Decomposing Parkinson's Disease

Using Supervised Learning and Statistical Analysis

Héctor Manuel Ortiz-Mena, Bachelor of Fine Arts[1], Rohan Shroff, Bachelor of Science[2], Vincent Gacutan, Bachelor of Science in Electronics and Communications Enginneering[3]
[1]Georgia Institute of Technology, Atlanta, GA, U.S.A.

Abstract

This report presents the results of utilizing supervised learning and statistical analysis to decompose and understand, Parkinson's disease, which has no proper diagnosis and progresses differently from patient to patient. The goal is to achieve a better understanding about the factors contributing to the diagnosis of this disorder to be able to construct an efficient metric to diagnose patients with high probability of success. The research focuses on feature selection using randomized decision trees and its success was measured on supervised learning classifiers. Furthermore, we explored the distributions of the found features using basic statistics to determine how they are related and if any contribute to a positive diagnosis. Our results show that with 0.995 accuracy and AUC, the data set can be classified for each patient as control or having Parkinson's based on whether the patient produces non-zero values for NHY. Other features were also found to be relevant.

Introduction

Motivation

Parkinson's disease is a motor function disorder. This disease leads to the deterioration of motor functions and it is not easily diagnosed. By using machine learning we might discover a solution to this issue.

Symptoms of Parkinson's disease can include difficulty with motor movements and facial expressions. To express the symptoms, doctors look to see if the patient's expression is animated. Their arms are observed for tremor, which is present either when they are at rest, or extended. Is there stiffness in their limbs or neck? Can they rise from a chair easily? Do they walk normally or with short steps, and do their arms swing symmetrically? How quickly are they able to regain their balance[1]?

To diagnose this illness, the doctor will administer an exam. However, there is no standardized test, and doctors will devise their own way of administering their version of this examination. By using machine learning, we might be able to standardize and even create a basic detection system.

This research uses a data set composed of Biomarkers. A Biomarker is a sample of biological fluids which can include plasma, serum and cerebrospinal fluid[2]. After retrieving these samples from the data sets provided, the subgroups can then be divided into clinical, biochemical, genetic and imaging. By combining these subgroups it can be administered within the ML program and a true synopsis can be made. This can then lead to a more accurate method of diagnosing.

Related Works

"One of the major problems in early recognition of its symptoms is that proper diagnosis is unavailable... The cause of PD is still not known exactly and research is being carried out all over the globe.[3]" Researchers at The Australian National University(ANU) are currently using machine learning to help determine the progression of Parkinson's disease. "Currently when someone is diagnosed with Parkinson's disease it is difficult to determine what type of Parkinson's they have or how quickly the condition will progress.[3]"

Dr. Deborah Apthrop of the ANU Research School of Psychology states that "the thing about Parkinson's disease is that some people can be OK for quite a long time, while others progress more rapidly." She adds that "there are different types of Parkinson's that can look similar at the point of onset, but they progress very differently. We are hoping the information we collect will differentiate between these different conditions. It is her hope that "ultimately we'd like doctors to be able to conduct tests that can predict how the disease is likely to progress.[3]" Her research involves measuring brain imaging, eye tracking, visual perception and postural sway.

Other research on Parkinson's disease involving Machine Learning include the work done by Indrajit Mandal and N. Sairam of the School of Computing at SASTRA University. "Their research presents an enhanced prediction accuracy of diagnosis of Parkinson's disease (PD) to prevent the delay and misdiagnosis of patients using a proposed robust inference system.

