Parkinson’s Disease Classification of mPower Walking Activity Participants

Benjamin Pittman, Reza Hosseini Ghomi, and Dong Si*, Member IEEE

Abstract— **Motion sensor data collected using Sage Bionetwork's mPower application on the Apple iPhone to record participant activities is analyzed to classify samples as positive or negative for Parkinson's Diagnosis. Pre-processing of the data showed differences in the time and frequency dimensions for features derived from Apple Core motion data. Several classic machine learning classification algorithms were trained on seventy-seven derived data points for best precision, recall, and F-1 score. Accuracy as high as ninety-two percent were achieved, with the best results attained from decision tree and multi-layered artificial neural network algorithms. This research shows that motion data produced on the Apple iPhone using the mPower application shows promise as an accessible platform to classify participants for presence of Parkinson's Disease signs.**

I. INTRODUCTION

Parkinson's Disease (PD) has no known cure and is characterized by physical signs such as bradykinesia, tremor, and postural instability [1]. The goal of this research is to show these signs are represented in sensor data and accurately predict a positive professional Parkinson's Diagnosis. Accuracy of clinical diagnosis of PD has been shown to be approximately 73.8% by nonexperts and 82.7% by movement disorders experts when compared to pathological examination as a gold standard [2]. This paper outlines a supervised learning process for binary classification of the data using grid search and cross-validation to find and train the best parameters for a set of binary classification machine learning algorithms.

mPower is an ongoing research and data collection endeavor managed by Sage Bionetwork's on their Synapse platform and framework [3]. The mPower application for Apple iPhone was created to leverage sensors on the iPhone to record the movements of participants during prompted walking activities though in much less formal surroundings than previous similar studies. Overall participation in the study consisted of 14% with PD and skewed younger with 72% of participants between the ages of eighteen and forty-four. Participants were recruited via open enrollment by downloading the app in the app store.

Participants are recorded during four activities — walking, voice, tapping, and memory, three times per day (before, after, and another time related to medication timing or at any time if a control subject). This paper focuses on the walking activities. These activities are a thirty-second walking period, and a thirty-second standing still period. During these activity periods the participants are instructed to place the iPhone in their front pant pocket in a specific orientation where y is the vertical axis, x is the horizontal axis, and z is the axis along which the participant is to walk.

II. LITERATURE REVIEW

Similar studies have shown to classify PD patients from controls with great accuracy using low-cost sensors and data mining, all in supervised environments. Tucker, et al achieved .95 accuracy across 2,743 patients with a K-Nearest-Neighbors algorithm leveraging complex gait detection equipment [4]. Invasive and supervised study environments such as this have shown to hinder participation [5].

Previous research performed using similar technology for feature extraction, Android smart-phones, showed accuracy of 0.82 and 0.81, respectively, for hand-resting tremor and gait difficulty detection [6]. This study was comprised of forty PD patients and required them to wear the Android phone using specialized gear, in specific orientation, and in a supervised environment. This papers approach is similar as a framework, but features are derived in unsupervised environments and with the iPhones orientation far less restricted. Finally, our goal is classification of PD for screening and diagnosis.

III. METHOD

Data for this project was provided by Sage Bionetwork's mPower qualified research program via their website, synapse.org. Data is accessed and downloaded in a quasi-REST architecture using the synapseclient library in python and SQL commands. Participant data is hosted on the synapse.org site. The data set was collected from March to September of 2015 and consists of 35,410 samples from 3,101 unique participants. First released in early 2016, this is the first set of data to be released from the mPower application. To complete the programming necessary for processing the models and plot the results, python was used with the numpy, sklearn, and matplotlib libraries [7]. The Keras Python API [8] on top of the open-sourced machine learning framework Tensorflow [9] was used for the artificial neural network. In the next sections of this paper our methods for data pre-processing, feature extraction, model selection, and model evaluation are discussed, with a final discussion of the results.

