Predictive Ability of Cerebrospinal Fluid Biomarkers in Diagnosing and Evaluating Parkinson's disease

by

Michelle J. Wang

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Masters of Engineering in Electrical Engineering and Computer Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2014

Author . . . . . .
Department of Electrical Engineering and Computer Science
May 23, 2014

Certified by . . . . . .
Timothy Denison
Director of Core Technology, Medtronic Inc., Thesis Co-Supervisor
May 23, 2014

Certified by . . . . . .
John Guttag
Dugald C. Jackson Professor, Thesis Co-Supervisor
May 23, 2014

Accepted by . . . . . .
Professor Albert R. Meyer
Chairman, Masters of Engineering Thesis Committee

Predictive Value of Cerebrospinal Fluid Biomarkers in Diagnosing and Evaluating Parkinson's disease

by Michelle J. Wang
Submitted to the Department of Electrical Engineering and Computer Science

May 23, 2014
In Partial Fulfillment of the Requirements for the Degree of Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

Currently, there are a variety of clinical assessments and rating scales used in the research and treatment of Parkinson's disease (PD). Despite the widespread use and reliance on these scales, they do not offer a uniform, objective measure. Many previous studies have indicated promising relationships between various biomarkers and Parkinsonian symptoms that could lead to objective measures by using statistical methods and providing p-values. However, we could not find any literature that uses machine learning or directly tests predictive value. The goal of this thesis was to determine whether or not cerebrospinal fluid (CSF) biomarker data could predict incidence of Parkinson's with a high degree of accuracy and differentiate between patients with varying levels of severity. We used various supervised machine learning algorithms on the Parkinson's Progression Markers Initiative (PPMI) baseline data set provided by the Michael J. Fox Foundation, and reported the percentage of patients correctly diagnosed by each algorithm on an isolated test data set. The best classifier averaged 69% accuracy in distinguishing human controls from PD patients. While this does indicate the presence of some predictive power, it is not clinically useful and we tentatively conclude a negative result. The data pertain to the CSF biomarkers available from PPMI at the end of October 2013.

Tim Denison
Director of Core Technology, Medtronic Inc.
Thesis Advisor

John Guttag
Dugald C. Jackson Professor
Thesis Advisor

Acknowledgements

First and foremost, I would like to express my deep gratitude to Timothy Denison and John Guttag for supervising my thesis and guiding my research with their experience and wisdom.

I also greatly appreciate all my advisors at Medtronic Inc. for the time and effort they took out of their own work schedules to assist me and edit my writing. I want to specifically acknowledge Eric Panken, Bill Kaemmerer, and Siddharth Dani, for their advice, mentorship, acceptance, and patience as I conducted my research over the summer and fall at Medtronic.

Thanks to the Michael J. Fox Foundation for providing invaluable data, a positive research environment, and open communication in both the Kaggle competition and Parkinson's Progression Markers Initiative.

I want to thank the MIT, the Course VI Department and the VI-A Program for supporting my studies and providing excellent resources for the past five years. I will never forget this place.

Finally, special thanks to my family and friends for love, support, and edits

Table of Contents

1 INTRODUCTION

Parkinson's disease (PD) is a degenerative neurological disorder that affects seven to ten million people worldwide, most of them over the age of 60 (Parkinson's Disease Foundation). The primary symptoms are manifested in the form of motor impairments while secondary non-motor symptoms such as dementia may also occur. Each patient experiences different combinations of symptoms at varying severity levels. Both the initial diagnosis and tracking of Parkinson's depends on accurately evaluating the motor and non-motor manifestations in an objective manner. This allows researchers and clinicians to comparatively measure the effect of different therapies and guide treatment techniques. Currently, the clinical assessments used are largely subjective in nature. While the assessment process and ratings are standardized, there is still room for variability based on which physician assesses which patient. In short, the challenge of defining measures with objective data is largely unsolved.

Among the many clinical assessments that currently exist for evaluating symptoms and following patient progress, the most commonly used and widely accepted rating scale is the Unified Parkinson Disease Rating Scale (UPDRS). The UPDRS consists of four sections and seeks to quantify motor and non-motor manifestations of the disease. Part 3 quantifies motor manifestations and measures many different aspects of a patient's movement including rigidity, tremor, posture, stability, and gait (Perimutter, 2009). Other commonly used metrics include the Hoehn and Yahr Scale and the Schwab and England ADL Scale. Note that these scales generally measure symptomatic severity, which can vary with medication, rather than the actual state of the disease. While it is important to discern the severity of the disease itself, being able to quantify symptoms is equally important in gauging the benefit of various therapies and tracking patient progress in the context of medication. We choose to focus on predicting symptomatic severity for the purposes of this study.

Making use of the UPDRS and Hoehn and Yahr scales as the current assessment standard, we investigated the viability of using newly proposed quantitative measures for objective evaluation by using machine learning and biological data to predict the recorded assessment scores of the corresponding patient. The data were analysis values of various biomarkers present in patient

cerebrospinal fluid samples. All data, including clinical assessment results, were taken from the Parkinson's Progression Markers Initiative (PPMI) database at http://ppmi-info.org provided by the Michael J. Fox Foundation. The initiative has enrolled almost seven hundred patients and plans to collect data over a period of several years. The data used in machine learning were a subset of the baseline data available at the end of October 2013. The values have since then been updated, and some of the initial baseline data may have been subject to assay batch variability (PPMI Steering Committee, 2014). However, this is the same data set as the one used in a report indicating promising CSF biomarkers published in JAMA Neurology (Kang, et al., 2013). As the study grows, more complete results, including longitudinal data, are expected in the future. At this point, we are only able to present a cross-sectional, early evaluation.