These methods include sparse multinomial logistic regression, rotation forest ensemble with support vector machines and principal component analysis, artificial neural networks and boosting methods.[4]"

There is also research focused on speech data such as the work done by Hanzan, Hilu, Manevitz, Ramming and Sapir and Zhang, Yang, Liu, Wang, Yin, Li, Qui, Zhu and Yan. Using two distinct data sets of healthy controls and patients with early or mild stages of Parkinson's disease, they showed that machine learning tools can be used for the early diagnosis of Parkinson's disease from speech data. Their study showed that early detection of PD from speech data seems to be feasible and accurate with results approaching the 90% mark in two different data sets. In addition they showed that "while the training phase of machine learning process from one country can be reused in the other; different features dominate in each country; presumably because of language differences.[5]"

The second speech related research involving Parkinson's and Machine Learning showed that "the use of speech data in the classification of Parkinson's disease has been to provide an effect, non-invasive mode of classification." "This study showed that the proposed method could improve PD classification when using speech data and can be applied to further studies seeking to improve PD classification methods.[6]

Method

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

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. The choice of evaluation metric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods[7].

Scikit-Learn provides several classes in the sklearn.feature_selection module than can be used for feature selection/dimensionality reduction on data sets to determine which features are relevant for classification. The feature method selection used in this project uses an ensemble of randomized decision trees using Gini impurity as the metric to determine which features are important.

Data set

The diagnosis of Parkinson's disease (PD) currently relies upon clinical criteria. No existing laboratory test diagnoses PD or gauges the effectiveness of a treatment on underlying disease processes. Molecular imaging techniques are helpful for diagnosis, but are costly and cannot distinguish among different parkinsonian syndromes. Current methods for diagnosing, treating, and prognosticating PD are inadequate and would be greatly improved by the discovery and validation of biomarkers, namely, those "characteristics that are objectively measured and evaluated as indicators of normal biologic processes, pathogenic processes or pharmacological response to a therapeutic intervention."

In general, biomarkers involve measurements of biological samples (e.g., plasma, serum, cerebrospinal fluid (CSF) and biopsy) or measurements using brain imaging techniques to decipher changes in brain structure and function. Biomarkers of PD are diverse and can be categorized into four main subgroups: clinical, biochemical, genetic and imaging. When one group is considered alone the utility of the biomarker is often limited, but when combined and considered collectively, biomarkers for PD may be more useful.

The data set we are going to use for this project is from Biofind (Fox Investigation for New Discovery of Biomarkers in Parkinson's Disease). BioFIND is an observational clinical study designed to discover and verify biomarkers of Parkinson's disease (PD). Biofind study addressed these needs by establishing a cohort of moderate to advanced PD patients, typical of those PD who would most likely be encountered in the clinical practice or a trial setting and developing a repository of standardized, rigorously collected clinical data and biospecimens for biomarker research. Figure 1 is a sample demographics and available Biospecimen from unique individuals based on the Biofind data set.

(A). Biomarkers

Biomarkers will be used in this research. A Biomarker, is a sample of biological fluids which can be divided into categories to narrow down the synopsis of this illness. The Biomarkers that will be used within this research in conjunction with a written algorithm will lead our result in this study.

The Biomarkers Definitions Working Group has defined a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention.[8]"

(B). Classification and Features:

The classifications used have a large effect with our outcome. Classifications can be connected to the biomarkers. Classifications of Parkinson's disease include: clinical, biochemical, genetic and imaging.

The data that has been provided on the classification set will be derived from the hundreds of features given in the BioFind data. The features which in turn describe classifications, are housed within a tree structure. Classifications will include everything described earlier; blood, saliva, csf and biopsy and hundreds of others that can be used to find and algorithmic diagnosis for PD.

Overall, the features that are listed, will help the research put the information gathered into different classifications. It is also evident that the data influences the selections made by the algorithms and trees that are in place. The next section will describe the use of the data sets and the algorithms that come into play.

(C). BioFind Data Sets

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

Pipeline

Figure 2 shows our pipeline for discovering the factors contributing to the diagnosis of Parkinson's. The process begins by merging Biofind data sets stored as CSV files in Hadoop Hive from categories such as Family History, General Medical History, Neurological Exams, Cognitive Assessment, Biochemical origin, REM Sleep Disorder Questionnaire , Screening Demographics and Socio Economic data. Next, the output of this step was imported into Hadoop Pig for preprocessing and to generate an ETL in SVMLight format and other secondary data required to visualize and understand our results. Once the data is preprocessed by PIG, we began model testing using Scikit-Learn and Python. All 2371 features and 222 patients were divided into 117 patients for training and 45 patients for testing.