The first step in unlocking the data for our use is to understand how the data is collected. Due to the tools and methods provided by Synapse we begin with a proficient data set of known provenance. There are.JavaScript Object Notation (JSON) files that contain both the raw accelerometer and gyroscope data as well as files with that data pre-processed using sensor fusion technology within the

*Research supported by University of Washington.

B. Pittman and D. Si are with the University of Washington-Bothell Department of Computing and Software Systems, Bothell, WA 98011 USA (corresponding e-mail: dongsi@uw.edu).

R. Hosseini Ghomi is with the University of Washington Department of Psychiatry and Behavioral Sciences, Seattle, WA 98195 USA

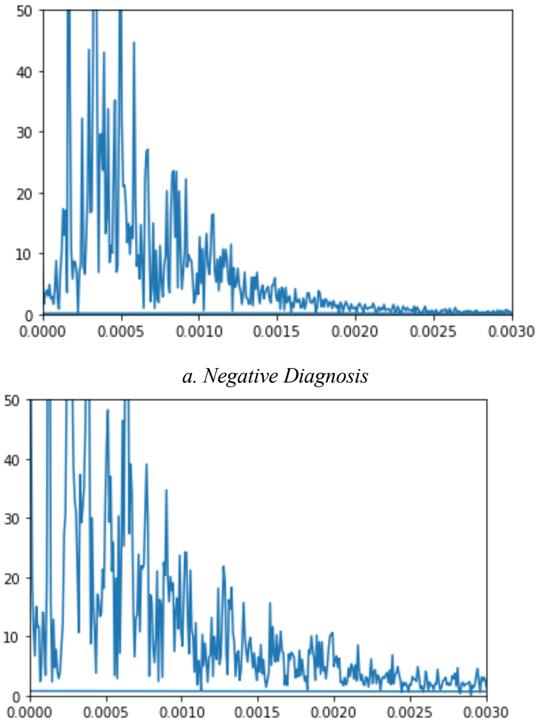
iPhone software and hardware. Additionally, there are collections of demographic data — including age, diagnosis data, and other relevant data. The Apple Core Motion object is the sensor fusion processed data that uses the six axes of the accelerometer and gyroscope to separate gravity from the acceleration vectors and eliminates the noise and drift that is common in gyroscopes (the exact algorithm is proprietary to Apple, Inc). For this paper, we used this processed Apple Core Motion data instead of the raw accelerometer and gyroscope data [10].

The processed data within each JSON file is a dictionary of lists that contain a list of timestamps (approximately every 1/100 of a second) and the following lists indexed to those time stamps – accelerometer recordings for acceleration in units of gravitational acceleration (x, y, z), and for gravity (x, y, z); a list representing the rotational acceleration of the device in radians; a list for the attitude of the device (a quaternion). From these lists we derive the features outlined in Table I. These features are extracted for a 10-second sampling of a walking period and a 10-second sampling of a standing still period, the participants age is then added to form each record's set of seventy-seven features.

Prior to the features extraction the data set is cleaned of records that do not contain samples of adequate length for the walking or standing periods or those that are missing demographic/diagnosis data. In addition to a True/False field-for whether a record has a professional diagnosis, there is a field denoting the sampling period in-reference-to the patient's time of medication, or that no such time point exists. We filter out the records that took medication immediately prior to the sampling period as we want to make a positive or negative assertion of diagnosis on the records and PD medication is known to temporarily suppress disease signs [11].

TABLE I. MOTION DATA EXTRACTED FEATURES

Description	**Number**
Means Of Gravity	3
Absolute Means of Gravity	3
Means of Acceleration	6
Absolute Means of Acceleration	3
Moments: Kurtosis, Skew	12
Sums of Variances	4
Sums of Standard Deviations	4
Zero Cross Rate	3
Total	**38**



a. Negative Diagnosis

b. Positive Diagnosis

Figure 2. Sample frequency-based graphs over the sampling period for two participants – one without and one with PD. The x-axis represents different frequencies of movement, while the y-axis represents the number of times those frequencies were witnessed over the sampling period.

