learning techniques. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 123–130, 2012.

[56] Murtadha D. Hssayeni, Joohi Jimenez-Shahed, Michelle A. Burack, and Behnaz Ghoraani. Wearable sensors for estimation of parkinsonian tremor severity during free body movements. *Sensors*, 19(19):4215, 2019.

[57] Hyoseon Jeon, Woongwoo Lee, Hyeyoung Park, Hong Ji Lee, Sang Kyong Kim, Han Byul Kim, Beomseok Jeon, and Kwang Suk Park. Automatic classification of tremor severity in Parkinson's disease using a wearable device. *Sensors*, 17(9):2067, 2017.

[58] George Rigas, Alexandros T. Tzallas, Markos G. Tsipouras, Panagiota Bougia, Evanthia E. Tripoliti, Dina Baga, Dimitrios I. Fotiadis, Sofia G. Tsouli, and Spyridon Konitsiotis. Assessment of tremor activity in the Parkinson's disease using a set of wearable sensors. *IEEE Transactions on Information Technology in Biomedicine*, 16(3):478–487, 2012.

[59] Satyabrata Aich, Jinyoung Youn, Sabyasachi Chakraborty, Pyari Mohan Pradhan, Jin-han Park, Seongho Park, and Jinse Park. A supervised machine learning approach to detect the on/off state in Parkinson's disease using wearable based gait signals. *Diagnostics*, 10(6):421, 2020.

[60] Murtadha D. Hssayeni, Michelle A. Burack, and Behnaz Ghoraani. Automatic assessment of medication states of patients with Parkinson's disease using wearable sensors. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6082–6085, 2016.

[61] Vignesh Ramji, Murtadha Hssayeni, Michelle A. Burack, and Behnaz Ghoraani. Parkinson's disease medication state management using data fusion of wearable sensors. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 193–196, 2017.

[62] Ulrich Hoppe, Thomas Hocke, and Heinrich Iro. Age-related decline of speech perception. *Frontiers in Aging Neuroscience*, 14:891202, 2022.

[63] Working Group on Speech Understanding and Aging. Speech understanding and aging. *The Journal of the Acoustical Society of America*, 83(3):859–895, 1988.

[64] Gill Livingston, Andrew Sommerlad, Vasiliki Orgeta, Sergi G. Costafreda, Jonathan Huntley, David Ames, Clive Ballard, Sube Banerjee, Alistair Burns, Jiska Cohen-Mansfield, et al. Dementia prevention, intervention, and care. *The Lancet*, 390(10113):2673–2734, 2017.

[65] Frank R. Lin. Hearing loss and cognition among older adults in the united states. *The Journals of Gerontology: Series A*, 66A(10):1131–1136, 2011.

[66] Muhammad Ilyas, Alice Othmani Hiring Postdocs, and Amine Naït-ali. Computer-aided prediction of hearing loss based on auditory perception. *Multimedia Tools and Applications*, 79:15765–15789, 2020.

[67] Kenneth Vaden, Lois J. Matthews, Mark A. Eckert, and Judy R. Dubno. Longitudinal changes in audiometric phenotypes of age-related hearing loss. *Journal of the Association for Research in Otolaryngology*, 18:371–385, 2017.

[68] Gregory M. Ellis and Pamela E. Souza. Using machine learning and the national health and nutrition examination survey to classify individuals with hearing loss. *Frontiers in Digital Health*, 3:723533, 2021.

[69] Karen J. Cruickshanks, David M. Nondahl, Mary E. Fischer, Carla R. Schubert, and Ted S. Tweed. A novel method for classifying hearing impairment in epidemiological studies of aging: the wisconsin age-related hearing impairment classification scale. *American Journal of Audiology*, 29(1):59–67, 2020.

[70] Yanxia Zhao, Jingsong Li, Meibian Zhang, Yao Lu, Hongwei Xie, Yu Tian, and Wei Qiu. Machine learning models for the hearing impairment prediction in workers exposed to complex industrial noise: a pilot study. *Ear and Hearing*, 40(3):690–699, 2019.

[71] Maryam Farhadian, Mohsen Aliabadi, and Ebrahim Darvishi. Empirical estimation of the grades of hearing impairment among industrial workers based on new artificial neural networks and classical regression methods. *Indian journal of occupational and environmental medicine*, 19(2):84—89, 2015.

[72] Mohsen Aliabadi, Maryam Farhadian, and Ebrahim Darvishi. Prediction of hearing loss among the noise-exposed workers in a steel factory using artificial intelligence approach. *International archives of occupational and environmental health*, 88(6):779—787, 2015.

[73] Nazri Mohd Nawi, Syed Muhammad Rehman Gillani, and Mohd Imran Ghazali. Noise-induced hearing loss prediction in Malaysian industrial workers using gradient descent with adaptive momentum algorithm. *International Review on Computers and Software*, 6:740—749, 2011.

[74] Brandon Greenwell, Anthony Tvaryanas, and Genny Maupin. Risk factors for hearing decrement among U.S. air force aviation-related personnel. *Aerospace Medicine and Human Performance*, 89:80–86, 2018.

[75] Dan Bing, Jun Ying, Jun Miao, Lan Lan, Dayong Wang, Lidong Zhao, Zifang Yin, Lan Yu, Jing Guan, and Wang Qiuju. Predicting the hearing outcome in sudden sensorineural hearing loss via machine learning models. *Clinical Otolaryngology*, 43(3):868–874, 2018.

[76] Young-Soo Chang, Heesung Park, Sung Hwa Hong, Won-Ho Chung, Yang-Sun Cho, and Il Joon Moon. Predicting cochlear dead regions in patients with hearing loss through a machine learning-based approach: a preliminary study. *PLOS ONE*, 14(6):1–13, 2019.

[77] Keon Vin Park, Kyoung Ho Oh, Yong Jun Jeong, Jihye Rhee, Mun Soo Han, Sung Won Han, and June Choi. Machine learning models for predicting hearing prognosis in unilateral idiopathic sudden sensorineural hearing loss. *Clinical and Experimental Otorhinolaryngology*, 13:148–156, 2020.

[78] Hantai Kim, JaeYeon Park, Yun-Hoon Choung, Jeong Hun Jang, and JeongGil Ko. Predicting speech discrimination scores from pure-tone thresholds—a machine learning-based approach using data from 12,697 subjects. *PLOS ONE*, 16(12):1–18, 2022.

[79] Stephanie Haro, Christopher J. Smalt, Gregory A. Ciccarelli, and Thomas F. Quatieri. Deep neural network model of hearing-impaired speech-in-noise perception. *Frontiers in Neuroscience*, 14:588448, 2020.

[80] Ingrid Yeend, Elizabeth Francis Beach, and Mridula Sharma. Working memory and extended high-frequency hearing in adults: diagnostic predictors of speech-in-noise perception. *Ear and hearing*, 40(3):458—467, 2019.

[81] Marta Lenatti, Pedro A. Moreno-Sánchez, Edoardo M. Polo, Maximiliano Mollura, Riccardo Barbieri, and Alessia Paglialonga. Evaluation of machine learning algorithms and explainability techniques to detect hearing loss from a speech-in-noise screening test. *American Journal of Audiology*, 31(3S):961–979, 2022.

[82] Aravindakshan Parthasarathy, Sandra Romero Pinto, Rebecca M. Lewis, William Goedicke, and Daniel B. Polley. Data-driven segmentation of audiometric phenotypes across a large clinical cohort. *Scientific Reports*, 10(1):6704, 2020.

[83] Paul D. Allen and David A. Eddins. Presbycusis phenotypes form a heterogeneous continuum when ordered by degree and configuration of hearing loss. *Hearing Research*, 264(1):10–20, 2010.

[84] Raul Sanchez-Lopez, Torsten Dau, and William M. Whitmer. Audiometric profiles and patterns of benefit: a data-driven analysis of subjective hearing difficulties and handicaps. *International Journal of Audiology*, 61(4):301–310, 2022.

[85] John J. Geraghty, Joseph Jankovic, and Walter J. Zetusky. Association between essential tremor and Parkinson's disease. *Annals of Neurology*, 17(4):329–333, 1985.

[86] Eng-King Tan, Seng-Swim Lee, Fook-Chong S., and Sau-Ying Lum. Evidence of increased odds of essential tremor in Parkinson's disease. *Movement Disorders*, 23(7):993–997, 2008.

[87] Alfredo Berardelli, A.F. Sabra, and M. Hallett. Physiological mechanisms of rigidity in Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, 46(1):45–53, 1983.

[88] Matteo Bologna, Giorgio Leodori, Paola Stirpe, Giulia Paparella, Donato Colella, Daniele Belvisi, Alfonso Fasano, Giovanni Fabbrini, and Alfredo Berardelli. Bradykinesia in early and advanced Parkinson's disease. *Journal of the Neurological Sciences*, 369:286–291, 2016.

[89] Roberto Cantello, M. Gianelli, D. Bettucci, C. Civardi, M.S. De Angelis, and Roberto Mutani. Parkinson's disease rigidity: magnetic motor evoked potentials in a small hand muscle. *Neurology*, 41(9):1449–1456, 1991.

[90] Chatkaew Pongmala, Stiven Roytman, Miriam van Emde Boas, Olivia Dickinson, Prabesh Kanel, and Nicolaas I Bohnen. Composite measures of motor performance and self-efficacy are better determinants of postural instability and gait difficulties than individual clinical measures in Parkinson's disease. *Parkinsonism & Related Disorders*, 107:105251, 2023.

[91] Meg E. Morris, Robert Iansek, Thomas A. Matyas, and Jeffery J. Summers. The pathogenesis of gait hypokinesia in Parkinson's disease. *Brain*, 117(5):1169–1181, 1994.

[92] Olumide Sofuwa, Alice Nieuwboer, Kaat Desloovere, Anne Marie Willems, Fabienne Chavret, and Ilse Jonkers. Quantitative gait analysis in Parkinson's disease: comparison with a healthy control group. *Archives of Physical Medicine and Rehabilitation*, 86:1007–1013, 2005.

[93] Anke H. Snijders, Inge Leunissen, Maaike Bakker, Sebastiaan Overeem, Rick C. Helmich, Bastiaan R. Bloem, and Ivan Toni. Gait-related cerebral alterations in patients with Parkinson's disease with freezing of gait. *Brain*, 134(1):59–72, 2010.

[94] Heidemarie Zach, Arno M. Janssen, Anke H. Snijders, Arnaud Delval, Murielle U. Ferraye, Eduard Auff, Vivian Weerdesteyn, Bastiaan R. Bloem, and Jorik Nonnekes. Identifying freezing of gait in Parkinson's disease during freezing provoking tasks using waist-mounted accelerometry. *Parkinsonism & Related Disorders*, 21:1362–1366, 2015.

[95] Xiang Gao, Honglei Chen, Michael A. Schwarzschild, Giancarlo Logroscino, and Alberto Ascherio. Perceived imbalance and risk of Parkinson's disease. *Movement Disorders*, 23(4):613–616, 2008.

[96] Lisa M. Shulman, Ann L. Gruber-Baldini, Karen E. Anderson, Christopher G. Vaughan, Stephen G. Reich, Paul S. Fishman, and William J. Weiner. The evolution of disability in Parkinson disease. *Movement Disorders*, 23:790–796, 2008.

[97] Margaret M. Hoehn and Melvin D. Yahr. Parkinsonism: onset, progression, and mortality. *Neurology*, 17:427–442, 1967.

[98] Beth Anne Sieber, Story Landis, Walter Koroshetz, Randall Bateman, Andrew Siderowf, Wendy R. Galpern, John Dunlop, Steven Finkbeiner, Margaret Sutherland, Hao Wang, Virginia M.Y. Lee, Harry T. Orr, Katrina Gwinn, Kip Ludwig, Anna Taylor, Christine Torborg, and Thomas J. Montine. Prioritized research recommendations from the national institute of neurological disorders and stroke Parkinson's disease 2014 conference. *Annals of Neurology*, 76:469–472, 2014.

