Studying the Importance of Biomarkers for the Detection of Early-Stage Parkinson's Disease

Subhadeep Sengupta
Institute for Artificial Intelligence
University of Georgia
Athens, GA
ss45938@uga.edu

***Abstract*—Parkinson's disease is a neurodegenerative disease that requires early detection because treatment options are severely limited if found in advanced stages. Thus, finding reliable biomarkers that can provide valuable insights is imperative. Based on previous research, the most relevant features include protein concentrations from the Cerebrospinal Fluid, olfactory dysfunction, and REM sleep behavior disorder. This project also used additional features based on feature importances obtained from a boosted tree algorithm. Two sets of features, handpicked and those with high feature importance, were used on six algorithms for prediction focusing on maintaining optimal computational efficiency without the tradeoff of accuracy. Upon experimentation, various algorithms performed well depending on the features used and whether the data was balanced with oversampling. Tests administered by medical experts seem to take priority over biomarkers like Cerebrospinal fluid and striatal ratios.**

***Index Terms*—feature selection, data mining, classification, imbalanced data**

I. Introduction

Parkinson's disease (PD) is a neurodegenerative disease that is hard to detect in its early stages. By the time significant symptoms are displayed, around 60-80% of dopaminergic neurons are already impaired. The medical community has been actively studying Parkinson's disease (PD) to find ways to identify the disease early through potential biomarkers and changes in dopaminergic pathways in the brain. The stage between the onset of PD and the appearance of symptoms is called the prodromal phase and is an area targeted in current research. During the prodromal phase, subjects display minor symptoms like olfactory dysfunction and Rapid Eye Movement (REM) sleep disorders. The issue with these symptoms is that they may signify not just PD but many other similar diseases, such as Essential Tremor or Multiple System Atrophy. Some other relevant biomarkers have been spotted in specific protein concentrations in the Cerebrospinal fluid (CSF) and in the dopamine transporter imaging of subjects suspected of having a PD risk. Another aspect to consider is how model performances change based on the features used. This project will demonstrate the same.

II. Related Work

Zhang et al. [1] compiled relevant works on diagnosing PD by sorting them into categories based on the specific dataset features used for their prediction. Z Yu et al. [2] use a multivariate logistic regression model that includes the α-synuclein protein, olfactory disorder questionnaire, age, and gender features to discriminate between PD and healthy control. Prashant et al. [3] used the Wilcoxon rank sum test to find the statistical significance of features obtained from various modalities, such as non-motor features and interpreted ratios from imaging. They conclude their study by demonstrating impressive results from four algorithms, with SVMs generally performing the best. El Maachi et al. [4] used a 1D Convolutional neural network to classify patients based on gait data. They also predicted the severity of PD with an accuracy of 85.3%.

This project mentions many clinical motor and non-motor tests as features in the PPMI dataset. The REM(Rapid Eye Movement) sleep behavior disorder screening questionnaire [7] was introduced upon identifying sleep disorders as one of the indicators of the early onset of PD. The Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale [8] was a test that checked both motor and non-motor aspects of a patient suspected to have PD. A few non-motor features useful in the project are anxious mood, apathy, and urinary problems. Some tests developed early continue to have an impact. For instance, the Symbol digit modalities test [9], developed in 1973, tested the cognitive processing speed of the test-takers by having them substitute digits for abstract symbols. The State-trait anxiety inventory [10], introduced in 1983, was primarily designed to diagnose anxiety and differentiate its symptoms from that of depressive syndromes. The Hopkins Verbal Learning Test [12] tests word memorization right after learning and after a short time gap. The SCales for Outcomes in Parkinson's disease -AUTonomic [12] is a test that checks for nervous system disorders. The project found two valuable features with this test relating to urinary and sexual dysfunction disorders. The Montreal Cognitive Assessment [13] is a screening test that checks for mild cognitive impairment by testing the test-taker's concentration, memory, language, and orientation. The test results used in this project were adjusted for the patients' educational backgrounds. The Benton Judgement of Line Orientation [14] test checks the visuospatial ability of patients, something that is associated with the functioning of the parietal lobe of the right hemisphere.

The project also uses the XGBoost library [15] to help find feature importance coefficients by fitting the training data on the XGBClassifier model.

III. Dataset Summary

The Parkinson's Progression Marker's Initiative [6] is a study that has created an open-access dataset to study biological markers of Parkinson's disease and its progression. The data consists of around 1700 patient records collected through various modalities. Initially the combined curated data has about 111 features. Some of the relevant features are listed below.