One of the first features we aimed to extract was the participants total displacement along each axis of movement — how much distance the device traveled in each dimension. In so doing, we also found the total displacement of gravity in each dimension. If the greatest gravitational displacement is not along the y-axis this means the device was not oriented in the persons pocket as prescribed in the activities participant instructions. A quick look at this extracted gravitational displacement shows that around 30% of devices were not held in the position shown in the data wiki provided by synapse. As this would have limited our final data set to around 11,000 records we decided to use those records and keep our derivation of features from the base data simplified to limit this being a factor in our study. The final data set consists of 60% with a positive professional diagnosis for PD, and 40% without.

The first breakthrough for the feature extraction came when looking at acceleration over the time series and the corresponding Fourier Transforms. As is pictured in the example in Figure 2, there is greater activity in the higher frequencies of samples with a positive diagnosis for PD. This was consistent with the expected result and consistent with clinical observation. Normally, an action would be precise evident by a narrow frequency band demonstrating a non-Parkinsonian person using controlled muscle activation. In PD, however, a person has imprecise control of their movement and therefore a frequency representation would demonstrate a wide range of accelerations utilized to accomplish the same task. This feature is represented in the final data set by finding the skew and Kurtosis of each frequency-transformed series for, and about, each axis of motion in the data.

The Fisher-Pearson coefficient of skewness is zero for a normal distribution, and this value is greatest when the data is 'skewed' away from the center of the possible distribution of values. As shown in Figure 2, the positive diagnosis has more area under the curve toward the extremes and this will result in a different value being returned than the negative diagnosis.

Similarly, while the Pearson coefficient of Kurtosis is 3 for a normal distribution this value increases the more data is located away from the center; and decreases erstwhile. This is a measure of the peakedness of a distribution or set of values and will result in a lower value for Figure 2(a) than Figure 2(b)

IV. RESULTS

The first step in modeling the derived data was to standardize the data. Standardization was chosen over normalization as it maintains more of the information for outlier data. Our process for standardization is simply $Y_i = \frac{Y_i - \bar{Y}}{s}$ on each set of features, where $\bar{Y}$ is the mean and s is the standard deviation.

For this study, we primarily use measurements of recall, precision, and F-1 score (the harmonic mean of recall and precision) to explore and rate the data set and models. All models were cross validated using 10-fold cross validation with an 80/20 train/test split. All but Artificial Neural Networks were trained using the sklearn library and GridSearchCV class that allows us to pass in arrays of parameters for the model to train on and perform cross validation. The mean and standard deviation of each grid search and k-fold cross validation is returned. The best performing algorithm from each class is then validated on the test split. These results are shown in Table II.

Logistic Regression (LR) performed well on the data, returning a combined accuracy score of 79.8% for true-positive/true-false results. The focus in this project was on extracting features that are representative of the signs of PD so it is expected to have greater scoring parameters for the models in the positive classification. The best parameters for LR were returned using a Stochastic Average Gradient solver that always converges toward the dot product of the first two derivatives, a tolerance of 1e-2, and inverse regularization strength of 0.1. The training accuracy and standard deviation for the mean training scores of the grid search were 81.2%. Test results are shown in Table II and we show high recall for positive diagnoses with low recall for negative diagnoses, while precision scores are slightly better for false classifications.

The grid search for Decision Tree (DT) classification returned accuracy results as high as 86.9%. Our data set is imbalanced toward positive diagnoses at approximately 60% positive versus 40% negative. This explains some of the imbalance across the classifications, but also seems to intersect previous research that shows models trained on motor-disabled patients versus a control showed great inaccuracy when tested on the other group [4]. In short, features that are highly correlated with negative PD patients need to be extracted to increase accuracy of false predictions. The best model parameters for decision tree classifier used entropy to rate the quality of a data split, maximum depth of 14, minimum records for a split of 10, and minimum records for a leaf of 10. Decision tree classifier is our second most accurate model for predicting the presence of PD signs.