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

[100] Erika Rovini, Carlo Maremmani, and Filippo Cavallo. How wearable sensors can support Parkinson's disease diagnosis and treatment: a systematic review. *Frontiers in Neuroscience*, 11:555, 2017.

[101] Alejandro Rodríguez-Molinero, Albert Samà Monsonís, Carlos Pérez, Daniel Rodríguez-Martín, Sheila Alcaine, Berta Mestre, Paola Quispe, Benedetta Giuliani, Gabriel Vainstein, Patrick Browne, Dean Sweeney, Leo Quinlan, J Manuel, Moreno Arostegui, Angels Bayés, Hadas Lewy, Alberto Costa, Roberta Annicchiarico, Timothy Counihan, and Joan Cabestany. Analysis of correlation between an accelerometer-based algorithm for detecting parkinsonian gait and UPDRS subscales. *Frontiers in Neurology*, 8:431, 2017.

[102] Ervin Sejdić, Kristin A. Lowry, Jennica Bellanca, Mark S. Redfern, and Jennifer S. Brach. A comprehensive assessment of gait accelerometry signals in time, frequency and time-frequency domains. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(3):603–612, 2014.

[103] Silvia Del Din, Alan Godfrey, and Lynn Rochester. Validation of an accelerometer to quantify a comprehensive battery of gait characteristics in healthy older adults and Parkinson's disease: toward clinical and at home use. *IEEE Journal of Biomedical and Health Informatics*, 20:838–847, 2016.

[104] Ryan P. Hubble, Geraldine A. Naughton, Peter A. Silburn, and Michael H. Cole. Wearable sensor use for assessing standing balance and walking stability in people with Parkinson's disease: a systematic review. *PLOS ONE*, 10:1–22, 2015.

[105] Ruirui Lu, Yan Xu, Xiaohui Li, Yongli Fan, Weiqi Zeng, Yang Tan, Kang Ren, Wenwu Chen, and Xuebing Cao. Evaluation of wearable sensor devices in Parkinson's disease: a review of current status and future prospects. *Parkinson's Disease*, 2020:1–8, 2020.

[106] Caroline Moreau, Tiphaine Rouaud, David Grabli, Isabelle Benatru, Philippe Remy, Ana Raquel Marques, Sophie Drapier, Louise Laure Mariani, Emmanuel Roze, David Devos, Gwendoline Dupont, Matthieu Bereau, and Margherita Fabbri. Overview on wearable sensors for the management of Parkinson's disease. *npj Parkinson's Disease*, 9:1–16, 2023.

[107] Asma Channa, Nirvana Popescu, and Vlad Ciobanu. Wearable solutions for patients with Parkinson's disease and neurocognitive disorder: a systematic review. *Sensors*, 20:2713, 2020.

[108] Alberto J. Espay, Paolo Bonato, Fatta B. Nahab, Walter Maetzler, John M. Dean, Jochen Klucken, Bjoern M. Eskofier, Aristide Merola, Fay Horak, Anthony E. Lang, Ralf Reilmann, Joe Giuffrida, Alice Nieuwboer, Malcolm Horne, Max A. Little, Irene Litvan, Tanya Simuni, E. Ray Dorsey, Michelle A. Burack, Ken Kubota, Anita Kamondi, Catarina Godinho, Jean-Francois Daneault, Georgia Mitsi, Lothar Krinke, Jeffery M. Hausdorff, Bastiaan R. Bloem, Spyros Papapetropoulos, and on behalf of the Movement Disorders Society Task Force on Technology. Technology in Parkinson's disease: challenges and opportunities. *Movement Disorders*, 31(9):1272–1282, 2016.

[109] Ken J. Kubota, Jason A. Chen, and Max A. Little. Machine learning for large-scale wearable sensor data in Parkinson's disease: concepts, promises, pitfalls, and futures. *Movement Disorders*, 31(9):1314–1326, 2016.

[110] Charalampos Sotirakis, Zi Su, Maksymilian A. Brzezicki, Niall Conway, Lionel Tarassenko, James J. FitzGerald, and Chrystalina A. Antoniades. Identification of motor progression in Parkinson's disease using wearable sensors and machine learning. *npj Parkinson's Disease*, 9:1–8, 2023.

[111] Francesco Castelli Gattinara Di Zubiena, Greta Menna, Ilaria Mileti, Alessandro Zampogna, Francesco Asci, Marco Paoloni, Antonio Suppa, Zaccaria Del Prete, and Eduardo Palermo. Machine learning and wearable sensors for the early detection of balance disorders in Parkinson's disease. *Sensors*, 22(24):9903, 2022.

[112] Andrea C. Albán-Cadena, Fernando Villalba-Meneses, Kevin O. Pila-Varela, Alejandro Moreno-Calvo, Carlos P. Villalba-Meneses, and Diego A. Almeida-Galárraga. Wearable sensors in the diagnosis and study of Parkinson's disease symptoms: a systematic review. *Journal of Medical Engineering Technology*, 45(7):532–545, 2021.

[113] Huma Mughal, Abdul Rehman Javed, Muhammad Rizwan, Ahmad S. Almadhor, and Natalia Kryvinska. Parkinson's disease management via wearable sensors: a systematic review. *IEEE Access*, 10:35219–35237, 2022.

[114] Navleen Kour, Sunanda Gupta, and Sakshi Arora. Sensor technology with gait as a diagnostic tool for assessment of Parkinson's disease: a survey. *Multimedia Tools Applications*, 82(7):10211–10247, 2022.

[115] Aron S. Buchman, Sue E. Leurgans, Aner Weiss, Veronique VanderHorst, Anat Mirelman, Robert Dawe, Lisa L. Barnes, Robert S. Wilson, Jeffrey M. Hausdorff, and David A. Bennett. Associations between quantitative mobility measures derived from components of conventional mobility testing and Parkinsonian gait in older adults. *PLOS ONE*, 9(1):1–8, 2014.

[116] Aner Weiss, Talia Herman, Meir Plotnik, Marina Brozgol, Inbal Maidan, Nir Giladi, Tanya Gurevich, and Jeffrey M. Hausdorff. Can an accelerometer enhance the utility of the Timed Up & Go Test when evaluating patients with Parkinson's disease? *Medical Engineering Physics*, 32:119–125, 2010.

[117] Aner Weiss, Anat Mirelman, Aron S. Buchman, David A. Bennett, and Jeffrey M. Hausdorff. Using a body-fixed sensor to identify subclinical gait difficulties in older adults with IADL disability: maximizing the output of the Timed Up and Go. *PLOS ONE*, 8:68885, 2013.

[118] Siddharth Arora, Vinayak Venkataraman, Sean Donohue, Kevin M. Biglan, Earl R. Dorsey, and Max A. Little. High accuracy discrimination of Parkinson's disease participants from healthy controls using smartphones. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3641–3644, 2014.

[119] Johannes C. M. Schlachetzki, Jens Barth, Franz Marxreiter, Julia Gossler, Zacharias Kohl, Samuel Reinfelder, Heiko Gassner, Kamiar Aminian, Bjoern M. Eskofier, Jürgen Winkler, and Jochen Klucken. Wearable sensors objectively measure gait parameters in Parkinson's disease. *PLOS ONE*, 12(10):1–18, 2017.

[120] Radim Krupicka, Slavka Viteckova, Vaclav Cejka, Ondrej Klempir, Zoltan Szabo, and Evzen Ruzicka. Bradykan: a motion capture system for objectification of hand motor tests in Parkinson disease. In *2017 E-Health and Bioengineering Conference (EHB)*, pages 446–449, 2017.

[121] Mohammadreza Abtahi, Seyed Bahram Borgheai, Roohollah Jafari, Nicholas Constant, Rassoul Diouf, Yalda Shahriari, and Kunal Mankodiya. Merging fNIRS-EEG brain monitoring and body motion capture to distinguish Parkinson's disease. *IEEE Transactions Neural Systems and Rehabilitation Engineering*, 28(6):1246–1253, 2020.

[122] Yor J. Castaño-Pino, Maria C. González, Valentina Quintana-Peña, Jaime Valderrama, Beatriz Muñoz, Jorge Orozco, and Andrés Navarro. Automatic gait phases detection in Parkinson disease: a comparative study. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 798–802, 2020.

[123] Dung Phan, Malcolm Horne, Pubudu N. Pathirana, and Parisa Farzanehfar. Measurement of axial rigidity and postural instability using wearable sensors. *Sensors*, 18(2):495, 2018.

[124] Silvia Del Din, Morad Elshehabi, Brook Galna, Markus A. Hobert, Elke Warmerdam, Ulrike Suenkel, Kathrin Brockmann, Florian Metzger, Clint Hansen, Daniela Berg, Lynn Rochester, and Walter Maetzler. Gait analysis with wearables predicts conversion to Parkinson disease. *Annals of Neurology*, 86(3):357–367, 2019.

[125] Erika Rovini, Carlo Maremmani, Alessandra Moschetti, Dario Esposito, and Filippo Cavallo. Comparative motor pre-clinical assessment in Parkinson's disease using supervised machine learning approaches. *Annals of Biomedical Engineering*, 46:2057–2068, 2018.

[126] Rana Zia Ur Rehman, Silvia Del Din, Jian Qing Shi, Brook Galna, Sue Lord, Alison J. Yarnall, Yu Guan, and Lynn Rochester. Comparison of walking protocols and gait assessment systems for machine learning-based classification of Parkinson's disease. *Sensors*, 19(24):5363, 2019.

[127] Christopher Buckley, Brook Galna, Lynn Rochester, and Claudia Mazzà. Quantification of upper body movements during gait in older adults and in those with Parkinson's disease: impact of acceleration realignment methodologies. *Gait & Posture*, 52:265–271, 2017.

[128] Guoen Cai, Weikun Shi, Yingqing Wang, Huidan Weng, Lina Chen, Jiao Yu, Zhonglue Chen, Fabin Lin, Kang Ren, Yuqi Zeng, Jun Liu, Yun Ling, and Qinyong Ye. Specific distribution of digital gait biomarkers in Parkinson's disease using body-worn sensors and machine learning. *Journals of Gerontology A, Biological Sciences and Medical Sciences*, 78(8):1348–1354, 2023.

[129] Luka Jovanovic, Robertas Damaševičius, Rade Matić, Milos Kabiljo, Vladimir Simic, Goran Kunjadic, Milos Antonijevic, Miodrag Zivkovic, and Nebojsa Bacanin. Detecting Parkinson's disease from shoe-mounted accelerometer sensors using convolutional neural networks optimized with modified metaheuristics. *PeerJ Computer Science*, 10:e2031, 2024.

[130] Luigi Battista and Antonietta Romaniello. A new wrist-worn tool supporting the diagnosis of parkinsonian motor syndromes. *Sensors*, 24(6):1965, 2024.

[131] Peng Yue, Ziheng Li, Menghui Zhou, Xulong Wang, and Po Yang. Wearable-sensor-based weakly supervised Parkinson's disease assessment with data augmentation. *Sensors*, 24(4):1196, 2024.

[132] Gaia Bailo, Francesca Lea Saibene, Virginia Bandini, Pietro Arcuri, Anna Salvatore, Mario Meloni, Anna Castagna, Jorge Navarro, Tiziana Lencioni, Maurizio Ferrarin, and Ilaria Carpinella. Characterization of walking in mild Parkinson's disease: reliability, validity and discriminant ability of the six-minute walk test instrumented with a single inertial sensor. *Sensors*, 24(2):662, 2024.