- Olfactory dysfunction obtained from the University of Pennsylvania Smell Identification Test.
- REM Sleep Behavior Disorder Questionnaire (RBDSQ)
- Cerebrospinal fluid (CSF) protein concentrations
- SCOPA-AUT (SCales for Outcomes in Parkinson's disease - Autonomic Dysfunction)
- STAI (State-Trait Anxiety Inventory)
- MCI (Mild Cognitive Impairment)
- Semantic Fluency Score
- MDS-UPDRS (Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale)
- HVLT (Hopkins Verbal Learning Test)
- MoCA (Montreal Cognitive Assessment)
- Categorical Hoehn and Yahr Scale
- Symbol Digital Modalities Score
- Benton Judgement of Line Orientation Score
- Striatum, Putamen, and Caudate measurements

IV. Data Preprocessing

The train test split was 80-20, and adding stratification using the Sklearn parameter ensured the ratio of classes remained the same in the training and testing dataset. The Imblearn package offers functions to resample the training set to check whether similar target class value ratios affected the model's accuracy. Given the relatively small size of the dataset, oversampling was chosen over undersampling. Missing values were hard to deal with, given the nature of the features, i.e., requiring specific domain knowledge. Also, both training and testing sets would need imputations. Thus, removing those features with more than 200 missing values left about 77 functional features. Among those features, the selection was made based on two approaches.

A. Feature Selection

The first was to handpick features highlighted by Prashant et al. [2], namely the various protein concentrations and RBDSQ scores. The second was to evaluate the training subset on the XGBoost classifier, a boosted tree-based classifier. Then, the feature importances were obtained. Feature Importance calculates how the performance measure improves for an attribute if the tree splits at that attribute. [5] Selecting twenty features with the highest feature importance allows for targeted focus and faster runtime. These ended up being slightly different from the handpicked features.

Following are the descriptions of the twenty selected features below.

- 'fampd_new': Whether the patient has a family member with PD. The new indicates more detailed intervals separating families within first-degree and non-first-degree family members with PD.
- 'HISPLAT': Whether the patient is of Hispanic ethnicity
- 'rem': RBDSQ score
- 'SDMTOTAL': Symbol Digit Modalities Score
- 'NP1APAT': MDS-UPDRS Part 1 Apathy
- 'stai_trait': STAI (State-Trait Anxiety Inventory) State Subscore
- 'hy': Categorical Hoehn & Yahr
- 'age_cat': Age of patients at enrollment arranged into intervals.
- 'MCI_testscores': MCI (Mild Cognitive Impairment) decided based on at least two cognitive test scores being 1.5 below the standardized mean.
- 'bjlot': Benton Judgement of Line Orientation score.
- 'scopa_ur': SCOPA-AUT (Autonomic Symptom test) Urinary Score
- 'updrs2_score': MDS-UPDRS Part 2 Score
- 'scopa_sex': SCOPA-AUT Sexual Dysfunction Score
- 'VLTVEG': Semantic Fluency Score - Vegetable Subscore
- 'td_pigd_old': Tremor dominant (TD) and postural instability and gait difficulty (PIGD) disorder.
- 'hvlt_immediaterecall': HVLT (Hopkins Verbal Learning Test) Immediate/Total Recall
- 'hvlt_retention': HVLT Retention
- 'hvlt_discrimination': HVLT Discrimination
- 'NP1ANXS': MDS-UPDRS Part 1 Anxious Mood
- 'moca': MOCA (Montreal Cognitive Assessment) Score (with a note adjusting for education)

