Gait Model Analysis of Parkinson's Disease Patients under Cognitive Load

James Alexander Hughes
St. Francis Xavier University
Computer Science
Antigonish, Nova Scotia, Canada
jhughes@stfx.ca

Sheridan Houghten
Brock University
Computer Science
St. Catharines, Ontario, Canada
shoughten@brocku.ca

Joseph Alexander Brown
Innopolis University
AI in Games Development Lab
Innopolis, Republic of Tatarstan, Russia
j.brown@innopolis.ru

***Abstract*—Parkinson's disease is a neurodegenerative disease that affects close to 10 million with various symptoms including tremors and changes in gait. Observing differences or changes in an individual's manifestations of gait may provide a mechanism to identify Parkinson's disease and understand specific changes.**

In this study, timeseries data from both Control subjects and Parkinson's disease patients was modelled with symbolic regression and extreme gradient boosting.

Model effectiveness was analyzed along with the differences in the models between modelling strategies, between Control subjects and Parkinson's disease patients, and between normal walking and walking while under a cognitive load. Both modelling strategies were found to effective. The symbolic regression models were more easily interpreted, while extreme gradient boosting had higher overall accuracy. Interpretation of the models identified certain characteristics that distinguished Control subjects from Parkinson's disease patients and normal walking conditions from walking while under a cognitive load.

***Index Terms*—Cognitive Load; Gait; Genetic Programming; Parkinson's Disease; Symbolic Regression; Time series; XGBoost.**

I. Introduction

Parkinson's disease (PD) affects approximately 10 million people globally with neurodegenerative effects [3]. Symptoms including tremors, depression, hallucinations, cognitive decline, falls, and changes in gait [17]. Gait has been used as a primary diagnosis factor as the disease affects the rhythm, speed, and stride [4], [5], [10], [18], [23].

There are problems with gait being used as a diagnostic tool, namely symptom similarity with other disorders [16], [6], [2], [15] and the creation and maintenance of a clear *data-driven* record. Observational diaries, while a popular method to record symptom frequency and effects, are subjective, inconsistent, and error prone; objective recording methods should be preferred for data collection and monitoring [11].

The cognitive load that a person is under during the act of walking can affect how an individual walks (however, some mental actions, such as singing, have been found to PD symptoms that affect gait [8]). The data set used in this study reflects subjects under a cognitive load, and therefore is more reflective of gait outside of a laboratory setting [23].

This study builds upon previous work studying different modelling techniques for PD patient gait data [13], [14]. In this study the focus is on two of the more effective and explainable modelling techniques, namely, *Symbolic Regression* and *Extreme Gradient Boosting*. These are explainable both in terms of how the models are generated and the explainability of the resulting models. As in previous work, we compare the models generated for Control subjects and PD patients. However, unlike previous work, we also include a comparison of models generated from data recorded from subjects walking normally and while under a cognitive load. As a result, a deeper analysis into model effectiveness is performed along with a much deeper analysis of model feature importance.

Details on the data used in this study are presented in Section II and a summary of the algorithms and methodology can be found in Section III. A summary of the effectiveness of the generated models, an analysis of feature importance within the models, and the differences between Control and PD patients and normal and cognitive load walking models are presented in Section IV. Section V discusses the main conclusions and presents possible future directions for the long-term project.

II. Data

Data was obtained from *PhysioNet*, an open access collection of various physiologic data [7]. We use data from the *Gait in Parkinson's Disease* project [9], which contains data from a collection of studies on PD gait [23], [10], [4], [5].

For all studies, ground force throughout the foot was measures in Newtons with an *Ultraflex Computer Dyno Graphy* device with 8 sensors placed under each foot (refer to Figure 1 for the approximate sensor locations). Subjects were instructed to walk at a self-selected pace on level ground and data was recorded for 2min at a frequency of 100Hz for a total of $12,000$ time points per recording.

For this particular study we focus on the data from Yogev *et al.* [23] as it is the only project from *PhysioNet* that contains two recordings for some subjects — one recording of a subject walking normally, and a second recording of a subject walking while under a cognitive load (a serial-7s subtraction task[1]). Of the data recorded from this study that we obtained, there are a total of six Control subjects (2 male, 4 female) and 21

[1] Counting backwards from some number (typically 100) by sevens.