TABLE II. PREDICTION RESULTS BY USING DIFFERENT MODELS

Algorithm	State	Precision	Recall	F-1
Logistic Regression	False	0.82	0.65	0.73
	True	0.79	0.90	0.84
Decision Tree	False	0.82	0.90	0.84
	True	0.86	0.88	0.87
K Nearest Neighbors	False	0.75	0.52	0.62
	True	0.73	0.88	0.80
Support Vector Classification	False	0.90	0.62	0.73
	True	0.78	0.95	0.86
Multi-Layer Neural Network	False	0.89	0.80	0.84
	True	0.87	0.93	0.90

K Nearest Neighbors (KN) returned very tightly grouped means across all the grid search parameters. Manhattan (or block) distance and distance weighting returned highest accuracy of 75.4%, the lowest for this research. Full model results show that the models ability to recall false diagnoses is barely better than a coin flip while positive predictions score much better at 88%. We continue to see the same pattern as in our previous models where model results are much better for positive diagnoses. Overall, this is the worst performing model.

Support Vector Classification (SVC) showed the greatest variance within the grid search parameters. The polynomial and linear kernel performed similarly for high values of the C parameter. This value tells the model how sensitive we want to be for misclassification and chooses a smaller-margin hyperplane on the data. The model returns accuracy of 82.4% with a linear kernel and C = 1000 and 62.5% with C = 1 and gamma = 1e-3 suggesting that our data may be linearly separable. As shown in Figure 4, this model returned better positive diagnosis result than the Decision Tree model with precision of 78% and recall of 95% but was much worse at choosing negative diagnoses from the data with recall of 62%.

The Artificial Neural Network (ANN) performed similarly for average accuracy as the other top performing model, Decision Tree Classifier, but performed better at accurately classifying negative diagnoses (Figure 4). A collection of multi-layer fully-connected neural networks were trained using k-fold cross validation and found the best on our data set was a three-layer neural network of 40-21-9-1 neurons with a 25% drop-out rate in the first layer — each training step randomly chooses 25% of the neurons and skips training them for that iteration. A rectifier activation function was used at all but the output layer as this enables us to keep a greater proportion of the outlier data information than other functions (See Figure 3). The dropout layer helps to reduce over-fitting in the model by randomly choosing 25% of nodes to leave out of training at each iteration.

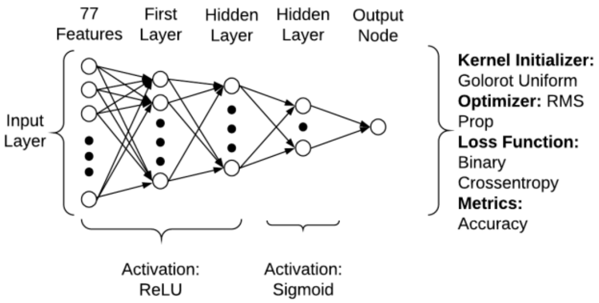


Figure 3. Artificial Neural Network topology

The output layer uses a sigmoid function for final binary classification of the output.

V. CONCLUSION

The results of the models clearly show that there is a bias in the models toward positive diagnoses. The approach taken with the data — we look for symptoms of a disease not a lack thereof — effects this as we're not able to derive features that show someone as negative. Another possibility, and worth further research for this data set and study is if there is a familial history of PD for participants as this could cause skew in the data for non-positive diagnoses due to a greater interest in being a participant in this study due to the familial connection. These models could therefore pick up early onset of the disease and mis-train the data; there is evidence of a genetic component to PD [12]. This feature should be tested against and possibly added to the model if it shows bias in the results. Given their performance on this feature set, the models to continue testing larger feature sets are Decision Tree and Artificial Neural Networks. Random Forests should also be considered.

The primary take-away from this study is that the mPower iPhone application walking and resting activity data show great promise for binary classification of the presence of PD signs. Exploration of the other mPower activities should be performed in isolation and then in concert with this study's features and to form a more accurate and robust model. Tracking PD symptoms using a smartphone is a novel model to further explore due to the wide adoption of smart phones, the capabilities they present for collecting, uploading, and processing data, and their ability to monitor participants in natural environments.

ACKNOWLEDGMENT

Data was contributed by users of the Parkinson mPower mobile application as part of the mPower study developed by Sage Bionetworks and described in Synapse [doi:10.7303/syn4993293].