[133] Parvathy Nair, Maryam Shojaei Baghini, Gita Pendharkar, and Hoam Chung. Detecting early-stage Parkinson's disease from gait data. *Proceedings of the Institution of Mechanical Engineers H*, 237(11):1287–1296, 2023.

[134] Sana M. Keloth, Rekha Viswanathan, Beth Jelfs, Sridhar Arjunan, Sanjay Raghav, and Dinesh Kumar. Which gait parameters and walking patterns show the significant differences between Parkinson's disease and healthy participants? *Biosensors*, 9(2):59, 2019.

[135] Dante Trabassi, Mariano Serrao, Tiwana Varrecchia, Alberto Ranavolo, Gianluca Coppola, Roberto De Icco, Cristina Tassorelli, and Stefano Filippo Castiglia. Machine learning approach to support the detection of Parkinson's disease in IMU-based gait analysis. *Sensors*, 22(10):3700, 2022.

[136] Frédéric Marin, Elke Warmerdam, Zoé Marin, Khalil Ben Mansour, Walter Maetzler, and Clint Hansen. Scoring the sit-to-stand performance of Parkinson's patients with a single wearable sensor. *Sensors*, 22(21):8340, 2022.

[137] Khadija Gourrame, Julius Griškevičius, Michel Haritopoulos, Donatas Lukšys, Dalius Jatužis, Rūta Kaladytė-Lokominienė, Ramunė Bunevičiūtė, and Gabrielė Mickutė. Parkinson's disease classification with CWNN: using wavelet transformations and IMU data fusion for improved accuracy. *Technology and Health Care*, 31(6):2447–2455, 2023.

[138] Lin Meng, Jun Pang, Yifan Yang, Lei Chen, Rui Xu, and Dong Ming. Inertial-based gait metrics during turning improve the detection of early-stage Parkinson's disease patients. *IEEE Transactions Neural Systems and Rehabilitation Engineering*, 31:1472–1482, 2023.

[139] Carlotta Caramia, Diego Torricelli, Maurizio Schmid, Adriana Muñoz-Gonzalez, Jose Gonzalez-Vargas, Francisco Grandas, and Jose L. Pons. IMU-based classification of Parkinson's disease from gait: a sensitivity analysis on sensor location and feature selection. *IEEE Journal of Biomedical and Health Informatics*, 22(6):1765–1774, 2018.

[140] Nikhil Mahadevan, Charmaine Demanuele, Hao Zhang, Dmitri Volfson, Bryan Ho, Michael Erb, and Shyamal Patel. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. *npj Digital Medicine*, 3:5, 2020.

[141] Zhirong Lin, Houde Dai, Yongsheng Xiong, Xuke Xia, and Shi-Jinn Horng. Quantification assessment of bradykinesia in Parkinson's disease based on a wearable device. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 803–806, 2017.

[142] Luigi Borzì, Ivan Mazzetta, Alessandro Zampogna, Antonio Suppa, Fernanda Irrera, and Gabriella Olmo. Predicting axial impairment in Parkinson's disease through a single inertial sensor. *Sensors*, 22(2):412, 2022.

[143] Gent Ymeri, Dario Salvi, Carl Magnus Olsson, Myrthe Vivianne Wassenburg, Athanasios Tsanas, and Per Svenningsson. Quantifying Parkinson's disease severity using mobile wearable devices and machine learning: the ParkApp pilot study protocol. *BMJ Open*, 13(12):e077766, 2023.

[144] Vladislava Bobić, Milica Djurić-Jovičić, Nataša Dragašević, Mirjana B. Popović, Vladimir S. Kostić, and Goran Kvaščev. An expert system for quantification of bradykinesia based on wearable inertial sensors. *Sensors*, 19(11):2644, 2019.

[145] Yi Han, Xiangzhi Liu, Ning Zhang, Xiufeng Zhang, Bin Zhang, Shuoyu Wang, Tao Liu, and Jingang Yi. Automatic assessments of parkinsonian gait with wearable sensors for human assistive systems. *Sensors*, 23(4):2104, 2023.

[146] Margot Heijmans, Jeroen Habets, Christian Herff, Jos Aarts, An Stevens, Mark Kuijf, and Pieter Kubben. Monitoring Parkinson's disease symptoms during daily life: a feasibility study. *npj Parkinson's Disease*, 5:21, 2019.

[147] Zhuang Wu, Xu Jiang, Min Zhong, Bo Shen, Jun Zhu, Yang Pan, Jingde Dong, Pingyi Xu, Wenbin Zhang, and li Zhang. Wearable sensors measure ankle joint changes of patients with Parkinson's disease before and after acute levodopa challenge. *Parkinson's Disease*, 2020:2976535, 2020.

[148] Xiaolong Wu, Lin Ma, Penghu Wei, Yongzhi Shan, Piu Chan, Kailiang Wang, and Guoguang Zhao. Wearable sensor devices can automatically identify the ON-OFF status of patients with Parkinson's disease through an interpretable machine learning model. *Frontiers in Neurology*, 15:1387477, 2024.

[149] R. Jamil Pugh, Rosemary Higgins, Hua Min, Clinton Wutzke, and Andrew Guccione. Turns while walking among individuals with Parkinson's disease following overground locomotor training: a pilot study. *Clinical Biomechanics*, 114:106234, 2024.

[150] Mehdi Delrobaei, Sara Memar, Marcus Pieterman, Tyler W. Stratton, Kenneth McIsaac, and Mandar Jog. Towards remote monitoring of Parkinson's disease tremor using wearable motion capture systems. *Journal of the Neurological Sciences*, 384:38–45, 2018.

[151] Sara Memar, Mehdi Delrobaei, Marcus Pieterman, Kenneth McIsaac, and Mandar Jog. Quantification of whole-body bradykinesia in Parkinson's disease participants using multiple inertial sensors. *Journal of the Neurological Sciences*, 387:157–165, 2018.

[152] Lazzaro di Biase, Susanna Summa, Jacopo Tosi, Fabrizio Taffoni, Massimo Marano, Angelo Cascio Rizzo, Fabrizio Vecchio, Domenico Formica, Vincenzo Di Lazzaro, Giovanni Di Pino, and Mario Tombini. Quantitative analysis of bradykinesia and rigidity in Parkinson's disease. *Frontiers in Neurology*, 9:121, 2018.

[153] Mehdi Delrobaei, Navid Baktash, Greydon Gilmore, Kenneth McIsaac, and Mandar Jog. Using wearable technology to generate objective Parkinson's disease dyskinesia severity score: possibilities for home monitoring. *IEEE Transactions Neural Systems and Rehabilitation Engineering*, 25(10):1853–1863, 2017.

[154] Mehar Singh, Prithvi Prakash, Rachneet Kaur, Richard Sowers, James Robert Brašić, and Manuel Enrique Hernandez. A deep learning approach for automatic and objective grading of the motor impairment severity in Parkinson's disease for use in tele-assessments. *Sensors*, 23(21):9004, 2023.

[155] Prithvi Prakash, Rachneet Kaur, Joshua Levy, Richard Sowers, James Brašić, and Manuel E. Hernandez. A deep learning approach for grading of motor impairment severity in Parkinson's disease. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4, 2023.

[156] Rene Peter Bremm, Lukas Pavelka, Maria Moscardo Garcia, Laurent Mombaerts, Rejko Krüger, and Frank Hertel. Sensor-based quantification of MDS-UPDRS III subitems in Parkinson's disease using machine learning. *Sensors*, 24(7):2195, 2024.

[157] Weiguang Huo, Paolo Angeles, Yen F Tai, Nicola Pavese, Samuel Wilson, Michele T. Hu, and Ravi Vaidyanathan. A heterogeneous sensing suite for multisymptom quantification of Parkinson's disease. *IEEE Transactions Neural Systems and Rehabilitation Engineering*, 28(6):1397–1406, 2020.

[158] Fujian Yan, Jiaqi Gong, Qiang Zhang, and Hongsheng He. Learning motion primitives for the quantification and diagnosis of mobility deficits. *IEEE Transactions Biomedical Engineering*, pages 1–10, 2024.

[159] Tibor Zajki-Zechmeister, Mariella Kögl, Kerstin Kalsberger, Sebastian Franthal, Nina Homayoon, Petra Katschnig-Winter, Karoline Wenzel, László Zajki-Zechmeister, and Petra Schwingenschuh. Quantification of tremor severity with a mobile tremor pen. *Heliyon*, 6(8):e04702, 2020.

[160] Vassilis Tsakanikas, Adamantios Ntanis, George Rigas, Christos Androutsos, Dimitrios Boucharas, Nikolaos Tachos, Vasileios Skaramagkas, Chariklia Chatzaki, Zinovia Kefalopoulou, Manolis Tsiknakis, and Dimitrios Fotiadis. Evaluating gait impairment in Parkinson's disease from instrumented insole and IMU sensor data. *Sensors*, 23(8):3902, 2023.

[161] Taylor Chomiak, Wenbiao Xian, Zhong Pei, and Bin Hu. A novel single-sensor-based method for the detection of gait-cycle breakdown and freezing of gait in Parkinson's disease. *Journal of Neural Transmission*, 126:1029–1036, 2019.

[162] Catalina Punin, Boris Barzallo, Roger Clotet, Alexander Bermeo, Marco Bravo, Juan Pablo Bermeo, and Carlos Llumiguano. A non-invasive medical device for Parkinson's patients with episodes of freezing of gait. *Sensors*, 19(3):737, 2019.

[163] Bo Li, Yuqian Zhang, Liang Tang, Chao Gao, and Dongyun Gu. Automatic detection system for freezing of gait in Parkinson's disease based on the clustering algorithm. In *2018 2nd IEEE advanced information management, communicates, electronic and automation control conference (IMCEC)*, pages 1640–9, 2018.

[164] Juli Camps, Albert Sam, Mario Martn, Daniel Rodrguez-Martn, Carlos Prez-Lpez, Joan M. Moreno Arostegui, Joan Cabestany, Andreu Catal, Sheila Alcaine, Berta Mestre, Anna Prats, Maria C. Crespo-Maraver, Timothy J. Counihan, Patrick Browne, Leo R. Quinlan, Gearid Laighin, Dean Sweeney, Hadas Lewy, Gabriel Vainstein, Alberto Costa, Roberta Annicchiarico, ngels Bays, and Alejandro Rodrguez-Molinero. Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit. *Knowledge -Based Systems*, 139:119–131, 2018.

[165] Albert Samà, Daniel Rodríguez-Martín, Carlos Pérez-López, Andreu Català, Sheila Alcaine, Berta Mestre, Anna Prats, M. Cruz Crespo, and Àngels Bayés. Determining the optimal features in freezing of gait detection through a single waist accelerometer in home environments. *Pattern Recognition Letters*, 105:135–143, 2018.

[166] Marianna Capecci, Lucia Pepa, Federica Verdini, and Maria Gabriella Ceravolo. A smartphone-based architecture to detect and quantify freezing of gait in Parkinson's disease. *Gait & Posture*, 50:28–33, 2016.

[167] Tal Reches, Moria Dagan, Talia Herman, Eran Gazit, Natalia A. Gouskova, Nir Giladi, Brad Manor, and Jeffrey M. Hausdorff. Using wearable sensors and machine learning to automatically detect freezing of gait during a FOG-provoking test. *Sensors*, 20(16):4474, 2020.

[168] Thuy T. Pham, Steven T. Moore, Simon John Geoffrey Lewis, Diep N. Nguyen, Eryk Dutkiewicz, Andrew J. Fuglevand, Alistair L. McEwan, and Philip H.W. Leong. Freezing of gait detection in Parkinson's disease: a subject-independent detector using anomaly scores. *IEEE Transactions Biomedical Engineering*, 64(11):2719–2728, 2017.