We used both supervised and unsupervised machine learning methods in evaluating predictive power. Supervised learning was used to determine the data's value in re-affirming the results seen in clinical assessments while unsupervised methods were used to determine whether the data exhibited patterns beyond those of the clinical assessments.

The main contribution of this work for the Parkinson's research community is to provide a foray into using machine learning to identify the practicality of potential objective measures. Works that have been published thus far (Shi, et al., 2011) (Tokuda, et al., 2006) (Gerlach, et al., 2012) have only reported relationships in the form of correlation coefficients or p-values representing the significance of the difference between biomarkers values of PD patients and control subjects. While this is a useful first step for flagging potential markers, it does not directly indicate how clinically valuable the biomarker may be when it comes to actual diagnosis or disease staging. We provide further evidence as to whether the correlations described in previous works can actually lead to a quantitative marker by reporting the machine learning results on the PPMI data set that was available to us at the end of October 2013. We also discuss the strengths and weaknesses of each algorithm used. This could aid future research that may seek to re-use similar methods in further analyses of the updated PPMI data. The negative results expose some limits of previously used analysis techniques in showing that statistical relationships are suggestive but not conclusive. However, as the PPMI study is still in early stages of data collection, we hold back on making lasting conclusions about the viability of CSF biomarker predictors. It is

possible that future data that has been adjusted for assay batch variability (PPMI Steering Committee, 2014) or longitudinal patient data may provide different results.

The contents of this thesis are structured as follows. Chapter two presents background information. Chapter three presents related research and previous work in objective measures for PD. Section four describes the methods and algorithms used for investigation. Chapter five presents the results and of testing the machine learning models on a hold out set. A deeper analysis of these results and the implications of our experiments are discussed in chapter six.

2 BACKGROUND

2.1 Biomarkers

A biomarker can be any objectively measurable characteristic of the patient associated with the incidence or progression of a disease. Optimally, a biomarker should be easily accessible, inexpensive, and validated. In this study, we choose to examine biomarkers that are found in Cerebrospinal Fluid (CSF). Since PD is a disease affecting the central nervous system, CSF markers, which surround the brain and spinal cord, have a higher relevance to the pathology of the disease than other bodily fluids. Unlike other static biomarkers such as DNA, CSF markers have the capacity to change with neurodegeneration, making it a sensible choice for tracking progression. Much of the longitudinal biomarker work thus far has been in brain imaging (McGhee, et al., 2013). CSF is collected through lumbar punctures, which while uncomfortable, are less costly than imaging techniques.

2.2 Current Measures for Parkinson's

To investigate the use of certain biomarkers in diagnosing and tracking Parkinson's disease, we wanted to show that these biomarkers can accurately reflect or predict the results of current best practices in these areas. In this study, we chose first to investigate whether the biomarkers could predict incidence of Parkinson's disease by distinguishing between PD patients and control subjects. Further analysis involves using different progressive measures for validation such as the UPDRS and Hoehn and Yahr.

Though the consistency of the UPDRS still depends on patient reporting and examiner skill, it has gained the most acceptance as a rating scale used in assessing Parkinson's disease. The most current version is a revision called the MDS-UPDRS. It was released in 2007 in response to limitations of the original UPDRS reported by a Movement Disorder Society task force. The MDS-UPDRS is comprised of four subscales and includes a thorough evaluation of non-motor symptoms in Subscale 1. Subscale 2 includes questions pertaining to mobility in daily life. Subscale 3 is a monitored motor examination that is scored by a clinician, and Subscale 4 details motor complications. Ratings in all of these subscales range from 0 to 4, with 0 indicating a control subject and 4 indicating severe PD. These ratings do enforce rank order, but they are not

necessarily linear. For instance, a rating of 4 does not indicate twice the severity of a 2. In investigating cerebrospinal fluid biomarkers, we mainly looked for relationships between the bio-specimen analysis data and patient scores from Subscale 3. Values in subscale 3 range from 0 to 108 points, and a healthy control subject typically scores less than 9 (Perimutter, 2009).

The Hoehn and Yahr Rating Scale is a much simpler measure of Parkinson's patients ranging in value from 0 to 5. It is also scored by a clinician after patient examination. The rating is meant to incorporate both motor and daily impairments into a score that serves as a general reflection of disease progression. However, unlike UPDRS, it is not necessarily reflective of symptomatic severity meaning that the rating is uninfluenced by factors such as medication and how a patient is feeling on a certain day. Patients at the same stage on the Hoehn and Yahr scale could have very different symptoms at varying severity levels (Perimutter, 2009). While the extensive, detailed nature of UPDRS makes it best used for longitudinal tracking of a specific patient, Hoehn and Yahr is best used for a broad description of a patient group. The stages are:

0 - Healthy, control subject
1- Unilateral disease with minimal disability
2- Bilateral or midline disease involvement without balance impairment
3- Bilateral disease with mild to moderate disability and postural instability
4- Severely disabling disease; can still walk or stand unassisted
5- Confinement to bed or wheelchair unless assisted