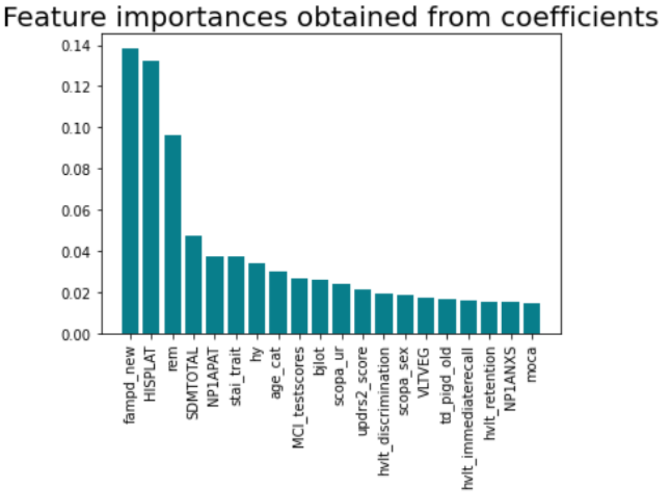


Fig. 1. Features selected using the second method

One of the features, namely, 'HISPLAT,' appeared to be problematic and was removed due to a lack of significant connection to the diagnosis of PD. The feature identified whether the patient was Hispanic or not. It also appeared to have a high correlation with the target class, as shown in Figure 1. If it stayed in the dataset, it would be a bias.

V. Model Development Process

The Kaggle Kernel was the primary tool for experimentation. It uses the NVIDIA Tesla GPU P100 with 12 GB memory. The script was written in Python and used the Sklearn, Numpy, Pandas, Imblearn, and Tabulate packages. Model testing was done on the preprocessed version of the dataset to verify and compare model performance with previous existing works, namely with the work done by Prashant et al. [3].

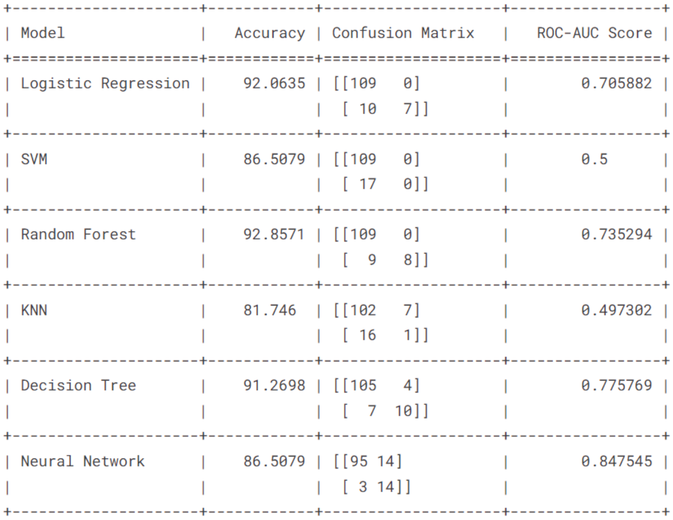
VI. Results

Based on feature importance, the algorithms preferred the data obtained from various motor and non-motor tests that check cognitive competencies over biomarkers like protein concentrations. The authors in [3] used biosamples as part of their dataset to check their viability. Experimentation done with four sets of results was obtained based on the dataset used. Upon testing the imbalanced data with handpicked features (Figure 2), Multi-layer Neural Networks appeared to give the best result with the highest ROC-AUC score, with Random Forests giving the best accuracy but falling short on the ROC-AUC score. With the balanced data and handpicked features (Figure 3), Logistic Regression performs the best with the highest ROC-AUC score, while Random Forest gives the highest accuracy. With the imbalanced data and feature importance selected features (Figure 4), Neural Networks performed the best with the highest ROC-AUC score and accuracy. With the balanced data and feature importance selected features (Figure 5), KNN gave the highest ROC-AUC score, while Random Forest and Neural Networks had the highest accuracy. Comparing the results with [3], SVM could have performed better. One of the possible reasons for the poor performance is a lack of clarity on the parameter specifications for SVM. The only mention is that the authors used the radial basis function kernel. The authors also insisted that individual model performances could not reliably be said to be superior to one another. Instead, it was better to focus on the features that helped arrive at these predictions. Given the present data, medical tests still prevail over biosample-based markers. This implies a hindrance towards models that rely on automating the diagnosis of PD and minimizing the manual efforts from healthcare providers.

