Classification of Parkinson’s Disease and Its Stages Using Machine Learning

John Michael Templeton (jtemplet@nd.edu)
University of Notre Dame

Christian Poellabauer
Florida International University

Sandra Schneider
Saint Mary's College

Research Article

Keywords:

Posted Date: February 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1273825/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Classification of Parkinson’s Disease and its Stages using Machine Learning

John Michael Templeton¹,*, Christian Poellabauer², and Sandra Schneider³

¹University of Notre Dame, Department of Computer Science and Engineering, Notre Dame, IN, 46556, USA
²Florida International University, Department of Computing and Information Sciences, Miami, FL, 33199, USA
³Saint Mary’s College, Department of Communicative Sciences and Disorders, Notre Dame, IN, 46556, USA
*jtemplet@nd.edu

ABSTRACT

As digital health technology becomes more pervasive, machine learning (ML) provides a robust way to analyze and interpret the myriad of collected features. The purpose of this work is to use ML classification to assess the benefits and relevance of features from clinically relevant sources (e.g., tablet-based assessments, self-reported metrics, and functional movement assessments), as they relate to Parkinson’s Disease (PD) and its stages (Hoehn and Yahr (H&Y) Stages 1 - 5). Further, this work aims to compare perceived versus actual neurocognitive abilities. In this study, 75 participants (n = 50 PD; n = 25 control) completed 14 tablet-based neurocognitive functional tests (e.g., motor, memory, speech, executive, and multi-function), functional movement assessments (e.g., Berg Balance Scale), and standardized health questionnaires. Significant features were identified using decision tree classification. Device magnitude of acceleration was significant in 12 of 14 tests (85.7%), regardless of test type. For classification between diagnosed and control populations, 17 motor, 9 accuracy, and 8 timing features were significant. For classification between early (H&Y Stages 1 and 2) and advanced (H&Y Stages 3, 4, and 5) stages of PD, 7 motor, 12 accuracy, and 14 timing features were significant. Perceived functionality of individuals in early-stage PD was 21.6% lower than actual scores with notable perceived deficits in memory and executive function. However, individuals in advanced stages had elevated perceptions (1.57x) for executive and behavioral functions compared to early-stage populations. Machine learning in digital health systems allows for a more comprehensive understanding of neurodegenerative diseases and their stages. ML may also depict new features that influence the ways digital health technology should be configured. Updates to these systems will ultimately aid clinicians, diagnosed populations, and caretakers in monitoring and assessing neurodegenerative diseases.

Introduction

As the prevalence of digital health technology and the subsequent collection of large amounts of complex health data increases, machine learning methods provide epidemiologists with a robust way to analyze and interpret relevant patterns¹,². The increased availability of electronic health data and the use of machine learning models presents major opportunities in the healthcare space for discovery, improving patient safety, and the quality of care for individuals with neurodegenerative diseases³-⁵. Moreover, this digital health technology has become increasingly widespread in the area of neurocognitive assessments⁶. Mobile device capabilities allow for the collection of more objective information (e.g., digital biomarkers) than is currently achievable using pen-and-paper style tests⁷. In addition, digital health technology allows for the implementation of standardized health screenings and patient reported outcomes for individual evaluation⁸. The combination of objective metrics and patient reported outcomes allows for the collection of relevant health information and monitoring of all functional areas of neurocognition (e.g., motor, memory, speech, language, executive function, autonomic function, sensory, behavior, and sleep)⁹. This paper focuses on individuals with Parkinson’s Disease (PD) as they may demonstrate impaired functionality across each of these functional areas of neurocognition¹⁰.

The objective of this preliminary work was to use supervised machine learning classification for the assessment of novel features gathered from specifically-designed, tablet-based, digital neurocognitive assessments (e.g., digital biomarkers) as they relate to Parkinson’s Disease and its stages (Hoehn and Yahr Stages 1 - 5)¹¹. In addition, commonly used self-reported metrics (e.g., from specifically designed questionnaires) and clinically-relevant functional movement assessments were included in the supervised machine learning process. Decision tree classification was used to gather significant, objective features (e.g., novel digital biomarkers) for the depiction of whether an individual has PD, and if so, what stage of PD they are in. Further, this work explored individuals’ perceived neurocognitive capabilities (e.g., responses from subjective questionnaires) in comparison to actual neurocognitive functionality (e.g., completion of functional assessments) between groups with and without PD.
Related Work

The use of digital health technology and its capabilities allows for the collection of far more information and objective metrics than we ever could achieve using pen-and-paper style tests, while also assessing all functional areas of neurocognition\(^9,12\). Previous work has identified that device-based sensors and/or user-device interactions used in digital assessments (e.g., accelerometry based gait assessments or speech recognition systems for healthcare) enhances the utility and quality of collected data\(^7\).

