**Assisting Physician Diagnosis of Parkinson’s: Predicting the Disease by Avoiding Subjectivity**

Dean D’Souza

Department of Data Science, University of Wisconsin – Green Bay

DS785: Capstone

Dr. Tracy Bibelnieks

April 28, 2024

**Abstract**

This case study aims to predict Parkinson’s disease using curated data from the Micheal J. Fox Foundation. The goal is to assist physicians by creating a reliable and reproducible tool through machine learning models. Unique to this paper, commonly used subjective predictors like the UDPRS are removed in favor of non-biased objective metrics. Data consisted of demographic, clinical, biological, and genetic subgroups. Validation through training and test splits reveal Random Forest as the best classifier of the disease. A model accuracy of 85% beats current estimations of physician diagnosis. Model predictors are relatively accessible and support the function of a rudimentary initial disease screening. Caudate region-related variables from DaTscan data heavily influence model prediction while BMP derivatives offer further research opportunities. Steps forward involve medical integration and advisement from diagnosing physicians. Alternative designs consider time-series modeling.

Keywords: PPMI, Parkinson’s, machine learning, binary classification

**Table of Contents**

[List of Tables 6](#_Toc165044808)

[List of Figures 7](#_Toc165044809)

[Assisting Physician Diagnosis of Parkinson’s: Predicting the Disease by Avoiding Subjectivity 8](#_Toc165044810)

[Problem Background 9](#_Toc165044811)

[Overall Project Goal 10](#_Toc165044812)

[Case Study Purpose 10](#_Toc165044813)

[Project Objectives and Business Significance 10](#_Toc165044814)

[Definition of Terms 11](#_Toc165044815)

[Assumptions and Limitations 13](#_Toc165044816)

[Conclusion 14](#_Toc165044817)

[Chapter 2: Literature Review 16](#_Toc165044818)

[Introduction 16](#_Toc165044819)

[PPMI Studies 16](#_Toc165044820)

[Previous Methodologies 18](#_Toc165044821)

[General Frameworks 18](#_Toc165044822)

[Conclusion 20](#_Toc165044823)

[Chapter 3: Methodology 22](#_Toc165044824)

[Introduction 22](#_Toc165044825)

[Overview of Methodology 22](#_Toc165044826)

[Data Modality 22](#_Toc165044827)

[Data Filtering 23](#_Toc165044828)

[Missing Data 23](#_Toc165044829)

[Feature Selection 24](#_Toc165044830)

[Class Imbalance 26](#_Toc165044831)

[Model Algorithms 27](#_Toc165044832)

[Data Collection 27](#_Toc165044833)

[Data Preparation 28](#_Toc165044834)

[Data Loading and Variable Removal 28](#_Toc165044835)

[Data Cleaning 28](#_Toc165044836)

[Missing Data 28](#_Toc165044837)

[Multicollinearity 29](#_Toc165044838)

[Variance 30](#_Toc165044839)

[Data Training/Test Split and SMOTE 31](#_Toc165044840)

[Model Generation and Deployment 32](#_Toc165044841)

[Conclusion 32](#_Toc165044842)

[Chapter 4: Results 34](#_Toc165044843)

[Introduction 34](#_Toc165044844)

[Results 34](#_Toc165044845)

[Best Model 36](#_Toc165044846)

[Feature Importance 37](#_Toc165044847)

[Top Predictors Discussion 38](#_Toc165044848)

[Project Objectives 40](#_Toc165044849)

[Physician Accuracy 40](#_Toc165044850)

[Non-Subjective Initial Screening Tool 41](#_Toc165044851)

[Confirm Possible Biomarkers 41](#_Toc165044852)

[Conclusion 42](#_Toc165044853)

[Chapter 5: Discussion 43](#_Toc165044854)

[Introduction 43](#_Toc165044855)

[Summary of Approach 43](#_Toc165044856)

[Project Objective Conclusions 46](#_Toc165044857)

[Steps Forward 47](#_Toc165044858)

[Conclusion 49](#_Toc165044859)

[Conclusion 50](#_Toc165044860)

[References 51](#_Toc165044861)

[Appendix A Alternative Feature Selection 57](#_Toc165044862)

[Appendix B Code 58](#_Toc165044863)

# List of Tables

[Table 1*Definition of Key Terms* 11](#_Toc164872015)

[Table **2** *Near Zero Variance Variables* 30](#_Toc164872016)

[Table 3 *SMOTE Imbalance changes* 31](#_Toc164872017)

[Table 4 *Model Metrics* 34](#_Toc164872018)

# List of Figures

[Figure 1 *Histogram and Pattern of Missing Data* 29](#_Toc164966254)

[Figure 2 *Correlation Plot after Dropping Colinear Variables* 30](#_Toc164966255)

[Figure 3 *Graph of Model ROC’s with AUC Values* 34](#_Toc164966256)

[Figure 4 *Shapley Value Feature Importance on Random Forest Model* 37](#_Toc164966257)

[Figure 5 *Boxplot of Caudate Levels and Parkinson’s Family History* 40](#_Toc164966258)

# Assisting Physician Diagnosis of Parkinson’s: Predicting the Disease by Avoiding Subjectivity

Parkinson’s disease is a rare neurological illness affecting 1% of the population at age 60 (Hauser, 2023). Its key indicators are the reduction of dopamine in the brain and the presence of alpha-synuclein proteins, commonly referred to as Lewy bodies. A progressive and debilitating disease, Parkinson’s is expected to rise in prevalence and economic burden (Ohene, 2018). The projected cost of medical services, productivity, lost job opportunities, and more combine to reach almost $80 billion in 2037 from $52 billion in 2017 (Yang et al., 2020). Individuals with the disease suffer from a wide range of motor impairments, including tremors, reduced gait, and muscle stiffness. The disease has no known cure. Treatment involves dopamine-boosting medication such as levodopa or strength training through physical therapy. These remedies can reduce Parkinson’s symptoms in its early stages, making disease detection critical. Traditional blood tests cannot predict Parkinson’s. The search for potential biomarkers of the disease is ongoing. Manual diagnosis involves repeated meetings with physicians and various tests to rule out other disorders. A formal exam introduces expensive neurological tests and brain scans (Ayaz et al., 2023). A different approach necessitates affordable, reliable, and reproducible Parkinson’s predictions.

           Data Science, specifically machine learning methods, can assist in this disease detection process. Models created from various predictors can classify whether a subject falls into a healthy control or a person with Parkinson’s. The success of this classifier depends on performance metrics such as AUC/ROC. A successful model implementation would facilitate a cheaper, reliable, and reproducible tool to assist Parkinson’s predictions. This case study aims to supplement the model by removing subjective predictors commonly used to predict and assist physician diagnosis.

## Problem Background

           Aside from predicting Parkinson’s disease, this case study adds an additional element. It also tackles subjectivity in Parkinson’s diagnosis. Diagnosing physicians often rely on subjective tests such as the Unified Parkinson’s Disease Rating Scale (UDPRS) to aid in diagnosing and progressing the disease. This test is a questionnaire separated into motor and non-motor sections. While this has been beneficial in determining and evaluating Parkinson’s, it might not be as reliable in its progression. A high amount of error variance was found in the change in UDPRS scores, casting doubt on whether reliance on subjective tests is the way forward (Evers et al., 2019). Furthermore, this test is infrequent by design and, as a result, misses rapid changes in the patient. Depending on the medication and dosage, the test results can vary and must be considered. These are challenges physicians must navigate to ascertain an accurate diagnosis and progression of the disease.

           Including a machine learning implementation that avoids the UDPRS and other subjective tests has two applications and many benefits. The first use case functions as a non-subjective aid to assist physicians in their diagnosis. Relying on concrete data through non-clinical categories reduces possible bias from both the patient and the physician. The resulting model does not intend to replace physician input entirely but rather support or question their affirmations. The second use case is as an initial screening tool that secures a starting point for diagnosis. Due to the reproducibility of models and the focus on reliable metrics, both use cases could save time and money for the patient and the physician. Tests could be performed frequently without wait times or necessarily accounting for medication fluctuations and the need for subjective tests.

           This proof-of-concept machine learning model allows physicians to remove subjectivity from their decisions while automating aspects of Parkinson’s diagnosis. It could streamline the diagnostic process.

## Overall Project Goal

           This project aims to create a model that yields reliable and reproducible test results for physicians. If the model can accurately supplement physician diagnosis and display robustness through cross-validation, it can prove its reliability and replicability indifferent to the patient. By removing subjective metrics such as the UDPRS, the model may act as an initial screening tool. Furthermore, it may help discover or confirm influential Parkinson’s predictors besides the UDPRS.

