SAS® GLOBAL FORUM 2021

Paper 1200-2021

Prediction of Parkinson's Disease with a Regression-based Model Using Vertical Ground Reaction Force Data

Jin Ge, Tianchen Xu, Jia Guo, and Mengyu Zhang, Columbia University

INTRODUCTION

Parkinson's disease (PD) is a long-term degenerative disorder of the central nervous system. Approximately 60,000 Americans are diagnosed with PD each year. The Parkinson's Foundation Prevalence Project estimated that 930,000 people in the United States were living with PD by the year 2020. This number is predicted to rise to 1.2 million by 2030. Patients with PD may experience movement or motor related difficulties such as tremor, slowness of movement (bradykinesia), rigidity, and postural instability. Our goal is to investigate the most influential features, including demographics and motor related biomedical signals, for PD prediction. PD gold standard for diagnosis is based on subjective clinical evaluation. An objective and quantitative gait analysis system could potentially improve the current practice in diagnosis, symptom monitoring, therapy management, rehabilitation and fall risk assessment and prevention in PD patients [Biase et al., 2020].

There are two types of technologies that can be used for quantitative gait analysis, which are wearables and non-wearables. Force sensors are the wearable sensors that were used to generate the database investigated in this paper. Force sensors measure the ground reaction force (GRF) under the foot and return a current or voltage proportional to the pressure measured [Parkinson's Foundation, 2021]. The advantage of using force sensors is it is easily integrated into instrumented shoes. For non-wearables, floor sensors, and image processing-based methods are commonly used [Colyer et al., 2018; Derawi et al., 2011].

Nicolas Khoury et al. [2019] found that supervised classification methods are effective to detect PD patients, including K-nearest neighbor (K-NN), decision tree (DT), random forest (RF), Naïve Bayes (NB), support vector machine (SVM) and unsupervised classification methods such as K-means and the Gaussian mixture model (GMM). Perumal et al. [2016] used Linear Discriminant analysis (LDA) algorithms to distinguish between PD subjects and healthy subjects and found that stance, swing phase and step distance are the variables that are efficient for classification. In the paper, we applied logistic regression that is a more simple and interpretable approach to predict the status of PD and achieved satisfactory results with appropriated feature engineering from the time series data.

DATA COLLECTION

With well-developed sensor, physiological signals such as gait and force of PD patients can be recorded digitally. In order to stimulate the research and investigation in the study of complex biomedical signals such as physiological signals for PD patients, under the auspices of National Center for Research Resources of the National Institutes of Health, Research Resource for Complex Physiologic Signals was created. It has three components PhysioBank, PhysioToolkit and PhysioNet. The data we used is from the database in PhysioNet [Goldberger et al., 2000].

This database contains measures of gait from 92 patients with idiopathic PD and 73 healthy controls. Vertical ground reaction force (VGRF) was recorded by the 8 sensors attached under each foot when subjects walked at their usual, self-selected pace for 2 minutes on level ground. For each of these 16 sensors, 100 signals were digitized and recorded at per second. For each foot, there was also a signal to reflect the sum of the 8 sensor outputs (Ultraflex Computer Dyno Graphy, Infotronic Inc.). Hence, for each subject there were 18 time series measurements in total. Database also includes demographic information, measures of disease severity and other related measures.

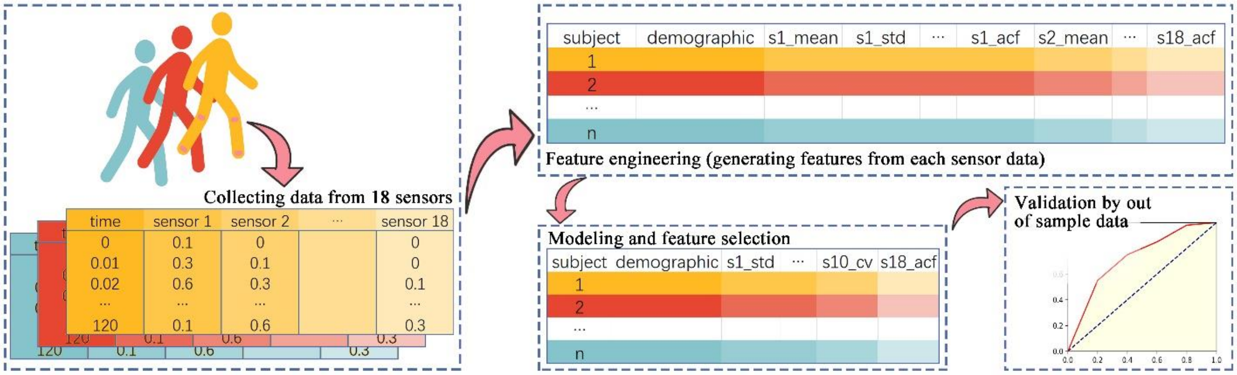

Figure 1. Data collection steps and pipeline of analysis.