[169] Spyroula Masiala, Willem Huijbers, and Martin Atzmüller. Feature-set-engineering for detecting freezing of gait in Parkinson's disease using deep recurrent neural networks. *arXiv preprint arXiv:1909.03428*, 2019.

[170] Martina Mancini, Vrutangkumar V. Shah, Samuel Stuart, Carolin Curtze, Fay B. Horak, Delaram Safarpour, and John G. Nutt. Measuring freezing of gait during daily-life: an open-source, wearable sensors approach. *Journal of NeuroEngineering and Rehabilitation*, 18(1):1, 2021.

[171] Andrea Marcante, Roberto Di Marco, Giovanni Gentile, Clelia Pellicano, Francesca Assogna, Francesco Ernesto Pontieri, Gianfranco Spalletta, Lucia Macchiusi, Dimitris Gatsios, Alexandros Giannakis, Maria Chondrogiorgi, Spyridon Konitsiotis, Dimitrios I. Fotiadis, and Angelo Antonini. Foot pressure wearable sensors for freezing of gait detection in Parkinson's disease. *Sensors*, 21(1):128, 2021.

[172] Luis Sigcha, Nélson Costa, Ignacio Pavón, Susana Costa, Pedro Arezes, Juan Manuel López, and Guillermo De Arcas. Deep learning approaches for detecting freezing of gait in Parkinson's disease patients through on-body acceleration sensors. *Sensors*, 20(7):1895, 2020.

[173] Thomas Bikias, Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, and Leontios J. Hadjileontiadis. DeepFoG: an IMU-based detection of freezing of gait episodes in Parkinson's disease patients via deep learning. *Frontiers in Robotics and AI*, 8:537384, 2021.

[174] Scott Pardoel, Gaurav Shalin, Julie Nantel, Edward D. Lemaire, and Jonathan Kofman. Early detection of freezing of gait during walking using inertial measurement unit and plantar pressure distribution data. *Sensors*, 21(6):2246, 2021.

[175] Bohan Shi, Arthur Tay, W. L. Au, Dawn M. L. Tan, Nicole S. Y. Chia, and Shih-Cheng Yen. Detection of freezing of gait using convolutional neural networks and data from lower limb motion sensors. *IEEE Transactions Biomedical Engineering*, 69(7):2256–2267, 2022.

[176] Linda Rennie, Niklas Löfgren, Rolf Moe-Nilssen, Arve Opheim, Espen Dietrichs, and Erika Franzén. The reliability of gait variability measures for individuals with Parkinson's disease and healthy older adults – the effect of gait speed. *Gait & Posture*, 62:505–509, 2018.

[177] Peter S. Myers, Marie E. McNeely, Kristen A. Pickett, Ryan P. Duncan, and Gammon M. Earhart. Effects of exercise on gait and motor imagery in people with Parkinson disease and freezing of gait. *Parkinsonism & Related Disorders*, 53:89–95, 2018.

[178] Assaf Zadka, Neta Rabin, Eran Gazit, Anat Mirelman, Alice Nieuwboer, Lynn Rochester, Silvia Del Din, Elisa Pelosin, Laura Avanzino, Bastiaan R. Bloem, Ugo Della Croce, Andrea Cereatti, and Jeffrey M. Hausdorff. A wearable sensor and machine learning estimate step length in older adults and patients with neurological disorders. *npj Digital Medicine*, 7:142, 2024.

[179] Nooshin Haji Ghassemi, Julius Hannink, Nils Roth, Heiko Gaßner, Franz Marxreiter, Jochen Klucken, and Björn M. Eskofier. Turning analysis during standardized test using on-shoe wearable sensors in Parkinson's disease. *Sensors*, 19(14):3103, 2019.

[180] An Nguyen, Nils Roth, Nooshin Haji Ghassemi, Julius Hannink, Thomas Seel, Jochen Klucken, Heiko Gaßner, and Bjoern Eskofier. Development and clinical validation of inertial sensor-based gait-clustering methods in Parkinson's disease. *Journal of NeuroEngineering and Rehabilitation*, 16:77, 2019.

[181] Piergiuseppe Liuzzi, Ilaria Carpinella, Denise Anastasi, Elisa Gervasoni, Tiziana Lencioni, Rita Bertoni, Maria Chiara Carrozza, Davide Cattaneo, Maurizio Ferrarin, and Andrea Mannini. Machine learning based estimation of dynamic balance and gait adaptability in persons with neurological diseases using inertial sensors. *Scientific Reports*, 13:8640, 2023.

[182] Emma Stack, Veena Agarwal, Rachel King, Malcolm Burnett, Fatemeh Tahavori, Balazs Janko, William Harwin, Ann Ashburn, and Dorit Kunkel. Identifying balance impairments in people with Parkinson's disease using video and wearable sensors. *Gait & Posture*, 62:321–326, 2018.

[183] Deborah Jehu and Julie Nantel. Fallers with Parkinson's disease exhibit restrictive trunk control during walking. *Gait & Posture*, 65:246–250, 2018.

[184] Xiaodan Zhang, Yulan Jin, Mateng Wang, Chengcheng Ji, Zhaoying Chen, Weinv Fan, Timothy Rainer, Qiongfeng Guan, and Qianyun Li. The impact of anxiety on gait impairments in Parkinson's disease: insights from sensor-based gait analysis. *Journal of NeuroEngineering and Rehabilitation*, 21:68, 2024.

[185] Robbin Romijnders, Elke Warmerdam, Clint Hansen, Julius Welzel, Gerhard Schmidt, and Walter Maetzler. Validation of IMU-based gait event detection during curved walking and turning in older adults and Parkinson's disease patients. *Journal of Neuroengineering and Rehabilitation*, 18(1):28, 2021.

[186] Donatas Lukšys and Julius Griškevičius. Application of nonlinear analysis for the assessment of gait in patients with Parkinson's disease. *Technology and Health Care*, 30(1):201–208, 2022.

[187] Stefano Negrini, Mauro Serpelloni, Cinzia Amici, Massimiliano Gobbo, Clara Silvestro, Riccardo Buraschi, Alberto Borboni, Diego Crovato, and Nicola Francesco Lopomo. Use of wearable inertial sensor in the assessment of Timed-Up-and-Go Test: influence of device placement on temporal variable estimation. In *Wireless Mobile Communication and Healthcare*, pages 310–317, 2017.

[188] W.R. Gibb and A.J. Lees. The relevance of the lewy body to the pathogenesis of idiopathic Parkinson's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 51:745–752, 1988.

[189] Marzieh Keshtkarjahromi, Danielle S. Abraham, Ann L. Gruber-Baldini, Katrina Schrader, Stephen G. Reich, Joseph M. Savitt, Rainer von Coelln, and Lisa M. Shulman. Confirming Parkinson disease diagnosis: patterns of diagnostic changes by movement disorder specialists. *Parkinson's Disease*, 2022, 2022.

[190] Shadi Ghourchian, Ann L. Gruber-Baldini, Sunita Shakya, John Herndon, Stephen G. Reich, Rainer von Coelln, Joseph M. Savitt, and Lisa M. Shulman. Weight loss and weight gain in Parkinson disease. *Parkinsonism & Related Disorders*, 83:31–36, 2021.

[191] Lisa M. Shulman, Ann L. Gruber-Baldini, Karen E. Anderson, Paul S. Fishman, Stephen G. Reich, and William J. Weiner. The clinically important difference on the unified Parkinson's disease rating scale. *Archives of Neurology*, 67:64–70, 2010.

[192] Mark D. Miller, Cynthia F. Paradis, Patricia R. Houck, Sati Mazumdar, Jacqueline A. Stack, A. Hind Rifai, Benoit Mulsant, and Charles F. Reynolds. Rating chronic medical illness burden in geropsychiatric practice and research: application of the cumulative illness rating scale. *Psychiatry Research*, 41(3):237–248, 1992.

[193] Michael A. Smyer Gerda G. Fillenbaum. The development, validity, and reliability of the OARS multidimensional functional assessment questionnaire. *Journal of Gerontology*, 36:428–434, 1981.

[194] David Cella, Seung W. Choi, David M. Condon, Ben Schalet, Ron D. Hays, Nan E. Rothrock, Susan Yount, Karon F. Cook, Richard C. Gershon, Dagmar Amtmann, Darren A. DeWalt, Paul A. Pilkonis, Arthur A. Stone, Kevin Weinfurt, and Bryce B. Reeve. Promis® adult health profiles: efficient short-form measures of seven health domains. *Value Health*, 22:537–544, 2019.

[195] Ziad S. Nasreddine, Natalie A. Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L. Cummings, and Howard Chertkow. The montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53:695–699, 2005.

[196] Christopher G. Goetz. The unified Parkinson's disease rating scale (UPDRS): status and recommendations. *Movement Disorders*, 18:738–750, 2003.

[197] Christopher G. Goetz, Werner Poewe, Olivier Rascol, Cristina Sampaio, Glenn T. Stebbins, Carl Counsell, Nir Giladi, Robert G. Holloway, Charity G. Moore, Gregor K. Wenning, Melvin D. Yahr, and Lisa Seidl. Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: status and recommendations. *Movement Disorders*, 19:1020–1028, 2004.

[198] Rainer von Coelln, Robert J. Dawe, Sue E. Leurgans, Thomas A. Curran, Timothy Truty, Lei Yu, Lisa L. Barnes, Joshua M. Shulman, Lisa M. Shulman, David A. Bennett, Jeffrey M. Hausdorff, and Aron S. Buchman. Quantitative mobility metrics from a wearable sensor predict incident parkinsonism in older adults. *Parkinsonism & Related Disorders*, 65:190–196, 2019.

[199] Robert J. Dawe, Sue E. Leurgans, Jingyun Yang, Joshua M. Bennett, Jeffrey M. Hausdorff, Andrew S. Lim, Chris Gaiteri, David A. Bennett, and Aron S. Buchman. Association between quantitative gait and balance measures and total daily physical activity in community-dwelling older adults. *Journals of Gerontology A Biological Sciences and Medical Sciences*, 73:636–642, 2018.

[200] David A. Bennett, Aron S. Buchman, Patricia A. Boyle, Lisa L. Barnes, Robert S. Wilson, and Julie A. Schneider. Religious orders study and Rush memory and aging project. *Journal of Alzheimer's Disease*, 64(s1):S161–S189, 2018.

[201] Mandy Miller Koop, Sarah J. Ozinga, Anson B Rosenfeldt, and Jay L Alberts. Quantifying turning behavior and gait in Parkinson's disease using mobile technology. *IBRO Reports*, 5:10–16, 2018.

[202] Richard L. Burden and J. Douglas Faires. *Numerical Analysis (9th ed.)*. Brooks/Cole, Boston, MA, 2015.

[203] Arash Salarian, Cris Zampieri, Fay B. Horak, Patricia Carlson-Kuhta, John G. Nutt, and Kamiar Aminian. Analyzing 180 degrees turns using an inertial system reveals early signs of progression of Parkinson's disease. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 224–227, 2009.

[204] Stephen Butterworth. On the theory of filter amplifiers. *Experimental Wireless and the Wireless Engineer*, 7:536–541, 1930.

[205] Nicholas R. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39(2):447–462, 1976.

[206] Jeffrey D. Scargle. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, 1982.

[207] Maryam Banitalebi Dehkordi, Abolfazl Zaraki, and Rossitza Setchi. Feature extraction and feature selection in smartphone-based activity recognition. *Procedia Computer Science*, 176:2655–2664, 2020.

[208] Aditya Mehta, Sai Krishna Vaddadi, Vinay Sharma, and Prateek Kala. A phase-wise analysis of machine learning based human activity recognition using inertial sensors. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–7, 2020.

