2017

Big data and Parkinson's: Exploration, analyses, data challenges and visualization

Mahalakshmi Senthilarumugam Veilukandammal
Iowa State University

Big data and Parkinson's: Exploration, analyses, data challenges and visualization

by

Mahalakshmi Senthilarumugam Veilukandammal
(Mahalakshmi S V)

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Information Systems

Program of Study Committee:
Sree Nilakanta, Major Professor
Baskar Ganapathysubramanian
Vellareddy Anantharam
James A Davis

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2017

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

NOMENCLATURE

PPMI	Parkinson's Progression Markers Initiative
PD	Parkinson's disease
AD	Alzheimer's disease
MDS	Movement Disorder Society
UPDRS	Unified Parkinson Disease Rating Scale
AUC	Area Under the Curve
ICDM	International Conference on Data Mining
HC	Healthy Control
GENPD	Genetic Cohort Parkinson's disease
GENUN	Genetic Cohort Unaffected
REGPD	Genetic Registry Parkinson's disease
REGUN	Genetic Registry Unaffected
PRODROMA	Prodromal
EDA	Exploratory Data Analysis
PCA	Principal Component Analysis

ACKNOWLEDGEMENT

I would like to acknowledge Dr. Sree Nilakanta, my major professor. This thesis would not have been possible without his supervision, advice, and guidance from the very early stage to the end.

At the same time, I am very grateful to Dr. Vellareddy Anantharam, Dr. Baskar Ganapathysubramanian, Dr. James A Davis, Dr. Anumantha Kanthasamy and Dr. Auriel A Willette. I am deeply indebted for their insightful suggestions and guidance.

My sincere appreciation to the cybiz team Judy Eyles - Director, Armita Ahmadzadehhosseini -MS Finance ,Kennan Davis – Undergrad Finance, Avantika Ram – MBA, Naveen Dhanpal – MBA, Alexandre Andrade – MBA . I would like to extend my sincere appreciation to Dr. David Jiles and Joseph Boldrey, Graduate Student in Electrical and Computer Engineering.

In addition, I would also like to thank my friends, the department faculty, and staff for making my time at Iowa State University a wonderful experience. Last but not least my deepest gratitude goes to my family.

This research was supported by Presidential Initiative for Interdisciplinary Research at Iowa State University.

ABSTRACT

In healthcare, a tremendous amount of clinical, laboratory tests, imaging, prescription and medication data are collected. Big data analytics on these data aim at early detection of disease which will help in developing preventive measures and in improving patient care. Parkinson disease is the second-most common neurodegenerative disorder in the United States. To find a cure for Parkinson's disease biological, clinical and behavioral data of different cohorts are collected, managed and propagated through Parkinson's Progression Markers Initiative (PPMI). Applying big data technology to this data will lead to the identification of the potential biomarkers of Parkinson's disease. Data collected in human clinical studies is imbalanced, heterogeneous, incongruent and sparse. This study focuses on the ways to overcome the challenges offered by PPMI data which is wide and incongruent. This work leverages the initial discoveries made through descriptive studies of various attributes. The exploration of data led to identifying the significant attributes. This research project focuses on data munging or data wrangling, creating the structural metadata, curating the data, imputing the missing values, using the emerging big data analysis methods of dimensionality reduction, supervised machine learning on the reduced dimensions dataset, and finally an interactive visualization. The simple interactive visualization platform will abstract the domain expertise from the sophisticated mathematics and will enable a democratization of the exploration process. Visualization build on D3.Js is interactive and will enable manual exploration of traits that correlate with the disease severity.

CHAPTER 1. INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disorder, and millions of people suffer from it all over the world. The incidence of PD increases with the age growth, about 6.3 million people are suffering from this disease. Notably, in a developed country, the number of patients with PD has increased significantly in recent years. However, there are no methods which can measure the PD progression efficiently and accurately in its early stages. The last known drug for Parkinson's disease was found in 1967.

Common symptoms of PD are

- Muscular rigidity (inflexibility of muscles)
- Shivering (vibration in upper and lower limbs or jaws)
- Speech problem,
- Expressionless face
- Bradykinesia (slow movements)
- Lethargy
- Postural instability (depression and emotional changes)
- Involuntary movements
- Dementia (loss of memory)
- Thinking inability
- Sleeping disorders

Various stages of Parkinson's disease are,

- Primary - Due to unknown reasons
- Secondary - Dopamine deficiency
- Hereditary- Genetic origin

- Multiple system atrophy - Degeneration of parts other than midbrain

Movement Disorder Society-sponsored Unified Parkinson Disease Rating Scale (MDS-UPDRS) is traditionally used for PD assessment. To better understand PD progression and to identify potential biomarkers Parkinson's Progression Markers Initiative (PPMI) was set up. PPMI data is collected from clinical sites in the United States, Europe, Israel and Australia. The Michael J. Fox Foundation funds the PPMI. PPMI collects clinical, biological and imaging data from multiple sites and disseminates it. This data can be used to diagnose, track and predict PD and its progression.

Parkinson's data possess all the characteristics of big data, which are characterized by volume, variety, velocity, veracity, and value. From the context of Parkinson's data, these five Vs are,

- Volume – With more and more attributes being collected for the Parkinson's research and with the increase in participation of different cohorts through various initiatives, the volume of the data is growing.
- Variety –Parkinson's disease contains structured, text, images, audio and semi-structured data collected from the various smart fitness tracking devices.
- Velocity-Velocity is depicted by the speed in which data is created, stored and processed. Nowadays real-time processing systems aid in real-time decision making.
- Veracity- Veracity deals with the integrity of data. Data quality issues and reliability of the information are the key elements in veracity. Parkinson's data is heterogeneous, multi-source, incomplete, incongruent and sparse.
- Value- Extracting value from the data is the goal of big data analytics.

The goal of working with the Parkinson's data from the public databases is to find potential biomarkers thereby finding a cure for the disease. This research has data from human clinical, GWAS, proteomics, RNAomics, metabolomics and other databases associated with Parkinson's disease (PD). Data is downloaded from PPMI (Parkinson's Progression Markers Initiative) and PDBP (Parkinson's Disease Biomarker Program). Cleaning and curating the data, to discover patterns from it is very challenging.

The main contribution of this study is to identify significant attributes related to PD. This study tries to answer the following questions.

- Is it possible to use various machine learning algorithms to help in early detection of Parkinson's disease?
- Which data needs to be analyzed to discover the biomarkers of Parkinson's disease?
- How can we develop an interactive visualization that helps physicians understand the relations between various attributes that correlate with Parkinson's disease?
- Is it feasible to scale the visualization for many user inputs? Does it yield the same result as the initial visualization with the training set?

This study focuses on understanding the attributes and creating a metadata of the attributes. Descriptive studies, creating a master sheet, dimensionality reduction and building predictive models with various supervised machine learning algorithms are the next steps in the study. Significant attributes from the large set of attributes are identified and mapped in a simple interactive visualization to understand the correlation between these attributes and Parkinson's disease.

CHAPTER 2. LITERATURE REVIEW

Big data in healthcare has always been an exciting field. With the tremendous amount of data in healthcare applying big data technology is very useful for early detection of disease and focused treatment. Preventive care and early detection of diseases will help improve lifestyle and will reduce the economic burden of the nation.

Big Data Challenges in Healthcare

There is no single method or a unique way of applying big data technology. The complexity is because of the volume, variety, and veracity of the data sets (Amiri, Clarke and Clarke 2015). The authors have discussed clustering technique which helps in identifying natural classes within a dataset. Clustering is an unsupervised learning technique. Genomic data has all the characteristics of big data that is high dimensional and mostly categorical. The most challenging task is inferring the result of the high dimensional data. Curse of dimensionality impedes from understanding the result. However, the authors have created ensemble clustering package in R, statistical tool. The advantage of this ensembling method is to create more clearly separated clusters from these categorical data.

In cancer research, high-throughput genomic and proteomic data has been used (Clarke, Ressom, Wang, Xuan, et al. 2008). The authors have built predictive models for diagnosis of cancer and to identify therapeutic targets for drug development. These technologies presented by investigators helped in extracting meaningful statistical and biological information from high-dimensional data space. Each sample had hundreds or thousands of measurements. The researchers endorse the idea about the properties of high dimensionality data which was

stated above and agree that high dimensional data are often poorly understood or overlooked in data modeling and analysis.

The complexity (volume, variety, and velocity) of biomedical data have tremendously increased. Biomedical field is providing petabytes of new neuroimaging and genetics data every year (Dinov, Petrosyan, Liu, et al. 2014). The authors discussed big data's challenges and the role of big data technology in the biomedical field. Many computational algorithms are being developed at the same time. Software tools and services to use these algorithms on those petabytes of data are also being designed. Users are expecting to have intuitive and quick access to data, software tools, and a high-speed infrastructure. The challenges in data analysis are exponential with the explosion of information, scientific techniques, computational models. The pipeline environment overcomes the problems posed by biomedical data analysis. The pipeline is the crowd-based distributed solution for data management.The data is heterogeneous in nature. The pipeline enables multiple (local) clients and (remote) servers to connect, exchange data, control the process, monitor the activities of different tools or hardware, and share complete protocols as portable XML workflows. As stated by this paper Laboratory of Neuro Imaging (LONI) is one such pipeline environment for Parkinson's big data research. LONI aims to improve the understanding of the brain health and disease.

Parkinson's Disease

Parkinson disease (PD) is the second-most common neurodegenerative disorder in the United States. The lack of proper treatment options for PD progression, with an increasingly elderly population, portends a rising economic burden for the Nation (Kowal, Dall, Chakrabarti, et al. 2013). The financial hardship seems to grow exponentially both in direct and indirect ways. The study substantiated the lack of innovative methods to detect the disease in the early stage.

Earlier detection leads to preventive cure and better medication. Their argument to set up a national disease registry accompanied with an extensive database to store all the information collected seems reasonable. PPMI is one such initiative, and this has led to many innovative researches attempting to find a cure for Parkinson's.

To bolster the fact mentioned above about Parkinson's disease, the report of Alzheimer's disease (AD) also illustrates the effect of neurodegenerative disease and the need to identify the onset of disease as early as possible. The report provides the details about Alzheimer's disease. Incidence and prevalence, mortality rates, health expenditures and costs of care, are some of the attributes of the report. It explores not only the challenges of patients but also of the caregivers. An estimated 5.2 million Americans have an AD. 200,000 young people less than 65 years are also affected with the AD; 5 million comprise the older-onset AD population. In the coming decades, the number of people with an AD will increase by ten million. Every 68 seconds someone develops AD. It is estimated that someone will develop an AD every 33 seconds. The Alzheimer disease is the sixth leading cause of death in the United States. Many Americans age 65 years or older die because of the AD. Between 2000 and 2010, there was a decrease in the number of deaths caused by heart disease, stroke, and prostate cancer, whereas the proportion resulting from AD increased by 68%. In 2012, it was estimated that $216 billion was spent by AD caregivers. Medicare payments to beneficiaries age 65 years and older with the AD and other dementias are three times greater than amounts for beneficiaries with other diseases. Costs related to all these neurodegenerative diseases are continually increasing as their no cure or prevention as of now.

