# Decomposing Parkinson's Disease Using Supervised Learning and Statistical Analysis

HÉCTOR MANUEL ORTIZ-MENA, BACHELOR OF FINE ARTS,

ROHAN SHROFF, BACHELOR OF SCIENCE,

VINCENT GACUTAN, BACHELOR OF SCIENCE IN ELECTRONICS AND COMMUNICATIONS ENGINEERING

GEORGIA INSTITUTE OF TECHNOLOGY, ATLANTA, GA, U.S.A.

# Abstract: A concise understanding of our research topic

- This presentation demonstrates the results of utilizing supervised learning and statistical analysis.

**Objective:**

- To decompose and understand, Parkinson's disease, which has no proper diagnosis and progresses differently from patient to patient.
- To achieve a better understanding about the factors contributing to the diagnosis of this disorder.
- To be able to construct an efficient metric to diagnose patients with high probability of success.

**Focus:**

- Feature selection using randomized decision trees and its success was measured on supervised learning classifiers.
- Supervised Classification
- We explored the distributions of the found features using basic statistics to determine how they are related and if any contribute to a positive diagnosis.

**Summary of Achievements:**

- Our results show that with 0.995 accuracy and AUC, the data set can be classified for each patient as control or having Parkinson's based on whether the patient produces non-zero values for **NHY**.
- Other features were also found to be relevant such as: NP3RIGLL, NP3PRSPL, NP3RTCON, NP3FTAPL, NP3GAIT, PN3RIGRL and NP3SPCH

# What is Parkinson's Disease?

- Parkinson's Disease is a progressive disease of the nervous system marked by tremor, muscular rigidity, and slow, imprecise movement, chiefly affecting middle-aged and elderly people. It is associated with degeneration of the basal ganglia of the brain and a deficiency of the neurotransmitter dopamine (Google).

# Motivation

- Disease leads to the deterioration of motor functions and it is not easily diagnosed.
- There is no standardized test, and doctors will devise their own way of administering their version of this examination.
- "Currently when someone is diagnosed with Parkinson's disease it is difficult to determine what type of Parkinson's they have or how quickly the condition will progress."
- "There are different types of Parkinson's that can look similar at the point of onset, but they progress very differently. We are hoping the information we collect will differentiate between these different conditions. It is her hope that "ultimately we'd like doctors to be able to conduct tests that can predict how the disease is likely to progress."

  - Dr. Deborah Apthrop of the ANU Research School of Psychology

# Related Works:  Machine Learning and Parkinson's Disease

- Dr. Deborah Apthrop of the ANU Research School of Psychology
  - Her research involves measuring brain imaging, eye tracking, visual perception and postural sway.
- Mandal I, Sairam N. New machine-learning algorithms for prediction of Parkinson's disease.
  - "Their research presents an enhanced prediction accuracy of diagnosis of Parkinson's disease (PD) to prevent the delay and misdiagnosis of patients using a proposed robust inference system. These methods include sparse multinomial logistic regression, rotation forest ensemble with support vector machines and principal component analysis, artificial neural networks and boosting methods.
- Hazan H, Hilu D, Manevitz L, Ramig L, Sapir S. Early diagnosis of Parkinson's disease via machine learning on speech data. Early Diagnosis of Parkinson's Disease via Machine Learning on Speech Data
  - Their study showed that early detection of PD from speech data seems to be feasible and accurate with results approaching the 90% mark in two different data sets.
- Zhang H-H, Yang L, Liu Y, Wang P, Yin J, Li Y, et al. Classification of Parkinson's disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples.
  - "This study showed that the proposed method could improve PD classification when using speech data and can be applied to further studies seeking to improve PD classification methods.

# Our Approach

- The research examples described above focuses on classification.
- None seem to point to which factors have the most influence on the diagnosis of Parkinson's.
- Determining these factors is crucial to find insight on what causes Parkinson's with the purpose of preventing or determining future cases with high probability of success and accuracy.

# Method:  The data set

- The diagnosis of Parkinson's disease (PD) currently relies upon clinical criteria.
- No existing laboratory test diagnoses PD or gauges the effectiveness of a treatment on underlying disease processes.
- Current methods for diagnosing, treating, and prognosticating PD are inadequate and would be greatly improved by the discovery and validation of biomarkers.
- The data set we are going to use for this project is from Biofind (Fox Investigation for New Discovery of Biomarkers in Parkinson's Disease)
- BioFIND is an observational clinical study designed to discover and verify biomarkers of Parkinson's disease (PD).