[209] Alexander Weiss, Talia Herman, Meir Plotnik, Marina Brozgol, N. Giladi, and J.M. Hausdorff. An instrumented timed up and go: the added value of an accelerometer for identifying fall risk in idiopathic fallers. *Physiological Measurement*, 32:2003, 2011.

[210] Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer, New York, 1998.

[211] Aron S. Buchman, Robert J. Dawe, Sue E. Leurgans, Thomas A. Curran, Timothy Truty, Lei Yu, Lisa L. Barnes, Jeffrey M. Hausdorff, and David A. Bennett. Different combinations of mobility metrics derived from a wearable sensor are associated with distinct health outcomes in older adults. *Journals of Gerontology A*, 75(6):1176–1183, 2019.

[212] Victoria N. Poole, Robert J. Dawe, Melissa Lamar, Michael Esterman, Lisa Barnes, Sue E. Leurgans, David A. Bennett, Jeffrey M. Hausdorff, and Aron S. Buchman. Dividing attention during the Timed Up and Go enhances associations of several subtask performances with MCI and cognition. *PLOS ONE*, 17:1–13, 2022.

[213] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, Berlin, 2nd edition, 2009.

[214] Mark J. Van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6:25, 2007.

[215] John Ashworth Nelder and Robert W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, 135:370–384, 1972.

[216] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[217] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[218] Tianqi Chen and Carlos Guestrin. XGboost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[219] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[220] Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics Data Analysis*, 90:15–35, 2015.

[221] Angelo Canty and Brian Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2024. R package version 1.3-30.

[222] Anthony Davison and D. Hinkley. *Bootstrap methods and their applications*. Cambridge University Press, 1997.

[223] Charles H. Adler, Thomas G. Beach, Joseph G. Hentz, Holly A. Shill, John N. Caviness, Erika Driver-Dunckley, Marwan N. Sabbagh, Lucia I. Sue, Sandra A. Jacobson, Christine M. Belden, and Brittany N. Dugger. Low clinical diagnostic accuracy of early vs advanced Parkinson disease: clinicopathologic study. *Neurology*, 83:406–412, 2014.

[224] Marta Isabel A.S.N. Ferreira, Fabio Augusto Barbieri, Vinícius Christianini Moreno, Tiago Penedo, and João Manuel R.S. Tavares. Machine learning models for Parkinson's disease detection and stage classification based on spatial-temporal gait parameters. *Gait & Posture*, 98:49–55, 2022.

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

[226] Sanghee Moon, Hyun-Je Song, Vibhash D. Sharma, Kelly E. Lyons, Rajesh Pahwa, Abiodun E. Akinwuntan, and Hannes Devos. Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach. *Journal of Neuroengineering and Rehabilitation*, 17:125, 2020.

[227] Natasa Kleanthous, Abir Jaafar Hussain, Wasiq Khan, and Panos Liatsis. A new machine learning based approach to predict freezing of gait. *Pattern Recognition Letters*, 140:119–126, 2020.

[228] Anat Mirelman, Mor Ben Or Frank, Michal Melamed, Lena Granovsky, Alice Nieuwboer, Lynn Rochester, Silvia Del Din, Laura Avanzino, Elisa Pelosin, Bastiaan R. Bloem, Ugo Della Croce, Andrea Cereatti, Paolo Bonato, Richard Camicioli, Theresa Ellis, Jamie L Hamilton, Chris J. Hass, Quincy J. Almeida, Maidan Inbal, Avner Thaler, Julia Shirvan, Jesse M. Cedarbaum, Nir Giladi, and Jeffrey M. Hausdorff. Detecting sensitive mobility features for Parkinson's disease stages via machine learning. *Movement Disorders*, 36:2144–2155, 2021.

[229] Md Faisal Kabir and Simone Ludwig. Enhancing the performance of classification using super learning. *Data-Enabled Discovery and Applications*, 3:5, 2019.

[230] Marie Demonceau, Anne-Françoise Donneau, Jean-Louis Croisier, Eva Skawiniak, Mohamed Boutaayamou, Didier Maquet, and Gaëtan Garraux. Contribution of a trunk accelerometer system to the characterization of gait in patients with mild-to-moderate Parkinson's disease. *IEEE Journal of Biomedical and Health Informatics*, 19(6):1803–1808, 2015.

[231] Janet M. T. Van Uem, Stefan Walgaard, Erik Ainsworth, Sandra E. Hasmann, Tanja Heger, Susanne Nussbaum, Markus A. Hobert, Encarnación M. Micó-Amigo, Rob C. Van Lummel, Daniela Berg, and Walter Maetzler. Quantitative Timed-Up-and-Go parameters in relation to cognitive parameters and health-related quality of life in mild-to-moderate Parkinson's disease. *PLOS ONE*, 11(4):1–15, 2016.

[232] Nicholas Shawen, Megan O'Brien, Sanjeev Venkatesan, Luca Lonini, Tanya Simuni, Jamie Hamilton, Roozbeh Ghaffari, John Rogers, and Arun Jayaraman. Role of data measurement characteristics in the accurate detection of Parkinson's disease symptoms using wearable sensors. *Journal of NeuroEngineering and Rehabilitation*, 17:52, 2020.

[233] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9:307, 2008.

[234] Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *Statistics Computing*, 27:659–678, 2017.

[235] Sigurlaug Sveinbjornsdottir. The clinical symptoms of Parkinson's disease. *Journal of Neurochemistry*, 139:318–324, 2016.

[236] Ruiping Xia and Zhi-Hong Mao. Progression of motor symptoms in Parkinson's disease. *Neuroscience Bulletin*, 28:39–48, 2012.

[237] Ahmed A. Moustafa, Srinivasa Chakravarthy, Joseph R. Phillips, Ankur Gupta, Szabolcs Keri, Bertalan Polner, Michael J. Frank, and Marjan Jahanshahi. Motor symptoms in Parkinson's disease: a unified framework. *Neuroscience Biobehavioral Reviews*, 68:727–740, 2016.

[238] Dag Aarsland, Lucia Batzu, Glenda M. Halliday, Gert J. Geurtsen, Clive Ballard, K. Ray Chaudhuri, and Daniel Weintraub. Parkinson disease-associated cognitive impairment. *Nature Reviews Disease Primers*, 7:47, 2021.

[239] Diane Podsiadlo and Sandra Richardson. The Timed Up & Go: a test of basic functional mobility for frail elderly persons. *Journal of the American Geriatrics Society*, 39:142–148, 1991.

[240] Anne Shumway-Cook, Sandy Brauer, and Marjorie Woollacott. Predicting the probability for falls in community-dwelling older adults using the Timed Up & Go test. *Physical Therapy*, 80:896–903, 2000.

[241] Roisin C. Vance, Dan G. Healy, Rose Galvin, and Helen P. French. Dual tasking with the Timed Up & Go test improves detection of risk of falls in people with Parkinson disease. *Physical Therapy*, 95:95–102, 2015.

[242] Pilar Serra-Añó, José Francisco Pedrero-Sánchez, Marta Inglés, Marta Aguilar-Rodríguez, Ismael Vargas-Villanueva, and Juan López-Pascual. Assessment of functional activities in individuals with Parkinson's disease using a simple and reliable smartphone-based procedure. *International Journal of Environmental Research and Public Health*, 17:4123, 2020.

[243] Susan Morris, Meg E. Morris, and Robert Iansek. Reliability of measurements obtained with the Timed Up & Go test in people with parkinson disease. *Physical Therapy*, 81:810–818, 2001.

[244] Erin Smith, Lorcan Walsh, Julie Doyle, Barry Greene, and Catherine Blake. The reliability of the quantitative timed up and go test (QTUG) measured over five consecutive days under single and dual-task conditions in community dwelling older adults. *Gait & Posture*, 43:239–244, 2016.

[245] Jason K. Longhurst, John V. Rider, Jeffrey L. Cummings, Samantha E. John, Brach Poston, Elissa C. Held Bradford, and Merrill R. Landers. A novel way of measuring dual-task interference: the reliability and construct validity of the dual-task effect battery in neurodegenerative disease. *Neurorehabilitation and Neural Repair*, 36:346–359, 2022.

[246] Julie D. Ries, John L. Echternach, Leah Nof, and Michelle Gagnon Blodgett. Test-retest reliability and minimal detectable change scores for the Timed Up & Go test, the six-minute walk test, and gait speed in people with Alzheimer disease. *Physical Therapy*, 89:569–579, 2009.

[247] Carole L. Kimberlin and Almut G. Winterstein. Validity and reliability of measurement instruments used in research. *American journal of health-system pharmacy (AJHP)*, 65:2276–2284, 2008.

[248] Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–63, 2016.

[249] Thâmela Thaís Santos dos Santos, Amélia Pasqual Marques, Luis Carlos Pereira Monteiro, Enzo Gabriel da Rocha Santos, Gustavo Henrique Lima Pinto, Anderson Belgamo, Anselmo de Athayde Costa e Silva, André dos Santos Cabral, Szymon Kuliś, Jan Gajewski, Givago Silva Souza, Tacyla Jesus da Silva, Wesley Thyago Alves da Costa, Railson Cruz Salomão, and Bianca Callegari. Intra and inter-device reliabilities of the instrumented timed-up and go test using smartphones in young adult population. *Sensors*, 24(9):2918, 2024.

[250] Pernille Botolfsen, Jorunn L Helbostad, Rolf Moe-nilssen, and James C Wall. Reliability and concurrent validity of the expanded Timed Up-and-Go test in older people with impaired mobility. *Physiotherapy Research International*, 13(2):94–106, 2008.

[251] Denise McGrath, Barry R. Greene, Emer P. Doheny, David J. McKeown, Giuseppe De Vito, and Brian Caulfield. Reliability of quantitative TUG measures of mobility for use in falls risk assessment. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 466–469, 2011.

[252] Arash Salarian, Fay B. Horak, Cris Zampieri, Patricia Carlson-Kuhta, John G. Nutt, and Kamiar Aminian. iTUG, a sensitive and reliable measure of mobility. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18:303–310, 2010.

[253] Rob C. van Lummel, Stefan Walgaard, Markus A. Hobert, Walter Maetzler, Jaap H. van Dieën, Francisca Galindo-Garre, and Caroline B. Terwee. Intra-rater, inter-rater and test-retest reliability of an instrumented timed up and go (iTUG) test in patients with Parkinson's disease. *PLOS ONE*, 11(3):1–11, 2016.

[254] Christian Werner, Patrick Heldmann, Saskia Hummel, Laura Bauknecht, Jürgen M. Bauer, and Klaus Hauer. Concurrent validity, test-retest reliability, and sensitivity to change of a single body-fixed sensor for gait analysis during rollator-assisted walking in acute geriatric patients. *Sensors*, 20(17):4866, 2020.

[255] Sheau-Ling Huang, Ching-Lin Hsieh, Ruey-Meei Wu, Chun-Hwei Tai, Chin-Hsien Lin, and Wen-Shian Lu. Minimal detectable change of the Timed Up & Go test and the dynamic gait index in people with Parkinson disease. *Physical Therapy*, 91:114–121, 2011.

[256] Ellinor Nordin, Erik Rosendahl, and Lillemor Lundin-Olsson. Timed Up & Go test: reliability in older people dependent in activities of daily living— focus on cognitive state. *Physical Therapy*, 86:646–655, 2006.

[257] Lei Yang, Chengqi He, and Marco Yiu Chung Pang. Reliability and validity of dual-task mobility assessments in people with chronic stroke. *PLOS ONE*, 11:1–22, 2016.

[258] Talia Herman, Aner Weiss, Marina Brozgol, Nir Giladi, and Jeffrey M. Hausdorff. Gait and balance in Parkinson's disease subtypes: objective measures and classification considerations. *Journal of Neurology*, 261:2401–2410, 2014.