Now knowing the economic repercussion of Parkinson's disease, it becomes necessary to learn about the factors causing Parkinson's disease. Mutations in leucine-rich repeat kinase 2

(*LRRK2*), Rab7L1 are the genes associated with Parkinson's disease (Beilina, Cookson, 2015). The genetic basis of Parkinson's disease (PD) was progressing. The study led to the identification of genes that are inherited in PD or show robust association with the sporadic disease. They have discussed what those genes tell about the underlying biology of PD. The relationships between protein products of PD genes displayed the conventional links that include regulation of the autophagy-lysosome system.Autophagy-lysosome is an essential way by which cells recycle proteins and organelles. They also discuss whether all PD genes should be in the same way. In some instances, the gene relationships are closer, whereas in other cases the interactions are more distant and might be considered separate. The authors also review the links between genes for Parkinson's disease and the autophagy-lysosomal system. They proposed the hypothesis that many of the known PD genes are assigned to pathways that affect turnover of mitochondria via mitophagy turnover of several vesicular structures via macroautophagy and chaperone-mediated autophagy or general lysosome function.

Machine Learning Algorithms for Parkinson's Data

Researchers used the longitudinal data available from Parkinson's Progression Markers Initiative (PPMI) data for an accurate diagnosis and early detection of Parkinson's data (Nalls, McLean, Rick, et al.). This research has the potential to be of great benefit for researchers and clinical practice. Their research aimed to create an accurate classification model for the diagnosis of Parkinson's disease, which could be extended to longitudinal cohorts. The study was conducted on Parkinson's Progression Marker Initiative (PPMI) dataset with 367 PD patients and phenotypical imaging data and 165 controls without neurological disease.The logistic regression classification model had an olfactory function, genetic risk, family history of Parkinson's disease, age, and gender as the significant features. The test dataset had 825 subjects

with Parkinson's disease and 261 controls from five independent cohorts with varying recruitment strategies and designs.The data was collected from Parkinson's Disease Biomarkers Program (PDBP), the Parkinson Associated Risk Study (PARS), 23andMe, the Longitudinal and Biomarker Study in PD (LABS-PD), and the Morris K Udall Parkinson's Disease Research Center of Excellence cohort (Penn-Udall). The model also used to investigate patients who had imaging scans without evidence of dopaminergic deficit (SWEDD). Four of 17 SWEDD participants which this model classified as having Parkinson's disease converted to Parkinson's disease within one year. This model provided a potential new approach to distinguish participants with Parkinson's disease from controls. This research has future scope to also identify individuals with prodromal or preclinical Parkinson's disease in prospective cohorts that could lead to the identification of biomarkers.

With a high accuracy classification model, the research needs to be further expanded to image files also (Dinov, Van Horn, Lozev, et al., 2009). The authors studied in detail about the LONI Pipeline. LONI pipeline is a graphical environment for execution of advanced neuroimaging data analysis. LONI pipeline enables construction, validation, and implementation of neuroimaging analysis. LONI has a library for computational tools which allows automated data format conversion. LONI *Pipeline* performs better than the other workflow architectures for graphical analysis. LONI is a distributed Grid computing environment. Efficient tool integration, protocol validation, and new algorithms for neuroimages are the critical features of LONI environment. Integration of existing data and computational tools within the LONI *Pipeline* environment is straight forward and intuitive. The LONI *Pipeline* has several types of process submissions depending on the server infrastructure. The LONI pipeline is portable, computationally efficient, distributed and independent.The individual binary processes

in pipeline data-analysis workflows do not affect the performance of the LONI pipeline. Advanced computational algorithms with quantitative mapping of brain structure and function have been developed by LONI. This research has dealt with Alzheimer's data, and it can be extended to Parkinson's disease neuroimages.

Fuzzy k-nearest neighbor (FKNN) was another efficient and effective algorithm for the diagnosis of Parkinson's disease (PD) (Chen, Huang, Yu, et al. 2013). The proposed FKNN-based system is compared with the support vector machines (SVM) based approaches. To further improve the diagnostic accuracy for detection of PD, principal component analysis was employed. Experimental results have demonstrated that the KNN-based system dramatically outperforms SVM-based approaches and other methods in the literature. A 10-fold cross-validation on FKNN method had an accuracy of 96.07%. Promisingly, the proposed system might serve as a powerful tool for diagnosing PD.

The most common problem in classification is imbalanced data (Ramentol, Caballero, Bello, et al. 2012). This phenomenon is significant as the data in practice is imbalanced. Many techniques have been developed to handle the imbalanced training sets in supervised learning. Such methods were divided into two large groups: those at the algorithm level and those at the data level. Data level groups emphasized to balance the training sets by the elimination of samples in the larger class or by constructing new samples for the smaller class, known as undersampling and oversampling, respectively. A hybrid method was proposed in the study for preprocessing imbalanced data-sets.They intended to construct new samples, using the Synthetic Minority Oversampling Technique.This technique together with the application of an editing technique based on the Rough Set Theory and the lower approximation of a subset. The proposed

method was validated by an experimental study showing good results using C4.5 as the learning algorithm.

Several machine learning algorithms were applied to various Parkinson's dataset. C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART are among them. These are the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) (Wu, Kumar, Quinlan, 2008). The datamining research community values these top 10 algorithms as the most powerful and influential analytical models. Classification, clustering, statistical learning, association analysis, and link mining are covered by these 10 algorithms.

It's mandatory to review the performance of top 10 data mining algorithms on Parkinson data .An automatic machine learning approach was designed for the detection of the Parkinson disease by analyzing the speech/voice of a person (Rustempasic, Indira, Can. 2013). The fuzzy C-means clustering algorithm was used in the study to identify Parkinson's disease. The fuzzy C-means had 68.04% accuracy, 75.34% sensitivity, and 45.83% specificity. The study was extended by applying the boosting algorithms, and principal component analysis for data reduction. The authors have also investigated and performed the feature relevance analysis to calculate the score. UPDRS (Unified Parkinson Disease Rating Scale) helped in diagnosing the Parkinson accurately. Further enhancements were made using five different classification paradigms utilizing a wrapper feature selection scheme which can predict each of the class variables with estimated accuracy in the range of 72–92 (Armañanzas, Rubén, Bielza, et al. 2013). They also proposed an SVM and k-Nearest Neighbor (k-NN). PD patients were monitored by recording their voice at regular intervals. The age, gender, voice recordings were taken at baseline, after three months, and after six months are used as features in the machine

learning algorithms. Support Vector Machine was successful in detecting significant deterioration in UPDRS score of the patients. The feature selection process for random forest and support vector machine increased the accuracy to discriminate PD from healthy controls. Only ten dysphonia features led to the classification accuracy of 99%. They also proposed a nonlinear signal approach for a large dataset.

Many researchers worked on a subset of data to determine Parkinson's. The data considered for analysis in the above papers were limited to only one subset of data such as voice dataset and gait dataset. Complex, heterogeneous and incomplete data from multiple sources in Parkinson's Progression Markers Initiative (PPMI) were cleaned, curated, harmonized and various machine learning classification algorithms were applied (Dinov, Heavner, Tang, et al. 2016). The study not only focusses on Fuzzy K nearest neighbors but other classification techniques as well. This research paper compares model based and model free classification techniques on PPMI dataset. The study indicated there was a significant difference in cognitive scores of PD (Parkinson's disease), HC (Healthy Control) groups. Classification improved significantly when rebalanced data was used.

Visualization

After understanding the various types of research in Parkinson data, trying to understand the latest studies using big data techniques in other diseases might help in implementing those well-proven ideas to Parkinson as well. Diagnosis of ovarian cancer is very complicated and expensive (Cheong, Song, Park, 2012). The data collected from the medical equipment when analyzed had some specific pattern. Examining the medical data in MATLAB was difficult for clinicians. When data mining model was built, and the output was visually published clinicians found the tool very convenient. The research paper focuses on applying various big data

methodologies for earlier diagnosis of ovarian cancer. The research paper aimed to develop a software tool that helps in early biomarker discovery of ovarian cancer. Data for the research and development of customizable software tool was collected from Luminex equipment that generates a tremendous amount of complex medical data. The data from the Luminex machine was curated, logistic regression was applied. The output was visualized in the form of a histogram, scatter plot, ROC curve. Logistic regression algorithm was run on the output data from the Luminex machine. The customized visualization software led to reduced lead time to analyze the result of the Luminex machine.

Current visual analytics systems provide users with the means to explore trends in their data (Maciejewski, Hafen, Rudolph, et al. April 2011).The study results indicate that interactive displays and linked views provide great insights about events.The visual analytics performed in study help understand the correlations between various attributes such as people, events, and places. Analysts use visual analytics to work on their events of interest ; drill down into the data, and form hypotheses based on visualizations. Hotspots (high incedence of events) were identified using the visualization of the spatiotemporal data. Analysts want to predict the future hotspots. Forecasting was done using a predictive visual analytics linked spatiotemporal and statistical analytic views of data. These visual analytics tools allowed analysts to do a hypothesis testing and plan for resources to tackle the predicted threats.

This research will overcome the limitations posed by various classification methods and will holistically analyze PPMI data. The main contributions of this study include

- An approach to cleaning and curating the dataset
- Aggregating the vast collection of attributes
- Preprocessing and imputing the sparse dataset

- Applying dimensionality reduction technique such as principal component analysis
- Utilize the reduced principal components to perform supervised machine learning algorithms
- Visualizing the highly correlated features in each principal component.

CHAPTER 3. RESEARCH GOALS

Data collected from PPMI study consists of clinical, biological and imaging data of various patients. There are 2600 attributes, and the number is continuously increasing as it is an ongoing study. This paper addresses the general challenges of data curation, munging, aggregation, and preliminary descriptive analyses, dimensionality reduction and supervised machine learning algorithms with principal components obtained. This study provides the results of the analyses and also a simple interactive visualization of the features.

Merits of this research – This framework with a simple interactive visualization will abstract people from sophisticated mathematics to provide a simplified and understandable version of the disease to the lifestyle of an ordinary man.