## Case Study Purpose

This case study aims to assist physicians in diagnosing and treating Parkinson’s disease. A machine learning model’s cost, time, and savings benefit patients and physicians. Removing subjectivity ensures results dependent on collectible and reliable metrics. Defined objectives help elucidate the purpose and project goal.

## Project Objectives and Business Significance

First-time diagnoses of Parkinson’s disease by clinical physicians hover around 80% (Rizzo et al., 2016). The first objective of this study is to reach a reliable model accuracy above 70%. Equaling or surpassing 80% is the best outcome, directly competing with physician diagnosis. However, accuracy cannot be the sole determinant of a successful model classifier. When classes are imbalanced, such as in medical diagnosis, the area under the curve (AUC) of the receiver operating characteristic (ROC) gives a better evaluation. This metric considers the relationship between sensitivity and the false-positive rate to determine model performance. As such, a strong AUC coupled with a high accuracy, at least above 0.7, would confirm the model’s success. This would prove the model’s predictive ability and improve its standing as a tool to confirm diagnosis in the next objective.

           The second objective examines the model’s purpose as a non-subjective initial screening tool. The model can avoid biased data by removing most physician and patient input. Its feasibility as a screening tool somewhat depends on the predictors' obtainability. If most predictors are obtainable through common tests, that would be an additional bonus. Depending on the model's complexity, they could be direct inputs into the model’s diagnosis. Removing important subjective predictors and lowering model complexity is expected to hurt predictive power. As such, an accuracy lower than 80% might be acceptable if the tool offers this non-subjective input tradeoff.

           Finally, the third objective is to find or confirm influential predictors for classifying people with Parkinson’s disease. In this case, a predictor would have to strongly correlate with predicting Parkinson’s disease or be important in the model. This information could be important for further research or to assist physicians in focusing on specific predictors in their evaluations. It could also push researchers to reconsider lesser-known confirmed variables.

These objectives define the project not only as a method for predicting the disease but also for its applicability to physicians in a practical setting. It reinforces the overall project goal and case study purpose through measurable metrics and clear statements.

## Definition of Terms

Table one defines key terms presented throughout the paper.

**Table 1**Definition of Key Terms

| Term | Acronym | Definition |
| --- | --- | --- |
| Levodopa | - | Dopamine-boosting medicine given to Parkinson’s patients |
| Carbidopa | - | Agent that prolongs the delivery of levodopa to the brain |
| Caudate | - | Region in brain that processes movement, visuals, and memory |
| Receiver Operating Characteristic | ROC | Relationship between sensitivity and false positive rate |
| Area Under the Curve | AUC | Measure the model’s performance using the area under the ROC curve |
| Unified Parkinson’s Disease Rating Scale | UDPRS | Test to rate progression and initial onset of Parkinson’s disease |
| Dopamine | - | Chemical in the brain related to movement and indicator of neurological disorders such as Parkinson’s |
| Confusion matrix | - | Table that reports the number of true positives, false negatives, false positives, and true negatives |
| Parkinson’s Progression Markers Initiative | PPMI | Collaborative dataset of Parkinson’s patient data created by the Micheal J. Fox Foundation |
| Synthetic Minority Oversampling Technique | SMOTE | A technique that creates new samples from the minority class |
| K-fold Cross Validation | - | Method for evaluating models by splitting the data into k groups. |

## Assumptions and Limitations

           The UDPRS is often used in machine learning models to predict Parkinson’s disease (Gerraty et al., 2023). Furthermore, it can be an influential predictor of this disease. As such, the accuracy without this predictor is expected to be lower than in current research. Removing this variable in favor of reduced bias has its reliability limitations. However, as a tool designed to assist physicians, including such a test in the model with physician input would be circular. Physician input would directly impact the tool’s results through the subjective variables. For example, the UDPRS questionnaire has a numerical score scale that depends on the section. Choosing a high score could affect the model’s decision and confirm Parkinson’s, especially if the UDPRS score is a significant variable. With this intention in mind, this accuracy/bias tradeoff might be acceptable. This follows for other subjective metrics.

           There are data limitations in the PPMI. This study uses a curated dataset that is available upon request. However, other genetic, imaging, and x-ray data sources are not curated or require additional access. A more detailed study looking at a more significant number of predictors may provide more insights. This would require joining the PPMI dataset with their patient identifiers. Thus, patients need to be cross-listed or participate in the same studies. Furthermore, prior data cleaning performed by the PPMI maintainers of the curated data is assumed to be valid. Some of their decisions are presented in the data dictionary associated with the dataset.

Validation is also exclusive to the PPMI dataset. Studies mentioned in the Chapter 2 literature review utilize external validation as a more robust test. This offers a better test of models on unrelated data. Datasets like the Parkinson’s Disease Biomarker Program (PDBP) may provide this type of validation while maintaining primary patient identifiers across data. Likewise, a barrier to external validation is the need to stick within the PPMI ecosystem. Unique external tests would need to repeat the same tests on a new set of patients.

The extent of the work is limited to a project semester. Further work and steps forward in Chapter 5 offer continuations of the project. Particularly, time-related models better use the curated dataset’s multiple patient visit data and include more information for modeling.

## Conclusion

           The subsequent sections demonstrate previous attempts, model design and construction, and final findings. Overall, the model reached an acceptable accuracy and surpassed a clinical diagnosis of 80%. Its use case as a non-subjective tool without physician or patient input could be beneficial as a possible screening test or follow-up confirmation. Furthermore, the model confirmed influential predictors in predicting Parkinson’s disease. No newly unconfirmed predictors were significant. Some of the findings coincide with results from the literature review. The remaining chapter breakdown is as follows:

* Chapter 2—Literature Review: This section defines several previous attempts at modeling Parkinson’s disease diagnosis, as well as different types of classifiers, results, and limitations from previous findings.
* Chapter 3—Methodology: This section covers building the Parkinson’s prediction model. It encompasses data cleaning from the raw PPMI dataset, exploratory analysis, feature selection, and training and evaluating machine learning models.
* Chapter 4—Results: Final metrics are drawn from the methodological analysis. Model strength and validity are evaluated through accuracy and AUC. Correlated and important features are displayed through Shapley values. Project objectives are confirmed.
* Chapter 5—Discussion: This includes a summary of the design approach. Steps forward and possible improvements are discussed for future iterations.

# Chapter 2: Literature Review

## Introduction

           Parkinson’s disease is an ongoing research field spanning decades with many different obstacles to tackle. As such, research has involved experimentation, clinical trials, data analysis, and more. Regarding data science and machine learning, several attempts tried to predict Parkinson’s disease using vocal patterns, gait analysis, biometric data, novel blood tests, and existing data. For example, researchers who studied reduced gait built a model using data from motion capture cameras studying a patient's velocity, stride, and cadence (Ferreira et al., 2022). Meanwhile, vocal impairment, a key characteristic of people with Parkinson’s, can be studied to distinguish frequency variations classifying Parkinson’s from control. A popular dataset studying this includes the Max Little Phonation database, with models reaching very high AUC’s above 0.9 (Ayaz et al., 2023). Alternatively, others have turned to physical experiments to back up their models. Channa et al. (2021) developed wearable bracelets measuring hand and arm data. By training models testing for tremors or bradykinesia (slowed movement), they could diagnose patients with high accuracy. Others have tackled the problem of blood biomarkers. Diana et al. (2023) developed a neural network framework analyzing metabolic data from blood plasma to predict people with Parkinson’s successfully. While many fields mentioned here offer promise, this study applies machine learning and data science specifically for predicting Parkinson's using PPMI data.

## PPMI Studies

           A general review of machine learning within the PPMI by Gerraty et al. (2023) showcased the many methods applied to predict or reveal insights into the disease. The paper summarized results by examining other studies using the dataset that analyzed prediction diagnosis, symptoms, disease progression, dimensionality reduction, clustering, or a combination. Most papers implemented supervised classification or regression models, half of which attempted Parkinson’s diagnosis, making it the largest contributor. The review goes on to distinguish papers working on disease progression and symptom variability. However, the methods presented are insightful for general supervised learning within the PPMI. The most common models used were linear regression with regularization, support vector machines (SVM), random forests, and gradient boosting. Modality splits into clinical, neuroimaging, and genetic data. Clinical data comprises the motor and cognitive functions through tests such as the UDPRS and is the most used in models. This is followed by neuroimaging data through MRIs, diffusion tensor imaging (DTI), and DaTscan data. Genetic data such as DNA, RNA, or cerebrospinal fluid (CSF) are infrequent. The authors offer room for improvement for future research.