The first stage in our model testing process used three unoptimized supervised classification models: logistic regression, support vector machine and decision tree. These unoptimized model served as a raw guideline to get quick results and to confirm that our data preprocessing was successful. These produced results for the full training set and testing sets.

Next, began the process of optimizing and incorporating more models using GridSearchCV and 3 folds of cross-validation using AUC as the measure of classification performance. This process allowed us to use the best parameters for the included models producing the highest cross-validation AUC scores for each model. In this stage, we included an optimized logistic regression classifier, a support vector machine (SVC, LinearSVC and NuSVC), a decision tree and a random forest. Originally, we were going to select the best model out of these, but in stage 2 of this project, the final stage with more features from more data sets, we discovered that all tested model produced results that were almost equally successful based on our cross-validation AUC score. Therefore, we decided to continue our next steps with all models. We also tested these models after model optimization with the full training and testing prior to feature selection using accuracy, AUC, precision, recall and F1-score.

After model optimization, we began the process of reducing our feature set. Prior to this point, all our results were based on all available features to use these results as a control set to compare with our reduced feature set. To compute the feature importance score of all our features, we used the Extra Trees Classifier in Scikit-Learn which is composed of randomized decision trees using Gini impurity as the metric to define feature importance. In stage 1, our draft, we discovered 405 important features using the default threshold of the average feature importance. However, in stage 2, we decided to see how the AUC of one of our optimized models, the decision tree, was affected by reducing the feature from full 2371 to a number where the AUC of the optimized decision tree began to lower from least important to very important as reported by the randomized decision trees.

After feature selection, we output 9 of the most important features to R to perform Statistical Analysis on the features to explore their distributions and to see if any relationship or boundaries existed among them.

Finally, out of the statistical analysis, we came up with the results and conclusions described next. This process was done on three different local clusters in three different parts of North America.

Experiments and Results

As demonstrated on figure 3, all three, the logistic regression, support vector machine and decision tree produced perfect accuracy, AUC, precision, recall and f1-score using all features on the training set. Initially, we were anticipating overfitting. However, on figure 4, we noticed that our models are able to generalize on our testing set, as all unoptimized models also generated perfect test results using all the same metrics used on the training set. At this point, it seems that with all 2371 features, all our models produced perfect results.

During the model optimization stage, as seen on figure 5, we noticed that our cross-validation results using all features and no feature selection are slightly lower. Our theory for the reason that these are slightly lower it is that 3 folds is not producing enough data to construct the same results as shown previously. However, as shown in Figure 6, all our models were able to again produce perfect test results. At this point, we believe that our results might need more patients than are available to confirm our results.

Figure 7 shows what happens to one of our optimized models, the decision tree, as we reduced the data set from 2371 features to 1 feature. No change. Therefore, we concluded a this point that out of the 2371 features, only 1 is needed.

We then computed our classification metrics again with only 1 feature. Figure 8 shows that on the testing set all models are still able to produce perfect results. Therefore, this one feature is as important as using 2371 features.

Finally, we decided to confirm our findings by exploring the distributions of 9 of the most import features returned by the randomized decision tree in R. Figure 9 and 10 show that NHY, the single most important feature, is able to split the data into Parkinson's and control except for that one outlier. Other features such as NP3RIGLL, NP3PRSPL, NP3RTCON, NP3FTAPL, NP3GAIT, PN3RIGRL and NP3SPCH also seem to behave the same as NHY with slightly less success independently as shown in the accuracy column on figure 10.