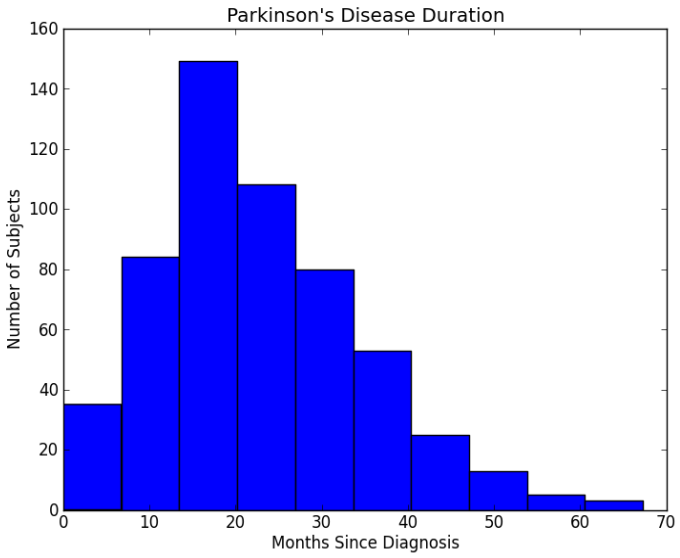
2.3 PPMI Dataset

The Parkinson's Progression Markers Initiative is a five year observational, clinical study funded by the Michael J. Fox Foundation for discovering new markers for measuring Parkinson's disease severity and progression. Clinics in the United States, Western Europe, and Australia are helping to compile the largest collection of clinical, imaging, and bio-specimen data in the Parkinson's community. The initiative consists of data from over 700 patients (216 Control, 453 PD, and 84 SWEDD). SWEDD patients are those that have started experiencing some clinical symptoms of Parkinson's but have not yet shown any loss of dopamine. Though there were over 700 patients enrolled, biomarker data was only available for a small subset (114) of these

patients at the time of the study (Parkinson's Progression Markers Initiative, 2014).

Of the patients enrolled in the PPMI study, the majority are still in the earlier stages of Parkinson's due to the long-term longitudinal goals. The patient age range provides a larger spread from 30-89. However, previous studies indicate that time since disease onset is more relevant to disease severity than age itself (Kempster, O'Sullivan, Holton, Revesz, & Lees, 2010).

Figure 1. Distribution of Patients based on Disease Duration (Months)



PPMI data is available to both industry and academic researchers. For this study, we used the Biospecimen Analysis Results, MDS UPDRS Part III, PD Features, and Patient Status data sheets last updated on October 14, 2013, for our analysis. The Hoehn and Yahr scores used were included in the MDS UPDRS results. Only baseline patient data was readily available and used in learning and prediction.

3 RELATED WORK

3.1 Biomarkers in PD Diagnosis and Progression

There are several papers that have looked at the potential of using cerebrospinal fluid biomarkers in evaluating Parkinson's disease. In a systematic review of biomarkers for Parkinson's disease progression, (McGhee, et al., 2013) reported that 29 of the 183 articles they reviewed were studies examining CSF. All 29 of these studies were cross-sectional. The majority of biomarker studies conducted on longitudinal patient data thus far have been with brain imaging. PPMI seeks to change this by collecting longitudinal CSF data as well. Common markers from previous CSF studies include alpha-synuclein (a-syn), DJ-1, amyloid beta peptide 1-42 (ABeta-42), and total tau (t-tau). Previous investigations looked to correlate these CSF markers with PD severity or progression, approximated by the UPDRS motor scores or other clinical assessments. Papers present a varying amount of detail when it comes to results. Most include p-values from significance testing. (Shi, et al., 2011) (Tokuda, et al., 2006). However, we were unable to find any studies that have gone further to investigate the predictive ability of CSF biomarkers in staging PD patients.

The first results from work done on the same PPMI baseline data set that was used in this study were published by in JAMA Neurology (Kang, et al., 2013). It presented significant lower levels of CSF biomarkers in subjects with PD compared with healthy controls. Specifically, PD patients were found to have lower concentrations of ABeta-42, T-tau, P-tau181, and A-Syn. While the correlation was statistically significant, there was still a marked overlap in the two groups. The paper concluded with a positive outlook on the prognostic and diagnostic potential of CSF biomarkers in early-stage PD but cited the need for further investigation to test the predictive performance of CSF biomarkers in PD progression.

3.2 Machine Learning in Objective Measures

Other initiatives by the Michael J. Fox Foundation to discover objective measures for Parkinson's include a Data Challenge hosted on Kaggle, a machine learning competition platform. This challenge focused not on biomarkers but on passively collected movement data. The competition provided movement data collected from the smartphone of 16 subjects (8

control, 8 PD).

Using the smartphone accelerometry data provided by the Michael J. Fox Foundation along with the phone's compass and GPS information, the winning entry (Brunato, Battiti, Pruitt, & Sartori, 2013) showed that Parkinson's patients could be successfully identified from control subjects with 100 percent accuracy. Acceleration features were extracted by using the compass and GPS information to throw away noisy data. A Support Vector Machine was then used to do a binary classification between controls and PD patients with each patient represented as a data point "cloud."