           Recommendations include integrating multimodal data into model building, improving validation methods, and using the longitudinal nature of PPMI data. Few models incorporated multimodal data. Leveraging underutilized predictors such as genetic data may be helpful for future models and robustness. Furthermore, there was sharp criticism of validation techniques. Thirteen of 97 studies had no validation and data leakage was common. Features selected through differentiating between patients with Parkinson’s and healthy controls were misused in model training and test data, increasing overfitting. This insufficient validation casts doubt on the replicability and authenticity of their models. Established validation techniques include Monte Carlo, K-fold cross-validation, or a separate validation set. Furthermore, the time-series, longitudinal, nature of PPMI data is underutilized. Only 26 studies made use of this disease progression in their models.

## Previous Methodologies

           Specific studies vary in their methodology and data preparation. Dadu et al. (2022) implemented unsupervised clustering via the Gaussian Mixture model (GMM) to break down Parkinson’s predictors into three subtypes: cognitive impairment, motor disturbance, and sleep disturbance. They then applied supervised learning of disease progression using Random Forest, LightGBM, and XGBoost to predict subtype progression and validated with 5-fold cross-validation and an external dataset. Meanwhile, Nalls et al. (2015) developed their models using stepwise logistic regression. After inspection, predictors with low Akiaike information criterion (AIC) were removed from the subsequent model. Notably, they combined data from the Pennsylvania Smell Identification Test (UPSIT), a strong identifier of neurodegeneration, with PPMI. Others worked with disease progression through differences in PPMI visits. Chahine et al. (2019) used the change from the first visit (baseline) to a one-year in their models. They selected predictors based on visit differences and built linear and non-linear time models.

## General Frameworks

           Three studies resemble the general framework used in this case study’s methodology. They vary in scope and complexity and contribute different types of datasets, preparation, feature engineering, and model selection and training. However, all include the goal of predicting Parkinson’s by using the PPMI.

Dou et al. (2022) worked exclusively within the PPMI dataset. In addition to predicting the disease, the authors also worked on predicting its progression. Their data preparation included dropping missing values, separating their predictors into clinical, CSF, and genetic groups, and tests for differences using Mann–Whitney–Wilcoxon and Chi-square. Backward stepwise logistic regression using a cutoff of P < 0.05 iteratively trimmed predictors from each group and combined into a final model. For example, after finding the best predictors from the clinical group, surviving predictors were added to the CSF group and repeated. Then, Spearman coefficients evaluated collinearity on the final model. CSF α-syn and a genetic variable contributed most to the model besides clinical predictors. Only internal validation was performed on the final model, characterized by calibration and optimism scores within acceptable ranges. This metric is an outlier among most papers mentioned here. The following studies utilize external validation. The authors mention the low-cost nature of the clinical model, citing practical applications of predicting Parkinson’s disease if needed.

Leger et al. (2020) attempted to classify Parkinson’s disease using non-motor clinical and biomarker predictors. They included analysis comparing Parkinson’s with scans without evidence of dopamine deficit (SWEDD), data outside this project's scope. Baseline data was filtered and the synthetic minority oversampling technique (SMOTE) handled imbalanced data. Classification accuracy was compared using logistic regression, general additive (GAM), decision tree, random forest, and XGBoost. AUC measured their performance along with sensitivity, specificity, accuracy, and Kappa static. Model-based feature elimination cut down predictors with low AUC. After feature elimination, a collinearity cutoff of 0.75 set the bar for the final model. This is to avoid issues with logistic regression coefficients. Instead of setting a defined value, the pROC package provided an optimized classification threshold. Models were cross-validated for performance, resulting in GAM and XGBoost with the highest AUCs. The previously mentioned smell test, UPSIT, was the most significant.

Finally, Makarious et al. (2022) used multimodal data across datasets and numerous models to predict Parkinson’s disease. They selected PPMI patients on their first baseline visit to better link predictors across similar datasets. An automated ML library specialized in genomics, GenoML, was chosen to work with genome sequencing (DNA sequence), transcriptomics (RNA transcripts), and clinical data. Furthermore, continuous features were normalized. Aside from the models mentioned in the previous study, this paper expanded to include adaptive boosting, gradient boosting, stochastic gradient descent, support vector machines, multi-layer perceptron neural networks, k-nearest neighbors, linear and quadratic discriminant analysis, and bagging. Likewise, AUC was maximized when deciding between models, but accuracy, sensitivity, and specificity were considered. Feature selection involved removing redundant features through a randomized trees classifier algorithm (extra trees). Hyper-parameter tuning and cross-validation revealed adaptive boosting as the model with the highest mean AUC of 0.9. External validation using the Parkinson’s Disease Biomarker Program (PDBP) also proved successful with an AUC of 0.85. Further class imbalance optimization improved accuracy using Youden’s J calculation. The Shapley Additive exPlanation (SHAP), a method to rank importance using game theory, found UPSIT again as the leading predictor for predicting Parkinson’s. Numerous highly ranked gene expressions showcased possible links to Parkinson’s, supporting future drug targets. The authors also caution against using this model to replace physician diagnosis, offering function as a screening tool instead.

## Conclusion

There are numerous approaches for working with PPMI data and different data preparation and modeling flavors. The general framework consists of data cleaning, feature engineering and selection, imbalance and collinearity checks, model building and validation, and predictor importance. From the overview, some form of validation (k-fold or external validation set) is a prerequisite to trustworthiness and replicability, and caution should be given to data leakage. Data modality includes three main areas: clinical, neuroimaging, and genetic. Many researchers combined PPMI with other datasets across these regions for more predictors, resulting in greater accuracy. Some standard algorithms include logistic regression, random forest, gradient boosting, and XGBoost. In relation to this study, the UDPRS was present in many studies modeling Parkinson’s disease. However, the UPSIT, a different type of multiple-choice subjective test, proved more conclusive as a significant predictor. No study removed both the UDPRS and UPSIT in their modeling.

# Chapter 3: Methodology

## Introduction

**As stated in previous sections, the main application of this case study is to assist physicians in diagnosing Parkinson’s disease without the need for entirely subjective tests like the UDPRS or UPSIT, predictors commonly used in the literature review for this purpose. Subsequently, the methodology in this section requires the removal of most subjective tests in the dataset. Predicting the disease is accomplished by building a machine-learning model. The methods applied here coincide with the summarized Chapter 1 project objectives:**

1. **The model accuracy must achieve at least 70%, aiming for a stated physician accuracy of 80%.**
2. **Without patient or physician input, examine the model’s ability as a tool or initial screening.**
3. **Find or confirm correlated predictors for predicting Parkinson’s or possible biomarkers.**

## Overview of Methodology

**This section overviews the reasoning behind the different techniques applied in this case study and the consideration of other tools and options.**

### Data Modality

**Referring back to the general review by Gerraty et al. (2023), a discussion of data modality using the PPMI dataset is the first design consideration. Data modality included primarily clinical, neuroimaging, and genetic data. One of their recommendations involved combining these categories in model building. However, clinical data comprises many subjective tests including the UDPRS and UPSIT. Per the case study’s objectives, most clinical data was removed from consideration. Exceptions were given to tests with binary, yes/no responses as they verge on static information instead of subjective answers. Aside from this concession, the dataset included demographic, neuroimaging, biologics, and genetic data. In essence, this study’s process focuses solely on empirical and demographic data to predict the disease.**

### Data Filtering

**The next design consideration involved the quantity of data used to predict Parkinson’s from the healthy control. The healthy controls are individuals subject to the same tests as Parkinson’s patients. While the PPMI includes several patient visit data, a simplistic approach following** Leger et al. (2020) in the literature review focused on only the first patient visit data, called the baseline (BL). This baseline approach offers advantages in model building and simplicity. Furthermore, data sparsity challenges increase with more patient visit data. Its major drawback is the loss of time-sensitive information, which may help better predict the disease. Collecting data on only the first visit may miss trends or misevaluate significant predictors over time. Ultimately, the baseline approach was chosen. The steps forward discuss an alternative time-based modeling approach.

### Missing Data

Missing data is a common problem with collected patient-visit data (Agiwal & Chaudhuri, 2024). This case study dropped most missing data from its rows to avoid bias and assumption handling with imputation methods. It also is a prerequisite format for model building involving predictions. Removal of empty values resulted in over 1000 rows dropped from the dataset. This follows the data preparation of Dou et al. (2022); however, many papers targeting specific subgroups (only clinical or biologic) did not encounter this problem since their data did not contain the depth of this curated dataset. The data simply holds many more categories and, thus, thousands of rows of extra data. Instead, alternative solutions to missing data could be considered for this specific dataset.