The code for the project in the form of a Kaggle notebook can be found on this website.(https://www.kaggle.com/code/alexmas0n/data-mining-project)

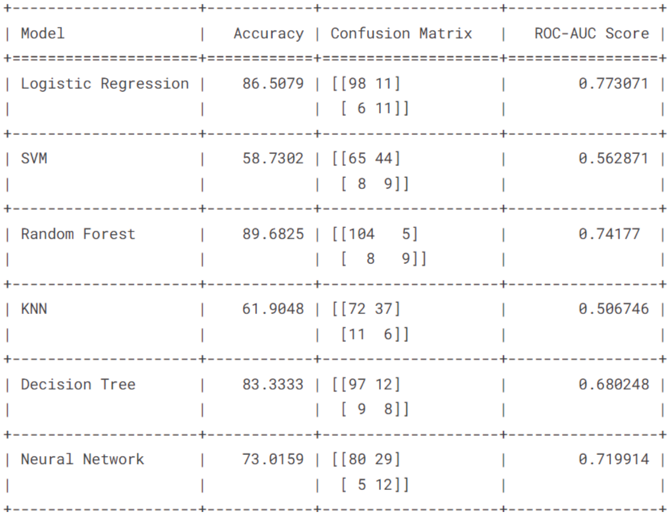
VII. Future Work

A significant hindrance to the project was missing values. Missing values could not simply be imputed, given that each



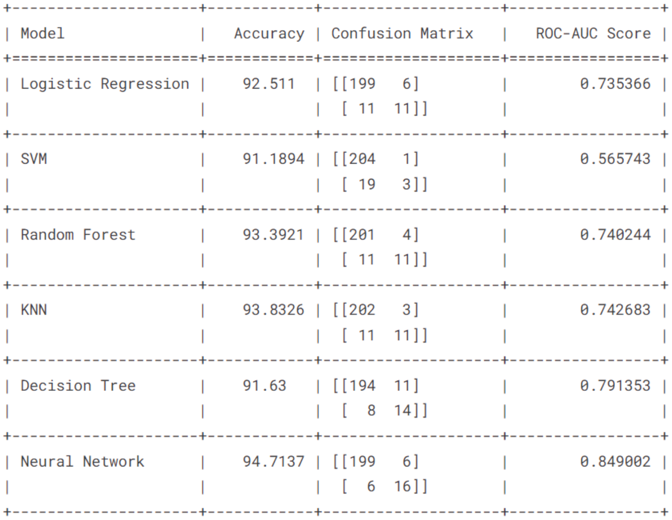
```
+---------------------+------------+--------------------+-----------------+
| Model               |   Accuracy | Confusion Matrix   |   ROC-AUC Score |
+=====================+============+====================+=================+
| Logistic Regression |    92.0635 | [[109   0]         |        0.705882 |
|                     |            |  [ 10   7]]        |                 |
+---------------------+------------+--------------------+-----------------+
| SVM                 |    86.5079 | [[109   0]         |        0.5      |
|                     |            |  [ 17   0]]        |                 |
+---------------------+------------+--------------------+-----------------+
| Random Forest       |    92.8571 | [[109   0]         |        0.735294 |
|                     |            |  [  9   8]]        |                 |
+---------------------+------------+--------------------+-----------------+
| KNN                 |    81.746  | [[102   7]         |        0.497302 |
|                     |            |  [ 16   1]]        |                 |
+---------------------+------------+--------------------+-----------------+
| Decision Tree       |    91.2698 | [[105   4]         |        0.775769 |
|                     |            |  [  7  10]]        |                 |
+---------------------+------------+--------------------+-----------------+
| Neural Network      |    86.5079 | [[95 14]           |        0.847545 |
|                     |            |  [ 3 14]]          |                 |
+---------------------+------------+--------------------+-----------------+
```

Fig. 2. Results on Handpicked Features (Imbalanced Data)