Method

PPMI data is complex, incomplete, inconsistent, incongruent and heterogeneous. The approach for data analysis starts with obtaining the required data from PPMI website. (http://www.ppmi-info.org/access-data-specimens/download-data/). Mining PPMI dataset is a very challenging problem. Data is analyzed, manipulated, cleaned, munged and aggregated. Data is categorized into six major categories for analysis. Metadata file containing the information of the curated dataset is created as well.

Data aggregation results in a large sparse dataset. Then the methods involve identification of missing patterns, data wrangling, and imputation. There are several methods available to handle missing values. The best method for this dataset was kNN based imputation. kNN based imputation method performed very well for the dataset.

The next major challenge after the sparsity of data is reducing the dimension of the aggregated dataset. The imputed data was normalized to apply the principal component analysis – an unsupervised technique for dimensionality reduction. The feature correlation in each principal component is analyzed. The result of the analysis is visualized using a circos graph. The principal components are further used to perform supervised machine learning classification algorithms such as logistic regression and support vector machines.

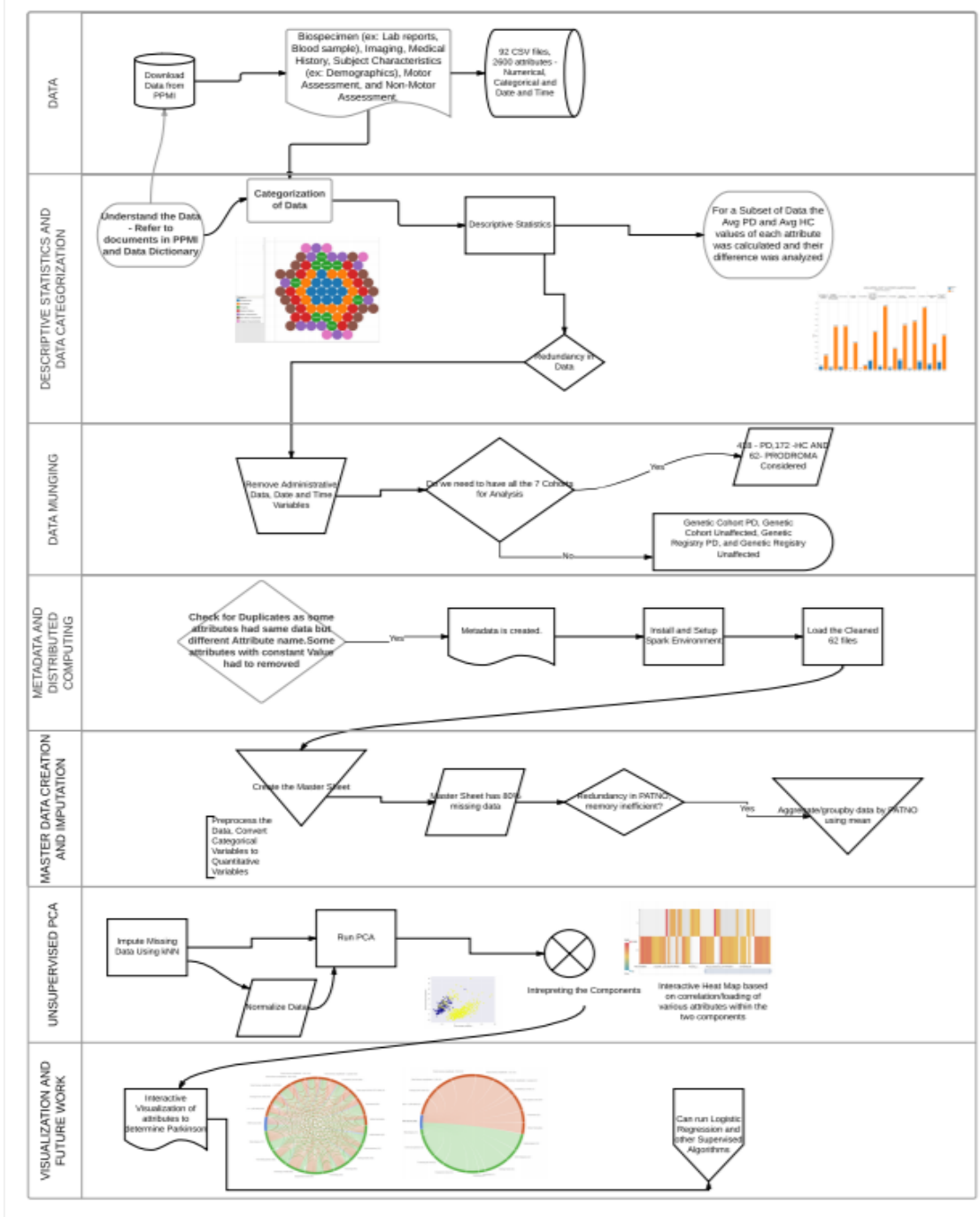

Figure 1. Top-level Schematic Representation of the End-to-End Protocol

CHAPTER 4. DATA DESCRIPTION

PPMI Bioinformatics Core disseminates the PPMI study dataset at the University of Southern California. This database includes clinical, biological and imaging data collected at various participating sites. PPMI also collects biologic specimens including urine, plasma, serum, cerebrospinal fluid, DNA, and RNA. The complete PPMI data set during May 2017 had 92 CSV data files. 12 files contained details about the administration and subject details included in the PPMI study. 80 files contained clinical information that can be used to find the biomarker of the disease. This study utilizes the dataset that was available during May 2017.

Data management is one of the most challenging and significant problems in medical studies.

Figure 2. Details about PPMI data files

The administrative data included in PPMI does not disclose the identity of the actual subject. Enrollment status of each subject, the clinical site details are included in the administrative data. Scripts update the data from clinical websites to the database every day. Transfer of imaging data to the database is a separate process. The database is updated every Sunday. After the high-level analysis (file-wise), the next step is a detailed analysis to understand the attributes and types.

Table 1. Details of Various Attributes in PPMI

Total	Total Number of Attributes	2600
Type	Numerical Attributes	779
	Categorical Attributes	1316
	Date	458
	Time	47
Redundancy	Total Number of Unique Attributes	1358
	Number of redundant Attributes	1242

The dataset had 2600 attributes as of May 2017. Only 30 % of the data are numerical. Remaining 70 % of the data are categorical, date and time. There are about 50 % redundant attributes as well.

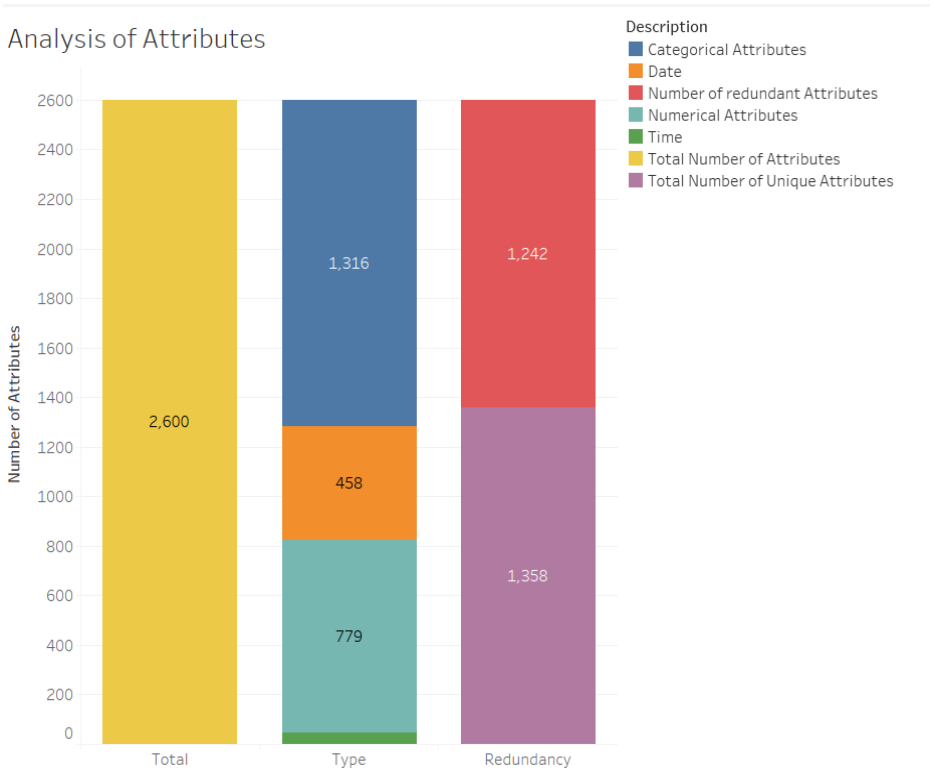

Figure 3. High-level Analysis of the Attributes

Study Participants

The dataset obtained from PPMI for our study consists of 1479 patients. PD are patients with Parkinson's disease. Our study data included 418 PD study participants. Healthy control (HC) is 172 in number. The study dataset consists of 418 PD, 172 HC and 62 Prodromal patients (In medicine, a prodromal is a set of signs and symptoms, which often indicate the onset of a disease before more diagnostic specific symptoms develop). The three categories of study participants total up to 652 out of 1479 patients.

Participants included in the study data are

- PD – Parkinson's disease
- HC – Healthy Control
- PRODROMA

Participants excluded from the current study are genetic cohort PD, genetic cohort unaffected, genetic registry PD and genetic registry unaffected.