An approach with the imputation package Multivariate Imputation by Chained Equations (MICE) offered promise, but almost 50% of data in some columns would be imputed. The more imputed data, the higher the chance of bias in the transformed dataset. Model predictions relying on these imputations should proceed with caution. Furthermore, imputation assumptions such as missing completely at random (MCAR) would necessitate knowing how each data was recorded in PPMI (van Buuren, 2018). Finally, a test of MICE on the dataset revealed errors in logistic regression convergence, indicating some form of feature selection or regularization was needed to fix multicollinearity or perfect separation. Perfect separation results when a variable perfectly predicts the response, in this case, Parkinson’s diagnosis, with high accuracy. Since MICE can use logistic regression in its imputation, imputed values would run into this error. Changing the imputation model to bypass this increases the complexity. This is unnecessary and overcomplicated. Due to these obstacles, imputation was not used.

### Feature Selection

Feature selection is the process of reducing irrelevant predictors in model building. This step avoids noise from predictors, which can hurt a model’s predictions. It is also relevant for regression-like methods where extra coefficients can impact the accuracy or overfit the data. The first type of feature selection through predictor removal comes from inspection and the project objectives. Subjective tests and variables unrelated to Parkinson’s are not helpful for this case study. The next option considered automated types of feature selection. From the literature review, Makarious et al. (2022) implemented an extraTrees classifier algorithm to reduce predictors. Due to the nature of tree-based algorithms, predictors are naturally trimmed down when model building and feature selection is performed through modeling. As a result, this case study experimented with a similar process through the *Boruta* package for random forest feature selection. However, the end model yielded only seven predictors (see Appendix A for *Boruta* variable importance graph). While this was helpful for logistic regression, it was unnecessary for other algorithms like xgboost and might remove too much information other algorithms can handle. Therefore, a simpler approach to feature selection was employed. Automated feature selection was not implemented.

Another option previously mentioned by Dou et al. (2022) is a form of stepwise logistic regression. This is where predictors are removed based on the significance of their coefficients depending on the type of step. The study from the literature review used backward stepwise elimination that starts with all possible variables. At each step, one variable is dropped based on the least significance until the final model contains only significant variables. This might have been satisfactory for the purposes of their study; however, there has been concern that stepwise forms of feature selection carry certain faults. A study on stepwise regression by Smith (2018) criticized the method, particularly for cases with many explanatory variables, such as in big data. Non-significant variables dropped by stepwise can still have causal effects on predictions. Furthermore, other variables may be coincidently significant but do not come up as important individually. The paper explains that this hurts out-of-sample model predictions. While many predictors were cut from this case study’s initial curated dataset, a large number remained for modeling. Therefore, including stepwise regression on a large number of predictors could have unnecessarily removed information.

A better method is to use penalized regularization to remove predictors (Flom, 2018). This could be lasso, elastic net, or ridge regression. This was a viable approach. Ultimately, this case study implemented regularization only on logistic regression in model building. Like with the automated feature selection through the *Boruta* package, this was done to keep as many variables as possible for other algorithms that do not necessarily require the reduction of predictors as much as regression methods.

Instead of a model-based selection approach with random forest, stepwise methods, or penalized regularization, predictors were cut down based only on multicollinearity and variance. This removed far fewer predictors while also helping logistic regression. Multicollinearity hurts regression because the predictors should be independent. Ignoring this leads to sensitive coefficients and untrustworthy p-values (Frost, 2017). A correlation matrix and set correlation coefficient cutoff determined which predictors would be cut. Also, predictors with near-zero variance were removed. Predictors with near zero variance have few unique values and are not significant. These methods were the limit of feature selection in the case study.

### Class Imbalance

Another challenge with medical diagnosis data is the inherent imbalance between the two classes. In this case, since it is a Parkinsons’s dataset, there is a heavy imbalance in favor of PD patients versus healthy control. Predictions made without accounting for this imbalance may lead to inaccurate results. If a large proportion of the data is already Parkinson’s patients, then the models must prove very little to show the correct result (Wu, 2022). A model predicting all PD patients would still be accurate. Instead, the SMOTE technique Leger et al. (2020) implemented in the literature review balances out the binary classification. SMOTE is an oversampling technique that synthesizes minority example data using K Nearest Neighbors (Awan, 2022). It has shown promise in tackling a wide variety of data dimensionality and is a popular method for class imbalance (Blagus & Lusa, 2013). Other options include oversampling or relying on classification matrix values to interpret results. It is important to note that SMOTE applies to the training data; application to the entire dataset introduces bias and invalidates results in the subsequent testing dataset.

### Model Algorithms

Model algorithms were chosen based on popularity in the literature review but also due to the variety of methods covered including regression, trees, and boosting models. It generally combines linear and non-linear models. The chosen models were logistic regression with regularization, linear discriminant analysis (lda), random forest, support vector machine with radial kernel (svm), k nearest neighbors (knn), and extreme gradient boosting (xgboost). Models were tested on the training data using five-fold cross-validation. The AUC metric helped compare models due to the initial imbalanced nature of the data and its popularity in previous papers. Accuracy was retained to compare to the project objective of 80% from physician diagnosis. Predictors were normalized to help model performance and correct estimate coefficients from logistic regression. Default parameters were chosen and iterated upon using the *caret* package.

## Data Collection

**The dataset used in this case study comes from the Michael J. Fox Foundation’s PPMI cohort. Its contributors synthesized a curated dataset combining many of the common datasets related to Parkinson’s. It was published on June 12, 2023, with the latest update to the data dictionary coming on December 12, 2023. It is grouped by patient identification visits and consists of derived and preprocessed variables. As a result, less data cleaning is necessary to analyze the data. The initial dataset has a total of 10,152 observations with 155 variables. This dataset and surrounding data are available upon request to the foundation.**

## Data Preparation

**This section details the steps to process the data before model implementation and expands upon the methods in the overview.**

### Data Loading and Variable Removal

**The curated dataset was loaded in R from a csv file. Irrelevant or derived variables directly inferring Parkinson's prediction were dropped by inspection. For example, the variable *PRIMDAIG* for primary diagnosis would significantly correlate with predicting Parkinson's because physicians directly input their diagnoses using that variable. This would not be helpful for modeling predictions nor help physicians beforehand. Next, variables related to subjective tests in the clinical subgroup were dropped per the objectives of this case study. Exceptions include static data from binary responses like *quip* disorder data. Since a simplistic approach was chosen, patient visits were filtered on the baseline value for their first visit. Furthermore, the response variable *CONCOHORT* was filtered to only Parkinson’s patients and the healthy controls. The resultant dataset is 1395 observations with 47 variables.**

### Data Cleaning

**Data cleaning involved converting categorical variables to factors in R. In doing so, some variables had empty categorical level values displayed as “.”. These levels and empty values were coerced to N/A during the factor conversion. For clarity, the response was renamed from numerical factors *1* and *2* to *PD* and *Healthy*.**

### Missing Data

**The *Vim* library in R helped tackle the issue of missing data in the resultant dataset. A histogram of missing data in Figure 1 helps showcase the sparseness of the data.**

**Figure 1   
  
Histogram and Pattern of Missing Data**

A graph of a pattern

Description automatically generated with medium confidence

**Due to a high amount of missing data, variables above 75% missingness were dropped from the dataset. The remaining variables with moderate missingness were kept to retain as much information as possible. Then, rows with missing data were dropped, resulting in a dataset of 265 observations and 47 variables. The significant decrease in observations is concerning but necessary compared to possible bias introduced with imputation methods replacing significant portions of data.**

### Multicollinearity

**A correlation matrix was created to find pairwise correlations between variables. A cutoff of 0.8 separated variables with high collinearity. The corresponding columns were dropped from the dataset using the *FindCorrelation* function from the *caret* library. Figure 2 shows the resulting correlation plot between the remaining numerical predictors.**

**Figure 2   
  
Correlation Plot after Dropping Colinear Variables**

A diagram of a number of bmi

Description automatically generated with medium confidence

### Variance

**Predictors with near zero variance were dropped using the *nearZeroVar* from *caret*. These predictors had few unique values and would not help classify the response. These were all categorical variables. Table 2 displays the columns dropped. The column *nzv* signals near zero variance. The final dataset contained 21 predictors.**

**Table 2   
  
Near Zero Variance Variables**

| Variables | freqRatio | percentUnique | zeroVar | nzv |
| --- | --- | --- | --- | --- |
| race | 36.14 | 1.51 | FALSE | TRUE |
| HISPLAT | 19.38 | 0.75 | FALSE | TRUE |
| AFICBERB | 131.50 | 0.75 | FALSE | TRUE |
| BASQUE | 52.00 | 0.75 | FALSE | TRUE |
| quip\_gamble | 36.86 | 0.75 | FALSE | TRUE |
| quip\_sex | 36.86 | 0.75 | FALSE | TRUE |
| quip\_buy | 25.50 | 0.75 | FALSE | TRUE |
| quip\_pund | 36.86 | 0.75 | FALSE | TRUE |
| quip\_walk | 131.50 | 0.75 | FALSE | TRUE |