Mobile devices allow for the implementation of digital versions of standardized assessments (e.g., Montreal Cognitive Assessment (MoCA)\(^13\); Mini Mental State Examination (MMSE)\(^14\)) and questionnaires (e.g., PDQ-39\(^15\)) for the collection of objective digital biomarkers and relevant patient reported outcomes\(^16\). Further, these mobile devices can utilize machine learning for the depiction of an individual’s neurocognitive functionality, quality of life, and quantification of disease progression\(^9,17\).

An important aspect of personal health monitoring is collected information from the patient themselves\(^16\). Patient reported outcomes (PROs) are commonly utilized to monitor patients’ thoughts or opinions on short-term (e.g., day to day) changes in their condition. Subsequently, this can lead to improved disease management by allowing the individual to recognize and understand their condition and to be aware of their symptoms and triggers\(^16,18\). The aim of these digital health systems should then be to increase the reliability and accuracy of patient reported data by combining it with objective data from mobile devices, as there may be individual variability and/or bias\(^19–21\).

As the volume of relevant health data increases, novel ways to interact with and extract meaning from the data emerge\(^1,22\). Machine learning is a key technique that has demonstrated the ability to translate these large health data sets into actionable knowledge\(^3,23,24\). Specifically, supervised machine learning of health data has shown potential in the area of disease prediction and classification\(^25,26\). In supervised machine learning problems, the utilization of clinically relevant and objective features is necessary as the performance of these algorithms is heavily dependent upon the quality of the input features\(^27\).

The use of decision trees is considered highly powerful in classification problems and there are many popular decision tree algorithms (e.g., CART, ID3, C4.5, CHAID, and J48)\(^28\). From these, an optimized version of the Classification and Regression Tree (CART) algorithm was used for this work. The CART algorithm constructs binary trees using the feature and threshold that yields the largest information gain at each node\(^29\). The CART algorithm uses the Gini Index as a metric to originate binary splits. The calculation of the Gini Index is depicted in Equation 1 where \(P_i\) is the probability of an object being classified to a particular class.

\[
\text{Gini Index} = 1 - \sum_{i=1}^{C} (P_i)^2
\]  

The Gini Index will always be between 0 and 1 where a value of 0.5 shows an equal distribution of elements over classes, and a value closer to 0 will depict a better binary split.

Methods

Hoehn and Yahr Scale

Parkinson’s Disease rating scales are a means of assessing the symptoms of the condition by providing information on the course of the condition and/or assessment of an individual’s quality of life. The Hoehn and Yahr Scale (H&Y) is an internationally used PD progression rating method for clinical practice\(^30,31\). It is used to measure how Parkinson’s symptoms progress and depict the level of clinical disability\(^11\).

- Stage 1: Symptoms are present on one side only (unilateral).
- Stage 2: Symptoms are present on both sides but no impairment of balance.
- Stage 3: Balance impairment and mild to moderate disease progression.
- Stage 4: Severe disability, but still able to walk or stand unassisted.
- Stage 5: Needing a wheelchair or bedridden unless assisted.

Cohort

Seventy-five adults between the ages of 50 and 85, divided into two groups- those with a confirmed diagnosis of Parkinson’s Disease and age-matched healthy controls participated in this study. The PD population included 50 individuals; with 22 being in confirmed early stages of PD (H&Y Stages 1 and 2), 9 being in confirmed advanced stages of PD (H&Y Stages 3, 4,
and 5), and the remaining 11 were unaware of what stage of the condition they were in (e.g., their respective stage was not communicated to them via a licensed clinician). A breakdown of the population is shown in Table 1. Of the group of individuals diagnosed with PD, slightly more than half were female (n = 26 or 52%). The age-matched control population included 25 individuals; with slightly less than half (n = 12 or 48%) being female. Participants were recruited through advertisements, designed rehabilitation programs, physician and clinician referrals, spouses or caretakers of the diagnosed population, and prior studies from our laboratory. As the mean onset age for PD in the Western world is early-to-mid 60s, recruitment efforts for this study were limited to individuals in the aforementioned groups aged 50 years or older. Participants were excluded from the current study if they were unable to provide informed consent or if they were unable to speak and/or understand English (as all instructions and tests were formatted in English). All methods in this study were performed in accordance with the relevant guidelines and regulations from the Institutional Review Board (IRB) for the protection of human subjects.