We also decided to reconfirm our findings by seeing if the distribution of ranges for each feature for Parkinson's and controls overlap. In an ideal setting, we would like to see no overlap by the control and Parkinson's distributions and indeed, without counting a few outliers, the data can be easily separated as Parkinson's or control by taking the min of each feature of the Parkinson's patient group and using it to classify the data manually as Parkinson's. For instances, as shown in Figure 9, if a patient has NHY greater than zero, then the patient has Parkinson's, similarly to the other features in descending order. Each of the features' success using this simple procedure is shown in the accuracy column. The threshold used on all of these was greater than zero for all Parkinson's patients. It would seem that for all of these features generally, scoring higher than zero is enough to be classified as Parkinson's, but in some rare cases a control patient can produce a value higher than zero. False positives?

Our findings show, that with 0.99549550 accuracy, one can manually classify the data as either Parkinson's or not. Alternatively, any of our models can perform the same using just the NHY feature with near positive metrics. This is an improvement to our previous AUC of 0.94 and 0.93 accuracy in our draft phase 1. More patients would need to be tested to confirm as we are aware that our testing set is only comprised of 45 patients.

Discussion

The experimental results shown above indicate that we have found a good selection of features for diagnosis Parkinson's disease. Previously, the data favored the decision tree, now all of our chosen models return similar near perfect results. It would seem that NHY is a good indicator alone for determining Parkinson's. The following NP3RIGLL, NP3PRSPL, NP3RTCON, NP3FTAPL, NP3GAIT, PN3RIGRL and NP3SPCH, 7 of the 9 features shown on figure 9 and 10, are also good indicators, however not as strong as NHY as determined by their feature importance value or our brute-force manual classification measure done in R.

We were not quite successful in reverse engineering a rule of a decision tree; it would seem that these features work independently. However, not enough experimentation was done to confirm this theory and it would seem that testing a value higher than zero for most of these features is enough to be diagnosed as Parkinson's on average. Again, NHY seem to be the most certain indicator.

To confirm our findings, we suggest for further research to include more patients. Another useful test would be to reduce the data set by first eliminating features that yield high importance values first rather than eliminating less important features and keeping the most important. That could reconfirm whether our found feature is indeed the most important. Due to our near perfect results, we believe that we have succeeded in finding a subset of features that work as strong indicators for diagnosis Parkinson's based on the limited amount of patients available to us.

Conclusion

The results of our investigation show how utilizing supervised learning, in particular, feature selection with randomized decision trees, and statistical analysis can help narrow down the factors contributing to the diagnosis of a medical disease, Parkinson's disease, which has no proper diagnosis and progresses differently from patient to patient. Our pipeline, composed of big data tools such as the Hadoop framework, Scikit-Learn and Python, is successful in discovering which features contribute to the diagnosis of Parkinson's. The new 112 features added since the draft phase improved the metric of all our chosen model to near perfection. Due to the results not consistently producing perfection, we claim that our method has an accuracy of 0.995. Out of the 2371 features, one was enough to classify the available patients as having or not having Parkinson's disease. Seven others were also found to be somewhat as strong indicators. Finally, we strongly believe that due to the successful results of all our models, there exist a function that maps features to classes and NHY is at its core.

References

1. Parkinson's Disease Foundation (PDF) [Internet]. Diagnosis | Parkinson's Disease Foundation (PDF). [cited 2017 Oct 15]. Available from: http://www.pdf.org/diagnosis

2. Delenclos M, Jones DR, McLean PJ, Uitti RJ. Biomarkers in Parkinson's disease: Advances and strategies [Internet]. Parkinsonism & related disorders. U.S. National Library of Medicine; 2016 [cited 2017 Oct 14]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5120398/

3. Machine learning to unlock Parkinson's disease mystery [Internet]. ANU. The Australian National University; 2016 [cited 2017 Oct 14]. Available from: http://www.anu.edu.au/news/all-news/machine-learning-to-unlock-parkinson%E2%80%99s-disease-mystery

4. Mandal I, Sairam N. New machine-learning algorithms for prediction of Parkinson's disease. New machine-learning algorithms for prediction of Parkinson's disease [Internet]. 2012 [cited 2017 Oct 14];:1–. Available from: https://www.researchgate.net/publication/234131546_New_machine-learning_algorithms_for_prediction_of_Parkinson%27s_disease