### Data Training/Test Split and SMOTE

**Data was split into 70% training and 30% test partitions for validation. The sampling technique SMOTE from the *performance estimation* library was appliedto the training data to correct the class imbalance. Table 3 shows the change in distribution. The majority class was down-sampled by a factor percent of 2, while the minority class was synthetically up-sampled by a factor percent of 2 using the *Smote* function from the same library.**

**Table 3   
  
SMOTE Imbalance changes**

| SMOTE | PD | Healthy |
| --- | --- | --- |
| Before | 0.73 | 0.27 |
| After | 0.57 | 0.43 |

## Model Generation and Deployment

**All algorithms were cross-validated and standardized using the preprocessing arguments in the *caret* library. ROC was set as the validation metric. The model consisted of the binary response *CONCOHORT* against all remaining predictors. Regularization through *glmnet* for logistic regression was applied to help avoid perfect separation and ensure convergence. All other algorithms were set to default parameters. AUC/ROC and accuracy were recorded for comparison in results.**

## Conclusion

**The methodology used in this section references the literature review to apply varying data preparation and modeling methods on the curated dataset from the PPMI. This case study’s objective necessitates the removal of a large portion of subjective test variables from the original dataset. Imputation through the MICE package was considered, but concerns of bias and the high percentage of missing data removed it from contention. Instead, rows with missing data were removed. Feature selection through the *Boruta* package with the random forest was tested, but a simplistic approach that kept more predictors by removing multicollinear and low-variance variables was chosen instead. The SMOTE technique handled the class imbalance problem between patients with Parkinson’s and the healthy control. A variety of models were tested using the *caret* package with ROC and accuracy was recorded. The next chapter follows the results of this analysis and interpretation of the best model.**

# Chapter 4: Results

## Introduction

**Following the methodology in Chapter 3, this chapter expands upon the model results and discusses their evaluation. A final model is chosen to predict Parkinson’s disease from the healthy control using accuracy and AUC. Feature importance is discussed through Shapley values. The overall project objectives are reviewed by comparing metrics and explaining predictors.**

## Results

**The methodology for model comparison used AUC/ROC as its deciding metric. Figure 3 compares the performance of each model using a ROC curve and the displayed AUC values. The area under the curve measures aggregate performance using different classification thresholds. The best model according to AUC is Random Forest with an area under the curve of 0.91. Xgboost follows with an AUC of 0.87. The worst performing model is k nearest neighbors.**

**Figure 3   
  
Graph of Model ROC’s with AUC Values**

A graph of a function

Description automatically generated with medium confidence

**Other metrics were recorded for model comparison and to validate project objectives. Table 4 includes accuracy, sensitivity, specificity, and F1 in addition to AUC. Random Forest performed the best according to all metrics except specificity. Xgboost had the second-best metrics, beating Random Forest in specificity. Logistic regression has a better specificity score than Random Forest but placed third in metrics. Unsurprisingly, K nearest neighbors supplemented its low AUC with the lowest test metrics.**

**Table 4   
  
Model Metrics**

| Models | Accuracy | Sensitivity | Specificity | F1 | AUC |
| --- | --- | --- | --- | --- | --- |
| rf | 0.85 | 0.90 | 0.71 | 0.90 | 0.91 |
| xgboost | 0.78 | 0.78 | 0.81 | 0.84 | 0.87 |
| glm | 0.77 | 0.74 | 0.86 | 0.83 | 0.84 |
| svm | 0.72 | 0.74 | 0.67 | 0.80 | 0.83 |
| lda | 0.71 | 0.69 | 0.76 | 0.78 | 0.83 |
| knn | 0.58 | 0.57 | 0.62 | 0.67 | 0.61 |

## Best Model

**Several considerations are involved when choosing the best model. Aside from accuracy and AUC, model complexity and interpretation are also important factors. The simplest model, logistic regression, is the easiest to interpret and could help determine trends or relationships for predicting Parkinson's. On the other hand, more complex ensemble models like Xgboost suffer the opposite, resulting in black-box answers with little explanation. There is also a tradeoff between sensitivity and specificity with Parkinson’s predictions.**

**Sensitivity is the model’s ability to predict Parkinson’s correctly. A high sensitivity results in fewer false negatives, meaning a low amount of Parkinson’s predictions which were healthy controls. Specificity is the model’s ability to predict healthy controls. A high specificity means low false positives. This mirrors sensitivity in that there would be a low amount of predicted healthy controls that were people with Parkinson’s. For this case study, it is more critical to predict people with Parkinson’s correctly; thus, a high sensitivity is the better comparison between models. However, concern should be given to predicting Parkinson’s for healthy controls. An inaccurate diagnosis can prove detrimental. The model Random Forest performs best in sensitivity while leveraging the second-best specificity.**

**Ultimately, the model choice is between Random Forest and Xgboost due to their accuracy and sensitivity/specificity trade-off. The best model must surpass or equal the project objective of 80% physician accuracy in diagnosing Parkinson’s disease, and these models came the closest or surpassed it. If there were a need to explain relationships between the binary class and the predictors, logistic regression would be a suitable option. However, with the objective goal of accuracy, Random Forest best predicts Parkinson’s for the purposes of this case study.**

## Feature Importance

**An objective of the study is to find possible influential predictors or biomarkers for predicting Parkinson’s. One way of accomplishing this is by ranking model predictors. Figure 4 displays one way of displaying feature importance through the** Shapley Test. Makarious et al. (2022) used this method in the literature review and is an alternative approach to finding the importance through gain used in the *caret* package. The Shapley method works by finding the contribution of each predictor to the final model using game theory. Features are randomly split and marginal predictor contributions of each split tell how much each feature affects the final prediction. The Shapley value is the average of these contributions (Gopinath & Kurokawa, 2021). This value is approximated in R with the *fastshap* library for computational reasons.

**Figure 4   
  
Shapley Value Feature Importance on Random Forest Model**

**A graph of a number of objects

Description automatically generated with medium confidence**

The Shapley values reveal that the most influential predictor is *ips\_caudate*, a DaTscan variable. It heavily controls the model's predictions, its value being almost double the Shapley value of the second predictor. The variables *fampd\_bin* (family history of Parkinson’s), *urate*, *APOE\_e4,* and *total\_di\_18\_1\_BMP* compose the following top five. Family history is a demographic variable, while *APOE\_e4* is a genetic variable signaling the number of E4 alleles in the APOE genotype. The remaining two are numerical biologics. Urate represents the amount of uric acid and *total\_di\_18\_1\_BMP* is the measure of a derivative of bis(monoacylglycerol)phosphate (BMP), a phospholipid collected from urine samples.

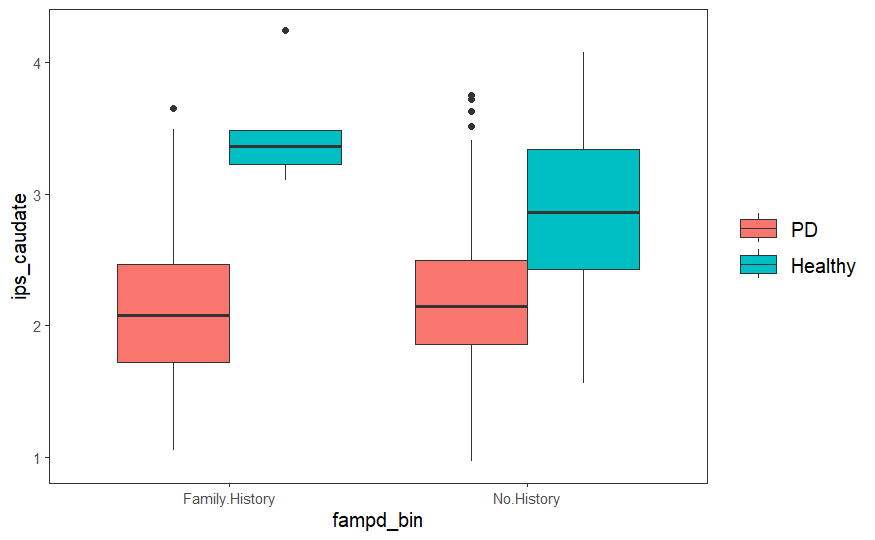
## Top Predictors Discussion

From the Shapley plot, the top model predictors for predicting Parkinson’s are a mix of demographic, DaTscan, genetic, and biological variables. The presence of the demographic variables age and, more significantly, family history may correlate to some aspects of Parkinson’s. The disease primarily affects people aged 60 or older. Furthermore, families with a history of Parkinson’s have a higher chance of developing the disease (Mayo Clinic, 2023). On the other hand, lowered urate levels have been associated with developing Parkinson’s and determining the disease’s severity (Danau et al., 2022). It can be collected through blood tests or urine samples. Similarly, the BMP derivative biological variable may function as a possible disease indicator. Elevated levels of both BMP derivatives have been found in mutations associated with Parkinson’s disease (Gomes et al., 2023). The data can be collected non-invasively through urine samples. The remainder of the predictors are also obtainable biologics or demographics except for the gene *APOE\_e4*. This requires additional patient genotyping. Variants of this gene are also related to Alzheimer’s risk (Legget, 2022). Furthermore, this specific variant has been tied to Parkinson’s through its effect on Lewy bodies (Ray, 2020).