<table>
<thead>
<tr>
<th>Group</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individuals With Confirmed Diagnosis</td>
<td>50</td>
</tr>
<tr>
<td>Individuals in Early Stages of PD</td>
<td>22</td>
</tr>
<tr>
<td>Individuals in Advanced Stages of PD</td>
<td>9</td>
</tr>
<tr>
<td>Individuals in Unknown Stages of PD</td>
<td>11</td>
</tr>
<tr>
<td>Age Matched Control Population</td>
<td>25</td>
</tr>
</tbody>
</table>

**Table 1.** Cohort Breakdown

**Qualitative Assessment Questionnaires**

All participants were given a set of questions commonly administered in clinical settings for aging populations in addition to questions specifically asked in the event of suspected neurodegenerative disease. Commonly administered questions included how the individual felt in general, their energy levels, their pain level, their sleep quality, in addition to rating their cognitive functions of memory, speech, motor, and executive functions. During data collection, patients were also asked to give information regarding when, relative to taking medication, their data was collected. The rationale for this parameter is that the assessment of the patient near trough levels will depict the most extreme effects of the PD without the effect of any medication. Finally, a specific quality of life questionnaire (e.g., PDQ-39) was used in which the individual assesses their mobility, activities of daily living, emotional well-being, stigma, social support, cognition, communication, and bodily discomfort. This standardized PDQ-39 begins each question with “Due to having Parkinson’s disease, how often during the last month have you...”. For the control group a modified version of the PDQ-39 (e.g., removing ‘Due to having Parkinson’s disease’ such that the question reads ‘How often during the last month have you...’) was given to understand quality of life over the same functional areas. In this modified version for the control population, a single question, specific only to the PD population (e.g., ‘How often during the last month have you felt you had to conceal your Parkinson’s from people?’), was removed.

**Functional Movement Assessments**

Participants in the PD group were also administered multiple functional movement assessments both subjective and objective in nature as part of regularly scheduled clinical measures. The Berg Balance Scale, Timed Up and Go (TUG), Sit to Stand (STS), and Six Minute Walk Test (6MWT) are all commonly used functional movement assessments administered by clinicians (e.g., physical trainers or therapists) to assess functional performance.

**Mobile Application Testing**

All participants were administered a tablet-based neurocognitive assessment specifically designed for individuals with Parkinson’s Disease that focused on user-device interactions for the collection of novel and objective metrics. Each participant completed mobile versions of 14 neurocognitive functional tests across the areas of motor, memory, speech, and executive function. Functional tests included single functional tests (e.g., having focus on only one area of neurocognition; motor or memory) and multi-functional tests (e.g., combining two or more single functional tests into one functional test). The 14 administered neurocognitive tests collected 208 objective tablet-based digital biomarkers for all participants. All test descriptions are listed. For a fine-motor tracing test the individual was instructed to use their index finger to trace a depicted shape. For a reaction test, the user was to tap on the screen to interact with a set of targets. For a memory test the user was to tap on depicted cards, in pairs, until all cards have been matched. In a trail making test the user was instructed to draw a line using their index finger connecting all shapes in increasing numerical order. For a set of speech based tests, the user was prompted to read a passage or sentence out loud or name prompted objects. An example of a multi-functional test includes a fine motor test (e.g., tracing an object) paired with a non-automatic speech test (e.g., listing the months of the year, aloud, in
reverse order; December to January). For an executive function/multi-functional test a digital version of the Stroop Word Color Test (SWCT) was utilized where the user was required to discern the difference between prompted colors and words and then speak the correct response. In an expanded multi-functional test approach (e.g., Narration Writer), the user was instructed to narrate a sentence (e.g., speech) while also writing (e.g., motor) word by word (e.g., writing the same word being said aloud) in the space provided (e.g., executive function).