# Method:  The data set (continued)

Further Explaining Datasets: BioFind Datasets

- The Data Sets recovered from BioFind have numerous rows and columns located in the .csv files. The team is restructuring the hundreds of features which are located under columns and merging them into a more organized fashion using Hadoop. By using Hadoop the Data can be processed quickly and also can be organized to reflect the classifications that are really needed. To put this in simple words, Hadoop is used as a filter to pinpoint the Biochemical Features to create the Classification data needed.

# Method: The data set (continued)

Further Understanding Classifications and Features

- The Classifications used have a large affect with our outcome. Classifications can be connected to the Biomarkers. Classifications of Parkinson's disease include, Clinical, Biochemical, Genetic, and Imaging. This study will focus on the Biochemical tributes of diagnosis.
- The Data that has been provided on the classification set will be derived from the hundreds of features given in the BioFind Data. The features which in turn describe Classifications, are housed within a tree structure. Classifications will include everything described earlier; blood, saliva, csf, and biopsy, and hundreds of others that can be used to find and algorithmic diagnosis for PD.
- Overall, the Features that are listed, will help the research put the information gathered into different classifications. It is also evident that the data influences the selections made by the algorithms and trees that are in place. The next section will describe the use of the data sets and the algorithms that come into play.

# Method: The data set (continued) Explaining Biomarkers:

- A Biomarker, is a sample of biological fluids which can be divided into categories to narrow down the synopsis of this illness. The Biomarkers that will be used within this research in conjunction with a written algorithm will lead our result in this study.
- To further explain the medical definition of a biomarker. A description is provided below from a local source.
- The Biomarkers Definitions Working Group [8] has defined a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention".
- The Biomarker this research compares will be of Biochemical origin. Biochemical biomarkers are data collected from Blood, Saliva, CSF, and Biopsy procedures. The Features used, link directly to the .csv files that provide the different categories that are needed to run the code.

**Project Pipeline**

Raw CSV files
(222 patients / 2371 features)

⇩

Hadoop Hive
(merge data sets)

⇩

Hadoop Pig
(Data Preprocessing and ETL (SVMLight) Construction)

⇩

Baseline Model Testing
(GridSearchCV in Scikit-Learn/Python)

⇩

Model Optimization
(GridSearchCV in Scikit-Learn/Python

⇩

Feature Selection (Using ExtraTree Classifier in Scikit-Learn/Python)
Output: 9 Features (NHY being the most important)

⇩

Statistical Analysis on Reduced Feature Set
(including manual classification in R)

⇩

Final Model Testing (Scikit-Learn/Python)

⇩

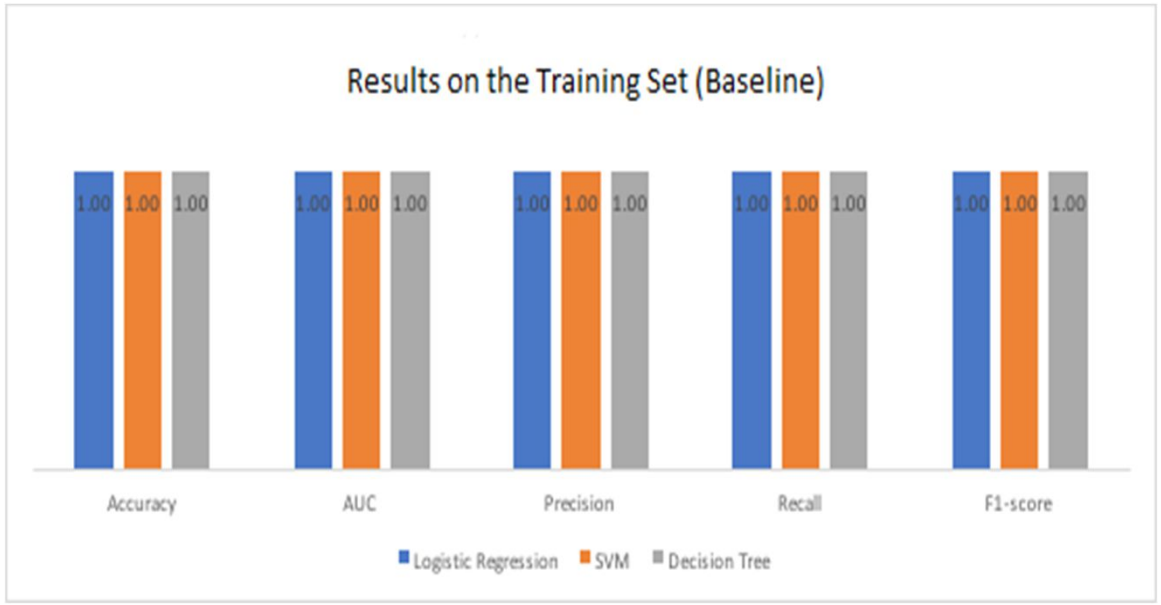
Factors Contributing to Parkinson's Disease

