A PREDICTIVE MODELLING APPROACH IN THE DIAGNOSIS OF PARKINSON'S DISEASE USING CEREBROSPINAL FLUID BIOMARKERS

Prathima Lakmala, Dr. Josette Jones, Patrick Lai

School of Informatics and Computing - IUPUI

IUPUI
SCHOOL OF INFORMATICS AND COMPUTING
INDIANA UNIVERSITY–PURDUE UNIVERSITY
Indianapolis

Abstract

The research in Parkinson's disease (PD) using biomarkers has long been dominated by measuring dopamine metabolites or alpha-Synuclein in cerebrospinal fluid. However, these markers do not allow early detection or monitoring of disease progression. In the recent years, metabolic profiling of body fluids has become powerful and promising tools in identification of the novel biomarkers in the diagnosis of the disease. While not much research has been done using machine learning techniques and predictive modeling to predict the severity of Parkinson's disease. The purpose of this project is to apply a predictive modelling approach in the diagnosis of Parkinson's disease using Cerebrospinal Fluid Biomarkers. The dataset for this study was collected from the PPMI website which comprises of 360 - Parkinson's patient, 220 - Control and 20 - SWEDD (Scans without evidence for dopaminergic deficit) . Various predictive model were developed in order to classify the disease based on its severity. The various machine learning algorithms used in this process are Decision tree, Random forest, Support Vector Machine (SVM), K- Nearest Neighbor (KNN), and Gradient boosting. Feature scaling and Mean normalization was applied to standardize the dataset. The above mentioned machine learning algorithms were applied on the Parkinson's Progression Markers Initiative (PPMI) data and accuracy for each algorithm was calculated. Out of all the models, Random forest and Gradient Boosting gave the best classification accuracy of 66.67%. In conclusion, the main factors that might have affected accuracy of the model are dataset size, missing data and number of features. To sum up, while the results shows some predictive power, we conclude negative results and hence these models are not Clinically significant.

Introduction

- Parkinson's disease (PD) is a neurodegenerative disease, affecting more than 1 % of the population over 60 years of age.
- Mostly the diagnosis of PD depends on the clinical history, physical examination and response to dopaminergic drugs, but misdiagnosis is possible in the early stages of disease [2].
- After the discovery of CSF hemoglobin, Abeta 42, CSF Alpha-Synuclein, P-tau181P and Total tau, major research efforts have been directed to the measurement of these proteins in body fluids, especially in CSF.

Hypothesis

- We hypothesized that there exists a correlation between different levels of CSF hemoglobin, Aß-42, T-tau, α-Synuclein and P- Tau$_{181}$ and the severity of Parkinson' s Disease.

Background

- Uric acid was proved to be a potential biomarker.
- Some studies showed inverse association between uric acid and risk related to PD.
- Some other potential biomarkers such as **CSF hemoglobin, Abeta 42, CSF Alpha-Synuclein, P-Tau181P and Total tau** can prove to be helpful in the early diagnosis of the disease.