METHODS

The data collection and analysis steps are summarized in **Figure 1**. After collecting time series data from each subject, we extracted relevant digital biomarkers from these raw sensor data for further downstream analysis. The procedure requires a combination of signal processing, data science and biological expertise in addition to extensive validation data [Goldsack, Chasse, Wood, 2019]. Softwares to preprocess and extract features from such data are becoming increasingly numerous and varied [Psaltos et al., 2019]. But without an established set of preprocessing techniques and features for capturing disease signal, it becomes necessary to make use of a broad spectrum of methods and tools when exploring potential biomarkers. Fortunately, With the help of `proc univariate` and `proc autoreg` in SAS, we were able to generate a variety of features easily. The following features were calculated for each of the 18 time series variables [Snyder et al., 2020]:

1) Location measures: mean value, mode, median, first/third quartile, 95th quantile;

2) Dispersion measures: standard deviation, interquartile range, coefficient of variation;

3) Shape measures: skewness, kurtosis;

4) Autocorrelation with lags 10, 30, 60, 100, 120.

After constructing the processing pipelines of raw data sourced from sensor measurements recorded, a feature set was obtained for each subject. Some subjects had more than one records, we utilized `proc means` to aggregate the features. Then the feature set could be merged with the demographic data and then used as the primary predictor set for analysis. Since not all features are related to the status of disease, we used `proc glmselect` to perform effect selection in the framework of general linear models. Good predictors for a logistic model could be identified and selected by `proc glmselect` when fitting a binary target [Robert Cohen, 2009]. The procedure selects a subset of features with a regression model using AIC as the criterion while forcing demographic variables (age, gender, study) to be included in the model as potential confounders.

Finally, we validated our selected model by calculating the AUC of the testing data. Specifically, we randomly separated our dataset into training data (70%) and testing data (30%) by `proc surveyselect`. Then A logistic model based on selected features was fitted on the training set using `proc logistic`, and AUC was obtained on the testing set. We repeated the same procedure for 100 times and got the out of sample AUCs.

RESULTS

FEATURE ENGINEERING

For each of the 18 time series variables, we considered 16 features, including location measures, dispersion measures, shape measures and autocorrelation, resulting in 288 features in total. By using a regression-based method with AIC as the model selection criteria, with age, gender and study group indicator being fixed in the model, 7 among 288 features were selected, including standard deviation of L1 sensor, skewness of L6 sensor, range of R1 sensor, median of R4 sensor, autocorrelation (ACF) with 30 lags of R4 sensor, range of R6 sensor, and coefficient of variation of R7 sensor. The ACF was rescaled by multiplying 100 due to its relatively small scale.

Table 1 showed the odds ratio estimates and 95% confidence intervals obtained from the logistic regression model which was performed on the entire dataset with 165 subjects. After adjusting for other covariates, age was not significantly associated with Parkinson disease status, but gender had a significant effect. The probability of developing Parkinson disease for a male is 5.793 times than that for a female, with 95% CI (1.704, 19.700). After adjusting for covariates, except the range of R1 sensor, other selected features were all significantly associated with Parkinson disease status. Standard deviation of L1 sensor was the most significant feature. Adjusting for other covariates, if a patient's standard deviation of L1 sensor was getting one unit larger, the probability that he/she develops Parkinson disease would be 0.963 times as before, with 95% CI (0.943, 0.984).

Table 1. Odds ratio estimates and 95% confidence intervals obtained from logistic regression model.

Effect	Odds ratio	Lower 95% CI	Upper 95% CI	Pr > ChiSq
Study group 1 vs 3	1.107	0.301	4.068	0.410
Study group 2 vs 3	3.105	0.626	15.406	0.122
Age	1.032	0.965	1.103	0.359
Gender male vs female	5.793	1.704	19.700	0.005
L1 std deviation	0.963	0.943	0.984	0.001
L6 skewness	0.145	0.021	0.997	0.050
R1 range	0.994	0.987	1.001	0.083
R4 median	1.054	1.016	1.093	0.005
100x (R4 ACF 30)	1.086	1.019	1.158	0.011
R6 range	0.991	0.982	0.999	0.037
R7 CV	0.949	0.905	0.995	0.050

PREDICTION OF PARKINSON DISEASE

To evaluate the prediction performance of the 7 selected features, we compared the final model with a baseline model. **Figure 2 (A)** showed the ROC curve of the final model including the 7 selected features as well as age, gender and group indicator. **Figure 2 (B)** showed the ROC curve of the baseline model which only included age, gender and group indicator. Both models were performed on the entire dataset. Evidently, our final model (AUC=0.9329) was considerably better than the baseline model (AUC=0.6134).