The best predictor, *ips\_caudate*, is obtained from DaTscan imaging data. The variable means ipsilateral caudate, meaning values were taken from the same side of the brain. The caudate is a region in the brain that processes visuals, movement, and memory. Decline in the region is associated with neurological disorders (Bellefonds, 2021). Likewise, DaTscan measurements in the caudate region may serve as a strong biomarker. A study by Pasquini et al. (2019) found caudate dopamine decreases in early Parkinson’s patients. It is believed that the degeneration of dopamine neurons, the aspect affecting movement in Parkinson’s, can reach the caudate region (Pena, 2019).

Figure 5 shows the relationship between the top two predictors by *CONCOHORT* class. There is a clear depiction that lower caudate levels correspond to the Parkinson’s class. Furthermore, this distinction is present in both categories of family history of the disease. However, it appears that family history exasperates the difference between Parkinson’s and healthy control.

**Figure 5**   
  
Boxplot of Caudate Levels and Parkinson’s Family History



These findings corroborate previous research and reinforce the significance of *ips\_caudate* as a possible biomarker and strong predictor.

## Project Objectives

**The section discusses the case study’s project objectives and evaluates their success.**

### Physician Accuracy

**The best model, Random Forest, surpasses physician diagnosis accuracy of 80%, reaching 85% accuracy. This metric is validated using a training and test set with 5-fold validation. While papers in the literature review have achieved higher accuracy, they incorporate clinical tests with subjectivity. Since this case study aims to predict Parkinson’s without subjectivity, the model passes this objective.**

### Non-Subjective Initial Screening Tool

**The next objective examines the model's function as a possible initial screening tool without physician input. The model methodology helps determine its validity as a test. Since almost all subjective tests and variables were dropped, the remaining predictors used in the model are all empirical genetic, biologics, or demographic data easily obtained from patients. Important predictors like urate and** BMP derivative biologicals **come from blood/urine samples,** while DaTscans are commonly performed as a test for possible Parkinson’s patients. Genetic data for *APOE\_e4* needs patient genotyping. The monetary costs of these tests are unclear and vary. Overall, due to this **ease of data collection for most variables and the inherent use of non-subjective metrics, this model can be successfully used as an initial screening tool to predict Parkinson's without patient or physician input. While more barriers and checks are present for implementing an initial screening tool, this model passes as a rudimentary initial test.**

### Confirm Possible Biomarkers

**One of the goals is to find or confirm possible biomarkers. Since the data came from a curated dataset with known Parkinson’s-related data, the possibility of a new unseen significant variable was low. Instead, the top five biological, genetic, and DaTscan variables were confirmed as important variables in the model. Other biologics were present but not as significant. Urine-collected urate and BMP derivatives were helpful predictors.** The variable, *APOE\_e4*, has been linked to diseases like Alzheimer's and Parkinsons. Ipsilateral caudate provided the most information for predicting the disease and was supported by studies.

In general, caudate-region-related variables might have more importance overall. Some caudate-like variables related to these predictors, like putamen, striatum, or their left and right separated brain regions, were dropped due to multicollinearity, meaning they provided similar information for model predictions. These regions may also be helpful for predicting Parkinson’s.

## Conclusion

**Chapter 4 discussed the results taken from the methodology. Model metrics were compared primarily using AUC and accuracy. Consideration was given to sensitivity and specificity among the top two models Random Forest and Xgboost. The Random Forest model best balanced project-objective accuracy, sensitivity, and specificity. Feature importance was determined with Shapley values and the predictor** ipsilateral caudate contributed most to model predictions across random splits of features. The model passed the three project objectives set out in Chapter 1. Model accuracy surpassed 80%, functioned as an initial screening tool, and confirmed possible biomarkers. Chapter 5 summarizes the case study, discusses the project objectives, and outlines the next steps.

# Chapter 5: Discussion

## Introduction

This chapter summarizes and concludes the case study. It offers room for improvement and steps forward. The text acknowledges the need for real-world physician input and mentions the shortcomings of the case study.

## Summary of Approach

Parkinson’s disease is a progressive and debilitating disease expected to afflict many people economically and physically. There have been numerous attempts at predicting or evaluating the progression of the disease through different types of tests. Researchers have successfully utilized machine learning methods to assist these attempts and other aspects of the disease.

This case study differs from previous approaches as it strives to build a model predicting the disease with limited subjective input. The overall project goal is to develop a model that produces dependable and reproducible test results for physicians. The case study's purpose is to use the model to assist physicians in diagnosing Parkinson’s disease. By eliminating possible bias from both the patient and physician, models can perform accurately based on reliable and obtainable metrics.

Combining the project goal and case study purpose, the project objectives give

concrete definitions for the project. The three objectives are repeated in concise form as follows:

1. **The model accuracy must achieve at least 70%, aiming for a stated physician accuracy of 80%.**
2. **Without patient or physician input, examine the model’s ability as a tool or initial screening.**
3. **Find or confirm correlated predictors for predicting Parkinson’s or possible biomarkers.**

These project objectives reflect the decisions in the subsequent methodology and results chapters. Model accuracy was a key metric, while model interpretation and variable importance were considered.

The literature review aimed to cover previous attempts at applying machine learning models to the disease. Starting broadly, there were several approaches to predicting Parkinson’s including vocal, gait, biometric, and historical data. Narrowing to the Parkinson’s Progression Marker Initiative provided the closest comparison to this case study. A general overview by Gerraty et al. (2023) delivered invaluable information on the state of machine learning models in the PPMI. The recommendations from this paper largely advised the general methodology. Looking into specific papers gave insight into modeling procedures. Researchers used many types of models to predict the disease. Three papers provided the basis for the specific decisions and reasonings for data cleaning, feature selection, and model building in Chapter 3.

Chapter 3 detailed the steps from the curated dataset to the final model. Data cleaning involved fixing variable formatting in R and correctly encoding categorical variables. Subjective and unrelated variables were dropped according to the project objectives. Feature selection narrowed down predictors using multicollinearity and variance checks. Data imputation through the MICE package and automated feature selection through the Boruta package were explored but discarded. Stepwise regression methods were shown to have flawed results. Penalized methods like lasso, elastic net, or ridge regression are acceptable alternatives. Regularization was deployed for logistic regression but was ultimately not used for feature selection. Instead of removing predictors, they were kept to retain as much information as possible for other models. Validation involved splitting data into training and test partitions. Before model testing, SMOTE balanced the classification of Parkinson’s versus healthy control in the training set. Five-fold cross-validation tested the models over different splits of data. The evaluation process involved testing the trained models on the previously reserved test dataset.

Chapter 4 evaluated the models through metrics and tradeoffs. The focus was on prioritizing AUC and accuracy while comparing sensitivity and specificity. Logistic regression provided the best interpretability. However, two models stood out above the rest. Random Forest and Xgboost carried the top values, with Random Forest having the better overall metrics. Random Forest was the best model according to the project objective of accuracy and the sensitivity/specificity tradeoff. Here, the focus could have shifted to better specificity for Xgboost.

Predictor importance helps determine which variable best classifies Parkinson’s. There are different ways of implementing this; however, this case study utilized Shapley values as its deciding criteria. The method, popular for dealing with black-box models, had precedence in the literature review. Shapley values were approximated for computational reasons. The fully calculated Shapley values may offer more precise results.

A DaTscan variable from the caudate region of the brain provided the best classification of Parkinson’s versus healthy controls. DaTscans are commonly deployed as tests for people with Parkinson’s. Furthermore, several papers corroborate this variable as a possible indicator of the disease (Bellefonds, 2021; Pasquini et al.,2019; Pena, 2019). This was replicated in Figure 5. The remaining five predictors were demographic family history, genetic *APOE\_E4*, and biological urate and BMP derivatives. It is likely that colinear variables in the brain's caudate region from DaTscan data are also equivalent predictors.