978-1-7281-6929-3/20/$31.00 ©2020 IEEE

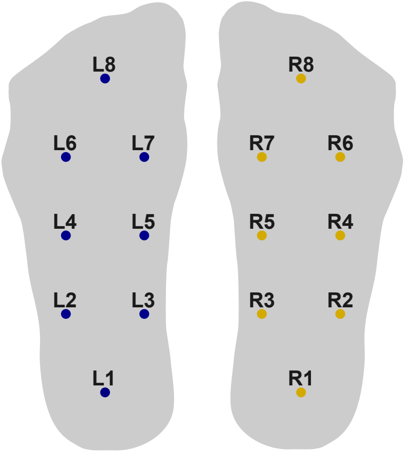

Fig. 1. Positions of the force sensors/devices placed on the soles of the subject's feet. Sensors/devices labeled with L are on the left foot and those with an R are on the right foot [13].

PD patients (17 male, 4 female) that performed both walking experiments (27 unique subjects and 54 total recordings).

It should be noted that some issues arise from this dataset as a result of the uneven number of male and female subjects and the small number of Control subjects. Throughout the paper we are careful to emphasize the constraints imposed by the dataset and its impact on statistical significance. Despite the limitations, we demonstrate the effectiveness of our pipeline and observe several trends. When additional data is available, further analysis can be performed which may be used to confirm and expand upon trends seen in the current data.

A. Preprocessing

Minimal preprocessing was done to the data. All data was z-score normalized (standard score) to make model generalization easier as the data may not have been scaled in any meaningful way. Each recording included aggregate values from all sensors on each foot, which were removed from the data before modelling. Lastly, each $2\mathrm{min}$ recording was divided into five equal parts of $24\mathrm{s}$ ($2,400$ time points). This was done to reduce modelling runtimes and to provide unseen data for simple testing. One subject's recording was slightly less than $2\mathrm{min}$, however their data was still divided into five equal parts that came out to less than $24\mathrm{s}$ each.

After all preprocessing, a total of 270 sets of data were produced (27 subjects total, two experiments per subject, five sets for each).

B. Data Terminology

Within this work we refer to different levels/resolutions of data with specific names. For each *subject*, we divided the recordings into five *sets*. We also refer to the normal walking and walking while under a cognitive load as the *experiment sets*. We use the word *cohort* to refer to a collection of the same type of subjects (*Control subjects* or *PD patients*), regardless of their experiment set. Table I summarizes how each cohort's experiment sets are labelled throughout this paper.

TABLE I
Explanation of Cohort and Experiment Sets

Label	Meaning
Control x *Normal*	Control subjects walking under normal conditions
Control x *Cog*	Control subjects walking under cognitive load
PD x *Normal*	PD patients walking under normal conditions
PD x *Cog*	PD patients walking under normal conditions

III. Algorithms and Methodology

Since we are performing regression analysis, we are looking to find some function $\hat{y} = f(X)$, where $\hat{y}$ is some *predictor* for a dependent variable y. In our case, we use sensor R_8 (very front of the right foot) as our dependent variable and all other sensors are the independent variables (X). The choice of R_8 was arbitrary.

Although we are generating a predictor, it should be noted that the motivation for this regression analysis is not truly prediction, but to build a temporarily independent symbolic model that describes how the sensors relate to one another as their values change. The resulting models can then be analyzed to gain insight about the underlying system. If the generated models are representative of the underlying system, then they should also be effective predictors. But it is to be noted that the true goal of the study is to generate models that will enable understanding of the system. Further, we use the error between the predicted and expected values as our measure of model accuracy. Overall this difference may seem subtle, but the difference is emphasized to frame the motivation for creating these models.

Unlike previous work [13], [14], we do not include a comparison to *Ordinary Least Squares* (OLS) and *Least Absolute Shrinkage and Selection Operator* (LASSO) regression. Although these regression techniques are capable of producing high-quality results that are the easiest to understand, we excluded them as they performed the worst of all modelling techniques in previous work and were limited to only linear models. Further, although *Artificial Neural Network* models were generated for this phase of the project, they were excluded from the analysis here since they not only performed worse than XGBoost but also are much less explainable than both SR and XGBoost. The authors suspect this is a consequence of the minimal amount of data currently available.

A. Genetic Programming Implementation

A custom built Genetic Programming (GP) system was used for Symbolic Regression (SR) [12]. The system was based on one designed by Schmidt *et al.* [22], which incorporates improvements for SR. These improvements include an *acyclic*

TABLE II
GP SYSTEM PARAMETERS.