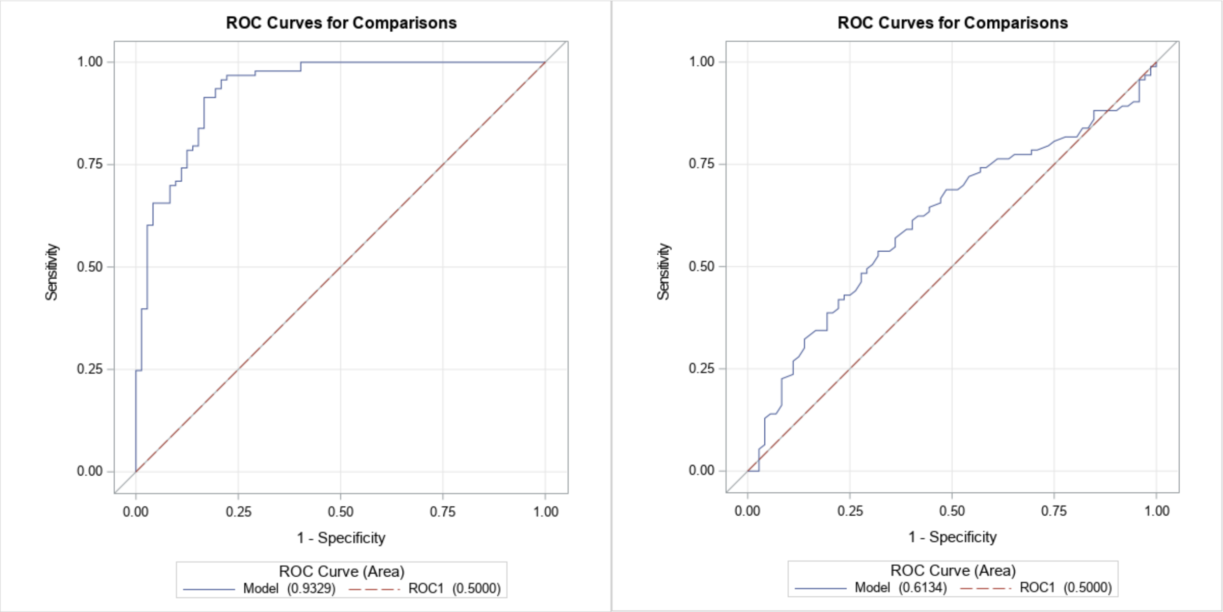

Figure 2. (A, left) ROC curve and AUC of final model with 7 selected features as well as age, gender and group indicator, on the entire dataset. (B, right) ROC curve and AUC of the baseline model with age, gender and group indicator, on the entire dataset.

We also compared the two models on the test sets using a resampling method that randomly split the dataset 100 times. **Table 2** showed the averaged AUCs and the standard deviations of the two models on test sets. The averaged AUC of our final model was 0.9039, while the averaged AUC of the baseline model was only 0.5303, which demonstrated that the selected features were highly important in terms of Parkinson disease prediction. It was also suggested that age, gender and group indicator only had limited predictability.

Table 2. Averaged AUCs and 95% confidence intervals of the final model and the baseline model on test sets, with 100 times repeated cross validation.

Model	**Mean of AUCs**	**Standard deviations of AUCs**
Final model	0.9039	0.0370
Baseline model	0.5303	0.0623

To further explore the selected features, we performed Principal component analysis (PCA) on all selected features, and visualized the data by using the first two principal components in **Figure 3**. It was clear that patients with Parkinson disease and healthy controls could be separated by the first two principal components, which further demonstrated the importance of our selected features.

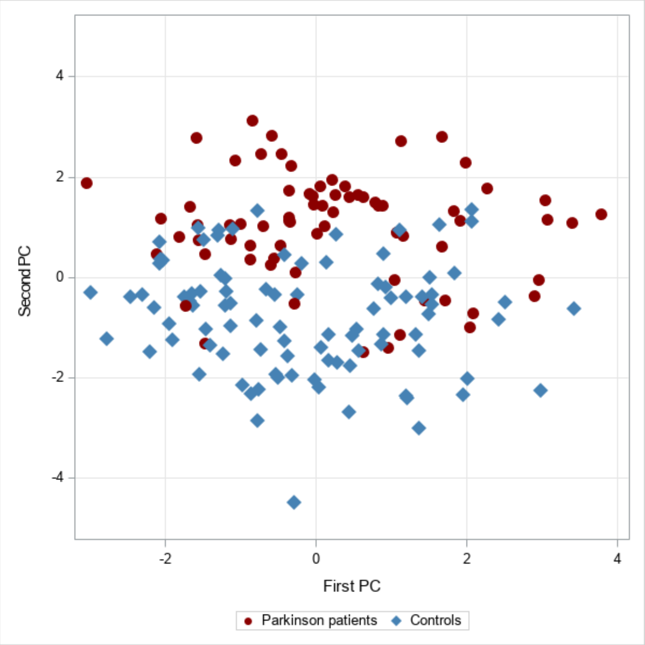

Figure 3. Visualization on the data by using the first two principal components obtained from the selected features.