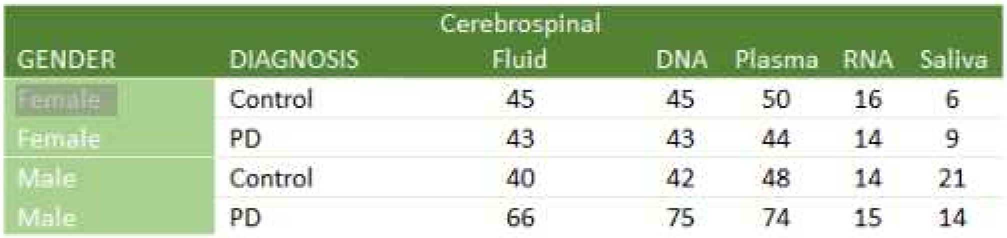
5. Hazan H, Hilu D, Manevitz L, Ramig L, Sapir S. Early diagnosis of Parkinson's disease via machine learning on speech data. Early Diagnosis of Parkinson's Disease via Machine Learning on Speech Data [Internet]. 2012 Nov [cited 2015 Oct 14];:1–. Available from:https://www.researchgate.net/publication/243632051_Early_diagnosis_of_Parkinson%27s_disease_via_machine_learning_on_speech_data

6. Zhang H-H, Yang L, Liu Y, Wang P, Yin J, Li Y, et al. Classification of Parkinson's disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples. BioMedical Engineering OnLine [Internet]. 2016;15(1):1–. Available from: https://www.researchgate.net/publication/243632051_Early_diagnosis_of_Parkinson%27s_disease_via_machine_learning_on_speech_data

7. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. An Introduction to Variable and Feature Selection [Internet]. 2003 Mar [cited 2017 Oct 14]; Available from: http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

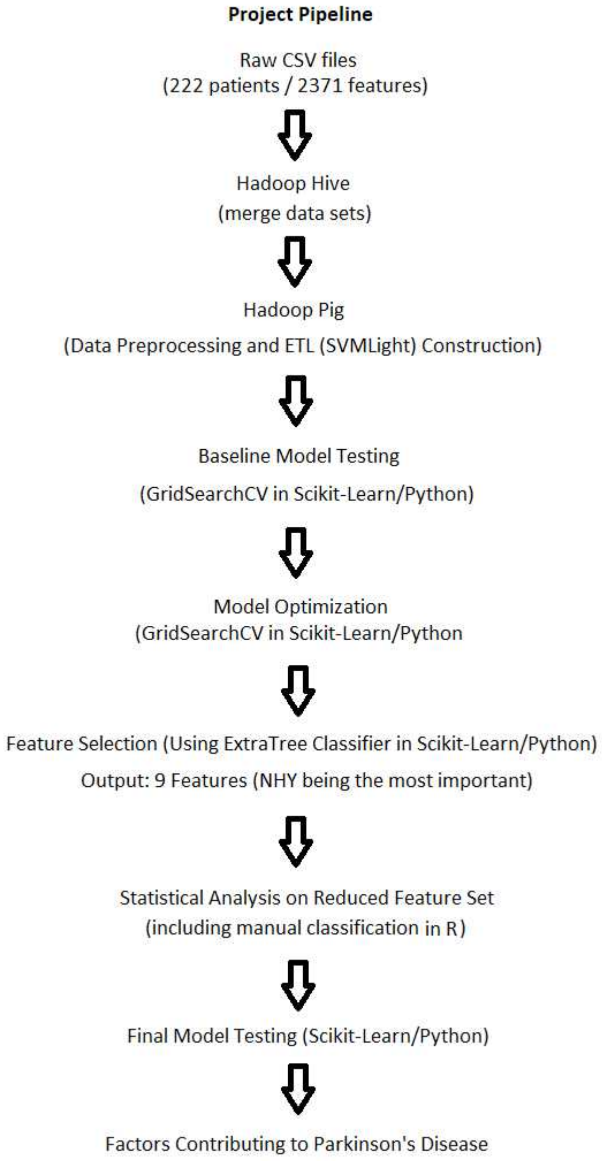
8.Biomarkers Definitions Working Group Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.Clin Pharmacol Ther. 2001;69:89–95. PubMed]