**Feature Normalization**

The standardization of feature values is necessary as many features are of unique type and have varying units. For this normalization, Z-scores were used. The calculation of a Z-score is depicted in Equation 2 where \( x \) is an individual’s score, \( \mu \) is the population mean and \( \sigma \) is the standard deviation of the population.

\[
Z = \frac{(x - \mu)}{\sigma}
\]  

(2)

The Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point’s score is identical to the mean. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

**Results**

**Disease Classification**

Decision tree classification between groups (e.g., individuals with PD and controls) was completed for both perceived (e.g., self reported outcomes) and new objective assessments (e.g., digital biomarkers from tablet-based functional assessments). This was done to gather further insights on which reported and objective features are significant in the classification of Parkinson’s Disease. Further, it was completed to give insights on an individual’s perceived versus actual neurocognitive functionalities in subsequent analysis.

**Patient Reported Outcomes**

Patient reported outcomes for disease classification, as seen in Table 2, come from general health questionnaires and the PDQ-39. Decision tree classification for both the general health questionnaire and PDQ-39 depicted the most significant questions to be:

- “Do you have any handwriting problems?”
- “What is your energy level?”
- “(Due to having Parkinson’s Disease) How often in the last month have you had difficulty writing clearly?”
- “(Due to having Parkinson’s Disease) How often in the last month have you had a fear of falling?”

<table>
<thead>
<tr>
<th>Assessment / Questionnaire</th>
<th>Feature</th>
<th>Value</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Questionnaire</td>
<td>Root Feature</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Handwriting Problems</td>
<td>( x \leq 0.5 )</td>
<td>0.404</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>( x \leq 8.5 )</td>
<td>0.492</td>
</tr>
<tr>
<td>PDQ - 39</td>
<td>Root Feature</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Difficulty Writing Clearly</td>
<td>( x \leq 1.5 )</td>
<td>0.404</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>( x \leq 1.5 )</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Table 2. Gini Index Values of Questionnaire Responses for PD Population versus Controls.
Objective Digital Biomarkers

Objective digital biomarker classification outcomes are shown in Tables 3 and 4. Objective features from digital versions of 14 functional assessments were used in the classification between groups. Decision trees were generated to discern what tests are the most relevant, while also identifying what objective digital features are the most significant within each test.