CONCLUSION

Vertical ground reaction force (VGRF) records in this study reflect the jitter and strength of patients' legs. The jitter frequency is associated with the status of PD. The result follows the common sense of PD in most cases. First, the less a patient's jitter dispersion and dispersion frequency are, the higher probability this patient may have PD. This is because PD patients invest much less force in striding than healthy controls, so that patients with high jitter frequency are not inclined to be PD. In addition, VGRF of PD patients is more likely to have a narrower range and be symmetric distributed due to the smaller skewness, when compared to healthy controls. The median of VGRF in PD patients tends to be higher than that in healthy controls. The autocorrelation across the time is higher in PD patients than that in the control group. This may because PD patients have a reduced stride length and a short average swing time, resulting in that the walking force of PD patients is relatively stable and predictable from the past records [Frenkel-Toledo et al., 2005].

Our study demonstrates a simple way to predict PD status from the walking condition of patients, which is usually obtained by accessible devices with force sensors. Specifically, we found that logistic regression, a conventional approach shows satisfactory results with a high predicting accuracy with AUC 0.9039 when 10 features are used. With the increasing trend of portable devices and digital health, we believe that our method with simple features can be widely used to predict clinical outcomes such as Parkinson disease that is potentially related to time series predictors.

REFERENCES

[1] Mirelman, A., Bonato, P., Camicioli, R., Ellis, T. D., Giladi, N., Hamilton, J. L., ... & Almeida, Q. J. (2019). Gait impairments in Parkinson's disease. The Lancet Neurology, 18(7), 697-708.

[2] Parkinson’s Foundation. https://www.parkinson.org/Understanding-Parkinsons/Statistics.

[3] Di Biase, L., Di Santo, A., Caminiti, M. L., De Liso, A., Shah, S. A., Ricci, L., & Di Lazzaro, V. (2020). Gait analysis in Parkinson’s disease: an overview of the most accurate markers for diagnosis and symptoms monitoring. Sensors, 20(12), 3529.

[4] Shah, J., Pillai, L., Williams, D. K., Doerhoff, S. M., Larson-Prior, L., Garcia-Rill, E., & Virmani, T. (2018). Increased foot strike variability in Parkinson's disease patients with freezing of gait. Parkinsonism & related disorders, 53, 58-63.

[5] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. circulation, 101(23), e215-e220.

[6] Goldsack, J., Chasse, R. A., & Wood, W. A. (2019). Digital endpoints library can aid clinical trials for new medicines-STAT. STAT https://www. statnews. com/2019/11/06/digital-endpoints-library-clinical-trials-drug-development.

[7] Psaltos, D., Chappie, K., Karahanoglu, F. I., Chasse, R., Demanuele, C., Kelekar, A., ... & Cai, X. (2019). Multimodal wearable sensors to measure gait and voice. Digital biomarkers, 3(3), 133-144.

[8] Snyder, P., Tummalacherla, M., Perumal, T., & Omberg, L. (2020). mhealthtools: A Modular R Package for Extracting Features from Mobile and Wearable Sensor Data. Journal of Open Source Software, 5(47), 2106.

[9] Frenkel-Toledo, S., Giladi, N., Peretz, C., Herman, T., Gruendlinger, L., & Hausdorff, J. M. (2005). Effect of gait speed on gait rhythmicity in Parkinson's disease: variability of stride time and swing time respond differently. Journal of neuroengineering and rehabilitation, 2(1), 1-7.

[10] Colyer S.L., Evans M., Cosker D.P., Salo A.I.T. A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. Sports Med. -Open. 2018;4:24.

[11] Derawi, M. O., Ali, H., & Cheikh, F. A. (2011). Gait recognition using time-of-flight sensor. BIOSIG 2011–Proceedings of the Biometrics Special Interest Group.

[12] Khoury, N., Attal, F., Amirat, Y., Oukhellou, L., & Mohammed, S. (2019). Data-driven based approach to aid Parkinson’s disease diagnosis. Sensors, 19(2), 242.

[13] Perumal, S. V., & Sankar, R. (2016, November). Gait monitoring system for patients with Parkinson's disease using wearable sensors. In 2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT) (pp. 21-24). IEEE.

ACKNOWLEDGMENTS

We appreciated the help from our friends Yun He, Zhiji Zhang, Kun Qian, Siqi Huang and advisor Prof. Yuanjia Wang.

RECOMMENDED READING

- *Base SAS® Procedures Guide*

- *SAS® For Dummies®*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jin Ge
Columbia University
jg4197@columbia.edu