Appendix



GENDER	DIAGNOSIS	Cerebrospinal Fluid	DNA	Plasma	RNA	Saliva
Female	Control	45	45	50	16	6
Female	PD	43	43	44	14	9
Male	Control	40	42	48	14	21
Male	PD	66	75	74	15	14

Figure 1. Biomarker data



Project Pipeline

Raw CSV files
(222 patients / 2371 features)

Hadoop Hive
(merge data sets)

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

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

Model Optimization
(GridSearchCV in Scikit-Learn/Python

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

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

Final Model Testing (Scikit-Learn/Python)

Factors Contributing to Parkinson's Disease

Figure 2. Pipeline

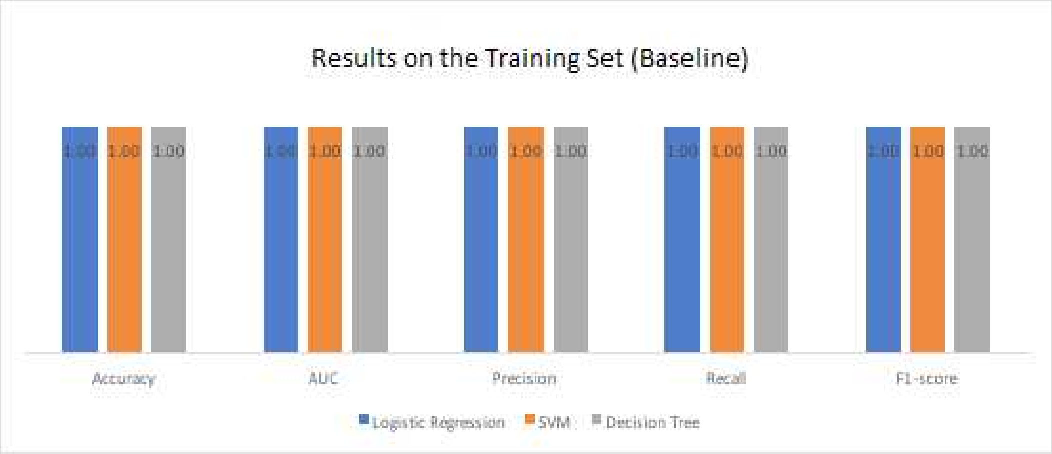


Figure 3. Baseline using full training set and no parameter optimization results (2371 features)

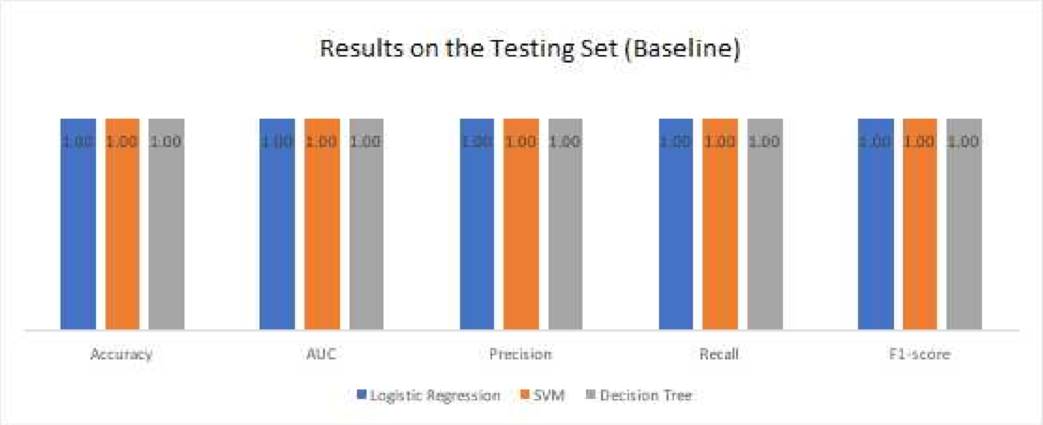


Figure 4. Baseline using full training/test sets and no parameter optimization results (2371 features)