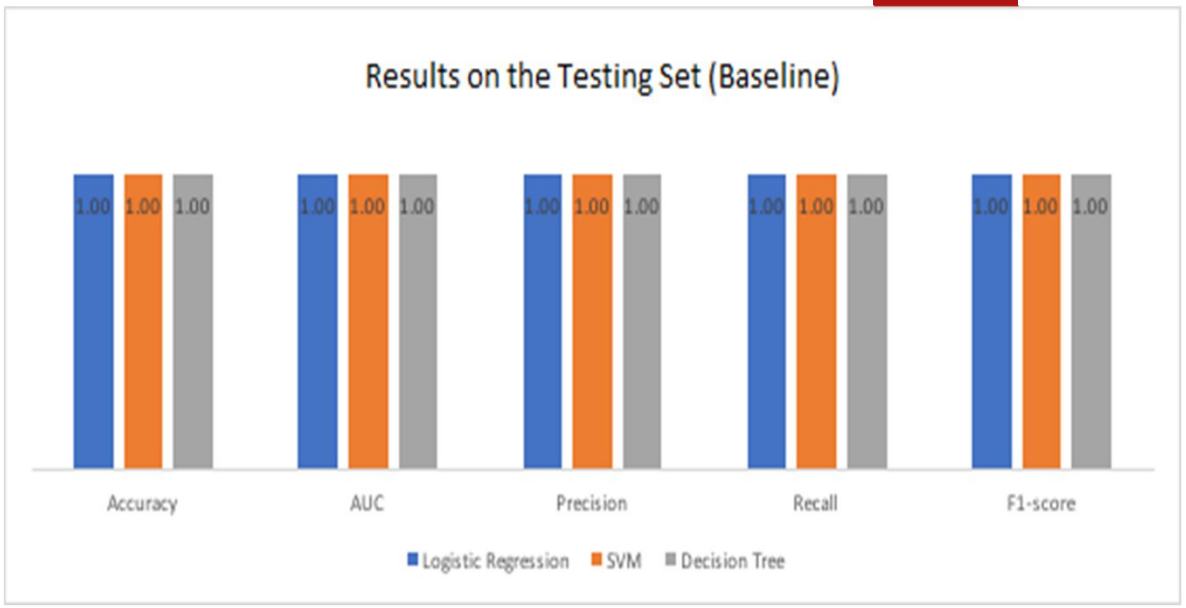
# Project Pipeline

**Total Features: 2371**
**Number of Patients: 222**
**Training Size: 117 patients**
**Testing Size: 45 patients**

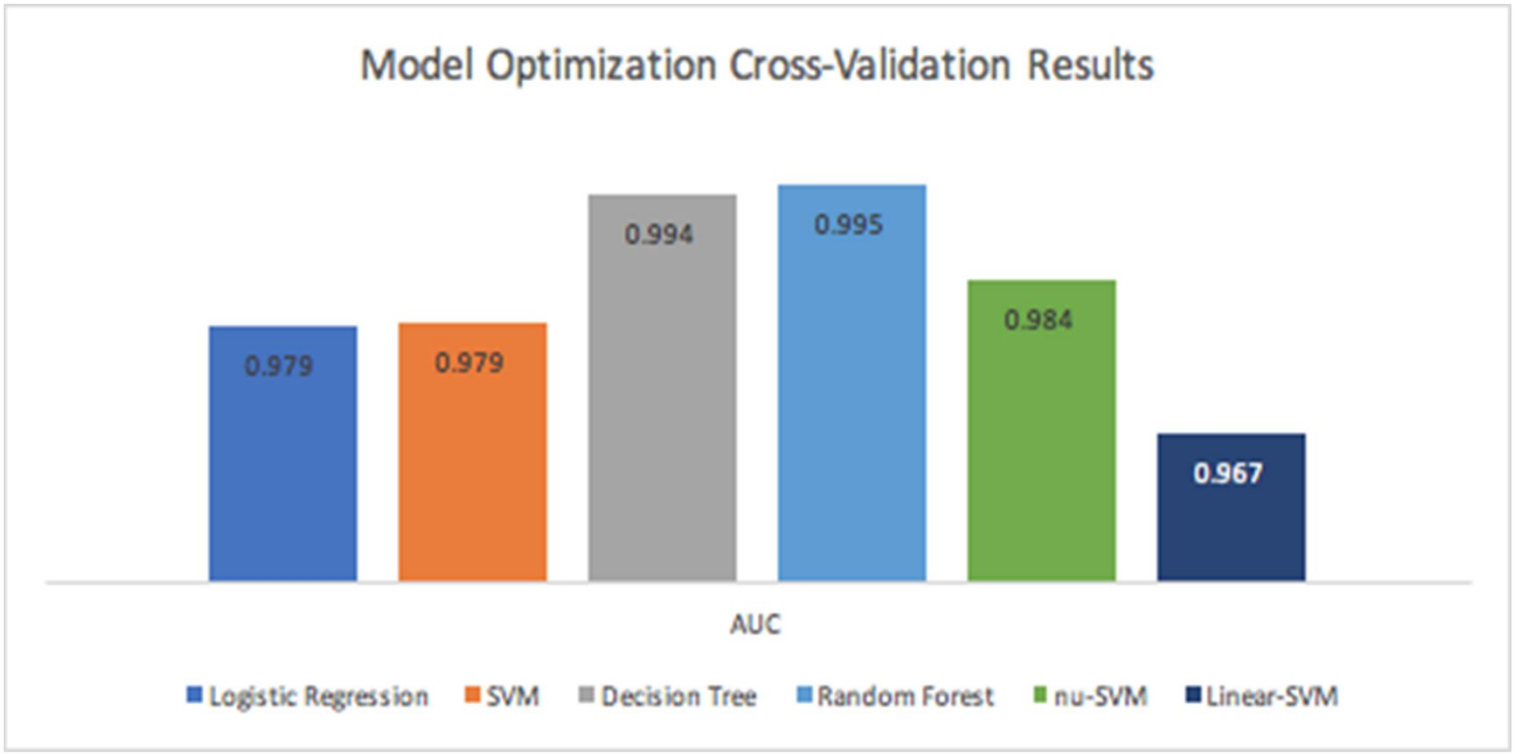
# Method: Code Pipeline

# Results:



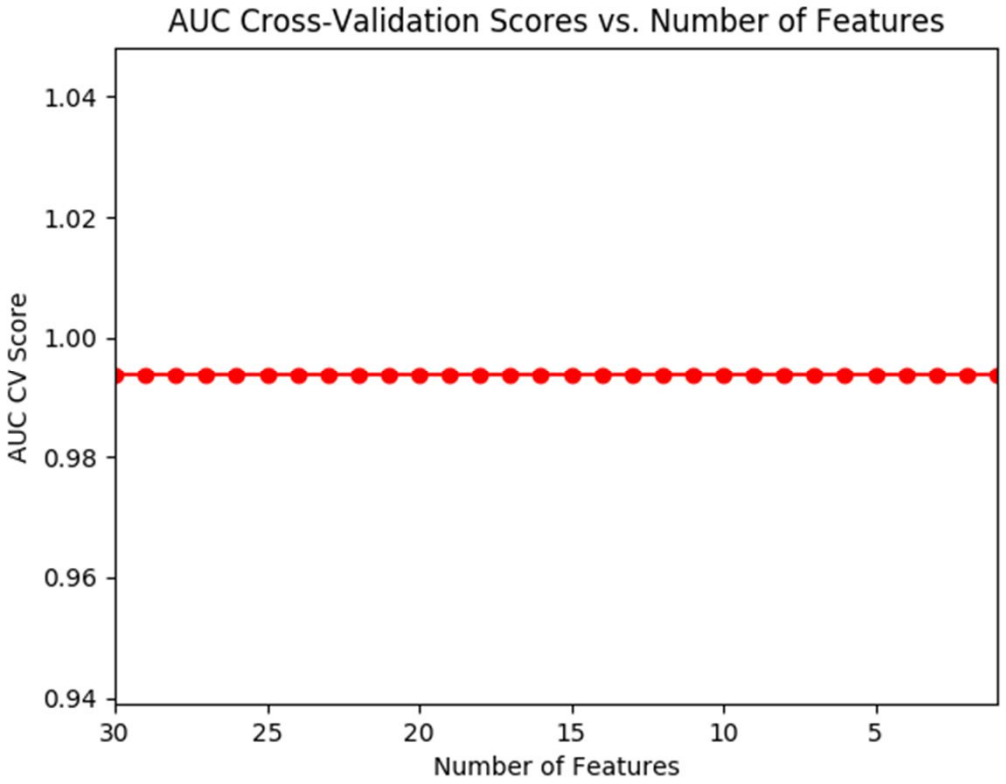

**Baseline using full training/test sets and no parameter optimization results (2371 features)**




**Model optimization (using cross validation (3 folds) from full