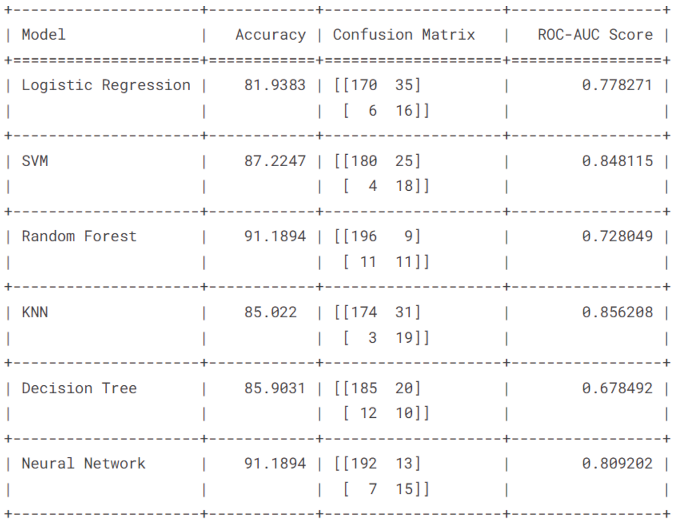
```
+---------------------+------------+--------------------+-----------------+
| Model               |   Accuracy | Confusion Matrix   |   ROC-AUC Score |
+=====================+============+====================+=================+
| Logistic Regression |    86.5079 | [[98 11]           |        0.773071 |
|                     |            |  [ 6 11]]          |                 |
+---------------------+------------+--------------------+-----------------+
| SVM                 |    58.7302 | [[65 44]           |        0.562871 |
|                     |            |  [ 8  9]]          |                 |
+---------------------+------------+--------------------+-----------------+
| Random Forest       |    89.6825 | [[104   5]         |        0.74177  |
|                     |            |  [  8   9]]        |                 |
+---------------------+------------+--------------------+-----------------+
| KNN                 |    61.9048 | [[72 37]           |        0.506746 |
|                     |            |  [11  6]]          |                 |
+---------------------+------------+--------------------+-----------------+
| Decision Tree       |    83.3333 | [[97 12]           |        0.680248 |
|                     |            |  [ 9  8]]          |                 |
+---------------------+------------+--------------------+-----------------+
| Neural Network      |    73.0159 | [[80 29]           |        0.719914 |
|                     |            |  [ 5 12]]          |                 |
+---------------------+------------+--------------------+-----------------+
```

Fig. 3. Results on Handpicked Features (Oversampling)



```
+---------------------+------------+--------------------+-----------------+
| Model               |   Accuracy | Confusion Matrix   |   ROC-AUC Score |
+=====================+============+====================+=================+
| Logistic Regression |    92.511  | [[199   6]         |        0.735366 |
|                     |            |  [ 11  11]]        |                 |
+---------------------+------------+--------------------+-----------------+
| SVM                 |    91.1894 | [[204   1]         |        0.565743 |
|                     |            |  [ 19   3]]        |                 |
+---------------------+------------+--------------------+-----------------+
| Random Forest       |    93.3921 | [[201   4]         |        0.740244 |
|                     |            |  [ 11  11]]        |                 |
+---------------------+------------+--------------------+-----------------+
| KNN                 |    93.8326 | [[202   3]         |        0.742683 |
|                     |            |  [ 11  11]]        |                 |
+---------------------+------------+--------------------+-----------------+
| Decision Tree       |    91.63   | [[194  11]         |        0.791353 |
|                     |            |  [  8  14]]        |                 |
+---------------------+------------+--------------------+-----------------+
| Neural Network      |    94.7137 | [[199   6]         |        0.849002 |
|                     |            |  [  6  16]]        |                 |
+---------------------+------------+--------------------+-----------------+
```

Fig. 4. Results on Significant Features (Imbalanced Data)



```
+---------------------+------------+-------------------+-----------------+
| Model               |   Accuracy | Confusion Matrix  |   ROC-AUC Score |
+=====================+============+===================+=================+
| Logistic Regression |    81.9383 | [[170  35]        |        0.778271 |
|                     |            |  [  6  16]]       |                 |
+---------------------+------------+-------------------+-----------------+
| SVM                 |    87.2247 | [[180  25]        |        0.848115 |
|                     |            |  [  4  18]]       |                 |
+---------------------+------------+-------------------+-----------------+
| Random Forest       |    91.1894 | [[196   9]        |        0.728049 |
|                     |            |  [ 11  11]]       |                 |
+---------------------+------------+-------------------+-----------------+
| KNN                 |    85.022  | [[174  31]        |        0.856208 |
|                     |            |  [  3  19]]       |                 |
+---------------------+------------+-------------------+-----------------+
| Decision Tree       |    85.9031 | [[185  20]        |        0.678492 |
|                     |            |  [ 12  10]]       |                 |
+---------------------+------------+-------------------+-----------------+
| Neural Network      |    91.1894 | [[192  13]        |        0.809202 |
|                     |            |  [  7  15]]       |                 |
+---------------------+------------+-------------------+-----------------+
```

Fig. 5. Results on Significant Features (Oversampling)

feature might have a large or small role in determining the diagnosis. Imputing would also need to be done for training and testing sets, leading to unreliable results. Datasets with a large number of features where each patient is thoroughly examined for all of these features still need to be more extensive. Further work could be applied in providing more detailed examinations of patients. One way the datasets may be expanded is by adding records of patients with diseases that often have similar symptoms, for example, Multiple System Atrophy. This might help focus and differentiate features and help models be more robust to real-world scenarios. Furthermore, adding a third target class for patients with advanced-stage Parkinson's might be used to check the model's generalizability and identify differences between the two stages of the disease.