Though the accuracy of diagnosis using this mobile data was perfect, there is a caveat in that the sample size was extremely small. Because of this, it is possible that the successful prediction is due more to the machine learning algorithm's ability to identify one patient from another rather than a healthy patient from a sick one. However, these results do encourage further study in the area of using motion data gathered from commercial devices to predict and track Parkinson's disease.

4 METHODS

4.1 Data Features and Labels

The CSF biomarkers provided in the PPMI data set were alpha-synuclein, amyloid beta-42, total tau, phosphorylated tau181, and hemoglobin. Before training, the biomarker values were normalized and centered based on the formula below to reduce potential bias from markers with intrinsically higher or lower levels in CSF.

$$\mu_{xi} = \text{average value of feature } X_i$$

$$\mathbf{X_i} = \frac{x_i}{||x_i||} - \mu_{xi}$$

One complete data point **X** is a vector of five elements, making **X** a point in five dimensional space:

$$\mathbf{X} = \langle x_1, x_2, x_3, x_4, x_5 \rangle$$

x_1 = p-tau181 level

x_2 = abeta42 level

x_3 = a-syn level

x_4 = hemoglobin level

x_5 = t-tau level

We chose to use all five of these features since they have all been previously found to be correlated with PD severity. In feature selection analysis, we also calculated correlation coefficients between all possible combinations of the normalized and mean-centered features and the target labels when using a linear estimator. We did this using the entire data set. Using all five features yielded the least residual training error of 0.4658 and the highest correlation coefficient of 0.732.

Figure 2. Correlation Coefficients for Combinations of CSF Features and UPDRS III Scores

One Feature

x_1	x_2	x_3	x_4	x_5
0.701	0.688	0.688	0.689	0.695

Two Features

x_1, x_2	x_1, x_3	x_1, x_4	x_1, x_5	x_2, x_3
0.717	0.701	0.709	0.703	0.691

x_2, x_4	x_2, x_5	x_3, x_4	x_3, x_5	x_4, x_5
0.696	0.699	0.710	0.696	0.708

Three Features

x_1, x_2, x_3	x_1, x_2, x_4	x_1, x_2, x_5	x_1, x_3, x_4	x_1, x_3, x_5
0.718	0.729	0.717	0.717	0.703

x_1, x_4, x_5	x_2, x_3, x_4	x_2, x_3, x_5	x_2, x_4, x_5	x_3, x_4, x_5
0.714	0.715	0.701	0.714	0.711

Four Features

x_1, x_2, x_3, x_4	x_1, x_2, x_3, x_5	x_1, x_2, x_4, x_5	x_1, x_3, x_4, x_5	x_2, x_3, x_4, x_5
0.731	0.719	0.729	0.717	0.716

Five Features

x_1, x_2, x_3, x_4, x_5
0.732

In supervised machine learning, a label is the "correct" or desired output **Y** to be predicted from **X.** In this study, the label was either incidence of PD, the UPDRS score value or Hoehn and Yahr stage.

Before any machine learning, the patients were sorted based on their Hoehn and Yahr staging and one third of the data set was randomly chosen and removed from each group. We employed the Hoehn and Yahr staging to ensure that the ratio of subjects at varying severity levels in the precluded data were representative of the entire data set. The remaining two thirds of the data were used in training the machine learning algorithm and developing a model for the relationship between the CSF biomarkers and incidence of PD. The removed data was then used as the test data for validating efficacy of each derived machine learning model.

4.2 Supervised Learning

All supervised machine learning algorithms were applied and evaluated within the same testing framework described below. Because of the limited amount of data available, it was seen that the partition of the data set in training and cross validation greatly affected the accuracy of each model. Because of this, rather than using cross validation to derive one "best" model for each algorithm we partitioned the data into different training/validation instances for each trial and judged the accuracy of each algorithm as an average over the collective models produced from different partitions.

Machine Learning Framework

1. Randomly set aside one third of the data as test data based on Hoehn and Yahr staging
2. Partition the rest of the data evenly into testing and validation.
3. Train a model and tune its parameters with the validation set
4. Use the model to predict outcome of test data
5. Record predictive accuracy of model on test data
6. Repeat 2-5 for 100 trials

The following machine learning algorithms from the Python scikit-learn package (Scikit-learn,

2013) were used to train different models:

4.2.1 K Nearest Neighbors (KNN)

The principle behind K Nearest Neighbors is that data points are likely to be similar in class to those nearest in distance to it. Classification of new data points are then predicted using the majority class of the K closest neighboring points. This method does not attempt to construct an internal model, but simply needs to store instances of the training data. Here, closeness or distance between data points can be defined using several different metrics. We used the following three:

Euclidean: $\sqrt{(x-y)^2}$

Manhattan: |x-y|

Weighted: weights each neighboring point's vote by the inverse of their distance (1/d) to the target point.

4.2.2 Support Vector Machine (SVM)

Support Vector Machines (SVM) classify data by representing data points in space and finding a separating hyperplane that results in the greatest gap between the classes. One of the greatest advantages of SVMs are their versatility. Kernel functions can map data into higher dimensions, allowing a hyperplane to separate data that wasn't originally linearly separable. SVMs are also computationally and memory efficient as only a subset of the training points, known as support vectors, are needed in calculating the decision function. A disadvantage of SVMs are that they do not provide probability estimates.