Elitism	1 (Single top candidate solution)
Population	101
Subpopulations	7
Generations	100,000 (1,000 per migration)
Migrations	100
Crossover	80%
Mutation	10% (x2 chances)
Fitness Metric	Mean Squared Error: $\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2$
Language	$+, -, ,*, /, exp, abs, sin, cos, tan$
Max # Graph Nodes	64
Predictors	10
Predictor Pop. Size	10% of whole dataset
Trainers	8

graph representation [19] and *fitness predictors* [21], [20]. Briefly, the acyclic graph representation is useful for SR as it provides a lightweight encoding, scales well, avoids bloat, and allows the search to reuse subexpressions. The fitness predictors reduce the cost of fitness evaluation by approximating the local search gradient by fitting to a small subset of data. Further, the subset changes through evolution which provides a mechanism to prevent overfitting and to focus the search on areas that need more improvement. For more information on these improvements, please see their respective sources.

The settings used for the GP system are the same as those used in previous work [13], [14]. These settings can be found in Table II. These values were determined empirically over multiple studies on modelling human gait. One point crossover and single point mutation were used as the genetic operators. Given the stochastic nature of GP, 50 models were generated for each set of data from each subject. This was done to improve the likelihood of generating more high-quality models (although, nearly all models generated were effective), to allow for an analysis of feature *importance* (see Section IV-B), and to improve the statistical analysis. From the 50 resulting models for each set, no significant model selection strategy was done other than the naïve approach of simply selecting the model with the lowest *training* error.

B. Extreme Gradient Boosting

XGBoost is a popular machine learning algorithm that generates an ensemble of gradient boosted decision trees [1]. We use the Python implementation provided by Chen *et al.* [1] in this work. The algorithm scales reasonably well and can produce models quickly, especially when compared to SR. Although the models generated by XGBoost are not as interpretable as a closed form mathematical expression like those generated by SR, one can still analyze the resulting XGBoost models relatively easily.

The same XGBoost settings used in the previous work studying PD patient gait were used here [14]. Each model was created with 200 trees (estimators), a learning rate of 0.05, training ratio of 0.75 (subsample), a subsample ratio of features used when building trees of 1 (colsample), and a maximum tree depth of 5. These values were determined empirically and produced high-quality results.

Since each regression of XGBoost would typically produce the same model, only one model was generated for each set of data for a total of 270 models.

IV. Results and Discussion

A. Model Quality

Table III presents summary statistics of the model effectiveness on various groupings of data. The groupings correspond to median model effectiveness when applied to the data the models were fit to (*Training*), unseen data from the same subject (*Testing*), all data from the same cohort and experiment set (Cohort), and all data from the same cohort, but alternative experiment set (*Other*) — for example, if models were fit to Control subjects walking normally, then for *Other*, the errors are those obtained when applying those models to data from Control subjects walking while under a cognitive load. Since the number of samples for the Control subjects was small, and we do not assume normality, we use median and interquartile range. The table includes a p-Value obtained by a Mann-Whitney U test comparing the distribution of errors obtained for the *Cohort* set against the *Other* set.

Figure 2 shows a collection of p-Value matrices comparing the distributions of error values obtained from the different cohorts, experiment sets, and modelling algorithms over the various groupings of data (*Training*, *Testing*, *Cohort*, and *Other*) presented in Table III; each matrix corresponds to a single row from Table 2. All p-Values were obtained with a Mann-Whitney U test.

The first and most obvious observation is that, regardless of the modelling strategy, as the generality of the groupings increases (*Training* to *Testing* to *Cohort*), the error values also increase. Given the amount of increase there are signs of overfitting the training data, however this is not atypical in terms of expectations and the error values on all sets are reasonable.

Table III shows that the XGBoost models performed much better than the models generated with SR. When referring to the p-Values in Figure 2 it can be seen that there is always a significant difference between the SR and XGBoost model performances, with the exception of SR models on *Control* x *Normal* versus XGBoost models on *Control* x *Cog*. It is noteworthy that in all other cases the XGBoost models on the *PD* x *Cog* data performed significantly better than the SR models on the more consistent *Control* x *Normal* data. In other words, XGBoost performed so well that it was capable of fitting data from the more inconsistent *PD* x *Cog* data better than SR fit the most consistent data, namely *Control* x *Normal*.

In a number of cases the SR models fit the PD data better (for both normal walking and cognitive load) than the Control subjects. This is not what one would expect since PD data should be more inconsistent, however the authors suspect this is a consequence of the small sample size of Control subjects. Further, this phenomenon is less noticeable for the XGBoost models and as the number of samples in the groupings being

TABLE III
MEDIAN MEAN ABSOLUTE ERROR VALUES WITH INTERQUARTILE RANGE FOR THE FOUR EXPERIMENT SETS.