Decision tree classification distinguished that the finger tapping test was the most significant of the administered tests in the separation of control and PD groups (Gini Index = 0.375 at the root). Other tests of importance in the separation of PD and control populations were the Grandfather Passage (Gini Index = 0.423 at the root), and multi-test assessments of fine motor tracing with speech (Gini Index = 0.429 at the roots). Finger tapping and multi-test assessments of fine motor tracing with speech had root features of device acceleration (e.g., the magnitude of acceleration of the device or how the user moves the device during the test) whereas the Grandfather Passage had an accuracy feature (e.g., the number of missed words in a speech test) as the root. In an expansion to include first order features, 17 motor, 9 accuracy, and 8 timing features are shown to be significant for the the classification between individuals diagnosed with PD and control populations. This is seen in Figure 1.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Feature</th>
<th>Value</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finger Tapping (Fine Motor)</td>
<td>Root Feature</td>
<td>x ≤ 0.871</td>
<td>0.375</td>
</tr>
<tr>
<td></td>
<td>Minimum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>x ≤ 35.5</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td>Total Small Circles Tapped</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grandfather Passage (Speech)</td>
<td>Root Feature</td>
<td>x ≤ 22.5</td>
<td>0.423</td>
</tr>
<tr>
<td></td>
<td>Number of Missed Words</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>x ≤ 1.001</td>
<td>0.165</td>
</tr>
<tr>
<td></td>
<td>Maximum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of Additional Words</td>
<td>x ≤ 17.5</td>
<td>0.202</td>
</tr>
<tr>
<td>Circle Tracing (Fine-Motor)</td>
<td>Root Feature</td>
<td>x ≤ 388.85</td>
<td>0.435</td>
</tr>
<tr>
<td></td>
<td>Maximum Drawing Speed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>x ≤ 0.998</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Minimum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Square Tracing (Fine-Motor)</td>
<td>Root Feature</td>
<td>x ≤ 454.433</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Maximum Drawing Speed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>x ≤ 0.997</td>
<td>0.387</td>
</tr>
<tr>
<td></td>
<td>Minimum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Narration (Speech)</td>
<td>Root Feature</td>
<td>x ≤ 0.5</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Phrase Said Incorrectly</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>x ≤ 2.569</td>
<td>0.251</td>
</tr>
<tr>
<td></td>
<td>Time from Start to First Word Said</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg Time between Words Said</td>
<td>x ≤ 0.324</td>
<td>0.278</td>
</tr>
<tr>
<td>Object Naming (Speech)</td>
<td>Root Feature</td>
<td>x ≤ 2.5</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Total Incorrect</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>x ≤ 1.003</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Minimum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum Magnitude of Acceleration</td>
<td>x ≤ 1.102</td>
<td>0.463</td>
</tr>
<tr>
<td>Card Matching (Memory)</td>
<td>Root Feature</td>
<td>x ≤ 0.953</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Minimum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>x ≤ 1.243</td>
<td>0.301</td>
</tr>
<tr>
<td></td>
<td>Maximum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Card Matching with Cognitive Interference (Memory)</td>
<td>Root Feature</td>
<td>x ≤ 1.135</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Minimum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>x ≤ 1.048</td>
<td>0.434</td>
</tr>
<tr>
<td></td>
<td>Maximum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Feature</th>
<th>Value</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circle Tracing with Speech (Multitest)</td>
<td>Root Feature</td>
<td>x ≤ 1.079</td>
<td>0.429</td>
</tr>
<tr>
<td></td>
<td>Maximum Magnitude of Acceleration First Order Feature(s)</td>
<td>x ≤ 0.924</td>
<td>0.444</td>
</tr>
<tr>
<td></td>
<td>Minimum Magnitude of Acceleration Total Distance Drawn</td>
<td>x ≤ 25268.62</td>
<td>0.137</td>
</tr>
<tr>
<td></td>
<td>Square Tracing with Speech (Multitest)</td>
<td>x ≤ 1.074</td>
<td>0.429</td>
</tr>
<tr>
<td></td>
<td>Maximum Magnitude of Acceleration First Order Feature(s)</td>
<td>x ≤ 2.897</td>
<td>0.444</td>
</tr>
<tr>
<td></td>
<td>Number of Outline Crossings</td>
<td>x ≤ 19.0</td>
<td>0.137</td>
</tr>
<tr>
<td>Narration Writer (Multitest)</td>
<td>Root Feature</td>
<td>x ≤ 0.919</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Minimum Magnitude of Acceleration First Order Feature(s)</td>
<td>x ≤ 1.003</td>
<td>0.313</td>
</tr>
<tr>
<td></td>
<td>Average Magnitude of Acceleration Total Elapsed Writing Time</td>
<td>x ≤ 27.251</td>
<td>0.32</td>
</tr>
<tr>
<td>Stroop Word Color (Executive Function)</td>
<td>Root Feature</td>
<td>x ≤ 1.002</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>Average Magnitude of Acceleration First Order Feature(s)</td>
<td>x ≤ 8.5</td>
<td>0.369</td>
</tr>
<tr>
<td>Visuospatial (Executive Function)</td>
<td>Root Feature</td>
<td>x ≤ 218.442</td>
<td>0.458</td>
</tr>
<tr>
<td></td>
<td>Average Drawing Speed First Order Feature(s)</td>
<td>x ≤ 1.002</td>
<td>0.305</td>
</tr>
<tr>
<td></td>
<td>Minimum Magnitude of Acceleration Time</td>
<td>x ≤ 8.456</td>
<td>0.355</td>
</tr>
<tr>
<td>Visuospatial with Cognitive Interference</td>
<td>Root Feature</td>
<td>x ≤ 186.201</td>
<td>0.458</td>
</tr>
<tr>
<td>(Executive Function)</td>
<td>Average Drawing Speed First Order Feature(s)</td>
<td>x ≤ 10.351</td>
<td>0.342</td>
</tr>
<tr>
<td></td>
<td>Average Distance to Correct Point</td>
<td>x ≤ 19.33</td>
<td>0.426</td>
</tr>
</tbody>
</table>


Stage Classification

The classification of disease stage (e.g., early (H&Y Stages 1 and 2) versus advanced-stage (H&Y Stages 3, 4, and 5) Parkinson’s Disease was completed across perceived (e.g., patient reported outcomes) and objective assessments (e.g., digital biomarkers), as well as clinically administered functional assessments (e.g., Berg Balance Scale, STS, TUG, and 6MWT) for all individuals with PD.