4.2.3 Linear Discriminant Analysis (LDA)

Linear discriminant analysis tries to find a linear combination of features that separates two or more classes. Benefits of LDA are that it includes an easily calculated closed-form solution, has no parameters to tune, and it is inherently multiclass. The decision boundary is found by modeling each class as a Gaussian based on the conditional distribution of the training data and calculating their intersection.

4.2.4 Naïve Bayes (NB)

Naive Bayes assumes features are independent and makes predictions based on an underlying probability model. It calculates prior probabilities from the training set and uses Bayes Theorem to output a prediction. Naïve Bayes is typically a very fast algorithm compared to other supervised methods. Decoupling of the features also means that the distribution of each feature can be estimated independently. However, this would not be a good method to use if it is suspected that features are correlated.

4.2.5 Adaboost

Adaboost is an ensemble learning method introduced by Freud and Schapire that combines a group of weaker classifiers into a stronger classifier (Freud & Schapire, 1997). Adaboost uses an iterative method that picks the best weak classifier for repeatedly reweighted versions of the training set on each round. The weights on the training data points are adjusted based on whether the classifier chosen in the previous round classifies that point correctly. Correctly classified training points are given reduced weight in future rounds while misclassified points are given a heavier weight. On the next round, the algorithm will prioritize the correct classification of points that are weighted more heavily. In the final classifier, each weak classifier gets a weighted "vote" for prediction. For this thesis, the weak classifiers used were the previously tested supervised methods described above. A weak classifier is defined as a classifier that performs only slightly better than random. Python code for our Adaboost adaption can be found in Appendix A.

4.3 Unsupervised Learning

Unsupervised learning was used to examine the potential of patterns in the data that weren't exposed in existing clinical assessments. The following unsupervised learning algorithms from the Python scikit-learn package were used on the CSF biomarker data. In this case, all data including the test set were used in learning. We first attempted to model the group as two distributions to represent PD and control subjects. It was seen that leaving out the test set did not largely influence results. Both algorithms tested were inclined to cluster the majority of the data into one group while selecting an outlier for the remaining cluster.

4.3.1 Gaussian Mixture Models (GMM)

GMMs group data points by modeling them as samples from one or more Gaussian distributions. GMMs can sometimes be difficult with smaller data sets as it provides limitations to calculating a covariance matrix.

4.3.2 K-Means

K-Means seeks to iteratively separate the data into K clusters (K must be specified). To begin, K centroids are initialized randomly and data points are grouped with the closest centroid. These sorted data points are then used to calculate a new centroid that better represents those points. This process continues iteratively until the centroids converge. Note that the results of K-Means are not necessarily deterministic due to the randomization at the beginning. However, with our data set, we ran the algorithm many times for K = 2 and achieved the same results each time.

5 RESULTS

5.1 Supervised Learning

Results are reported over 100 trials of model training on different partitions of the training data. Reported prediction accuracy indicates performance of the model on guessing the diagnosis of each patient in the isolated test data set. Note that most of the algorithms can do any better than randomly guessing whether each patient has Parkinson's or not. Adaboost achieves a slightly better result by combining the strengths of the previous algorithms.

5.1.1 K Nearest Neighbors

(K value determined individually for each model through cross validation)

Distance Metric	**Avg. Prediction Accuracy**	**Standard Deviation**
Manhattan Distance	0.507	0.058
Euclidean Distance	0.507	0.047
Weighted Euclidean	0.505	0.053

5.1.2 Support Vector Machine

Kernel	**Avg. Prediction Accuracy**	**Standard Deviation**
Linear	0.51	0.07
Radial	0.51	0.07

5.1.3 Linear Discriminant Analysis

Avg. Prediction Accuracy: 0.510

Standard Deviation: 0.080

5.1.4 Naive Bayes

Avg. Prediction Accuracy: 0.460

Standard Deviation: 0.141

5.1.5 Adaboost

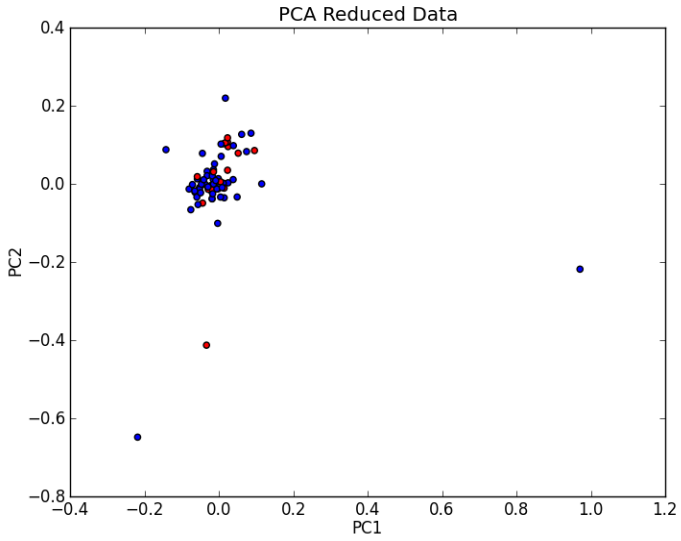
Avg. Prediction Accuracy: 0.697

Standard Deviation: 0.137

5.2 Unsupervised Learning

For better visualization of the unsupervised learning results, we first performed a Principal Component Analysis (PCA) Reduction of the data set into two dimensions. PCA is a process that aims to transform the axis of the data set to reduce dimensionality. It was also performed using tools in the Python scikit-learn package. From the figure below, it is already evident that there is no clear separation between the principal components.