		SR	SR (Cog)	XGBoost	XGBoost (Cog)
Control	Train	0.125 (± 0.016)	0.120 (± 0.035)	0.024 (± 0.006)	0.021 (± 0.013)
	Test	0.138 (± 0.036)	0.149 (± 0.064)	0.080 (± 0.028)	0.091 (± 0.063)
	Cohort	0.344 (± 0.267)	0.407 (± 0.250)	0.293 (± 0.166)	0.360 (± 0.168)
	Other	0.373 (± 0.203)	0.403 (± 0.265)	0.340 (± 0.157)	0.330 (± 0.177)
	p-Value	$1.0143 * 10^{-2}$	$\underline{1.512 * 10^{-1}}$	$3.504 * 10^{-13}$	$1.354 * 10^{-4}$
PD	Train	0.111 (± 0.044)	0.112 (± 0.033)	0.023 (± 0.006)	0.023 (± 0.007)
	Test	0.137 (± 0.062)	0.136 (± 0.048)	0.082 (± 0.039)	0.073 (± 0.028)
	Cohort	0.357 (± 0.196)	0.392 (± 0.210)	0.324 (± 0.149)	0.345 (± 0.161)
	Other	0.385 (± 0.199)	0.380 (± 0.209)	0.337 (± 0.162)	0.342 (± 0.153)
	p-Value	1.446^{-35}	$9.078 * 10^{-13}$	$1.619 * 10^{-20}$	$1.648 * 10^{-2}$

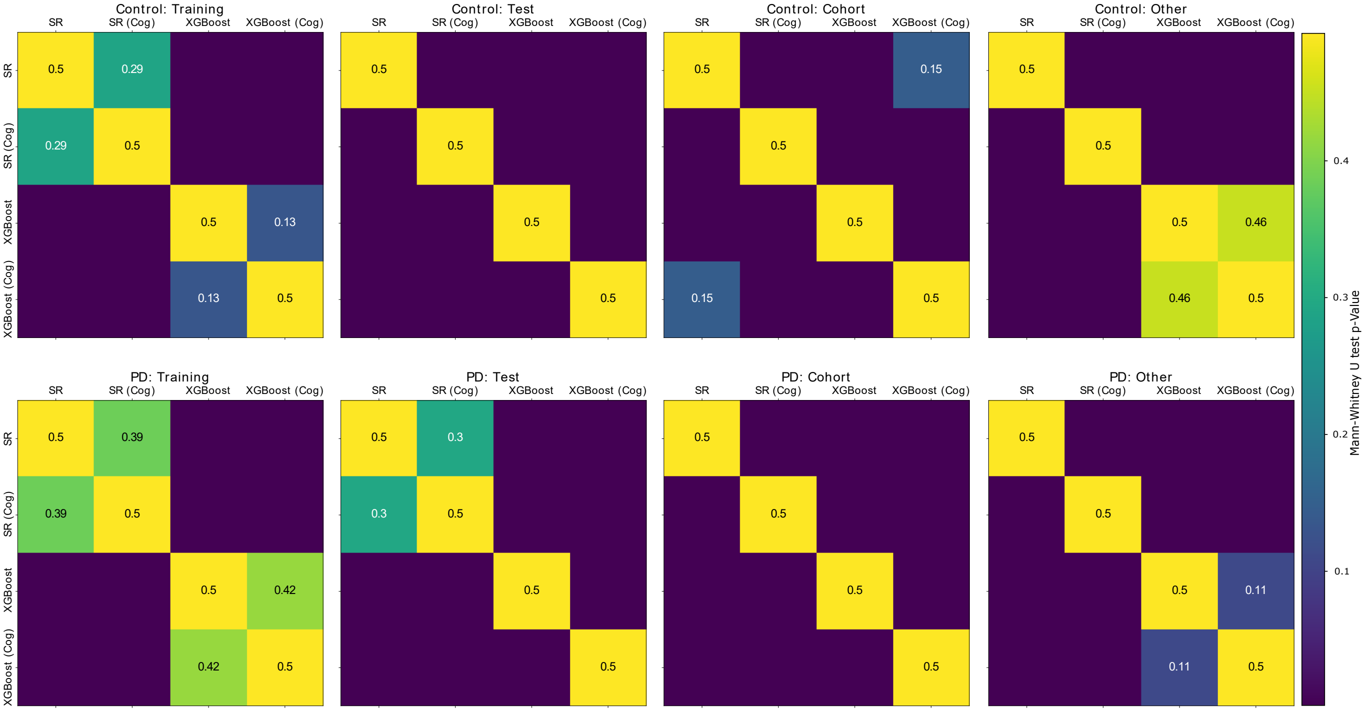

Fig. 2. Matrices of probability values obtained with a Mann-Whitney U test comparing the distributions of error values for various collections of data.

analyzed increases (for example, there are more Cohort data points than Training data points).