Patient Reported Outcomes and Functional Assessments

Patient reported outcomes and clinically administered functional assessment results for stage classification are seen in Table 5. Patient reported outcomes come from general health questionnaires and the PDQ-39, whereas functional assessment results come from the Berg Balance Scale, TUG, STS, and 6MWT. Decision tree classification for the patient reported outcomes depicted the most significant questions in the classification of PD stage to be:

- “Do you have any handwriting problems?”
- “What is your energy level?”
- “Due to having Parkinson’s Disease how often in the last month have you had difficulty with leisure activities?”
- “Due to having Parkinson’s Disease how often in the last month have you felt unable to communicate with others?”
- “Due to having Parkinson’s Disease how often in the last month have you felt unpleasently hot or cold?”
Classification of functional assessment features from the Berg Balance Scale depicted the most significant features to be standing with one foot in front and turning to look behind, whereas objective functional movement features included the distance traveled during the 6 Minute Walk Test and the individual’s speed during the Timed Up and Go.

<table>
<thead>
<tr>
<th>Assessment / Questionnaire</th>
<th>Feature</th>
<th>Value</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Questionnaire</td>
<td>Root Feature</td>
<td>Handwriting Problems</td>
<td>x ≤ 0.5</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>Energy Level</td>
<td>x ≤ 8.5</td>
</tr>
<tr>
<td>PDQ - 39</td>
<td>Root Feature</td>
<td>Difficulty with Leisure Activities</td>
<td>x ≤ 2.5</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>Felt Unable to Communicate</td>
<td>x ≤ 0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Felt Unpleasantly Hot or Cold</td>
<td>x ≤ 0.5</td>
</tr>
<tr>
<td>Berg Balance Scale</td>
<td>Root Feature</td>
<td>Standing With One Foot In Front</td>
<td>x ≤ 2.5</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td>Turning to Look Behind</td>
<td>x ≤ 3.5</td>
</tr>
<tr>
<td>Functional Movement</td>
<td>Root Feature</td>
<td>Six Minute Walk</td>
<td>x ≤ 494.5</td>
</tr>
<tr>
<td>Assessments</td>
<td>First Order Feature(s)</td>
<td>Timed Up and Go</td>
<td>x ≤ 9.305</td>
</tr>
</tbody>
</table>

Table 5. Gini Index Values of Functional Assessments and Questionnaire Responses for PD Populations in Early Stages (H&Y Stages 1 and 2) versus Advanced Stages (H&Y Stages 3, 4, and 5).

**Objective Digital Biomarkers**

Objective digital biomarker outcomes for stage classification can be seen in Tables 6 and 7. Similar to disease classification, the breakdown between stages (e.g., early and advanced-stage Parkinson’s Disease) was completed using collected objective features from digital versions of functional assessments.

Decision tree classification distinguished 11 of 14 administered tests as significant in the separation of early (H&Y Stages 1 and 2) and advanced stages (H&Y Stages 3, 4, and 5) of Parkinson’s Disease (Gini Index = 0.363 at the root). Tests with Gini Index higher than 0.363 include Trail Making Tests (e.g., executive function) and the Grandfather Passage with Gini Index values of 0.375 and 0.423, respectively. Root features from the 11 significant tests included 1 device acceleration feature (e.g., how the user moves the device during the test), 4 timing features (e.g., the total elapsed speaking time, or average time between non-match pair), and 6 accuracy features (e.g., the number of targets tapped, or total correct objects named). In an expansion to include first order features, 7 motor, 12 accuracy, and 14 timing features best separate groups based on stage; the inverse of diagnosis. This is also seen in Figure 1.

**Feature Normalization**

**Patient Reported Outcomes**

In the depiction of perceived neurocognitive functionalities between groups, normalized scores from general health questionnaires and the PDQ-39 were calculated for all functional areas of neurocognition. Z-scores were used in the standardization of these features. This standardization of feature values is necessary as many features are of a unique type and have varying units. The weighted Z-scores of perceived capabilities of controls, early-stage PD (H&Y Stages 1 and 2), and advanced-stage PD (H&Y Stages 3, 4, and 5) populations are shown in Figure 2. It should be noted that all executive function tests are inherently multi-functional in nature (e.g., an individual needs to move or speak to carry out the executive function) and therefore are a subset of the multi-functional test digital biomarker set (e.g., denoted by an * in Figure 3).

### Discussion

As Parkinson’s Disease is often described as a “designer disease”, meaning individuals with PD manifest different symptoms across the spectrum of disease characteristics, personalized medicine should be the goal and is required to optimize care\(^3\)\(^9\),\(^40\). However, to reach personalized medicine utilizing machine learning, relevant features need to be identified as the performance of given algorithms are heavily dependent upon the quantitative and quality of the extracted features\(^3\). Nearly 275 features were collected in this work from patient reported outcomes (e.g., from the PDQ-39), functional movement assessments (e.g., from the TUG, STS, and 6MWT), and novel objective digital biomarkers (e.g., from tablet-based assessments) across multiple neurocognitive tests. This work sought to identify new significant features in the classification of individuals with Parkinson’s Disease compared to controls (e.g., what features are the most important in discerning if an individual has PD or not), as well as the classification of different stages of PD (e.g., what features best aid the depiction of how far the disease has progressed).