Acknowledgment

The data used for this project may be accessed from the following website https://www.ppmi-info.org/accessdata-specimens/download-data. For up-to-date information on the study, visit www.ppmi-info.org.

PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including 4D Pharma, AbbVie Inc, AcureX Therapeutics, Allergan, Amathus Therapeutics, Aligning Science Across Parkinson's (ASAP), Avid Radiopharmaceuticals, Bial Biotech, Biogen, BioLegend, Bristol Myers Squibb, Calico Life Sciences LLC, Celgene Corporation, DaCapo Brainscience, Denali Therapeutics, The Edmond J. Safra Foundation, Eli Lilly and Company, GE Healthcare, GlaxoSmithKline, Golub Capital, Handl Therapeutics, Insitro, Janssen Pharmaceuticals, Lundbeck, Merck & Co., Inc., Meso Scale Diagnostics, LLC, Neurocrine Biosciences, Pfizer Inc., Piramal Imaging, Prevail Therapeutics, F. Hoffmann-La Roche Ltd and its affiliated company Genentech Inc., Sanofi Genzyme, Servier, Takeda Pharmaceutical Company, Teva Neuroscience, Inc., UCB, Vanqua Bio, Verily Life Sciences, Voyager Therapeutics, Inc., Yumanity Therapeutics, Inc.

The author would like to thank Dr. Kimberly Van Orman and Ms. Akhila Devabhaktuni for their guidance with the project planning, as well as for helping identify the important points of focus for the project.

References

[1] Zhang, Jing. ”Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of Parkinson's disease.” npj Parkinson's Disease 8.1 (2022): 1-15.

[2] Yu, Zhenwei, et al. ”Combining clinical and biofluid markers for early Parkinson's disease detection.” Annals of clinical and translational neurology 5.1 (2018): 109-114.

[3] Prashanth, R., et al. ”High-accuracy detection of early Parkinson's disease through multimodal features and machine learning.” International journal of medical informatics 90 (2016): 13-21.

[4] El Maachi, Imanne, Guillaume-Alexandre Bilodeau, and Wassim Bouachir. ”Deep 1D-Convnet for accurate Parkinson disease detection and severity prediction from gait.” Expert Systems with Applications 143 (2020): 113075.

[5] “Feature Importance and Feature Selection with XGBoost.” Notebook.community, https://notebook.community/minesh1291/MachineLearning/xgboost/feature_importance_v1

[6] Marek, Kenneth, et al. ”The Parkinson progression marker initiative (PPMI).” Progress in neurobiology 95.4 (2011): 629-635.

[7] Stiasny-Kolster, Karin, et al. ”The REM sleep behavior disorder screening questionnaire—a new diagnostic instrument.” Movement disorders 22.16 (2007): 2386-2393.

[8] Goetz, Christopher G., et al. ”Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): process, format, and clinimetric testing plan.” Movement disorders 22.1 (2007): 41-47.

[9] Smith, Aaron. Symbol digit modalities test. Los Angeles: Western psychological services, 1973.

[10] Spielberger, Charles D. ”State-trait anxiety inventory for adults.” (1983).

[11] Brandt, Jason. ”The Hopkins Verbal Learning Test: Development of a new memory test with six equivalent forms.” The clinical neuropsychologist 5.2 (1991): 125-142.

[12] Visser, Martine, et al. ”Assessment of autonomic dysfunction in Parkinson's disease: the SCOPA-AUT.” Movement disorders: official journal of the Movement Disorder Society 19.11 (2004): 1306-1312.

[13] Nasreddine, Ziad S., et al. ”The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment.” Journal of the American Geriatrics Society 53.4 (2005): 695-699.

[14] Spencer RJ, Wendell CR, Giggey PP, Seliger SL, Katzel LI, Waldstein SR. Judgment of Line Orientation: an examination of eight short forms. J Clin Exp Neuropsychol. 2013;35(2):160-6. doi: 10.1080/13803395.2012.760535. Epub 2013 Jan 28. PMID: 23350928; PMCID: PMC3668441.

[15] Chen, Tianqi, and Carlos Guestrin. ”Xgboost: A scalable tree boosting system.” Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.