[259] Rolf Moe-Nilssen and Jorunn L. Helbostad. Estimation of gait cycle characteristics by trunk accelerometry. *Journal of Biomechanics*, 37(1):121–126, 2004.

[260] James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

[261] Kenneth O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1:30–46, 1996.

[262] Kathleen Wyrwich and Fredric Wolinsky. Identifying meaningful intra-individual change standards for health-related quality of life measures. *Journal of Evaluation in Clinical Practice*, 6:39–49, 2000.

[263] Leslie G. Portney and Mary P. Watkins. *Foundations of clinical research: applications to practice*, volume 892. Pearson/Prentice Hall, 2009.

[264] J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327:307–310, 1986.

[265] Katrine Lyders Johansen, Rikke Derby Stistrup, Camilla Skibdal Schjøtt, Jacqueline Madsen, and Anders Vinther. Absolute and relative reliability of the Timed Up & Go test and 30second chair-stand test in hospitalised patients with stroke. *PLOS ONE*, 11:1–14, 2016.

[266] James Beyea, Chris A. McGibbon, Andrew Sexton, Jeremy Noble, and Colleen O'Connell. Convergent validity of a wearable sensor system for measuring sub-task performance during the timed up-and-go test. *Sensors*, 17(4):934, 2017.

[267] Margaret Schenkman, Richard A. Berger, Patrick O. Riley, Robert W. Mann, and W. Andrew Hodge. Whole-body movements during rising to standing from sitting. *Physical Therapy*, 70(10):638–648, 1990.

[268] Stephen Ashford and Lorraine De Souza. A comparison of the timing of muscle activity during sitting down compared to standing up. *Physiotherapy Research International*, 5(2):111–128, 2000.

[269] Wim G.M. Janssen, Hans B.J. Bussmann, and Henk J. Stam. Determinants of the sit-to-stand movement: a review. *Physical Therapy*, 82(9):866–879, 2002.

[270] Pamela J. Millington, Barbara M. Myklebust, Barbara M. Myklebust, and Georgia M. Shambes. Biomechanical analysis of the sit-to-stand motion in elderly persons. *Archives of Physical Medicine and Rehabilitation*, 73(7):609–617, 1992.

[271] Eduardo Tolosa, Alicia Garrido, Sonja W. Scholz, and Werner Poewe. Challenges in the diagnosis of Parkinson's disease. *Lancet Neurology*, 20:385, 2021.

[272] Arjun Tarakad and Joseph Jankovic. Essential tremor and Parkinson's disease: exploring the relationship. *Tremor and Other Hyperkinetic Movements*, 8:589, 2019.

[273] Johannes Levin, Alexander Kurz, Thomas Arzberger, Armin Giese, and Günter U. Höglinger. The differential diagnosis and treatment of atypical parkinsonism. *Deutsches Ärzteblatt International*, 113:61, 2016.

[274] Doug Everett Hobson. Clinical manifestations of Parkinson's disease and parkinsonism. *Canadian Journal of Neurological Sciences*, 30:S2–S9, 2003.

[275] Jeannette R. Mahoney, Joe Verghese, Roee Holtzer, and Gilles Allali. The evolution of mild parkinsonian signs in aging. *Journal of Neurology*, 261(10):1922–1928, 2014.

[276] Robert Caslake, Jolene Moore, J. Gordon, C. Harris, and Carl Counsell. Changes in diagnosis with follow-up in an incident cohort of patients with parkinsonism. *Journal of Neurology, Neurosurgery, and Psychiatry*, 79:1202–1207, 2008.

[277] Eduardo Tolosa, Gregor Wenning, and Werner Poewe. The diagnosis of Parkinson's disease. *Lancet Neurology*, 5:75–86, 2006.

[278] Maria Stathaki, Sophia Koukouraki, Panagiotis Simos, Iro Boura, Emmanouela Papadaki, Olga Bourogianni, Angeliki Tsaroucha, Nikolaos Kapsoritakis, Panayiotis Mitsias, and Cleanthe Spanaki. Is there any clinical value of adding 123i-metaiodobenzylguanidine myocardial scintigraphy to 123i-ioflupane (DaTscan) in the differential diagnosis of parkinsonism? *Clinical Nuclear Medicine*, 45(8):588—593, 2020.

[279] Anat Mirelman, Paolo Bonato, Richard Camicioli, Terry D. Ellis, Nir Giladi, Jamie L. Hamilton, Chris J. Hass, Jeffrey M. Hausdorff, Elisa Pelosin, and Quincy J. Almeida. Gait impairments in Parkinson's disease. *Lancet Neurology*, 18:697–708, 2019.

[280] Dustin A. Heldman, Alberto J. Espay, Peter A. LeWitt, and Joseph P. Giuffrida. Clinician versus machine: reliability and responsiveness of motor endpoints in Parkinson's disease. *Parkinsonism & Related Disorders*, 20:590–595, 2014.

[281] Cecilia Raccagni, Heiko Gaßner, Sabine Eschlboeck, Sylvia Boesch, Florian Krismer, Klaus Seppi, Werner Poewe, Bjoern M. Eskofier, Juergen Winkler, Gregor Wenning, and Jochen Klucken. Sensor-based gait analysis in atypical parkinsonian disorders. *Brain and Behavior*, 8:e00977, 2018.

[282] Heiko Gaßner, Cecilia Raccagni, Bjoern M. Eskofier, Jochen Klucken, and Gregor K. Wenning. The diagnostic scope of sensor-based gait analysis in atypical parkinsonism: further observations. *Frontiers in Neurology*, 10:5, 2019.

[283] Carlijn D.J.M. Borm, Florian Krismer, Gregor K. Wenning, Klaus Seppi, Werner Poewe, Maria Teresa Pellecchia, Paolo Barone, Erik L. Johnsen, Karen Østergaard, Tanya Gurevich, Ruth Djaldetti, Luisa Sambati, Pietro Cortelli, Igor Petrović, Vladimir S. Kostić, Hana Brožová, Evžen Růžička, Maria Jose Marti, Eduardo Tolosa, Margherita Canesi, Bart Post, Jorik Nonnekes, Bastiaan R. Bloem, Maria Stamelou, Vladimir S. Kostic, Thomas Klockgether, Richard Dodel, Michael Abele, Wassilios Meissner, Heinz Reichmann, Tim Lynch, Jaroslaw Slawek, Mag Klaus Seppi, Daniela Berg, Joaquim Ferreira, Henry Houlden, Niall P. Quinn, Håkan Widner, Alexander Gerhard, Karla Maria Eggert, Alberto Albanese, Francesca del Sorbo, Alfredo Berardelli, Carlo Colosimo, Jose Berciano, Latchezar Traykov, Nir Giladi, Olivier Rascol, Monique Galitzky, and Thomas Gasser. Axial motor clues to identify atypical parkinsonism: a multicentre European cohort study. *Parkinsonism & Related Disorders*, 56:33–40, 2018.

[284] Marianna Amboni, Carlo Ricciardi, Marina Picillo, Chiara De Santis, Gianluca Ricciardelli, Filomena Abate, Maria Francesca Tepedino, Giovanni D'Addio, Giuseppe Cesarelli, Giampiero Volpe, Maria Consiglia Calabrese, Mario Cesarelli, and Paolo Barone. Gait analysis may distinguish progressive supranuclear palsy and parkinson disease since the earliest stages. *Scientific reports*, 11:9297, 2021.

[285] Marian L. Dale, Carla Silva-Batista, Filipe Oliveira de Almeida, and Fay B. Horak. Balance and gait in progressive supranuclear palsy: a narrative review of objective metrics and exercise interventions. *Frontiers in Neurology*, 14:1212185, 2023.

[286] Chao Gao, Hanbo Sun, Tuo Wang, Ming Tang, Nico Bohnen, Martijn Muller, Talia Herman, Nir Giladi, Alexandr Kalinin, Cathie Spino, William Dauer, Jeffrey Hausdorff, and Ivo Dinov. Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. *Scientific Reports*, 8:7129, 2018.

[287] Jie Mei, Christian Desrosiers, and Johannes Frasnelli. Machine learning for the diagnosis of Parkinson's disease: a review of literature. *Frontiers in Aging Neuroscience*, 13:633752, 2021.

[288] Julian Varghese, Alexander Brenner, Michael Fujarski, Catharina Marie van Alen, Lucas Plagwitz, and Tobias Warnecke. Machine learning in the Parkinson's disease smartwatch (PADS) dataset. *npj Parkinson's Disease*, 10:1–11, 1 2024.

[289] Joomee Song, Junghyun Kim, Mi Ji Lee, Jong Hyeon Ahn, Dong Yeong Lee, Jinyoung Youn, Myung Jin Chung, Zero Kim, and Jin Whan Cho. Differential diagnosis between Parkinson's disease and atypical parkinsonism based on gait and postural instability: artificial intelligence using an enhanced weight voting ensemble model. *Parkinsonism & Related Disorders*, 98:32–37, 5 2022.

[290] Carlos Fernandes, Flora Ferreira, Rui L. Lopes, Estela Bicho, Wolfram Erlhagen, Nuno Sousa, and Miguel F. Gago. Discrimination of idiopathic Parkinson's disease and vascular parkinsonism based on gait time series and the levodopa effect. *Journal of Biomechanics*, 125:110214, 2021.

[291] Julian D. Loaiza Duque, Antonio J. Sanchez Egea, Theresa Reeb, Hernan A. Gonzalez Rojas, and Andres M. Gonzalez-Vargas. Angular velocity analysis boosted by machine learning for helping in the differential diagnosis of Parkinson's disease and essential tremor. *IEEE Access*, 8:88866–88875, 2020.

[292] Maarten De Vos, John Prince, Tim Buchanan, James J. FitzGerald, and Chrystalina A. Antoniades. Discriminating progressive supranuclear palsy from Parkinson's disease using wearable technology and machine learning. *Gait & Posture*, 77:257–263, 2020.

[293] Decho Surangsrirat, Chusak Thanawattano, Ronachai Pongthornseri, Songphon Dumnin, Chanawat Anan, and Roongroj Bhidayasiri. Support vector machine classification of Parkinson's disease and essential tremor subjects based on temporal fluctuation. *Annual*

International Conference of the IEEE Engineering in Medicine and Biology Society, pages 6389–6392, 2016.

[294] Xiao Wu, Bohan Yu, Peng Liu, Huaiyu Zhu, Junjie Li, Haotian Wang, Wei Luo, and Pan Yun. A wearable multi-sensor system for classification of multiple system atrophy and parkinson's disease. In *2022 10th International Conference on Bioinformatics and Computational Biology (ICBCB)*, pages 129–134, 2022.

[295] Michela Russo, Carlo Ricciardi, Marianna Amboni, Marina Picillo, Gianluca Ricciardelli, Filomena Abate, Maria Francesca Tepedino, Maria Consiglia Calabrese, Mario Cesarelli, and Maria Romano. Performing a short sway to distinguish parkinsonisms. In *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, pages 340–345, 2022.

[296] Shinuan Lin, Chao Gao, Hongxia Li, Pei Huang, Yun Ling, Zhonglue Chen, Kang Ren, and Shengdi Chen. Wearable sensor-based gait analysis to discriminate early Parkinson's disease from essential tremor. *Journal of neurology*, 270:2283–2301, 2023.

[297] Graeme Macphee. Diagnosis and differential diagnosis of Parkinson's disease. In *Parkinson's Disease in the Older Patient*, pages 43–77. Arnold, London, 2001.

[298] Weerasak Muangpaisan, Aju Mathews, Hiroyuki Hori, and David Seidel. A systematic review of the worldwide prevalence and incidence of Parkinson's disease. *Journal of the Medical Association of Thailand = Chotmaihet Thangphaet*, 94(6):749–755, 2011.