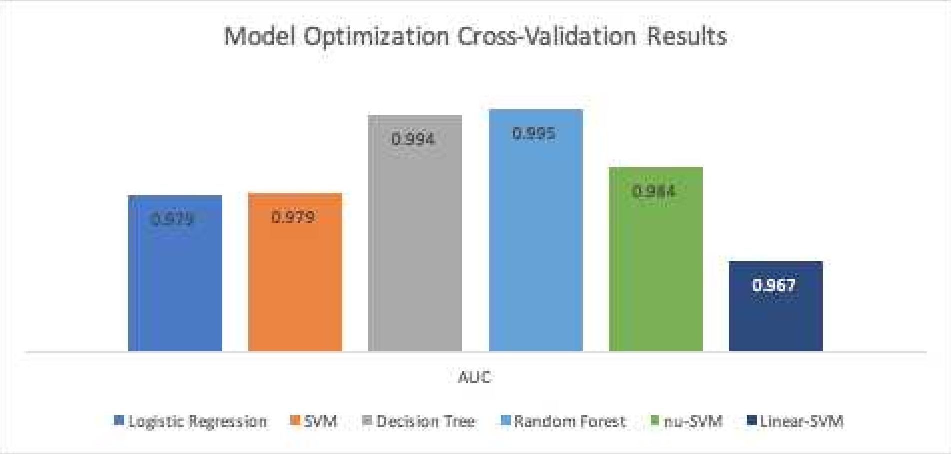
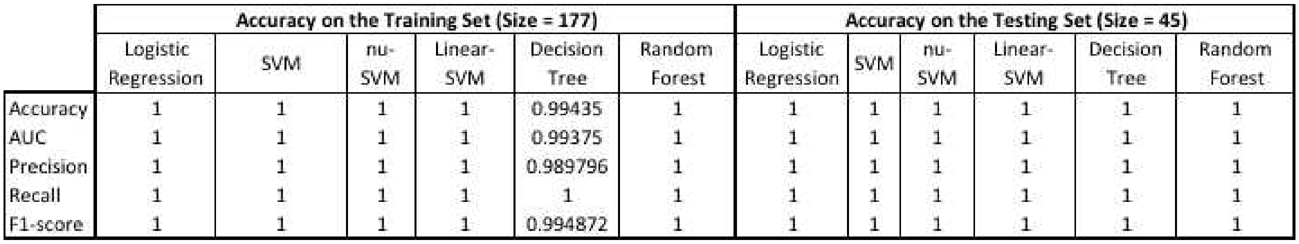


Figure 5. Model optimization (using cross validation (3 folds) from full training set and no feature selection) (2371 features)



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

Figure 6. Using full training/test sets and optimized models with no feature selection (2371 features)