## Project Objective Conclusions

The first objective aimed to equal or surpass physician accuracy of 80%. The random forest model surpassed this metric and provided robust AUC, sensitivity, and specificity scores. For the case study, this accuracy is a success. However, while this may comparatively prove its worth as a predictive model, greater accuracy is necessary for real-world implementation. As mentioned previously, misclassifying healthy controls as people with Parkinson’s can carry negative consequences. In this case, physicians would need to correct the evaluation with supplementary tests, defeating the purpose of the model. A higher accuracy compared not just to physicians but to other machine learning models lowers this possibility. As such, a higher objective accuracy would improve reliability and is necessary for any practical implementation. In contrast, the overall objective of the model should still be to assist and not replace diagnosing physicians. No matter the accuracy, model predictions should complement physician decisions.

The project objective with the biggest room for improvement is the model’s function as a non-subjective initial screening tool. As a rudimentary test, the model can provide a reliable prediction. The predictors are non-subjective; with a few exceptions, most appear to be easily collectible metrics or routine tests for Parkinson’s. However, jumping to a working screening tool requires advice from a neurologist and medical professionals diagnosing the disease. It also would need accurate integration within procedural diagnostic steps and existing medical infrastructure. For example, would it be appropriate for a first-visit screen or only applicable after visiting a medical professional? Could it be available to the general public? If so, how would the medical test results for the predictors be collected? The model’s implementation range is uncertain.

The model may also need adjustment to fit the real world practically. This is present in the predictors themselves. As mentioned in Chapter 4, the best predictor, a DaTscan variable, is correlated with other dropped predictors due to multicollinearity. It is possible that the dropped variables could provide similar information and give an easier, practical version for the real world. Some tests may also be more expensive than others. There must be a balance between practical predictor considerations and model accuracy.

One predictor might provide the newest research opportunities. The third objective confirmed genetic, biologic and DaTscan variables as possible predictors. This makes sense as DaTscans are commonly performed tests for Parkinson’s, and the other variables have supporting studies. While researchers have studied DaTscan, urate and APOE predictors, far fewer have connected BMP derivative biological data to Parkinson’s. This may present a unique opportunity for further research. The urine-collected variable can act as a surrogate for mutations associated with Parkinson’s. So far, companies like *Nextcea*, the patent holders of the BMP derivatives, have linked diagnostic capabilities and drug delivery to Parkinson’s (Nextcea, n.d.). The Micheal J. Fox Foundation has provided funding and researchers have looked into the area with promising initial studies (Keller, 2019). None have combined the BMP derivative with other variables for predictive modeling capabilities. Still, caution should be given to its influence. DaTscan variables provided the highest predictive ability and controlled the model’s decisions.

## Steps Forward

Several considerations exist for future case study iterations. Starting with the first filter of baseline patient visits in the curated data set, a different type of model design could instead include historical data. Utilizing the longitudinal nature of PPMI data is one of the literature review recommendations by Gerraty et al. (2023). Recall in Chapter 3 an alternative process for modeling the data. One suggestion for this process could be to take the difference between the baseline and a specific later patient visit across applicable predictors. Modeling would involve using the differences as predictors for Parkinson’s. For example, take patient visit one and visit six. For variables like the BMP derivatives, its value would be the difference between those visits. Demographic variables would remain static, assuming they do not change over that period, like *sex*, *family history*, or *race*. If other categorical predictors vary over time, they will need to be removed from the analysis. The advantage of this approach is that it keeps the model simple while incorporating changes in time, hopefully adding more information. One disadvantage is that model inputs would become differences between visits instead of just one baseline value.

Complex time-series models better fit the longitudinal nature of the PPMI. They are suited to finding changes in Parkinson’s and can also help predict the disease. However, it may not fit this case study’s goals. Many papers in the literature review used changes in UDPRS scores or other subjective metrics to determine their model’s success. Avoiding these scores is central to this study. An alternative approach with a different design philosophy using the new curated data set and including subjective metrics to predict Parkinson’s could use time-series modeling.

Some specific modeling changes could improve performance or reliability. In addition to the test validation set, an external data set like the Parkinson’s Disease Biomarker Program (PDBP) used by Makarious et al. (2022) could cross-validate the results. This would provide an honest test of the model separate from the curated dataset. However, in order to combine the datasets, the predictors and patient identification must be compatible with each other. The model may need further adjustments to fit the external dataset. Aside from validation improvements, the model could be improved through hyperparameter optimization, iterating over different model parameters. This was not done for fear of overfitting the training set. However, a practical implementation of the full dataset may need improved accuracy and lower misclassification errors.

## Conclusion

A discussion of previous chapter summaries showed an overview of the project. Project objectives were extended to reflect real-world applications and possible shifts in goals. The steps forward introduce alternative case study designs and recommend advice from diagnosing physicians. Modeling optimizations could increase metric precision.

# Conclusion

Predicting Parkinson’s disease using machine learning models is a worthwhile endeavor that offers reproducible and reliable results. With no cure, accurate diagnosis is paramount for early treatments and care. By removing subjectivity, the model adds function as an initial screening tool and removes bias from the patient and the diagnosing physician. The project objectives laid out in this case study combine the goal of achieving reliable model results and the purpose of assisting physicians.

This case study divided its post-introductory chapters into four sections: literature review, methodology, results, and discussion. The literature review provides the basis for the actions taken in the methodology. The final case study design considered different aspects of many papers. Data cleaning and feature selection introduced the methodology. These methods carried into the modeling process, where several algorithms were trained and validated using k-fold cross-validation. The results found that random forest best classified Parkinson’s disease primarily due to metrics like AUC and accuracy. Furthermore, all three project objectives passed. The model proved reliable with 85% accuracy, functioned as an initial screening tool, and confirmed possible predictors. A final chapter discussion overviewed the previous chapters and provided project objective improvements and next steps. These include types of time-series modeling as a prospective alternative for future studies. The model needs physician advisement and testing.

The model offers promise as an initial rudimentary test to assist physicians in diagnosing Parkinson’s. It takes a pragmatic approach by removing subjectivity from the diagnosis. Hope lies in future case study iterations that expand its applicability in the real world.

# References

Agiwal, V., & Chaudhuri, S. (2024). Methods and Implications of Addressing Missing Data in Health-care Research. *Current Medical Issues*, *22*(1), 60–62. https://doi.org/10.4103/cmi.cmi\_121\_23

Awan, A. (2022, November 29). *An Introduction to SMOTE*. KDnuggets. https://www.kdnuggets.com/2022/11/introduction-smote.html

Ayaz, Z., Naz, S., Khan, N. H., Razzak, I., & Imran, M. (2023a). Automated methods for diagnosis of parkinson’s disease and predicting severity level. *Neural Computing and Applications*, *35*(20), 14499–14534. https://doi.org/10.1007/s0052102106626y

Ayaz, Z., Naz, S., Khan, N. H., Razzak, I., & Imran, M. (2023b). Automated methods for diagnosis of parkinson’s disease and predicting severity level. *Neural Computing and Applications*, *35*(20), 14499–14534. https://doi.org/10.1007/s0052102106626y

Bellefonds, C. (2021, November 16). *Caudate Nucleus Function, Anatomy, and Definition | Body Maps*. Healthline. https://www.healthline.com/human-body-maps/caudate-nucleus

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-106

Chahine, L. M., Siderowf, A., Barnes, J., Seedorff, N., CaspellGarcia, C., Simuni, T., Coffey, C. S., Galasko, D., Mollenhauer, B., Arnedo, V., Daegele, N., Frasier, M., Tanner, C., Kieburtz, K., Marek, K., & The Parkinson’s Progression Markers Initiative. (2019). Predicting Progression in Parkinson’s Disease Using Baseline and 1Year Change Measures. *Journal of Parkinson’s Disease*, *9*(4), 4. https://doi.org/10.3233/JPD181518

Channa, A., Ifrim, R.-C., Popescu, D., & Popescu, N. (2021). A-WEAR bracelet for detection of hand tremor and bradykinesia in parkinson’s patients. *Sensors*, *21*(3), 981. https://doi.org/10.3390/s21030981