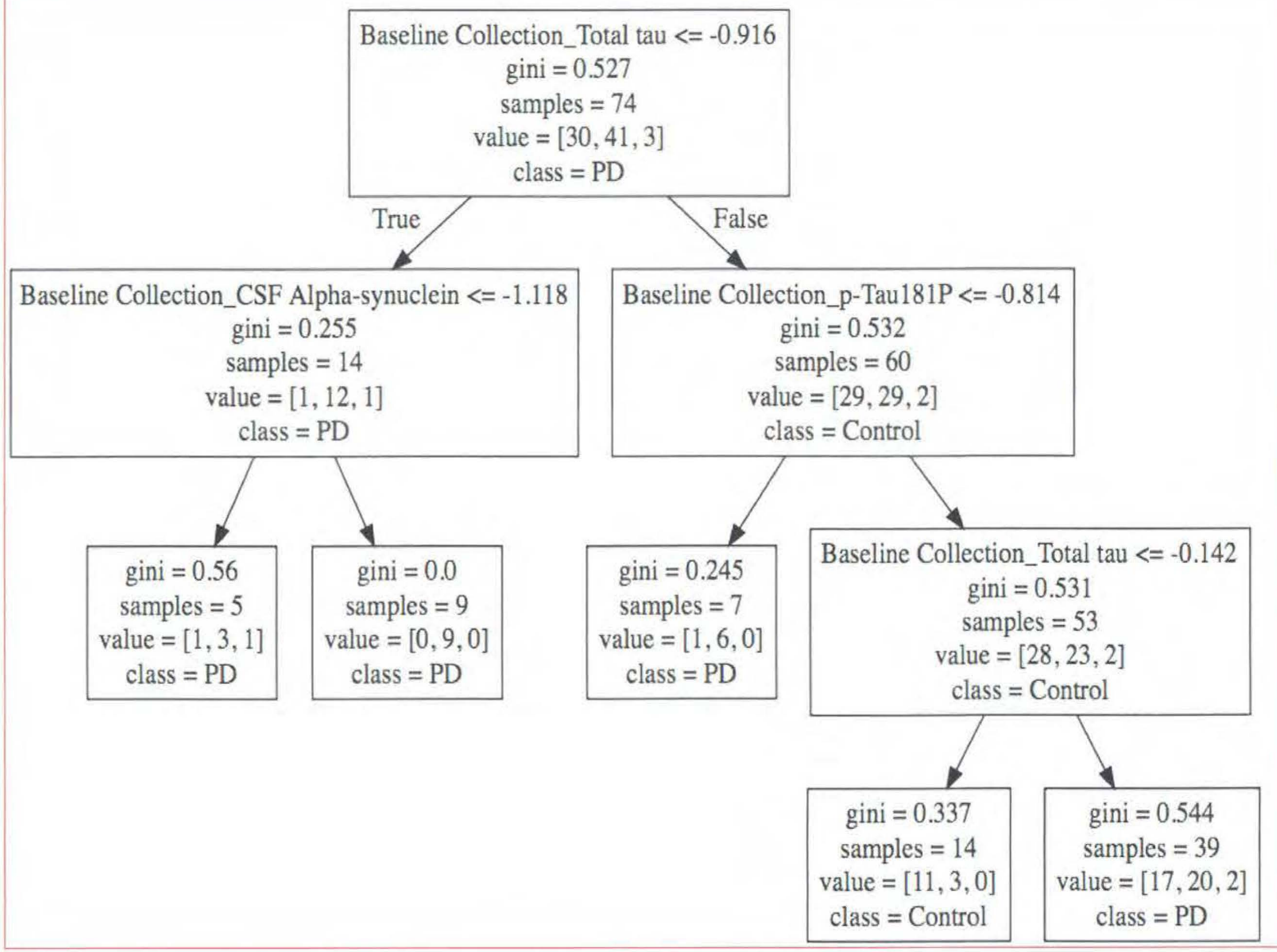
BIOMARKERS

- A measurable characteristic of a patient associated with incidence or progression of disease.
- An ideal biomarker is the one which should be easily accessible, validated and inexpensive.
- Since Parkinson's disease affects the central nervous system, CSF markers which surrounds the brain and spinal cord, will have high relevance in the pathology of the disease.

PPMI Dataset

- PPMI 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.
- **Dataset**: Control – 220
  PD – 360
  SWEDD - 20
- **Features**: Gender, Diagnosis, Baseline Collection_Total tau, Baseline Collection_ Abeta 42, Baseline Collection_ Alpha-Synuclein etc.