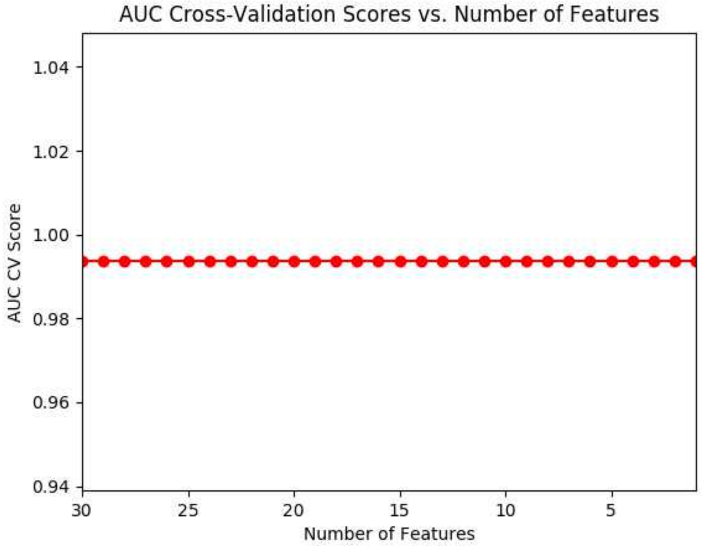
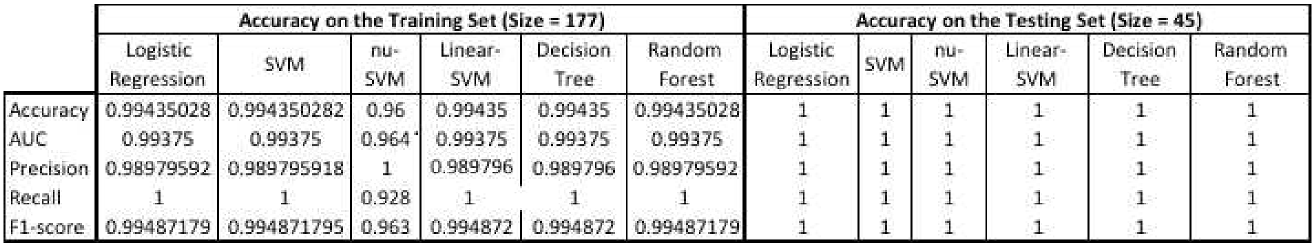


Figure 7. Effect on AUC on optimized decision tree model of reducing number of features



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

Figure 8. Using full training/test sets and optimized models with feature selection (1 feature)

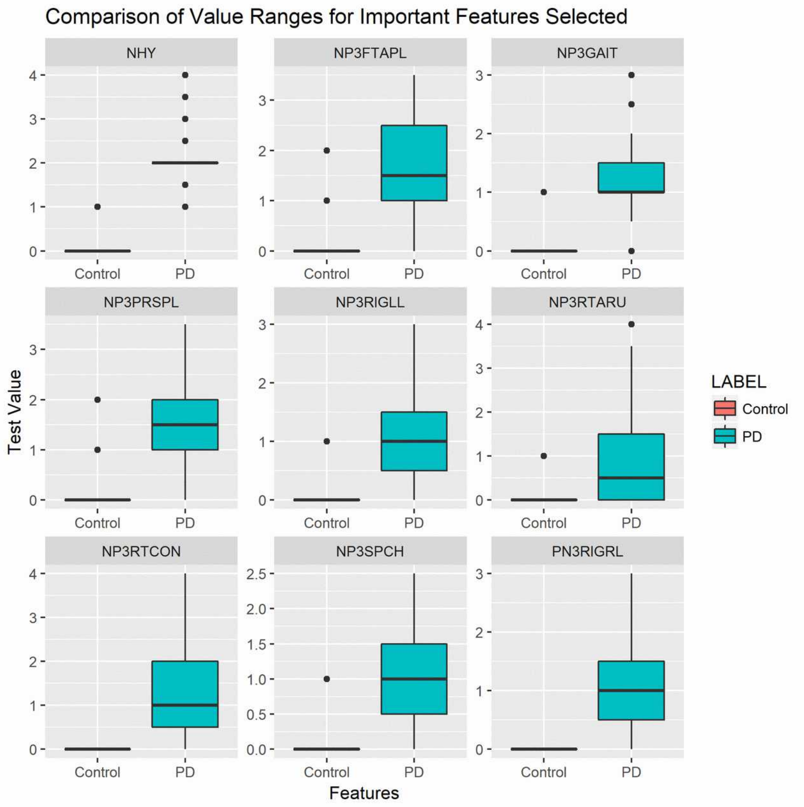


Figure 9. Ranges of 9 of the most important found features

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

Figure 10. Nine of the most important features (feature importance and manual classification accuracy)