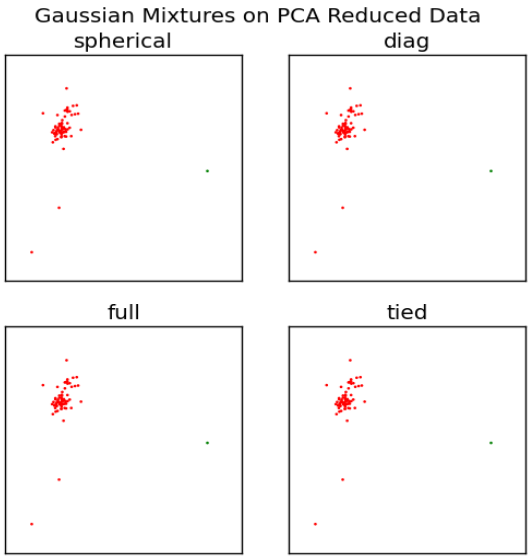
Figure 3. PCA Reduced CSF Biomarker Data Set



5.2.1 Gaussian Mixture Models

The results of modeling the data with Gaussian Mixtures showed the data belonging to one Gaussian regardless of the covariance matrix type.

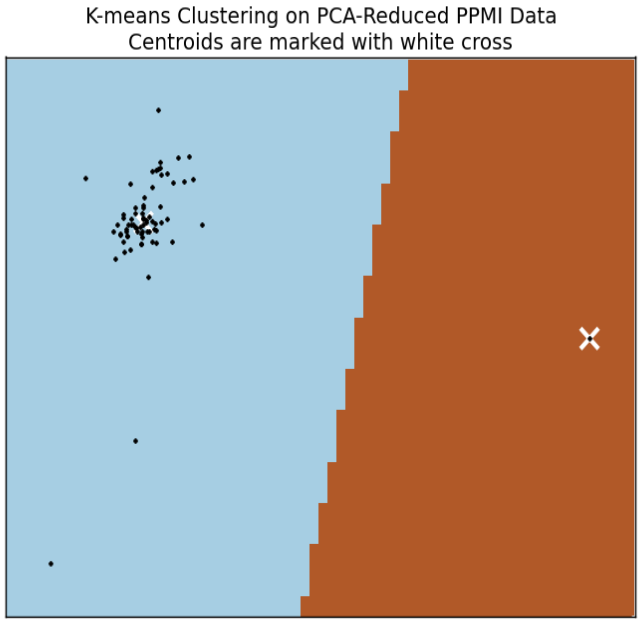
Figure 4. Gaussian Mixtures Results on PCA Reduced Data



5.2.2 K-Means

The results of using the K-Means algorithm on the principle components of the data are shown below. Because of the inseparability of the data, all but one outlier are in the same cluster when we choose K = 2.

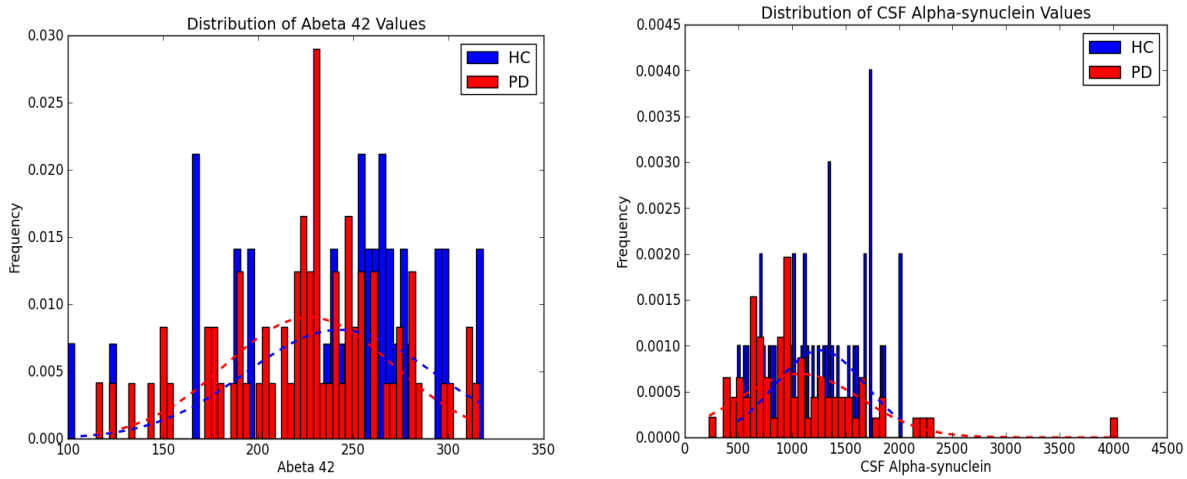
Figure 5. K-Means Clustering Results on PCA Reduced Data