Table 2. Details of Patient Status

Category of Study Participants	***Definition***	***Count***
HC	Healthy Control	172
PD	Parkinson's Disease	418
PRODROMA	Prodromal	62
GENPD	Genetic Cohort PD	181
GENUN	Genetic Cohort Unaffected	215
REGPD	Genetic Registry PD	191
REGUN	Genetic Registry Unaffected	240
Total		**1479**

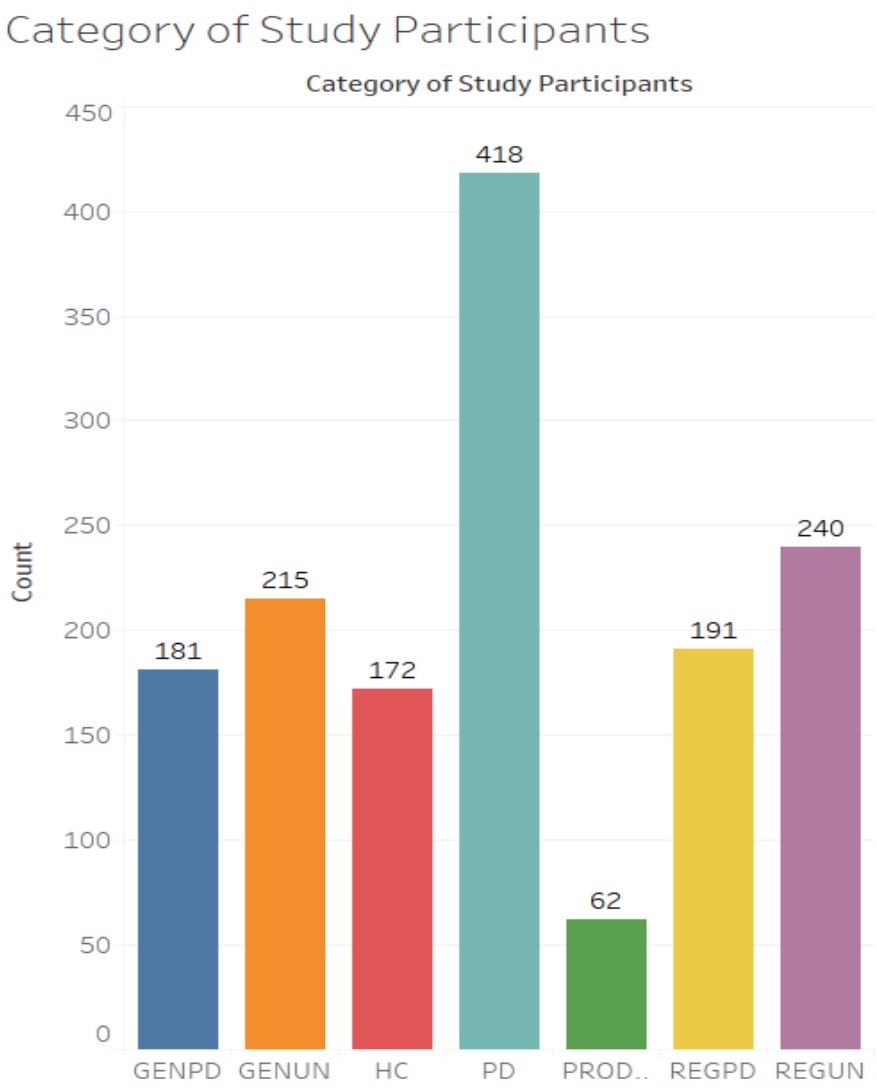

Figure 4. Category of Study Participants in PPMI

Data Categorization

Six major categories of the data files are

- Biospecimen (ex: Lab reports, Blood sample),
- Imaging (ex: DaTscan imaging, Magnetic Resonance Imaging)

- Medical History (ex: General medical history, General neurological exam, General physical exam, Pregnancy forms, Neurological exam cranial nerves)
- Subject Characteristics (ex: demographics, PPMI took place at clinical sites in the United States, Europe, Israel, and Australia)
- Motor Assessment (ex: assessment of tremor with bradykinesia, assessment of tremors in tongue, jaw, lower lip, hand or in the leg/foot. Movement Disorder Society (MDS) offers Unified Parkinson's Disease Rating Scale (UPDRS) which guides in the motor assessment)
- Non-Motor Assessment (ex: assessment of verbal learning, semantic fluency, and sleepiness scale is some of the non-motor assessment tests)

Categorization of data helps us understand the big picture, which data or type of test needs to be performed to determine Parkinson at the earliest. Each color code in the final visualization indicates the category of the attributes. The six categories of data for Parkinson disease is available in Parkinson Progressive Markers Initiative (PPMI). The categorization of data in our study is standardized and universal.

Table 3. Details of File Categorization

Data Category	**Files Included in Each Data Category**
Bio specimen	Biospecimen Analysis Results
	Blood Chemistry Hematology
	iPSC Blood Sample
	IUSM Biospecimen Cell Catalog
	IUSM Catalog
	Laboratory Procedures
	Laboratory Procedures with Elapsed Times
	Lumbar Puncture Sample Collection
	Pilot Biospecimen Analysis Results
	Skin Biopsy
	Whole Blood Sample Collection
Imaging	AV-133 Imaging
	AV-133 SBR Results
	DATScan Analysis
	DaTscan Imaging
	DaTSCAN SPECT Visual Interpretation Assessment
	DTI Regions of Interest
	FBB Analysis Data
	Magnetic Resonance Imaging
	MRI Imaging Data Transfer Information Source Document
	SPECT Scan Information Source Document
Medical History	Clinical Diagnosis and Management
	Concomitant Medications
	Diagnostic Features
	General Medical History
	General Neurological Exam
	General Physical Exam
	Initiation of PD Medication- incidents
	Neurological Exam Cranial Nerves
	PD Features
	Pregnancy Form
	Prodromal Diagnostic Questionnaire
	Surgery for Parkinson Disease
	Use of PD Medication
	Vital Signs
Motor	MDS UPDRS Part I
	MDS UPDRS Part I Patient Questionnaire
	MDS UPDRS Part II Patient Questionnaire
	MDS UPDRS Part III Post Dose

	Table 3. Details of File Categorization Continued
	MDS UPDRS Part IV
	Modified Schwab + England ADL
	PASE Household Activity
	PASE Leisure Time Activity
	TAP-PD Kinetics Device Testing
	TAP-PD OPDM Assessment
	TAP-PD OPDM Use Questionnaire
Non-Motor	Benton Judgment of Line Orientation
	Cognitive Categorization
	Epworth Sleepiness Scale
	Features of REM Behavior Disorder
	Geriatric Depression Scale-Short
	Hopkins Verbal Learning Test
	Letter Number Sequencing PD
	Montreal Cognitive Assessment MoCA
	Olfactory UPSIT
	QUIP Current Short
	REM Sleep Disorder Questionnaire
	SCOPA-AUT
	Semantic Fluency
	State-Trait Anxiety Inventory
	Symbol Digit Modalities
	University of Pennsylvania Smell ID Test
Subject Characteristics	Family History PD
	Patient Status
	Screening Demographics
	Socio-Economics

Table 4. Details of Munged and Aggregated date

Category	**Number of Files**
Biospecimen	11
Imaging	10
Medical History	14
Motor	11
Non-Motor	16
Subject Characteristics	4
Total	***66***

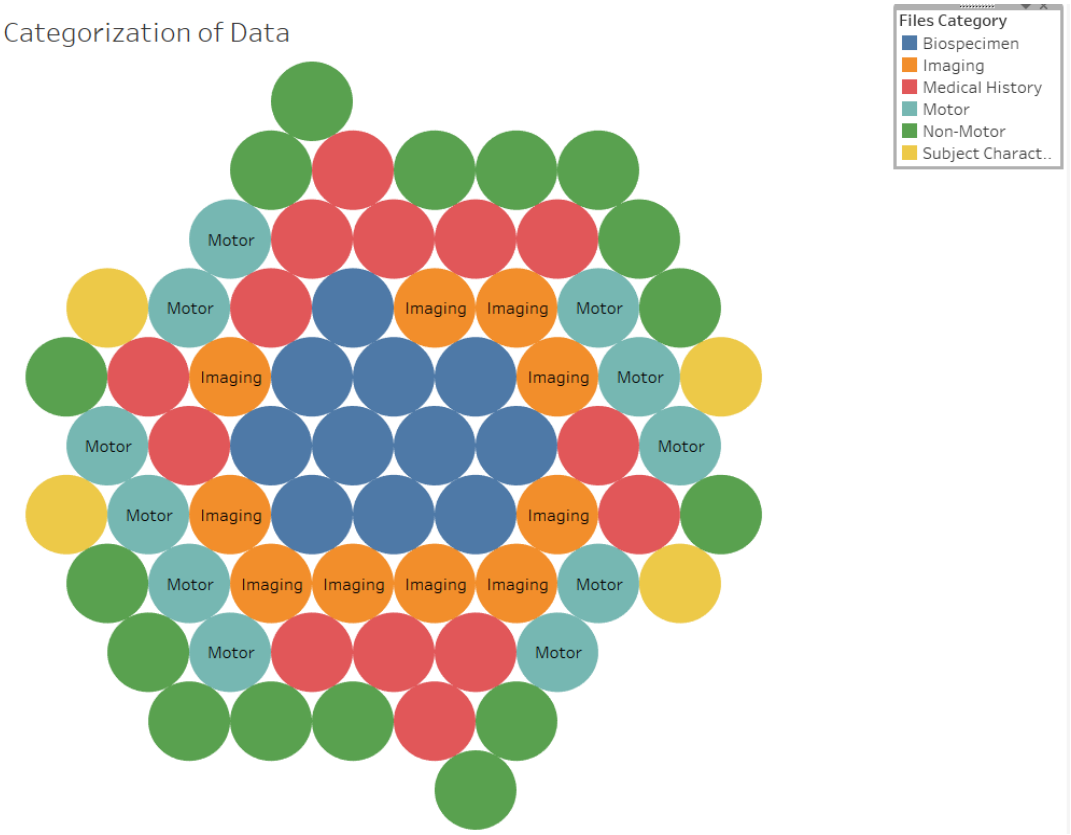

Figure 5. Categorization of Cleaned Data Files

Data Preparation

Cleaning and curating the data is the biggest challenge. Each file is analyzed in detail to remove the redundant and administrative data. Some files such as iUSM Biospecimen Cell Catalog, iUSM Catalog, MRI Imaging data, Olfactory UPSIT and SPECT Scan Information Source had the patient number listed as subject ID or specimen ID. Time and date attribute was causing a lot of duplication and aggregating them created a large sparse dataset. Aggregation process excludes the time and date attributes. Administrative data such as record id is ignored. The final aggregated dataset had 300543 row and 997 attributes. Pandas, a python package was used for the data aggregation. Pandas data frame is a two-dimensional data structure that helps to work with "relational" or "labeled" data. Loading this aggregated dataset is not memory

efficient. The aggregated dataset has 60% of categorical data. All medical studies are flooded with categorical and image data.

Categorical data are ubiquitous in the medical world. However, many powerful algorithms for numerical data do not work well on the categorical data. Having too many categorical data is a bottleneck in applying data mining techniques. Hence for our study, categorical data were converted to numerical data. The data preprocessing libraries in scikit learn in python was used to handle categorical data. One hot encoding feature of the preprocessing library was used to transform m possible categorical features to m binary features.

Imputation

PPMI data is curated for the biomedical research of Parkinson's disease, a large number of demographic and biomedical variables are collected, and missing values is inevitable in the data collection process. When data is harmonized and aggregated, 70 % of data is missing.

Missing data is a problem in data analysis. For the downstream statistical and data mining methods which require complete data matrix, imputation is a common and practical solution.

There are several software packages available to handle the missing data. Multicollinearity is a hindrance for imputing data. There are several packages in R which is used to handle missing data. Multivariate multiple imputations can be done using MICE package, MI, and Amelia. Since our curated dataset had several hundreds of feature with missing values, many of the packages were unable to impute it.

kNN, unsupervised machine learning method is very effective in missing value imputation. kNN based imputation is very intuitive and easily interpretable method used to impute in clinical trial data. For a missing value, this method seeks its K nearest variables or

subjects and imputes a weighted average of the observed values of the identified neighbors. In our study, VIM package in R was used to perform the kNN based imputation for a k value of 10.