[299] M.C. de Rijk, Christophe Tzourio, M.M. Breteler, J.F. Dartigues, L. Amaducci, S. Lopez-Pousa, J.M. Manubens-Bertran, A. Alpérovitch, and W. Rocca. Prevalence of parkinsonism and Parkinson's disease in Europe: the EUROPARKINSON collaborative study. European community concerted action on the epidemiology of Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, 62(1):10–15, 1997.

[300] Yaroslav Winter, Yury Bezdolnyy, Elena Katunina, Gagik Avakjan, Jens P. Reese, Jens Klotsche, Wolfgang H. Oertel, Richard Dodel, and Eugene Gusev. Incidence of Parkinson's disease and atypical parkinsonism: Russian population-based study. *Movement Disorders*, 25(3):349–356, 2010.

[301] Vanessa Fleury, Pauline Brindel, Nicolas Nicastro, and Pierre R. Burkhard. Descriptive epidemiology of parkinsonism in the Canton of Geneva, Switzerland. *Parkinsonism & Related Disorders*, 54:30–39, 2018.

[302] Jan Linder, Hans Stenlund, and Lars Forsgren. Incidence of Parkinson's disease and parkinsonism in northern Sweden: a population-based study. *Movement Disorders*, 25(3):341–348, 2010.

[303] Julián Benito-León, Félix Bermejo-Pareja, Javier Rodríguez, José-Antonio Molina, Rafael Gabriel, José-Manuel Morales, and for the Neurological Disorders in Central Spain (NEDICES) Study Group. Prevalence of PD and other types of parkinsonism in three elderly populations of central Spain. *Movement Disorders*, 18(3):267–274, 2003.

[304] Hamdy N. EL-Tallawy, Wafaa M. Farghaly, Ghaydaa A. Shehata, Tarek A. Rageh, Nabil M. Abdel Hakeem, Mohamed Abd Al Hamed, and Reda Badry. Prevalence of Parkinson's disease and other types of parkinsonism in Al Kharga district, Egypt. *Neuropsychiatric Disease and Treatment*, 9:1821–1826, 2013.

[305] Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A systematic review on imbalanced data challenges in machine learning: applications and solutions. *ACM Computing Surveys*, 52(4):1–36, 2019.

[306] Duraisamy Ramyachitra and Parasuraman Manikandan. Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, 5(4):1–29, 2014.

[307] Gregor K. Wenning, Iva Stankovic, Luca Vignatelli, Alessandra Fanciulli, Giovanna Calandra-Buonaura, Klaus Seppi, Jose Alberto Palma, Wassilios G. Meissner, Florian Krismer, Daniela Berg, Pietro Cortelli, Roy Freeman, Glenda Halliday, Günter Höglinger, Anthony Lang, Helen Ling, Irene Litvan, Phillip Low, Yasuo Miki, Jalesh Panicker, Maria Teresa Pellecchia, Niall Quinn, Ryuji Sakakibara, Maria Stamelou, Eduardo Tolosa, Shoji Tsuji, Tom Warner, Werner Poewe, and Horacio Kaufmann. The movement disorder society criteria for the diagnosis of multiple system atrophy. *Movement Disorders*, 37:1131–1148, 2022.

[308] Günter U. Höglinger, Gesine Respondek, Maria Stamelou, Caroline Kurz, Keith A. Josephs, Anthony E. Lang, Brit Mollenhauer, Ulrich Müller, Christer Nilsson, Jennifer L. Whitwell, Thomas Arzberger, Elisabet Englund, Ellen Gelpi, Armin Giese, David J. Irwin, Wassilios G. Meissner, Alexander Pantelyat, Alex Rajput, John C. van Swieten, Claire Troakes, Angelo Antonini, Kailash P. Bhatia, Yvette Bordelon, Yaroslau Compta, Jean Christophe Corvol, Carlo Colosimo, Dennis W. Dickson, Richard Dodel, Leslie Ferguson, Murray Grossman, Jan Kassubek, Florian Krismer, Johannes Levin, Stefan Lorenzl, Huw R. Morris, Peter Nestor, Wolfgang H. Oertel, Werner Poewe, Gil Rabinovici, James B. Rowe, Gerard D. Schellenberg, Klaus Seppi, Thilo van Eimeren, Gregor K. Wenning, Adam L. Boxer, Lawrence I. Golbe, Irene Litvan, Gregor K. Wenning, Günter U. Höglinger, Jan Kassubek, Jean Christophe Corvol, Jennifer L. Whitwell, Johannes Levin, John van Swieten, Kailash P. Bhatia, Keith A. Josephs, Klaus Seppi, Lawrence I. Golbe, Murray Grossman, Richard Dodel, Stefan Lorenzl, Thilo van Eimeren, Thomas Arzberger, Ulrich Müller, Werner Poewe, Wolfgang H. Oertel, Yaroslau Compta, and Yvette Bordelon. Clinical diagnosis of progressive supranuclear palsy: the movement disorder society criteria. *Movement Disorders*, 32:853–864, 2017.

[309] Ian G. McKeith, Bradley F. Boeve, Dennis W. Dickson, Glenda Halliday, John Paul Taylor, Daniel Weintraub, Dag Aarsland, James Galvin, Johannes Attems, Clive G. Ballard, Ashley Bayston, Thomas G. Beach, Frédéric Blanc, Nicolaas Bohnen, Laura Bonanni, Jose Bras, Patrik Brundin, David Burn, Alice Chen-Plotkin, John E. Duda, Omar El-Agnaf, Howard Feldman, Tanis J. Ferman, Dominic Ffytche, Hiroshige Fujishiro, Douglas Galasko, Jennifer G. Goldman, Stephen N. Gomperts, Neill R. Graff-Radford, Lawrence S. Honig, Alex Iranzo, Kejal Kantarci, Daniel Kaufer, Walter Kukull, Virginia M.Y. Lee,

James B. Leverenz, Simon Lewis, Carol Lippa, Angela Lunde, Mario Masellis, Eliezer Masliah, Pamela McLean, Brit Mollenhauer, Thomas J. Montine, Emilio Moreno, Etsuro Mori, Melissa Murray, John T. O'Brien, Sotoshi Orimo, Ronald B. Postuma, Shankar Ramaswamy, Owen A. Ross, David P. Salmon, Andrew Singleton, Angela Taylor, Alan Thomas, Pietro Tiraboschi, Jon B. Toledo, John Q. Trojanowski, Debby Tsuang, Zuzana Walker, Masahito Yamada, and Kenji Kosaka. Diagnosis and management of dementia with lewy bodies: fourth consensus report of the DLB consortium. *Neurology*, 89:88–100, 2017.

[310] Melissa J. Armstrong, Irene Litvan, Anthony E. Lang, Thomas H. Bak, Kailash P. Bhatia, Barbara Borroni, Adam L. Boxer, Dennis W. Dickson, Murray Grossman, Mark Hallett, Keith A. Josephs, Andrew Kertesz, Suzee E. Lee, Bruce L. Miller, Stephen G. Reich, David E. Riley, Eduardo Tolosa, Alexander I. Tröster, Marie Vidailhet, and William J. Weiner. Criteria for the diagnosis of corticobasal degeneration. *Neurology*, 80:496–503, 2013.

[311] Kailash P. Bhatia, Peter Bain, Nin Bajaj, Rodger J. Elble, Mark Hallett, Elan D. Louis, Jan Raethjen, Maria Stamelou, Claudia M. Testa, and Guenther Deuschl. Consensus statement on the classification of tremors. From the task force on tremor of the international parkinson and movement disorder society. *Movement Disorders*, 33:75–87, 2018.

[312] Hae Won Shin and Sun Ju Chung. Drug-induced parkinsonism. *Journal of Clinical Neurology*, 8:15–21, 2012.

[313] Wiebren Zijlstra and At L. Hof. Assessment of spatio-temporal gait parameters from trunk accelerations during human walking. *Gait & Posture*, 18(2):1–10, 2003.

[314] Thaiany Pedrozo Campos Antunes and Kirstin P. van Kesteren. Prediction of trip severity based on tri-axial accelerometry in healthy older adults. *Journal of Human Growth and Development*, 25(1):75–81, 2015.

[315] Marcel J.P. Toebes, Marco J.M. Hoozemans, Regula Furrer, Joost Dekker, and Jaap H. van Dieën. Associations between measures of gait stability, leg strength and fear of falling. *Gait & Posture*, 41(1):76–80, 2015.

[316] Tian-Yu Liu. Easyensemble and feature selection for imbalance data sets. In *2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, pages 517–520, 2009.

[317] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory under-sampling for class-imbalance learning. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, page 965–969, 2006.

[318] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Ltd, 1991.

[319] Gary M. Weiss. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newslette*, 6:7–19, 2004.

[320] Firuz Kamalov, Fadi Thabtah, and Ho Hon Leung. Feature selection in imbalanced data. *Annals of Data Science*, 10(6):1527–1541, 2023.

[321] Nutthaporn Junsomboon and Tanasanee Phienthrakul. Combining over-sampling and under-sampling techniques for imbalance dataset. In *Proceedings of the 9th International Conference on Machine Learning and Computing*, ICMLC '17, page 243–247, 2017.

[322] Ahmad Abu Shanab, Taghi M. Khoshgoftaar, Randall Wald, and Jason Van Hulse. Comparison of approaches to alleviate problems with high-dimensional and class-imbalanced data. In *2011 IEEE International Conference on Information Reuse & Integration*, pages 234–239, 2011.

[323] Jorma Laurikkala. Improving identification of difficult small classes by balancing class distribution. In *Artificial Intelligence in Medicine*, pages 63–66, 2001.

[324] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[325] Mansi Sharma, Ram kinker Mishra, Anna J. Hall, Jose Casado, Rylee Cole, Adonay S. Nunes, Gregory Barchard, Ashkan Vaziri, Alexander Pantelyat, and Anne Marie Wills. Remote at-home wearable-based gait assessments in progressive supranuclear palsy compared to Parkinson's disease. *BMC Neurology*, 23:1–9, 2023.

[326] Noriko Hatanaka, Kota Sato, Nozomi Hishikawa, Mami Takemoto, Yasuyuki Ohta, Toru Yamashita, and Koji Abe. Comparative gait analysis in progressive supranuclear palsy and Parkinson's disease. *European Neurology*, 75:282–289, 2016.

[327] Victoria Sidoroff, Cecilia Raccagni, Christine Kaindlstorfer, Sabine Eschlboeck, Alessandra Fanciulli, Roberta Granata, Björn Eskofier, Klaus Seppi, Werner Poewe, Johann Willeit, Stefan Kiechl, Philipp Mahlknecht, Heike Stockner, Kathrin Marini, Oliver Schorr, Gregorio Rungger, Jochen Klucken, Gregor Wenning, and Heiko Gaßner. Characterization of gait variability in multiple system atrophy and Parkinson's disease. *Journal of Neurology*, 268:1770–1779, 2021.

[328] David Bäckström, Gabriel Granåsen, Magdalena Eriksson Domellöf, Jan Linder, Susanna Jakobson Mo, Katrine Riklund, Henrik Zetterberg, Kaj Blennow, and Lars Forsgren. Early predictors of mortality in parkinsonism and Parkinson disease. *Neurology*, 91(22):e2045–e2056, 2018.

[329] S. Fielding, A.D. Macleod, and C.E. Counsell. Medium-term prognosis of an incident cohort of parkinsonian patients compared to controls. *Parkinsonism & Related Disorders*, 32:36–41, 2016.

[330] Frank Lin, Luigi Ferrucci, E. Metter, Yang An, Alan Zonderman, and Susan Resnick. Hearing loss and cognition in the Baltimore Longitudinal Study of Aging. *Neuropsychology*, 25(6):763, 2011.