Decision tree classification brought to light new metrics that expand the way digital versions of functional assessments should be configured. Although primary manifestations of PD include abnormalities of movement (e.g., akinesia, rigidity, and tremor), pen-and-paper style assessments cannot measure all of these abnormalities in the same manner as digital technology\(^9\),\(^12\),\(^41\). The way the individual moves the device during assessments (e.g., voluntarily and/or involuntarily) proves to be a significant feature in the classification between individuals with PD and control populations. Device magnitude of acceleration (e.g.,

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Feature</th>
<th>Value</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finger Tapping (Fine Motor)</td>
<td>Root Feature</td>
<td>x ≤ 32.5</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>Total Targets Tapped</td>
<td>x ≤ 1.1937</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Square Tracing (Fine-Motor)</td>
<td>Root Feature</td>
<td>x ≤ 0.907</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>First and Last Point Distance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average Drawing Speed</td>
<td>x ≤ 251.92</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td>First and Last Point Distance</td>
<td>x ≤ 182.21</td>
<td>0.375</td>
</tr>
<tr>
<td>Circle Tracing (Fine-Motor)</td>
<td>Root Feature</td>
<td>x ≤ 74.77</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Narration (Speech)</td>
<td>Root Feature</td>
<td>x ≤ 5.948</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>Total Elapsed Speaking Time</td>
<td>x ≤ 3.689</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time from Start to First Word Said</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Square Tracing (Fine-Motor)</td>
<td>Root Feature</td>
<td>x ≤ 1327.25</td>
<td>0.496</td>
</tr>
<tr>
<td></td>
<td>Minimum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum Magnitude of Acceleration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Object Naming (Speech)</td>
<td>Root Feature</td>
<td>x ≤ 10.5</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total Correct Objects</td>
<td>x ≤ 1.5</td>
<td>0.124</td>
</tr>
<tr>
<td></td>
<td>Average Time for Correct Response</td>
<td>x ≤ 2.52</td>
<td>0.444</td>
</tr>
<tr>
<td>Card Matching (Memory)</td>
<td>Root Feature</td>
<td>x ≤ 1.98</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>Average Time Between Non-Match Pair</td>
<td>x ≤ 1.372</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Card Matching with Cognitive Interference (Memory)</td>
<td>Root Feature</td>
<td>x ≤ 2.377</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>Average Time Between Non-Match Pair</td>
<td>x ≤ 0.949</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grandfather Passage (Speech)</td>
<td>Root Feature</td>
<td>x ≤ 1.325</td>
<td>0.423</td>
</tr>
<tr>
<td></td>
<td>Maximum Magnitude of Acceleration</td>
<td>x ≤ 0.584</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>First Order Feature(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average Time between Words Said</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
maximum, minimum, or average magnitude of acceleration) was a significant (e.g., root or first order) feature in 12 of 14 tests (85.7%), regardless of test type (e.g., motor, memory, speech, executive function or multi-functional tests). Magnitude of acceleration features are unable to be collected in pen-and-paper style assessments, nor are they currently being collected across all functional tests in implemented digital versions of clinical assessments. Conversely, variably present manifestations (e.g., dysarthria and difficulty performing simultaneous actions\textsuperscript{41}) are able to be seen in current pen-and-paper versions or other clinically implemented assessments. These difficulties can be depicted in timing and accuracy features. However, these depictions are more significant in the delineation between stages (e.g., early versus advanced-stage PD). Features of timing (10 of 14 tests; 71.43%) and accuracy (8 of 14 tests; 57.14%) are significant (e.g., root or first order) features in classification of stage, whereas only 5 of 14 tests (35.71%) contain significant device motion features. Mobile-based versions of neurocognitive assessments have the capability to expand clinically relevant functional assessments\textsuperscript{42,43}. Employing opportunistic approaches to monitoring (e.g., having device sensors on in the background regardless of test type) allows for the collection of novel objective features\textsuperscript{9}. An example shown in this work was the use of the device accelerometer across all functional tests. However, this premise could also include other device sensors such as microphones and cameras.