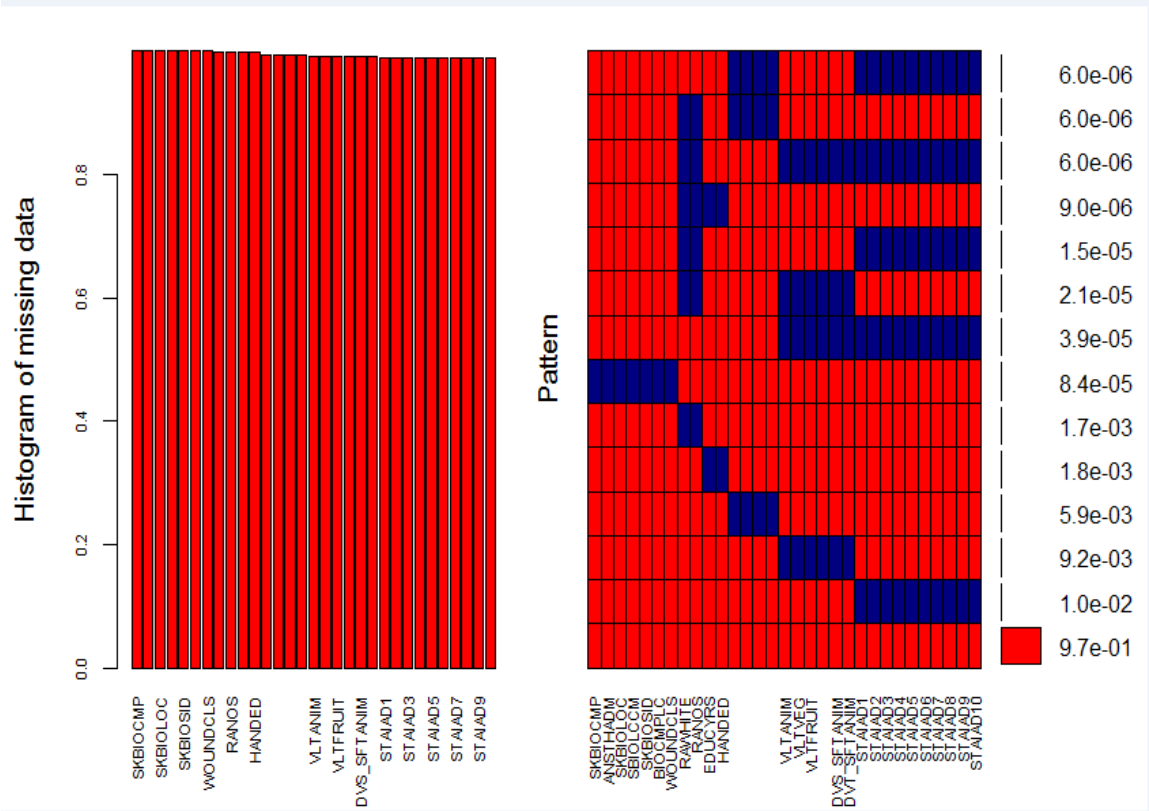

Figure 6. Histogram of Missing Values in the Aggregated Data Set

The computation time for imputing in the aggregated dataset was quite fast compared to imputation on the entire dataset with 300,000 rows with 70 % missing data.

CHAPTER 5. DATA ANALYTICS

Preliminary Exploratory Analysis

The main characteristics of PPMI dataset are summarized and visualized by initial exploratory data analysis (EDA). The average value of all the attributes for study subjects with Parkinson's disease and healthy control was computed. Bar graph represents the difference between the averages of PD and HC. The visualization helps us understand the critical features that contribute to discern Parkinson's patient from the healthy people.

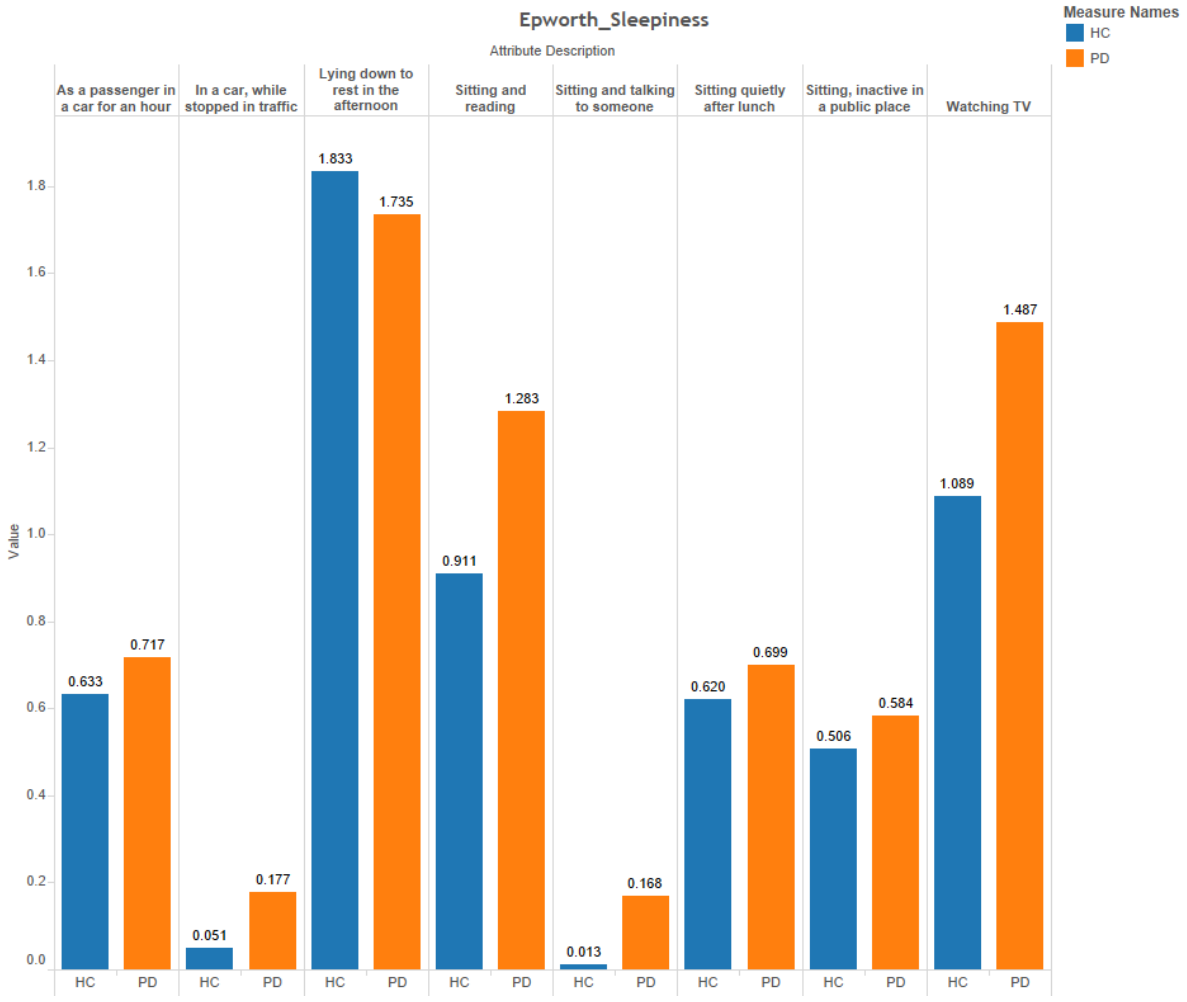

Figure 7. Comparison of Average HC and PD values

Assumptions in Data Aggregation

Following are the assumptions in the data aggregation process.

- Elimination of time and date attributes.
- Deduplication of attributes.
- Drop the administrative files from the data aggregation process.
- Remove constant attributes (do not have different values) from the aggregated data.
- Once the data from different cleaned files are merged, they are grouped by patient number using mean.
- Only Parkinson Disease (PD), Healthy Control (HC) and PRODROMA cohorts are considered for analysis.
- PRODROMA is considered as a Parkinson patient with the target variable "1".
- The target variable in the aggregated dataset is '1' for Parkinson's and '0' for Healthy Control.
- The shape of the final aggregated dataset is 594 x 759.

Principal Component Analysis

759 is a large number of variables; the matrix will be extensive to study and interpret as well. There are numerous pairwise correlations between the variables to consider. PPMI dataset is a perfect example of the curse of dimensionality. It is necessary to reduce the number of variables to a few, interpretable linear combinations of the data inorder to understand the data in a better way. Each linear combination will correspond to a principal component.

The result of principal component analysis (PCA) is a set of linearly uncorrelated variables represented by the principal components. Principal Component Analysis is a statistical procedure of orthogonal transformation to convert a set of correlated variables into the principal components. It is a dimensionality reduction technique. PCA is an unsupervised machine learning technique as it does not need marked labels to train the model. The number of principal components is always less than or equal to an original number of attributes. The largest possible variance is explained by the first principal component. The succeeding principal components explain the maximum variance possible under the constraint that it is orthogonal to the preceding component. PCA transforms the data and attempts to find out what features explain the most variance in the data.

Pros of Principal Component Analysis

- Dimensionality reduction and help us visualize our intuitions about the data.
- PCA aids in estimating probabilities in high dimensional data. Classification algorithms work better on Principal Components as the components are uncorrelated.

Cons of Principal Component Analysis

- Interpreting the components and tracing back to original features is difficult.
- Too expensive as it is cubic complexity.
- PCA does not work with fine-grained classes.

Python, scikit learn package was used to perform the PCA.