Decision Tree Classifier from Gini Index



Baseline Collection_Total tau <= -0.916
gini = 0.527
samples = 74
value = [30, 41, 3]
class = PD

True

False

Baseline Collection_CSF Alpha-synuclein <= -1.118
gini = 0.255
samples = 14
value = [1, 12, 1]
class = PD

Baseline Collection_p-Tau181P <= -0.814
gini = 0.532
samples = 60
value = [29, 29, 2]
class = Control

gini = 0.56
samples = 5
value = [1, 3, 1]
class = PD

gini = 0.0
samples = 9
value = [0, 9, 0]
class = PD

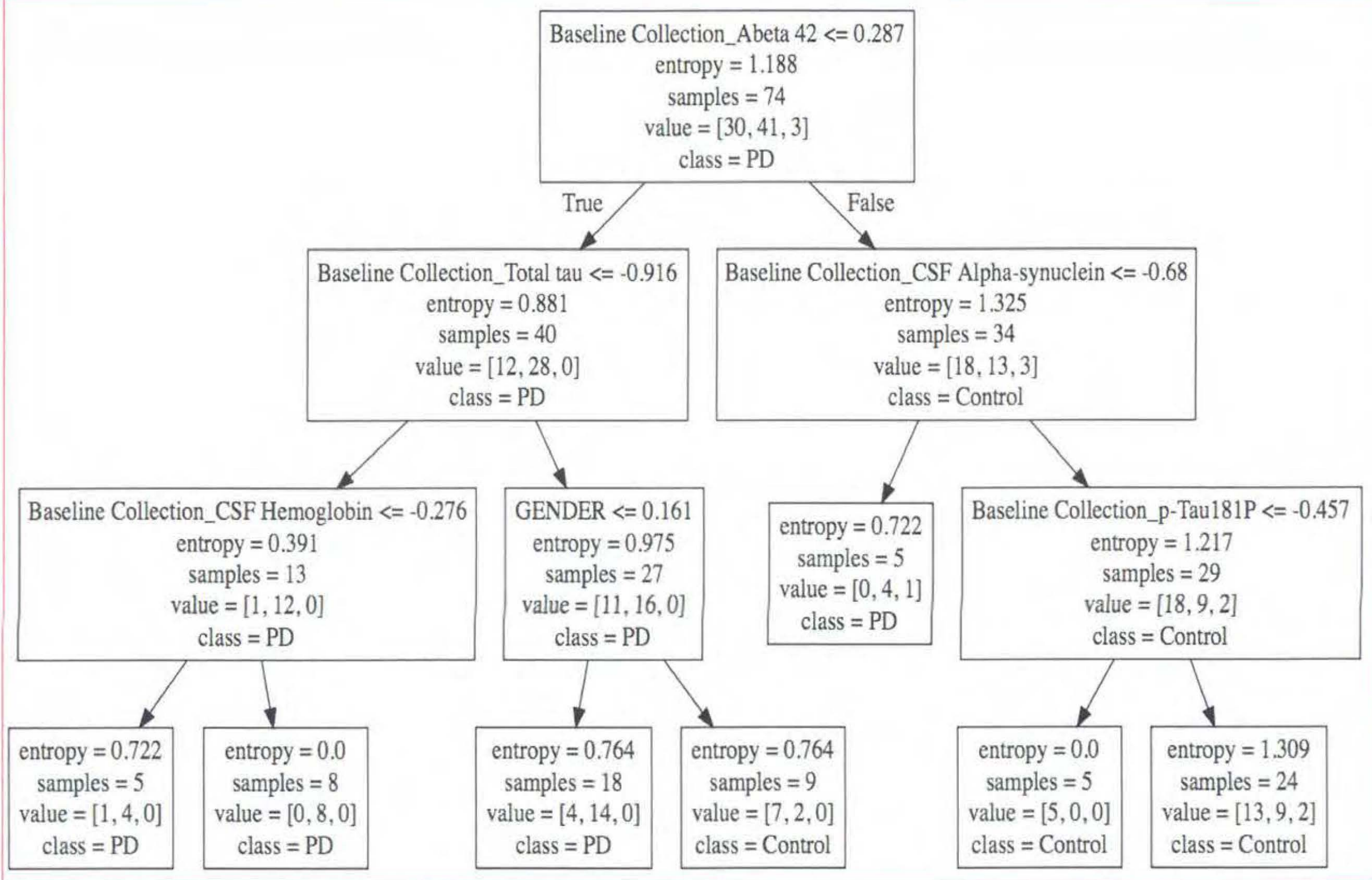
gini = 0.245
samples = 7
value = [1, 6, 0]
class = PD