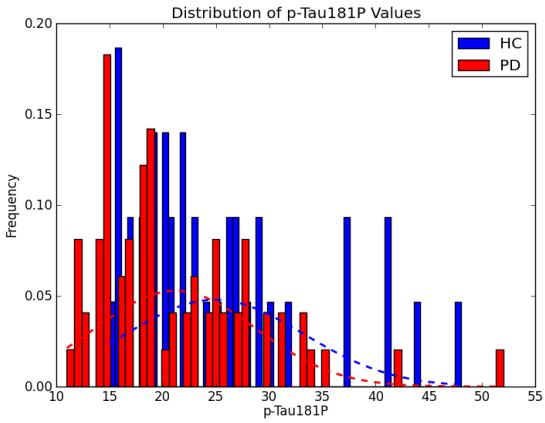


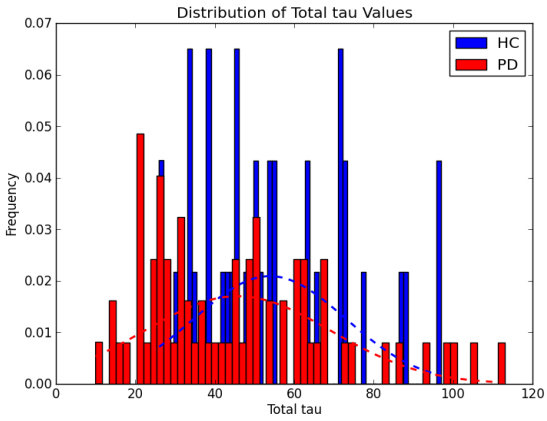
6 DISCUSSION

It can be seen from the results that we found very little predictive value in the baseline CSF biomarker data. In supervised learning, most of the results hovered around 50% accuracy, giving us no more insight to PD than a random guess. Several of the algorithms would guess that all the test patients were of the same class. Going further to examine the ability to predict disease severity does not make sense if diagnosis cannot successfully be predicted. The best supervised learning results came from using the Adaboost ensemble learning method. Since Adaboost is a composition of the other weak classifiers, its success depends on how well the different weak classifiers can lean on the strengths of some and make up for the weaknesses of other. In our case, the ensemble of weak learners was composed of various supervised learning methods that have been proven to work well on different types of data in the past. Because of this, it may be that each weak learner classifies different subsets of the data successfully and the combination of classifiers does a significantly better job of classifying the test set than any one classifier alone. Unsupervised learning results also struggled with the overlap in classes, finding the best results to contain all data points in a single cluster or Gaussian distribution. Examining the distributions of PD patients versus control subjects based on some of the CSF biomarkers indicates that the two groups may not be separable.

Figure 6. PD and Human Control Patient Distributions for CSF Features







While the results of this study are negative, both the research of objective measures and the PPMI database are in very early stages. There are several factors that may have contributed to the negative results and motivate further investigation when there is more data available.

The biggest factor that may have affected the results is assay batch variability between 2011 and 2013 in the research assays by the PPMI study centers. This was reported by PPMI in early May 2014 (PPMI Steering Committee, 2014)**,** and the data has since then been modified and updated to reflect this variability.

Another big factor affecting results may have been inter-patient variability. It is possible that there is far more variability between the cross-sectional data of two patients due to individual differences rather than the stage of the disease. This would makes it implausible to find patterns for progression in the baseline biomarker data. At the time of the study, there was no longitudinal data available. However, the recent update includes 300 longitudinal measures that can be used in further analysis of the predictive ability of CSF biomarkers on a patient by patient level. It's also possible that patient biomarker data will vary so much from patient to patient that a universal objective measure will be hard to achieve. Perhaps each patient will need to have their own calibrated scale based on their baseline measures.

The sample size and spread of the patient data used in machine learning was also small. We saw that sample size having a large effect on the results as different partitions of the training and validation data yielded various results. There was only biomarker data for about 100 patients available. With a third set aside for testing, training was done on around 70 data points. Though

over 700 subjects had already been enrolled in the PPMI, only this small portion of the biological analysis data was available for use. There were another 300 or so samples that contained incomplete biomarker levels for only a subset of the five CSF features. The patient distribution was also skewed towards earlier disease stages. While this skewed patient distribution potentially provides a good way of testing extrapolative ability of a model to predict later stages, the small range provides limitations on the efficacy of certain machine learning models and presents challenges in assessing validity. Overtime, more data will become available as the PPMI database becomes more complete and future tests in learning and prediction may yield different results.

7 CONCLUSION

This paper described the first exploration of using machine learning to evaluate the predictive ability of CSF biomarkers in staging PD. The biomarker data was obtained from the PPMI, a five year long longitudinal study. At the time of the research, only baseline biomarker data was available for a subset of the 700 patient study. Most of the PD patients were in an early stage of the disease.

Previous studies have indicated the potential of using CSF biomarkers to predict incidence and severity of PD. These studies reported a statistically significant relationship between these markers and diagnosis of PD, indicating that further research could be done to evaluate whether or not they could be used to predict disease severity. Using the levels of CSF a-syn, abeta-42, t-tau, p-tau181, and hemoglobin of 114 patients in the PPMI data set along with machine learning, we aimed to discover whether or not these biomarkers could lead to an objective measure.