training set and no feature selection) (2371 features)**

| | Accuracy on the Training Set (Size = 177) | | | | | | Accuracy on the Testing Set (Size = 45) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Logistic Regression | SVM | nu-SVM | Linear-SVM | Decision Tree | Random Forest | Logistic Regression | SVM | nu-SVM | Linear-SVM | Decision Tree | Random Forest |
| Accuracy | 1 | 1 | 1 | 1 | 0.99435 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| AUC | 1 | 1 | 1 | 1 | 0.99375 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Precision | 1 | 1 | 1 | 1 | 0.989796 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Recall | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| F1-score | 1 | 1 | 1 | 1 | 0.994872 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

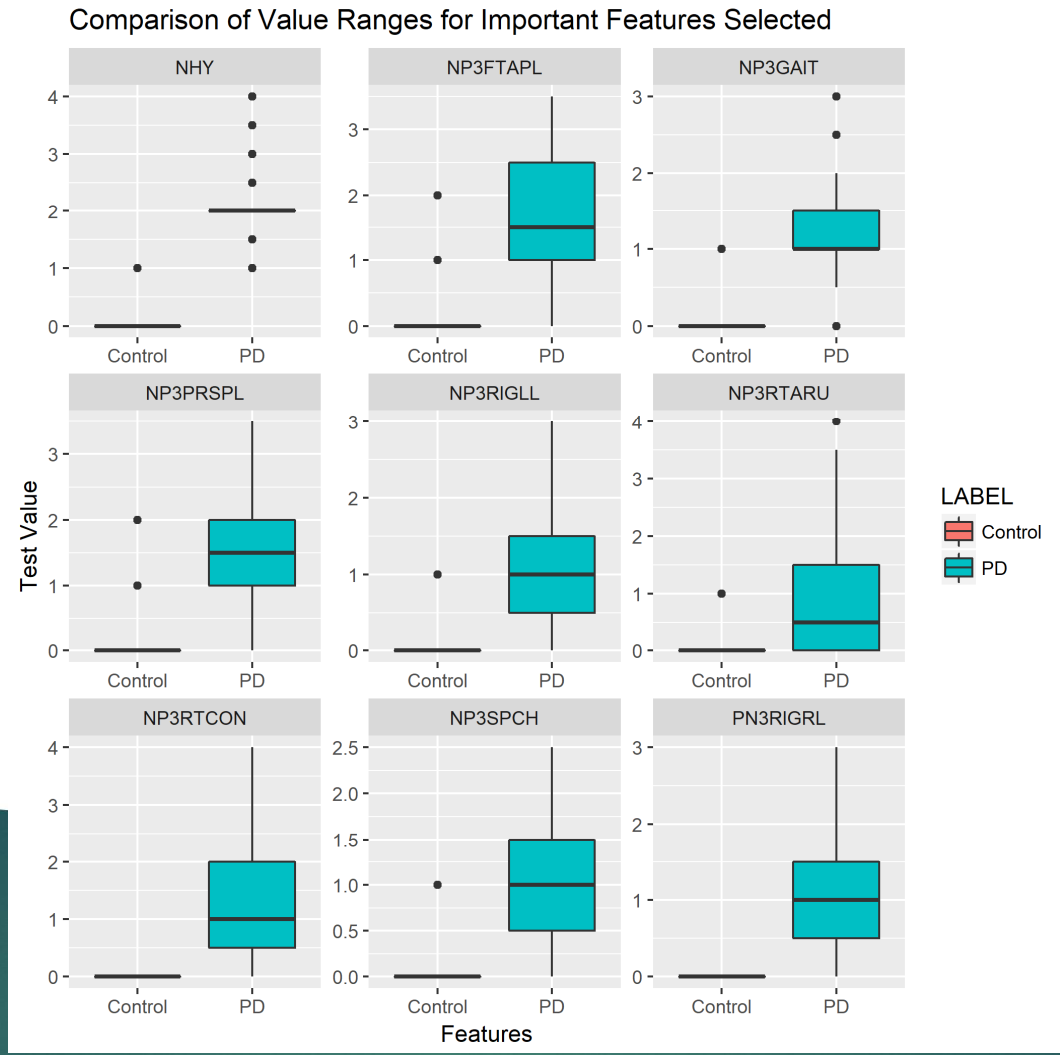
**Using full training/test sets and optimized models with no feature selection (2371 features)**