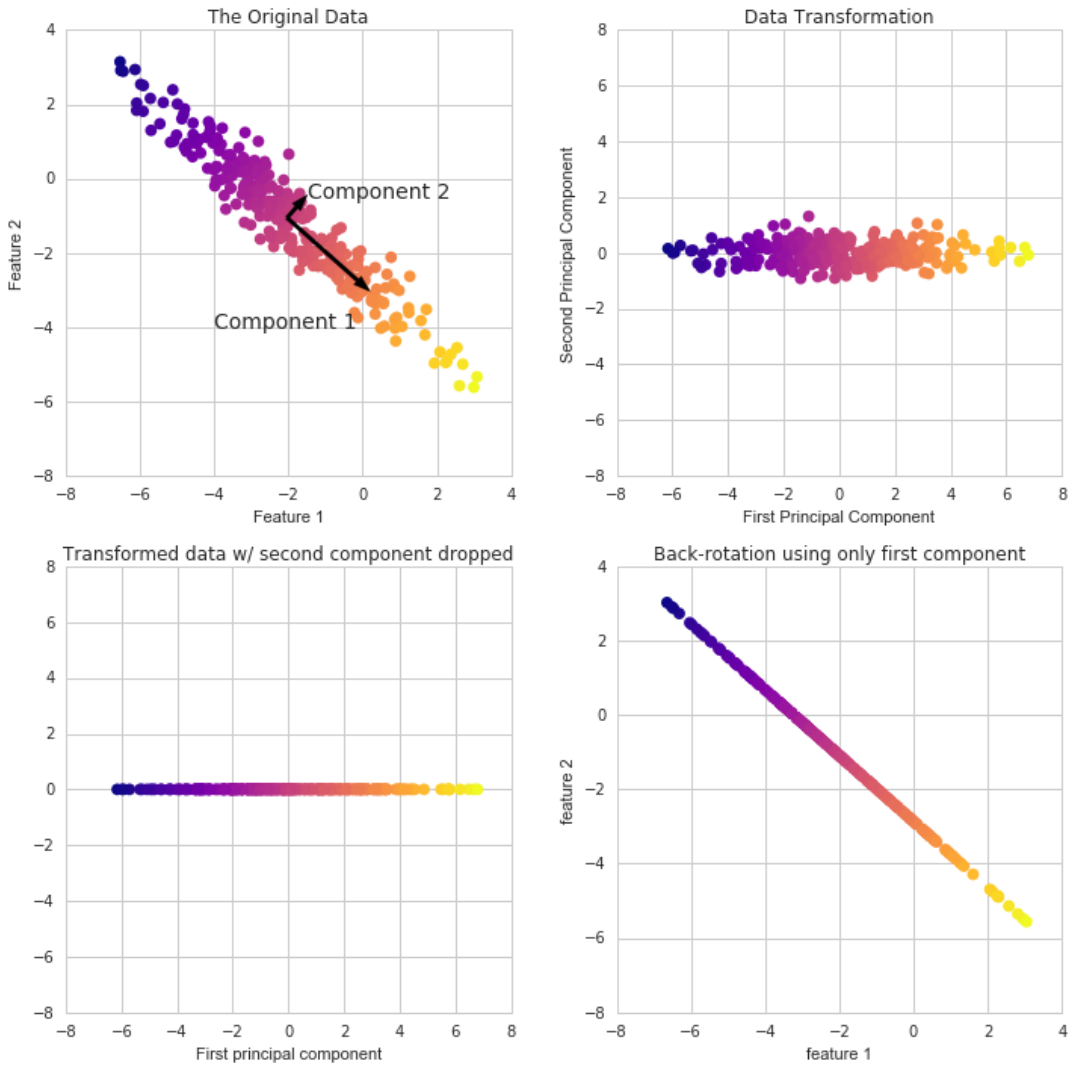

Figure 8. Transformation of Data Using PCA

Eigenvalues calculations obtain the number of principal components to consider in our algorithm.

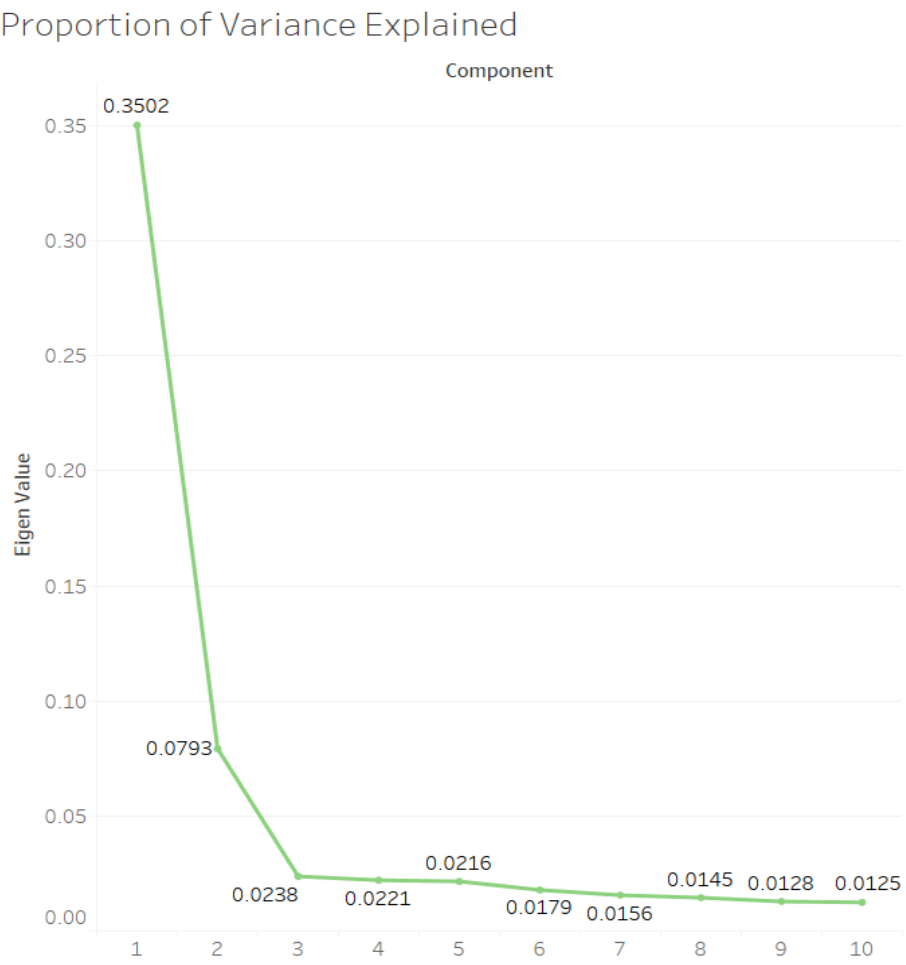

Figure 9. The Scree Plot of Covariance Matrix

In our data, 2 Principal components explain about 75% of the variance. 3 Principal Components explain 79% of the variance.

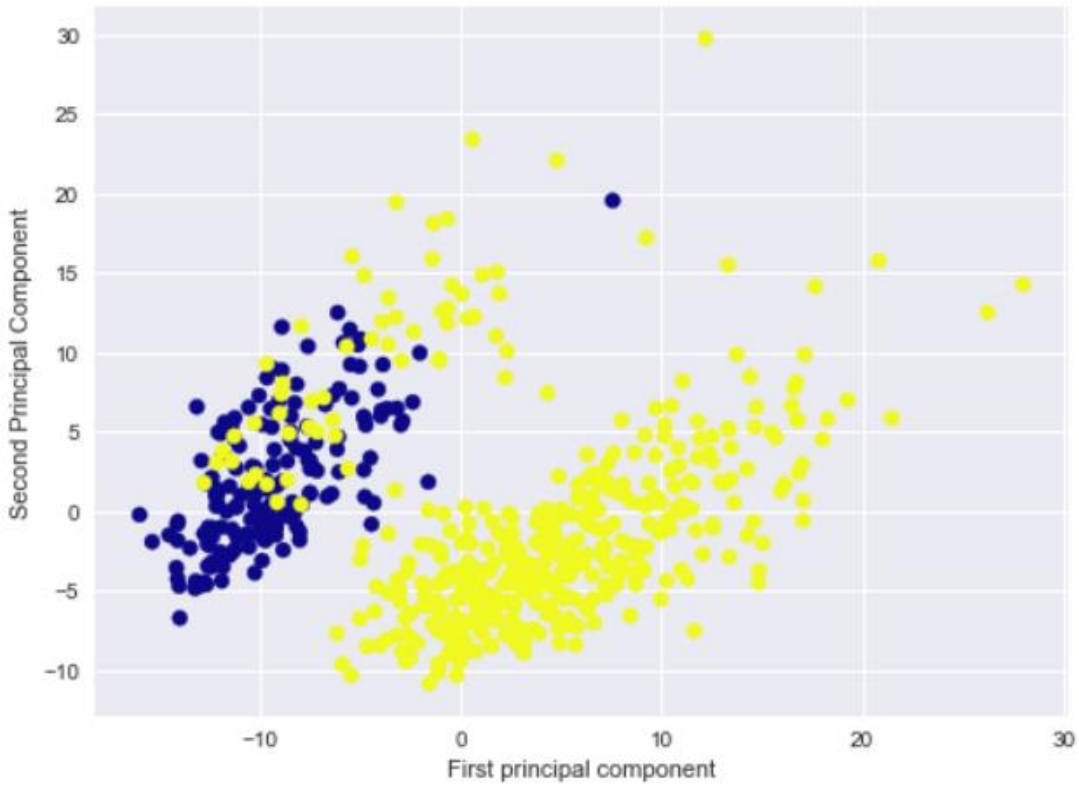

Figure 10. Plot of Principal Components

Principal Component Analysis was able to convert 759 dimensions to just two dimensions. There exists a clear pattern in the two dimensions that explains how the target variable is classified. Unfortunately, this great power of dimensionality reduction comes with the cost of being able to understand what these components represent easily. The components correspond to combinations of the original features. The components are stored as an attribute of the fitted PCA object in Python.

The heat map is used to visualize the relationship between the original feature and the principal component. This heat map and the color bar represent the correlation between the various features and the principal component itself.

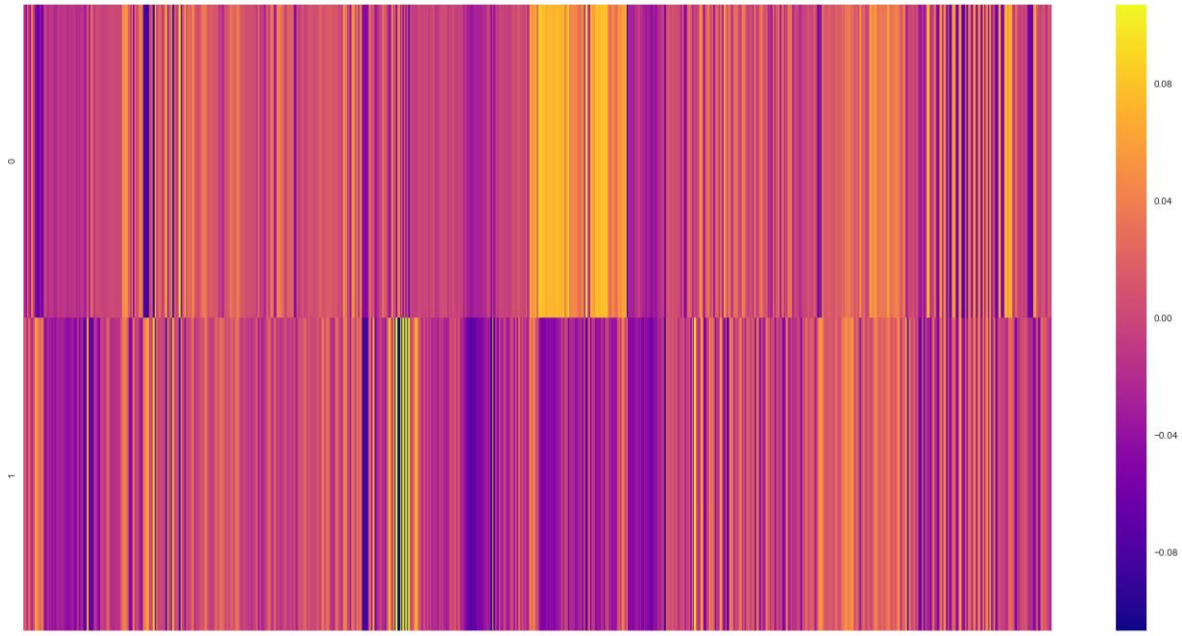

Figure 11. Correlation between various Features and the Principal Components

Since there are 759 original features, this heat map is too compressed and does not imply any meaning. So it is further required to drill down and split the heat map to get the list of variables that are highly correlated with each principal component.