[331] Sandra Gordon-Salant and Peter J. Fitzgibbons. Recognition of multiply degraded speech by young and elderly listeners. *Journal of Speech, Language, and Hearing Research*, 38(5):1150–1156, 1995.

[332] Laurie M. Heller, Benjamin Elizalde, Bhiksha Raj, and Soham Deshmukh. Synergy between human and machine approaches to sound/scene recognition and processing: an overview of ICASSP special session, 2023.

[333] Samira Anderson, Lindsay Devries, Ed Smith, Matthew Goupell, and Sandra Gordon-Salant. Rate discrimination training may partially restore temporal processing abilities from age-related deficits. *Journal of the Association for Research in Otolaryngology*, 23(6):771–786, 2022.

[334] Samira Anderson, Travis White-Schwoch, Alexandra Parbery-Clark, and Nina Kraus. A dynamic auditory-cognitive system supports speech-in-noise perception in older adults. *Hearing Research*, 300:18–32, 2013.

[335] Adam Dryden, Harriet A. Allen, Helen Henshaw, and Antje Heinrich. The association between cognitive performance and speech-in-noise perception for adult listeners: a systematic literature review and meta-analysis. *Trends in Hearing*, 21:2331216517744675, 2017.

[336] Sandra Gordon-Salant and Stacey Samuels Cole. Effects of age and working memory capacity on speech recognition performance in noise among listeners with normal hearing. *Ear and Hearing*, 37(5):593–602, 2016.

[337] Aravindakshan Parthasarathy, Sandra Pinto, Rebecca Lewis, William Goedicke, and Daniel Polley. Data-driven segmentation of audiometric phenotypes across a large clinical cohort. *Scientific Reports*, 10(1):6704, 2020.

[338] Raul Sanchez-Lopez, Michal Fereczkowski, Tobias Neher, Sébastien Santurette, and Torsten Dau. Robust data-driven auditory profiling towards precision audiology. *Trends in Hearing*, 24:2331216520973539, 2020.

[339] Saransh Jain, Vijaya Kumar Narne, N. P. Nataraja, Sanjana Madhukesh, Kruthika Kumar, and Brian C. J. Moore. The effect of age and hearing sensitivity at frequencies above 8 kHz on auditory stream segregation and speech perceptiona. *The Journal of the Acoustical Society of America*, 152(1):716–726, 2022.

[340] Accredited Standards Committee S3, Bioacoustics. American national standard specification for audiometers. Technical Report S3.6-2018, American National Standards Institute, June 2018.

[341] IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246, 1969.

[342] Mead C. Killion, Patricia A. Niquette, Gail I. Gudmundsen, Lawrence J. Revit, and Shilpi Banerjee. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 116(4):2395–2405, 2004.

[343] Tom W. Tillman and Raymond Carhart. An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University Auditory Test no. 6. Technical report, Northwestern Univ Evanston Il Auditory Research Lab, 1966.

[344] Philip David Zelazo, Jacob E. Anderson, Jennifer Richler, Kathleen Wallner-Allen, Jennifer L. Beaumont, Kevin P. Conway, Richard Gershon, and Sandra Weintraub. NIH toolbox cognition battery (CB): validation of executive function measures in adults. *Journal of the International Neuropsychological Society*, 20(6):620–629, 2014.

[345] Stuart Gatehouse and William Noble. The speech, spatial and qualities of hearing scale (SSQ). *International Journal of Audiology*, 43(2):85–99, 2004.

[346] Peter M. Rabinowitz, Deron Galusha, Martin D. Slade, Christine Dixon-Ernst, Kanta D. Sircar, and Robert A. Dobie. Audiogram notches in noise-exposed workers. *Ear and Hearing*, 27(6):742–750, 2006.

[347] Robert R. Coles, M.E. Lutman, and Jez Buffin. Guidelines on the diagnosis of noise-induced hearing loss for medicolegal purposes. *Clinical Otolaryngology & Allied Sciences*, 25(4):264–273, 2000.

[348] Amanda Sue Niskar, Stephanie M. Kieszak, Alice E. Holmes, Emilio Esteban, Carol Rubin, and Debra J. Brody. Estimated prevalence of noise-induced hearing threshold shifts among children 6 to 19 years of age: the third national health and nutrition examination survey, 1988–1994, United States. *Pediatrics*, 108(1):40–43, 2001.

[349] François Charih, Matthew Bromwich, Amy Mark, Renée Lefrançois, and James Green. Data-driven audiogram classification for mobile audiometry. *Scientific Reports*, 10:3962, 2020.

[350] Judy R. Dubno, Mark A. Eckert, Fu-Shing Lee, Lois J. Matthews, and Richard A. Schmiedt. Classifying human audiometric phenotypes of age-related hearing loss from animal models. *Journal of the Association for Research in Otolaryngology (JARO)*, 14:687–701, 2013.

[351] Eric Polley, Erin LeDell, Chris Kennedy, Sam Lendle, and Mark van der Laan. *SuperLearner: Super Learner Prediction*, 2021. R package version 2.0-28.

[352] R Core Team. *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[353] Torsten Dau Raul Sanchez-Lopez and William M. Whitmer. Audiometric profiles and patterns of benefit: a data-driven analysis of subjective hearing difficulties and handicaps. *International Journal of Audiology*, 61(4):301–310, 2022.

[354] Judy R. Dubno, Donald D. Dirks, and Amy B. Schaefer. Stop-consonant recognition for normal-hearing listeners and listeners with high-frequency hearing loss. II: articulation index predictions. *The Journal of the Acoustical Society of America*, 85(1):355–364, 1989.

[355] Alessandro Presacco, Jonathan Z. Simon, and Samira Anderson. Effect of informational content of noise on speech representation in the aging midbrain and cortex. *Journal of Neurophysiology*, 116(5):2356–2367, 2016.

[356] Ian T. Jolliffe. *Principal Component Analysis*. Springer New York, NY, 2002.

[357] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[358] Andrew Mead. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society Series D: The Statistician*, 41(1):27–39, 1992.

[359] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[360] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[361] Marc Raynal, Michel Kossowski, and Agnès Job. Hearing in military pilots: one-time audiometry in pilots of fighters, transports, and helicopters. *Aviation, Space, and Environmental Medicine*, 77(1):57–61, 2005.

[362] H.J. Cox and G.R. Ford. Hearing loss associated with weapons noise exposure: when to investigate an asymmetrical loss. *The Journal of Laryngology and Otology*, 109(4):291–295, 1995.

[363] David A. Lowe and Brian C. J. Moore. Audiometric assessment of hearing loss sustained during military service. *The Journal of the Acoustical Society of America*, 150(2):1030–1043, 2021.

[364] Tapio Pirilä, Martti Sorri, Katariina Jounio-Ervasti, Pckka Sipilä, and Heino Karjalainen. Hearing asymmetry among occupationally noise-exposed men and women under 60 years of age. *Scandinavian Audiology*, 20(4):217–222, 1991.

[365] Cara Sturman, Chris Frampton, and Wouter Cate. Hearing loss asymmetry due to chronic occupational noise exposure. *Otology & Neurotology*, 39(8):e627–e634, 2018.

[366] Liam Masterson, James Howard, Zi Liu, and John Phillips. Asymmetrical hearing loss in cases of industrial noise exposure: a systematic review of the literature. *Otology & Neurotology*, 37(8):998–1005, 2016.

[367] Tapio Pirilä, Katarüna Jounio-Ervasti, and Mariti Surri. Left-right asymmetries in hearing threshold levels in three age groups of a random population. *Audiology*, 31(3):150–161, 1992.

[368] Angkoon Phinyomark, Pornchai Phukpattaranont, and Chusak Limsakul. Feature reduction and selection for EMG signal classification. *Expert Systems with Applications*, 39(8):7420–7431, 2012.

[369] Cemil Altin and Orhan Er. Comparison of different time and frequency domain feature extraction methods on elbow gesture's EMG. *European Journal of Interdisciplinary Studies*, 5:35, 2016.

[370] Christer Sinderby, Lars Lindstrom, and Alejandro Grassino. Automatic assessment of electromyogram quality. *Journal of Applied Physiology*, 79(5):1803–1815, 1995.

[371] Lucia Pepa, Lucio Ciabattoni, Federica Verdini, Marianna Capecci, and Maria Ceravolo. Smartphone based fuzzy logic freezing of gait detection in Parkinson's disease. In *2014 IEEE/ASME 10th International Conference on Mechatronic and Embedded Systems and Applications (MESA)*, pages 1–6, 2014.

[372] Angkoon Phinyomark, Asan Nuidod, P. Phukpattaranont, and Chusak Limsakul. Improving class separability of frequency-domain features for EMG pattern recognition. In *Proceeding - Science and Engineering*, pages 124–130, 2012.

[373] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In *Pervasive Computing*, pages 1–17, 2004.

[374] Angkoon Phinyomark, Pornchai Phukpattaranont, and Chusak Limsakul. Fractal analysis features for weak and single-channel upper-limb EMG signals. *Expert Systems with Applications*, 39(12):11156–11163, 2012.

[375] Hazirah Hasni, Norashikin Yahya, Vijanth S. Asirvadam, and Munsif Ali Jatoi. Analysis of electromyogram (EMG) for detection of neuromuscular disorders. In *2018 International Conference on Intelligent and Advanced System (ICIAS)*, pages 1–6, 2018.

[376] Nagineni Sukumar, Sachin Taran, and Varun Bajaj. Physical actions classification of surface EMG signals using VMD. In *2018 International Conference on Communication and Signal Processing (ICCSP)*, pages 705–709, 2018.

[377] James F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 381–384, 1990.

[378] Thomas Penzel, Jan Kantelhardt, Ludger Grote, Jörg-Hermann Peter, and Armin Bunde. Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. *IEEE Transactions on Biomedical Engineering*, 50(10):1143–1151, 2003.

[379] Tomoyuki Higuchi. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, 31(2):277–283, 1988.

[380] Michael J. Katz. Fractals and the analysis of waveforms. *Computers in Biology and Medicine*, 18(3):145–156, 1988.

[381] Tilmann Gneiting, Hana Ševčíková, and Donald B. Percival. Estimators of fractal dimension: assessing the roughness of time series and spatial data. *Statistical Science*, 27(2):247–277, 2012.

[382] Jun C. Quiroz, Amit Bnerjee, Sergiu M. Dslu, and Sin Lun Lu. Feature selection for activity recognition from smartphone accelerometer data. *Intelligent Automation & Soft Computing*, 24(4):785–793, 2018.

[383] Ahmed Ayman, Omneya Attalah, and Heba Shaban. An efficient human activity recognition framework based on wearable IMU wrist sensors. In *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–5, 2019.

[384] Mouazma Batool, Ahmad Jalal, and Kibum Kim. Sensors technologies for human activity analysis based on SVM optimized by PSO algorithm. In *2019 International Conference on Applied and Engineering Mathematics (ICAEM)*, pages 145–150, 2019.

[385] Stefano Filippo Castiglia, Dante Trabassi, Carmela Conte, Alberto Ranavolo, Gianluca Coppola, Gabriele Sebastianelli, Chiara Abagnale, Francesca Barone, Federico Bighiani, Roberto De Icco, Cristina Tassorelli, and Mariano Serrao. Multiscale entropy algorithms to analyze complexity and variability of trunk accelerations time series in subjects with Parkinson's disease. *Sensors*, 23(10):4983, 2023.

[386] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Springer New York, New York, NY, 2002.

[387] Thomas M. Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley, New York, 1990.

[388] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, 2017.

[389] Andy Liaw and Matthew C. Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

[390] Marvin N. Wright and Andreas Ziegler. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

[391] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[392] Brian D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 2007.

[393] Brandon M. Greenwell and Bradley C. Boehmke. Variable importance plots—an introduction to the vip package. *The R Journal*, 12(1):343–366, 2020.