# Effect on AUC on optimized decision tree model of reducing number of features

| | Accuracy on the Training Set (Size = 177) | | | | | | Accuracy on the Testing Set (Size = 45) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Logistic Regression | SVM | nu-SVM | Linear-SVM | Decision Tree | Random Forest | Logistic Regression | SVM | nu-SVM | Linear-SVM | Decision Tree | Random Forest |
| Accuracy | 0.99435028 | 0.994350282 | 0.96 | 0.99435 | 0.99435 | 0.99435028 | 1 | 1 | 1 | 1 | 1 | 1 |
| AUC | 0.99375 | 0.99375 | 0.964 | 0.99375 | 0.99375 | 0.99375 | 1 | 1 | 1 | 1 | 1 | 1 |
| Precision | 0.98979592 | 0.989795918 | 1 | 1 | 0.989796 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Recall | 1 | 1 | 0.928 | 0.989796 | 1 | 0.98979592 | 1 | 1 | 1 | 1 | 1 | 1 |
| F1-score | 0.99487179 | 0.994871795 | 0.963 | 0.994872 | 0.994872 | 0.99487179 | 1 | 1 | 1 | 1 | 1 | 1 |

## Using full training/test sets and optimized models with feature selection (1 feature)




# Ranges of nine of the most important found features

| IDX | Feature Importance | Feature Name | Accuracy |
|---|---|---|---|
| 2038 | 0.07746535 | NHY | 0.99549550 |
| 2057 | 0.06717689 | NP3RIGLL | 0.91441440 |
| 2052 | 0.06573103 | NP3PRSPL | 0.90540540 |
| 2067 | 0.06549024 | NP3RTCON | 0.91441440 |
| 2042 | 0.05020073 | NP3FTAPL | 0.92342340 |
| 2066 | 0.03175642 | NP3RTARU | 0.78828830 |
| 2044 | 0.02791790 | NP3GAIT | 0.96396400 |
| 2164 | 0.02698589 | PN3RIGRL | 0.86936940 |
| 2068 | 0.02566674 | NP3SPCH | 0.89189190 |

Nine of the most important features (feature importance and manual classification accuracy)

# Final Results

- Final Results: 0.99549550 accuracy
- With this level of accuracy, one can manually classify the data as either Parkinson's or not. Alternatively, any of our models can perform the same using just the NHY feature with near positive metrics. This is an improvement to our previous AUC of 0.94 and 0.93 accuracy in our draft phase 1. More patients would need to be tested to confirm as we are aware that our testing set is only comprised of 45 patients.
- Found a good selection of features for diagnosing Parkinson's disease.
- NHY is a good indicator alone for determining Parkinson's
- NP3RIGLL, NP3PRSPL, NP3RTCON, NP3FTAPL, NP3GAIT, PN3RIGRL and NP3SPCH, 7 of the 9 features shown on figure 9 and 10, are also good indicators
- It would seem that these features work independently
- We believe that we have succeeded in finding a subset of features that work as strong indicators for diagnosis Parkinson's based on the limited amount of patients available to us

# Further Research

- Include more patients
- Reduce the data set by first eliminating features that yield high importance values first, rather than eliminating less important features and keeping the most important (reverse as we did for slide 15) → would reconfirm results

# Conclusion

- Results of our investigation show how utilizing supervised learning, in particular, feature selection with randomized decision trees, and statistical analysis can help narrow down the factors contributing to the diagnosis of a medical disease.
- Our pipeline, composed of big data tools such as the Hadoop framework, Scikit-Learn and Python, is successful in discovering which features contribute to the diagnosis of Parkinson's.
- The new 112 features added since the draft phase improved the metric of all our chosen model to near perfection.
- Due to the results not consistently producing perfection, we claim that our method has an accuracy of 0.995 for all models and manual classification.