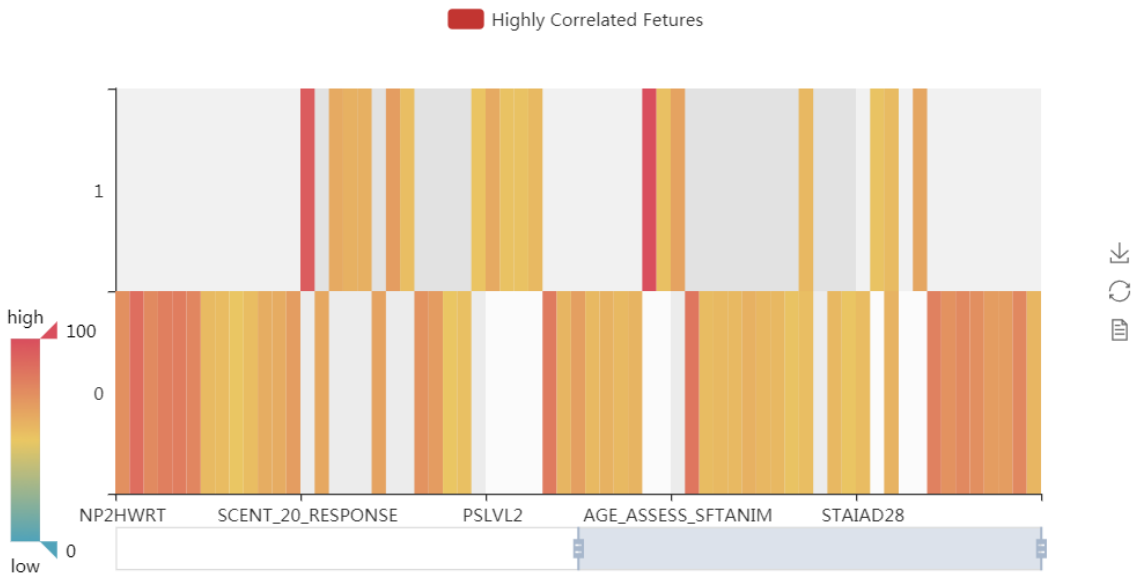

Figure 12. Interactive Heat Map

Pyechart library in Python is used to develop this interactive heat map. The list of important features in each component is determined. Pearson correlation between those critical features is calculated. Pearson correlation varies from -1 to 1. The closer ρ is to 1, the more increase in one variable associated with an increase in the other. On the other hand, the closer ρ is to -1, the increase in one variable would result in a decrease in the other.

Pairwise Correlation

	CELL_..	DFBRA..	DFPGD..	DFRIGI..	NHY	NP2DR..	NP2H..	NP2RI..	NP2TR..	NP2W..	NP3BR..	NP3FA..	NP3G..	NP3RI..	NP3SP..
CELL_TY..	1.000	-0.520	-0.400	-0.520	-0.500	-0.290	-0.240	-0.190	-0.390	-0.150	-0.410	-0.360	-0.250	-0.370	-0.270
DFBRAD..	-0.520	1.000	0.780	0.970	0.940	0.580	0.550	0.500	0.740	0.460	0.790	0.760	0.610	0.650	0.580
DFPGDIST	-0.400	0.780	1.000	0.780	0.750	0.540	0.490	0.430	0.610	0.380	0.670	0.690	0.660	0.530	0.550
DFRIGIDP	-0.520	0.970	0.780	1.000	0.920	0.580	0.530	0.480	0.730	0.440	0.790	0.750	0.590	0.670	0.550
NHY	-0.500	0.940	0.750	0.920	1.000	0.660	0.610	0.580	0.710	0.540	0.870	0.810	0.670	0.720	0.640
NP2DRES	-0.290	0.580	0.540	0.580	0.660	1.000	0.740	0.740	0.460	0.630	0.680	0.630	0.590	0.590	0.540
NP2HOBB	-0.240	0.550	0.490	0.530	0.610	0.740	1.000	0.710	0.560	0.660	0.610	0.540	0.540	0.480	0.480
NP2RISE	-0.190	0.500	0.430	0.480	0.580	0.740	0.710	1.000	0.410	0.730	0.610	0.560	0.590	0.480	0.500
NP2TRMR	-0.390	0.740	0.610	0.730	0.710	0.460	0.560	0.410	1.000	0.400	0.570	0.530	0.420	0.470	0.340
NP2WALK	-0.150	0.460	0.380	0.440	0.540	0.630	0.660	0.730	0.400	1.000	0.500	0.450	0.620	0.440	0.410
NP3BRA..	-0.410	0.790	0.670	0.790	0.870	0.680	0.610	0.610	0.570	0.500	1.000	0.850	0.720	0.730	0.710
NP3FACX..	-0.360	0.760	0.690	0.750	0.810	0.630	0.540	0.560	0.530	0.450	0.850	1.000	0.680	0.650	0.770
NP3GAIT	-0.250	0.610	0.660	0.590	0.670	0.590	0.540	0.590	0.420	0.620	0.720	0.680	1.000	0.570	0.540
NP3RIGLU	-0.370	0.650	0.530	0.670	0.720	0.590	0.480	0.480	0.470	0.440	0.730	0.650	0.570	1.000	0.480
NP3SPCH	-0.270	0.580	0.550	0.550	0.640	0.540	0.480	0.500	0.340	0.410	0.710	0.770	0.540	0.480	1.000
NUM_AV..	0.910	-0.410	-0.330	-0.410	-0.390	-0.240	-0.190	-0.160	-0.310	-0.120	-0.330	-0.290	-0.210	-0.290	-0.230
PARKISM	-0.470	0.880	0.700	0.860	0.850	0.500	0.490	0.430	0.670	0.400	0.710	0.660	0.560	0.560	0.540
PASSAG..	0.990	-0.520	-0.400	-0.520	-0.500	-0.290	-0.240	-0.190	-0.400	-0.150	-0.410	-0.370	-0.250	-0.370	-0.280
PLURIPO..	0.990	-0.510	-0.390	-0.510	-0.490	-0.290	-0.240	-0.190	-0.390	-0.150	-0.400	-0.360	-0.250	-0.370	-0.270
QMAT_O..	-0.290	0.420	0.370	0.420	0.460	0.410	0.410	0.370	0.370	0.370	0.490	0.450	0.410	0.420	0.420
SCENT_2..	0.480	-0.270	-0.190	-0.280	-0.250	-0.080	0.020	-0.010	-0.200	0.050	-0.190	-0.180	-0.060	-0.160	-0.060
VSINTRPT	0.520	-0.960	-0.770	-0.950	-0.920	-0.580	-0.540	-0.480	-0.730	-0.450	-0.780	-0.730	-0.600	-0.640	-0.570

Figure 13. Pairwise Correlation of the Important Features

Supervised Algorithms on the Reduced Dimension's Dataset

The original dataset of 759 features is reduced to 2 features by principal component analysis. The two principal components were powerful and explained the maximum variance in the dataset. The next step was to feed in the reduced component dataset into a classification algorithm. Logistic regression and support vector machine algorithms were run on the compressed component dataset to build a predictive model.

The performance of the supervised machine learning algorithms on the reduced dataset is illustrated below. The precision is intuitively the ability of the classifier not to label a Parkinson's patient as a healthy person. Recall score is the ability of the classification model to find all the Parkinson's patient.

Table 5. Performance of the Supervised Algorithms

Support Vector Machine

	Precision	Recall	f1-score	Support
0	0.81	0.85	0.83	46
1	0.95	0.93	0.94	133
average/total	0.91	0.91	0.91	179

Logistic Regression

	Precision	Recall	f1-score	Support
0	0.8	0.72	0.76	46
1	0.91	0.94	0.92	133
average/total	0.88	0.88	0.88	179

Both the supervised algorithm performed well and can predict the Parkinson's patient (our target variable) with the principal components as the independent variables. The computation of the model on the reduced dataset is fast. Scikit learn library in Python was used to build the predictive models.

CHAPTER 6. VISUALIZATION

There were 2600 features in the initial dataset, and our intuition that not all of them are equally important was correct. After all the data processing, aggregation, imputation, and dimensionality reduction, the focus is now only the highly correlated features in the principal components.

The correlation between different attributes was illustrated using D3.js. D3 (Data-Driven Documents) is a JavaScript library for producing dynamic, interactive data visualizations in web browsers. The output Circos visualization is an interactive tool with which we can visualize the relationship between one attribute and all other attributes, displaying the strength of the correlation as a measure of the width of the flare.