This work was supported by the Graduate Research Award from the Computing and Software Systems division of University of Washington Bothell and the startup fund 74-0525.

Dr. Hosseini Ghomi's work was supported by NIH R25 MH104159.

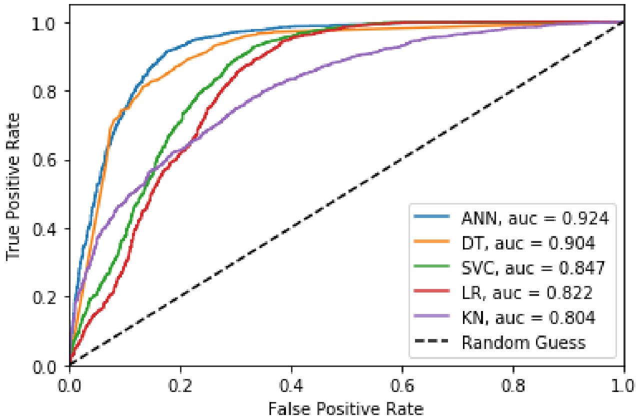


Figure 4. Receiver Operating Characteristic (ROC) Curves of Model Results

We thank David Atkins, PhD of the Behavioral Research in Technology and Engineering (BRiTE) Center at The University of Washington and Forrest Collman, PhD of the Allen Institute for Brain Science, Seattle, for their editorial review.

REFERENCES

[1] Uitti, R. J., Baba, Y., Wszolek, Z. K., & Putzke, D. J. (2005). Defining the Parkinsons disease phenotype: initial symptoms and baseline characteristics in a clinical cohort. *Parkinsonism & Related Disorders,11*(3), 139-145. doi:10.1016/j.parkreldis.2004.10.007

[2] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino, “Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis,” *Neurology*, vol. 86, no. 6, pp. 566–576, Feb. 2016.

[3] Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., . . . Trister, A. D. (2016). The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific Data, 3*, 160011. doi:10.1038/sdata.2016.11

[4] Tucker, C., Han, Y., Nembhard, H. B., Lee, W., Lewis, M., Sterling, N., & Huang, X. (2015). A data mining methodology for predicting early stage Parkinsons disease using non-invasive, high-dimensional gait sensor data. *IIE Transactions on Healthcare Systems Engineering,5*(4), 238-254. doi:10.1080/19488300.2015.1095256

[5] Albert, M. V., Toledo, S., Shapiro, M., & Kording, K. (2012). Using Mobile Phones for Activity Recognition in Parkinson’s Patients. *Frontiers in Neurology,3*. doi:10.3389/fneur.2012.00158

[6] Pan, D., Dhall, R., Lieberman, A., & Petitti, D. B. (2015). A Mobile Cloud-Based Parkinson’s Disease Assessment System for Home-Based Monitoring. *JMIR mHealth and uHealth,3*(1). doi:10.2196/mhealth.3956

[7] Jones, E. Oliphant, T., Peterson, P. & others (2001-). SciPy: Open source scientific tools for Python. http://www.scipy.org/ . Accessed on 2017-10-11.

[8] Chollet, F., & A. (2015). Keras: The Python Deep Learning library. https://keras.io/ . Accessed on 2017-11-1.

[9] Abadi, Martin & others (2015). Large-scale Machine Learning on Heterogeneous Distributed Systems. https://www.tensorflow.org/ . Accessed on 2017-11-1.

[10] Apple. Core Motion. https://developer.apple.com/documentation/coremotion. Accessed on 2017-10-11.

[11] Cenci, M. A. (2007). Dopamine dysregulation of movement control in l-DOPA-induced dyskinesia. *Trends in Neurosciences,30*(5), 236-243. doi:10.1016/j.tins.2007.03.005

[12] Payami, H., Zareparsi, S., James, D., & Nutt, J. (2002). Familial Aggregation of Parkinson Disease. *Archives of Neurology,59*(5). doi:10.1001/archneur.59.5.848