Commonly, patient reported outcomes are used to monitor an individual’s thoughts or opinions on changes in their condition which can lead to improved disease management in the recognition and understanding of their symptoms and triggers\textsuperscript{16,18}. However, this perceived information may be subject to individual variability and/or bias\textsuperscript{19–21}. This work depicts variability in perceived functionality scores (e.g., motor, memory, speech, and executive function) from actual scores for some groups. The perceived functionality for individuals in confirmed early-stage PD (H&Y Stages 1 and 2) across the areas of memory, speech, and executive function, differs by about 22% compared to their actual functionality as shown in Figures 2 and 3. Further, these figures show a relatively large perceived increase in executive function and behavioral abilities for individuals in advanced stages of PD (H&Y Stages 3, 4, and 5) compared to their early-stage counterparts. Therefore these digital health systems, with the ability to administer, collect, and subsequently analyze objective features, should be utilized in a way to allow individuals greater insights on their true capabilities.

A limitation of this preliminary work is that some participants in this study did not know what stage of PD they were in

Figure 1. Breakdown of actual scores across motor, memory, speech, executive function, and multi-test functional areas of neurocognition among groups of control, early (H&Y Stages 1 and 2), and advanced stages (H&Y Stages 3, 4, and 5) of PD.

(e.g., their respective stage was not communicated to them via a licensed clinician). This may be due to both the nature of the disease (e.g., being a “designer disease”) as well as the subjectivity in the staging criteria. Further, given the population size, stage classification had to occur on the basis of grouped stages of Parkinson’s (e.g., early (H&Y Stages 1 and 2) or advanced (H&Y Stages 3, 4, and 5)), rather than individual observed stages 1-5. Increased communication between clinicians and diagnosed populations on the advancement and severity of the disease is imperative for both rehabilitative efforts and disease monitoring. Future work in this domain will allow for the utilization of trained models to systematically classify individuals into different stages, and larger populations will allow for a more fine grained classification approach. Additionally, the work completed in this manuscript only shows a snapshot of Parkinson’s disease and their subsequent neurocognitive functionalities. Future work should analyze how different interventions (e.g., medical, pharmacological, dietary, physical, speech, and occupational therapies) affect significant features and overall functionality in comparison to controls. Finally, this work only shows a subset of commonly utilized clinical assessments. As this preliminary work includes novel digital features that were collected and analyzed via the supervised learning model, validation against other datasets was not possible. Future work should include further testing and metric collection across a much larger sample size for the provision of revised digital health system classification models. This is necessary for the classification of individuals with different neurodegenerative conditions (e.g., Alzheimer’s disease, ALS, multiple sclerosis, Huntington’s disease, and other types of dementia) and their respective stages.

Conclusions

Using digital health systems for monitoring individuals with neurodegenerative diseases allows for more comprehensive insights of these conditions and their progression. As digital health technology allows for the collection of large amounts of complex health data, ML provides a more efficient way to analyze and interpret patterns in the data. In addition, classification methods may also bring to light new features that expand the way digital versions of functional assessments should be administered. Utilizing this information to reevaluate and update standardized techniques and scales can truly allow for novel digital health
systems and disease monitoring. In turn, this can aid clinicians, diagnosed populations, and caretakers in monitoring all neurocognitive functions while also allowing for increased efficacy for both diagnostic and rehabilitative purposes.

References


Figure 3. Breakdown of actual functionality across motor, memory, speech, executive function, and multi-test functional areas of neurocognition among groups of control, early (H&Y Stages 1 and 2), and advanced stages (H&Y Stages 3, 4, and 5) of PD.


27. Domingos, P. review articles Tapping into the "folk knowledge" needed to advance machine learning applications. 55, DOI: 10.1145/2347736.2347755 (2012).


**Declarations**

All authors of this work declare that there are no conflicts of interest in the authorship or publication of this contribution. The work presented in this manuscript is part of an approved study by The University of Notre Dame Institutional Review Board (IRB) for the Protection of Human Subjects. All methods were performed in accordance with the relevant guidelines and regulations from the IRB. Written informed consent was collected from all participants included in this study. The collected data was authorized for disclosure as part of published works.

**Data Availability**

The datasets generated and/or analysed during the current study are not publicly available due to this data being part of a larger dataset for concurrent projects on mobile based neurocognitive assessments. However, the data can be made available from the corresponding author upon reasonable request.