Baseline Collection_Total tau <= -0.142
gini = 0.531
samples = 53
value = [28, 23, 2]
class = Control

gini = 0.337
samples = 14
value = [11, 3, 0]
class = Control

gini = 0.544
samples = 39
value = [17, 20, 2]
class = PD

Decision Tree Classifier from Entropy



Baseline Collection_Abeta 42 <= 0.287
entropy = 1.188
samples = 74
value = [30, 41, 3]
class = PD

True

False

Baseline Collection_Total tau <= -0.916
entropy = 0.881
samples = 40
value = [12, 28, 0]
class = PD

Baseline Collection_CSF Alpha-synuclein <= -0.68
entropy = 1.325
samples = 34
value = [18, 13, 3]
class = Control

Baseline Collection_CSF Hemoglobin <= -0.276
entropy = 0.391
samples = 13
value = [1, 12, 0]
class = PD

GENDER <= 0.161
entropy = 0.975
samples = 27
value = [11, 16, 0]
class = PD

entropy = 0.722
samples = 5
value = [0, 4, 1]
class = PD

Baseline Collection_p-Tau181P <= -0.457
entropy = 1.217
samples = 29
value = [18, 9, 2]
class = Control

entropy = 0.722
samples = 5
value = [1, 4, 0]
class = PD

entropy = 0.0
samples = 8
value = [0, 8, 0]
class = PD

entropy = 0.764
samples = 18
value = [4, 14, 0]
class = PD

entropy = 0.764
samples = 9
value = [7, 2, 0]
class = Control

entropy = 0.0
samples = 5
value = [5, 0, 0]
class = Control

entropy = 1.309
samples = 24
value = [13, 9, 2]
class = Control

Methods

The different machine learning techniques used in this project are

1. **Decision Tree Classifier**
   1.1. **Gini Index**
   1.2. **Entropy**
2. **Random Forest Classifier**
3. **Gradient Boosting**
4. **Support Vector Machine (SVM)**
5. **K – Nearest Neighbor (KNN)**

Results

S. No	M.L Technique		Accuracy
1.	Decision Tree Classifier	Gini Index	60.7%
		Entropy	60.7%
2.	Random Forest		66%
3.	Gradient Boosting		66%
4.	Support vector Machine (SVM)		57.5%
5	K- Nearest Neighbor (KNN)		60.6%

Conclusion

- Out of all the algorithms we attempted, Random Forest and Gradient Boosting methods performed best giving the accuracy of 66%.
- Based on Decision tree classifier, Baseline Collection_Total tau & Baseline Collection_ Abeta 42 were found to have a significant impact on diagnosis of PD patients.
- While it is not clinically significant, this attempt suggests the presence of valuable information in CSF biomarker data that can further be useful in diagnosis of the disease.
- The negative results of the study may be attributed to the small size of dataset, missing values and less number of features.

References

- Kang, J. (2013). Association of Cerebrospinal Fluid β-Amyloid 1-42, T-tau, P-tau181, and α-Synuclein Levels With Clinical Features of Drug-Naive Patients With Early Parkinson Disease. *JAMA Neurology*. doi:10.1001/jamaneurol.2013.3861.
- Havelund, J. F., Heegaard, N. H. H., Færgeman, N. J. K., & Gramsbergen, J. B. (2017). Biomarker research in Parkinson's disease using metabolite profiling. *Metabolites*, *7*(3). https://doi.org/10.3390/metabo7030042
- Wang, M. J., & Denison, T. (2014). Predictive Ability of Cerebrospinal Fluid Biomarkers in Diagnosing and Evaluating Parkinson ' s disease, MIT.