Employing a variety of machine learning algorithms, it was seen that most of them could not correctly predict diagnosis with results significantly better than random. Adaboost performed the best at around 69% accuracy. While this is not clinically useful, it does suggest the presence of some information in CSF biomarkers that may be useful for future objective measures. Unsupervised attempts to classify the data also struggled to identify any patterns. These negative results could be a product of limitations from the data set or inter-patient variability. Generally, there are still several avenues to be explored in using CSF biomarkers for objectively measuring PD. Immediately available ones include the newly updated PPMI data sets. Even then, it can be seen that more than a statistically significant relationship between biomarkers and clinical assessments is needed. Further analyses of predictive ability, similar to those performed in this study, will be necessary in the search for objectives biomarkers for the clinical purposes of determining disease stage or monitoring progression.

Bibliography

Brunato et al., M. (2013). *Supervised and unsupervised machine learning for the detection, monitoring, and management of Parkinson's disease from passive mobile phone data* . LIONSolver, Inc.

PPMI Steering Committee, (2014). *Cerebrospinal Fluid Data Posted to PPMI.* PPMI. Retrieved from http://www.ppmi-info.org/2014/05/cerebrospinal-fluid-posted-to-ppmi/

Gerlach et al., M. (2012). Biomarker candidates of neurodegeneration in Parkinson's disease for the evaluation of disease-modifying therapeutics. *J Neural Transm*, 39-52.

Kang et al., J.-H. M. (2013). Association of Cerebrospinal Vluid Beta-Amyloid 1-42, T-tau, P-tau181, and Alpha-Synuclein Levels With Clinical Features of Drug-Naive Patients With Early Parkinson Disease. *JAMA Neurology*.

Kempster et al. PA, O. S. (2010). Relationships between age and late progression of Parkinson's disease: a clinico-pathological study. *Brain*(133), 1755-1762.

McGhee et al., D. (2013). A systematic review of biomarkers for disease progression in Parkinson's disease. *BMC Neurology*, 13:35.

PPMI (2014). *Parkinson's Progression Markers Initiative.* Retrieved from http://www.ppmi-info.org/about-ppmi/study-faq/

Perimutter, J. S. (2009). *Assessment of Parkinson Disease Manifestations.* St. Louis: Curr Protoc Neurosci.

Shi, M. P. et al. (2011). Cerebrospinal Fluid Biomarkers for Parkinson Disease Diagnosis and Progression. *Ann Neurol*, 570-580.

Takahiko, T. et al. (2006). Decreased alpha syn-nuclein in ceregrospinal fluid of aged individuals and subjects with Parkinson's disease. *Biochemical and Biophysical Research Communications*, 162-166.

Appendix A: Adaboost.py

```
import numpy as np
import math
from scipy.stats import mode
from itertools import combinations

import machine_learning as ml
import plotting as pt
import lda
import naive_bayes
import svm
import knn


def computeAlpha(error):
	return (0.5) * np.log2((1.0-error)/error)

def getMisclassifiedSamples(model, targets, samples):
	misclassified_ind = []
	for i,s in enumerate(samples):
		if model.predict(s) != targets[i]:
			misclassified_ind.append(i)
	return misclassified_ind

def predictEnsemble(ensemble, sample):
	res = 0
	for alpha,classifier in ensemble:
		pred = alpha * classifier.predict(sample)
		res += pred
	if res > 0:
		return 1
	if res < 0:
		return (-1)

def createClassifiers(x,y):
	classifiers = []
	knn_m = knn.trainClassification(x,y,17,'manhattan')
	knn_e = knn.trainClassification(x,y,17,'euclidean')
	l = lda.trainClassification(x,y)
	svm_l = svm.trainClassification(x,y)
	svm_r = svm.trainClassification(x,y,'rbf')
	nb = naive_bayes.trainClassification(x,y)

	classifiers.append(knn_m)
	classifiers.append(knn_e)
	classifiers.append(l)
	classifiers.append(svm_l)
	classifiers.append(svm_r)
	classifiers.append(nb)

	return classifiers
```

```
def runAdaBoost(rounds, X, Y):

	num_val = 10
	classes = 2

	(trainX, trainY, validX, validY) = ml.getRandomSamples(X, Y, num_val)
	weak_classifiers = createClassifiers(trainX, trainY)

	weights = [1.0/len(trainX)] * len(trainX)

	final_classifier = []

	for i in range(rounds):

		best_classifier = None
		best_weighted_error = 1.0
		best_errors = None

		#--IDENTIFY BEST CLASSIFIER--#
		for classifier in weak_classifiers:

			errors = getMisclassifiedSamples(classifier, trainY, trainX)
			total_error = 0

			for wrong in errors:
				total_error += weights[wrong]

			if total_error < best_weighted_error:
				best_weighted_error = total_error
				best_classifier = classifier
				best_errors = errors

		#--CALCULATE ALPHA--#
		if best_weighted_error != 0.0:
			alpha = computeAlpha(best_weighted_error)
			#print "ALPHA: ", alpha

		#--REWEIGHT--#
			for k,weight in enumerate(weights):
				if k in best_errors:
					new_weight = weight / (2 * best_weighted_error)
				else:
					new_weight = weight / (2 * (1-best_weighted_error))

				weights[k] = new_weight

			final_classifier.append((alpha, best_classifier))
		else:
			final_classifier = [(1,best_classifier)]
			break

	return final_classifier
```