Dadu, A., Satone, V., Kaur, R., Hashemi, S. H., Leonard, H., Iwaki, H., Makarious, M. B., Billingsley, K. J., Bandres‐Ciga, S., Sargent, L. J., Noyce, A. J., Daneshmand, A., Blauwendraat, C., Marek, K., Scholz, S. W., Singleton, A. B., Nalls, M. A., Campbell, R. H., & Faghri, F. (2022). Identification and prediction of Parkinson’s disease subtypes and progression using machine learning in two cohorts. *Npj Parkinson’s Disease*, *8*(1). https://doi.org/10.1038/s41531-022-00439-z

Danau, A., Dumitrescu, L., Lefter, A., & Popescu, B. A. (2022). Serum Uric Acid Levels in Parkinson’s Disease: A Cross-Sectional Electronic Medical Record Database Study from a Tertiary Referral Centre in Romania. *Medicina*, *58*(2), 245–245. https://doi.org/10.3390/medicina58020245

Diana, Z. J., Xue, C., Kolachalama, Vijaya B, & Donald, W. A. (2023). Interpretable Machine Learning on Metabolomics Data Reveals Biomarkers for Parkinson’s Disease. *ACS Cent. Sci.*, *9*(5), 1035–1045. https://doi.org/10.1021/acscentsci.2c01468

Dou, K., Ma, J., Zhang, X., Shi, W., Tao, M., & Xie, A. (2022). Multipredictor modeling for predicting early Parkinson’s disease and nonmotor symptoms progression. *Frontiers in Aging Neuroscience*, *14*. https://www.frontiersin.org/articles/10.3389/fnagi.2022.977985

Evers, L. J. W., Krijthe, J. H., Meinders, M. J., Bloem, B. R., & Heskes, T. M. (2019). Measuring Parkinson’s disease over time: The real‐world within‐subject reliability of the MDS‐UPDRS. *Movement Disorders*, *34*(10), 1480–1487. https://doi.org/10.1002/mds.27790

Ferreira, M. I. A. S. N., Barbieri, F. A., Moreno, V. C., Penedo, T., & Tavares, J. M. R. S. (2022). Machine learning models for parkinson’s disease detection and stage classification based on spatial-temporal gait parameters. *Gait & Posture*, *98*, 49–55. https://doi.org/10.1016/j.gaitpost.2022.08.014

Frost, J. (2017). *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions*. Statistics by Jim. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Gerraty, R. T., Provost, A., Li, L., Wagner, E., Haas, M., & Lancashire, L. (2023a). Machine learning within the Parkinson’s progression markers initiative: Review of the current state of affairs. *ProQuest*, *15*(1076657). https://doi.org/10.3389/fnagi.2023.1076657

Gerraty, R. T., Provost, A., Li, L., Wagner, E., Haas, M., & Lancashire, L. (2023b). Machine learning within the Parkinson’s progression markers initiative: Review of the current state of affairs. *Frontiers in Aging Neuroscience*, *15*. https://www.frontiersin.org/articles/10.3389/fnagi.2023.1076657

Gomes, S., Garrido, A., Tonelli, F., Obiang, D., Tolosa, E., Martı́M. J., Ruiz‐Martínez, J., Vinagre‐Aragón, A., Hernandez-Eguiazu, H., Croitoru, I., Marshall, V. L., König, T., Hotzy, C., Hsieh, F., Sakalosh, M., Tengstrand, E., Padmanabhan, S., Merchant, K., Bruecke, C., & Pirker, W. (2023). Elevated urine BMP phospholipids in LRRK2 and VPS35 mutation carriers with and without Parkinson’s disease. *Npj Parkinson’s Disease*, *9*(1). https://doi.org/10.1038/s41531-023-00482-4

Gopinath, D., & Kurokawa, D. (2021, October 26). *The Shapley Value for ML Models*. Medium. https://towardsdatascience.com/the-shapley-value-for-ml-models-f1100bff78d1

Hauser, R. A. (2023, February 13). *Parkinson disease: Practice essentials, background, anatomy*. Medscape. https://emedicine.medscape.com/article/1831191-overview?0=reg=1&1=icd=login\_success\_email\_match\_norm#a1

Keller, D. (2019, November 4). *Urine Biomarker Correlates With Parkinson’s Risk Factor Gene*. Medscape. https://www.medscape.com/viewarticle/920818?form=fpf#vp\_2

Kursa, Miron B, & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *J. Stat. Soft.*, *36*(11), 113. https://doi.org/10.18637/jss.v036.i11

Leger, C., Herbert, M., & DeSouza, J. F. X. (2020). Non-motor Clinical and Biomarker Predictors Enable High Cross-Validated Accuracy Detection of Early PD but Lesser Cross-Validated Accuracy Detection of Scans Without Evidence of Dopaminergic Deficit. *Frontiers in Neurology*, *11*. https://doi.org/10.3389/fneur.2020.00364

Legget, H. (2022, May 31). *A rare mutation protects against Alzheimer’s disease, Stanford-led research finds*. Stanford Medicine News Center. https://med.stanford.edu/news/all-news/2022/05/gene-mutation-alzheimers.html

Makarious, M. B., Leonard, H. L., Vitale, D., Iwaki, H., Sargent, L., Dadu, A., Violich, I., Hutchins, E., Saffo, D., BandresCiga, S., Kim, J. J., Song, Y., Maleknia, M., Bookman, M., Nojopranoto, W., Campbell, R. H., Hashemi, S. H., Botia, J. A., Carter, J. F., & Craig, D. W. (2022). Multimodality machine learning predicting Parkinson’s disease. *Npj Parkinson’s Disease*, *8*(1), 35. https://doi.org/10.1038/s4153102200288w

Mayo Clinic. (2023, May 26). *Parkinson’s Disease*. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055

Nalls, M. A., McLean, C. Y., Rick, J., Eberly, S., Hutten, S. J., Gwinn, K., Sutherland, M., Martinez, M., Heutink, P., Williams, N. M., Hardy, J., Gasser, T., Brice, A., Price, T. R., Nicolas, A., Keller, M. F., Molony, C., Gibbs, J. R., Chen-Plotkin, A., & Suh, E. (2015). Diagnosis of Parkinson’s disease on the basis of clinical and genetic classification: a population-based modelling study. *The Lancet Neurology*, *14*(10), 1002–1009. https://doi.org/10.1016/s1474-4422(15)00178-7

Nextcea. (n.d.). *Parkinson’s disease Biomarker assay drug efficacy assessments and therapeutic uses*. Nextcea. https://nextcea.com/parkinsons-disease-lrrk2-biomarker/

Ohene, A. (2018, January 11). *Parkinson’s prevalence expected to increase by 18% in next seven years*. Parkinson’s Life; Parkinson’s Europe. http://parkinsonslife.eu/parkinsons-prevalence-expected-to-increase-by-18-by-2025/

Pasquini, J., Durcan, R., Wiblin, L., Morten Gersel Stokholm, Rochester, L., David James Brooks, Burn, D., & Pavese, N. (2019). Clinical implications of early caudate dysfunction in Parkinson’s disease. *J Neurol Neurosurg Psychiatry*, *90*(10), 1098. https://doi.org/10.1136/jnnp2018320157

Pena, A. (2019, May 15). *In Parkinson’s, Early Caudate Involvement Linked to Worse Prognosis*. Parkinsonsnewstoday.com. https://parkinsonsnewstoday.com/news/early-caudate-involvement-linked-worse-prognosis-parkinsons/

Ray, F. (2020, March 24). *APOE Variant Tied to Parkinson’s, Other Lewy Body Dementias in Studies*. Parkinsonsnewstoday.com; BIONEWS. https://parkinsonsnewstoday.com/news/apoe-variants-affect-alpha-synuclein-levels-in-lewy-body-dementias-2-studies-report/

Rizzo, G., Copetti, M., Arcuti, S., Martino, D., Fontana, A., & Logroscino, G. (2016). Accuracy of clinical diagnosis of parkinson disease. *Neurology*, *86*(6), 566–576. https://doi.org/10.1212/wnl.0000000000002350

van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259

Wu, Y. (2022, August 24). *7 Techniques to Handle Imbalanced Data*. KDnuggets. https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html

Yang, W., Hamilton, J. L., Kopil, C., Beck, J. C., Tanner, C. M., Albin, R. L., Dorsey, R., Dahodwala, N., Cintina, I., Hogan, P., & Thompson, T. (2020). Current and projected future economic burden of parkinson’s disease in the U.S. *Npj Parkinson’s Disease*, *6*(1), 15. <https://doi.org/10.1038/s4153102001171>

# Appendix A Alternative Feature Selection

A graph with different colored lines

Description automatically generated

# Appendix B Code

All code for this paper can be found in the following GitHub repository. The datasets are available upon request to the Micheal J. Fox Foundation and the PPMI.

Link:

<https://github.com/deands1073/Parkinson-s-Capstone.git>