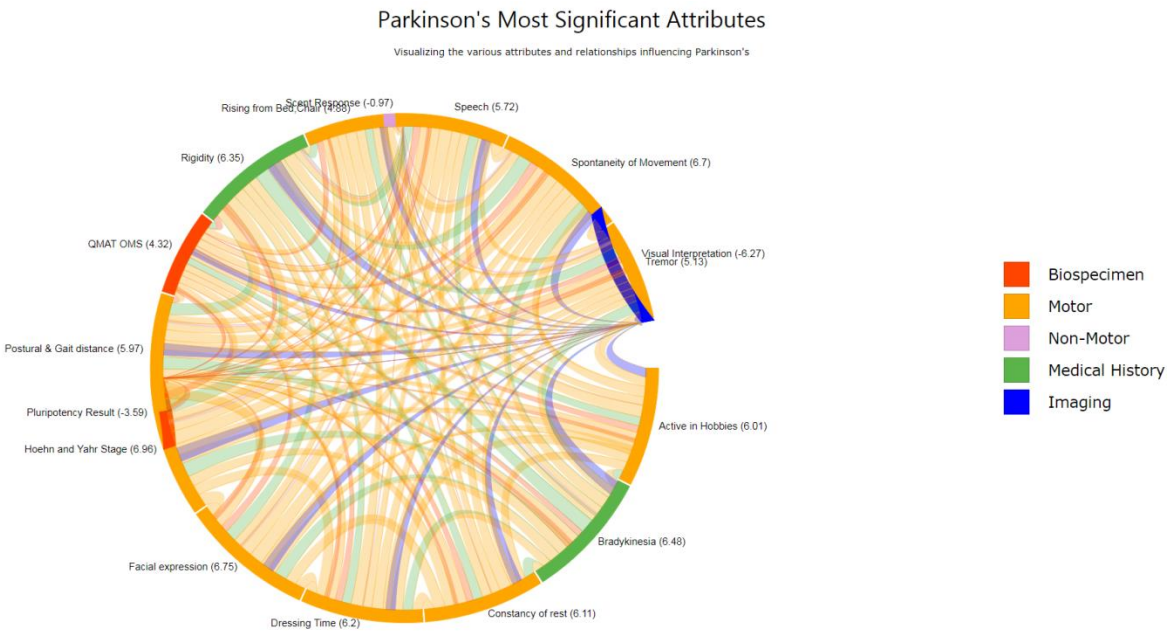

Figure 14. Parkinson's Most Significant Attributes

The above figure has been embedded in the thesis using Layar. Layar is a mobile browser based on augmented reality. When the picture is scanned using “ar” app, an icon of interaction appears on the top right corner. Clicking the icon enables the interactive content. A video demonstrating the interaction between the attributes plays automatically.

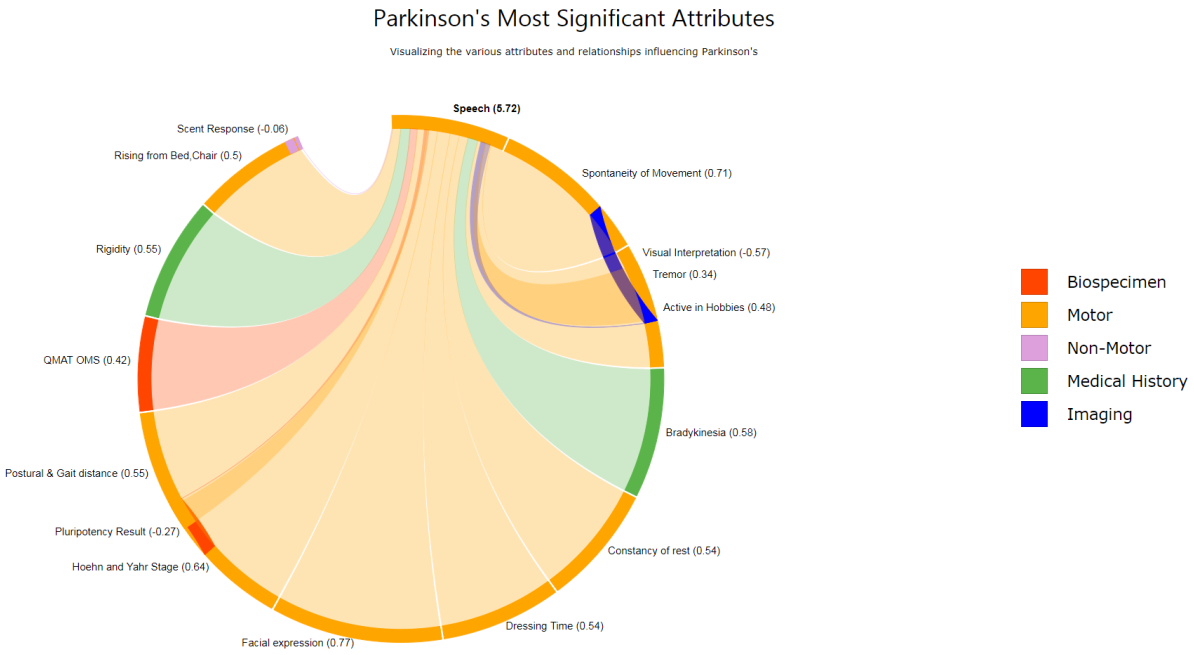

Figure 15. Interactions between Significant Attributes

CHAPTER 7. SUMMARY AND CONCLUSION

Summary of the Result

Data analysis procedure described in the study lead to the identification of the below mentioned important attributes in Parkinson's determination. Pluripotency result and QMAT OMS were the critical attributes of Biospecimen category that help in determining a Parkinson's onset. Freezing of gait, Responses to various scents were the crucial attributes of the Non-Motor class of Parkinson's disease.

All the primary motor symptoms of Parkinson's disease: tremor, rigidity, bradykinesia (slow movement), postural instability (balance problems), and walking/gait problems played a pivotal role in the principal components.

Table 6. Summary of the Result

Feature Description	***File Name***	***Category***
Pluripotency Result	iSUM Cell Catalog	Biospecimen
Hoehn and Yahr Stage	UPDRS Part III	Motor
Scent Response	Olfactory UPSIT	Non-Motor
Bradykinesia	Diagnostic Features	Medical History
Rigidity	Diagnostic Features	Medical History
Spontaneity of Movement	UPDRS Part III	Motor
Facial expression	UPDRS Part III	Motor
Active in Hobbies	UPDRS Part II: Patient Questionnaire	Motor
Visual interpretation	DaTSCAN Imaging	Imaging
Dressing Time	UPDRS Part II: Patient Questionnaire	Motor
Freezing of gait	REM Sleep Disorder Questionnaire	Non-Motor
Rising from Bed, Chair	UPDRS Part II: Patient Questionnaire	Motor
Tremor	UPDRS Part II: Patient Questionnaire	Motor
Speech	UPDRS Part III	Motor
QMAT OMS	iPSC Blood Sample	Biospecimen
Constancy of rest	UPDRS Part III	Motor
Postural and Gait distance	Clinical Diagnosis and Management	Medical History

Conclusion

Currently, data from clinical and behavioral studies of Parkinson's disease are proliferating and with little knowledge or coordination of attributes collected. Understanding the importance of each attribute collection to Parkinson's disease detection and treatment is essential.

Various tests performed to diagnose a disease cause a significant amount of stress. This study helps us understand the important attributes to be collected for diagnosing Parkinson rather than collecting several thousand feature data. The next step after a patient has tremor, rigidity, bradykinesia (slow movement), postural instability (balance problems), and walking/gait problems, is to get UPDRS Part II questionnaire, iPSC blood sample, iSUM cell catalog tests. DATScan imaging can be conducted to confirm our prediction of Parkinson's.

Data collection, management, and data quality are the most significant challenges. This study was done under certain assumptions in data cleaning and curation. As expected by the physicians the key findings of the study indicate that motor symptoms dominate and are the critical features in Parkinson's disease.

Limitation and Future Studies

This study has defined assumptions in data preprocessing, and only three cohorts are considered for the data analysis. The number of participants is less than the number of attributes collected in this study, which makes the data wide. An increase in study participants and data points can pave the way for further in-depth analysis. This study concentrated only on the correlation between attributes and correlation does not mean causation. Future work can be extended by allowing researchers to combine additional characteristics from image data and to discover the biomarkers. Another possible future direction is to look at the process improvement

and data management for clinical studies. Panel study can be an extension of the current research since the original dataset from PPMI has time and date variables.

REFERENCES

Alzheimer's disease facts and figures. (2013) *Alzheimer's & dementia,* 9(2):208-245.

Amiri, S., Clarke, B., Clarke, J. (2017) Clustering categorical data via ensembling dissimilarity matrices. *Journal of Computational and Graphical Statistics*

Beilina, A., Cookson, M.R. (2015) Genes associated with Parkinson's disease: regulation of autophagy and beyond. *Journal of Neurochemistry*

Bind, S., Tiwari, A.K., Sahani, A.K. (2015) A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction, *International Journal of Computer Science and Information Technologies'*. 6 (2): 1648-1655.

Cheong K.L., Song H.J., Park C.Y., et al. (2014) Biomarker discovery and data visualization tool for ovarian cancer screening. *International Journal of Bio-Science and Bio-Technology*,6 (2):169-178.

Clarke, R., Ressom, H.W., Wang, A., et al. (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37–49.

Dinov, I., Van Horn, J.D., Lozev, K.M., et al. (2009) Efficient, distributed and interactive neuroimaging data analysis using LONI pipeline. *Frontiers of Neuroinformatics,* 3(22):1–10.

Dinov, I.D., Petrosyan, P., Liu, Z., et al. (2014) The perfect neuroimaging-genetics-computation storm: the collision of petabytes of data, millions of hardware devices and thousands of software tools. *Brain Imaging and Behavior*, 8(2):311–22.

Hazan, H., Hilu, D., Manevitz, L., et al. (2014) Early diagnosis of Parkinson's disease via machine learning on speech data. *Software Science, Technology, and Engineering*.

Kowal, S.L., Dall, T.M., Chakrabarti, R., Storm, M.V., Jain, A. (2013) The current and projected economic burden of Parkinson's disease in the United States. *Movement Disorders*, 28 (3):311–8.

Maciejewski. R., Hafen, R., Rudolph, S., et al. (2011) Forecasting hotspots - A predictive analytics approach. *Visualization and Computer Graphics*, 17(4):440–53.

Nalls, M.A., McLean, C.Y., Rick, J., et al. (2015) Diagnosis of Parkinson's disease by clinical and genetic classification: population based modeling study. *The Lancet Neurology*.

Ramentol, E., Caballero, Y., Bello, R., Herrera, F. (2012) SMOTE-RSB*: a hybrid preprocessing an approach based on oversampling and under-sampling for high imbalanced data-sets Using SMOTE and rough sets theory. *Knowledge and Information Systems,* 33(2):245–65.

Rustempasic, Indira, & Can, M. (2013). Diagnosis of Parkinson's Disease using Fuzzy C-Means Clustering and Pattern Recognition.*South East Europe Journal of Soft Computing,* 2(1).

Wu, X., Kumar, V., Quinlan, J.R., et al. (2008) Top 10 algorithms in data mining. *Knowledge and Information Systems,* 14 (1):1– 37.

APPENDIX - PARKINSON'S DATASETS, PROGRAM FILES AND HEAT MAPS

Data Dictionary

The data dictionary of the initial set of 2600 attributes.

Data_Dictionary.xlsx

Original Dataset

PPMI data - 0523.zip

Cleaned and Curated Dataset

Curated Dataset - 0524.zip

Program Files

Program.zip

Interactive Heat Map

Interactive Heat Map.html

Parkinson's Most Significant Attributes

Parkinson Most Significant Attributes.html