One particularly interesting observation from Table III is that if one takes models fit to subjects walking normally, either Control or PD patients, regardless of the modelling algorithm, they performed significantly *worse* when applied to data recorded from the same subjects when walking while under a cognitive load. This is reasonable as one would expect the data recorded while subjects were under a cognitive load to be more inconsistent, which would lead to more errors.

Conversely, if one takes models fit to subjects when walking while under a cognitive load, either Control or PD patients, regardless of modelling strategy, they perform *better* when applied to data recorded from the same subjects when walking normally. These improvements were significant in all cases except for the SR models of *Control* x *Cog* data (p-Value of 0.15). The improved results are still far off the error values obtained when using models fit to normal walking data, however these results suggest that even the models fit to the less consistent cognitive load data are still high quality and can fit the less noisy normal walking data.

When focusing on Figure 2 we can see that there is effectively no difference between the training errors of the normal walking and cognitive load data regardless of modelling strategy and whether they were Control or PD patients. This is reasonable since all modelling strategies are known to be effective and there is no consideration for generalizability for the training results.

For the PD data there was similarity between the distribution of testing errors obtained for normal walking and cognitive load data for the SR models.

As already discussed for the Cohort grouping, there was a similarity between the errors obtained by the SR models for the *Control* x *Normal* data and the XGBoost models for the *Control* x *Cog* data. In most cases the XGBoost models fit the less consistent cognitive load data significantly better than SR was able to fit the more consistent *Control* x *Normal* data. This demonstrates how effective XGBoost is with this data.

For the Other grouping there were similarities between the distributions of the XBoost models' errors for both sets of data on both the Control and PD data. Again, this shows how effective XGBoost is at fitting the data and generalizing well to unseen data that was recorded under different parameters (normal walking vs. cognitive load walking).

Figure 3 presents the median MAE values obtained when a given subject's model (column) was applied to data from a specific subject (row). Note the relationships between Figure 3 and Table III: The *Test* row from Table III corresponds to the diagonals in Figure 3, the *Cohort* row corresponds to the four segments along the diagonal, and the *Other* row corresponds to the adjacent segment vertically. For example, the top left square labelled Con vs. Con is the *Cohort* group, and the cell immediately below (Con (Cog) vs. Con) is when the models fit *Control* x *Normal* data were applied to *Control* x *Cog* data (*Other* row).

Figure 3 provides a *proxy* view of subject data similarity. For example, if a model fit to a specific subject is capable of fitting data from a different subject reasonably well (and *vice versa*), then perhaps physical manifestations of walking are similar for those two subjects. Despite the fact that these models are proxies, there are a number of observations within Figure 3 that match expectations.

Although there is a small sample size for Control subjects, there is a clear difference in the error values between Control subjects' and PD patient data, regardless of experiment set (normal or cognitive load). The four top left squares, corresponding to the Control subject models applied to Control subject data, are much darker (smaller error values) than the top right four rectangles (PD patient models applied to Control data) and the bottom left four rectangles (Control subject models applied to PD patient data). This similarly applies to the four bottom right squares, that correspond to PD patient models applied to PD patient data.

In Figure 3, one can also observe, for both Control and PD data with both modelling techniques, the diagonals having low error values (when models fit to a specific subject were applied to data from the same subject) regardless of the experiment set. For example, observe the low error values for the case when models fit to *PD* x *Normal* data were applied to *PD* x *Cog* data. Table IV presents a comparison of the errors obtained when models were applied to data from the same cohort to the errors obtained when models were applied to data from the same subject, but while performing the different experiment set (normal walking vs. cognitive load). In other words, this compares a whole segment of the error matrix to the diagonal of the neighbouring segment. In all cases, models were better able to fit the data from the same subject on the other experiment set than models were able to fit data from all subjects within the same experiment set.

B. Model Feature Analysis

Figure 4 shows heatmaps of information that can be used as a proxy for feature importance within the models. The top heatmap corresponds to the percentage of times a given feature appeared within all SR models generated for each of the subjects. Given the stochastic nature of GP, and the fact that all final models generated are of high-quality, if a given feature appears more often, then it is likely more important. The bottom heatmap shows the F-scores for each feature in all XGBoost models. F-score — the number of times a given feature was split on within the XGBoost model — can be used as a measure of feature importance for our purposes.

Despite having two heatmaps of different proxies for feature importance, there are similarities within Figure 4. The most notable similarity is that sensors R_4 – R_7 (also R_8 if including the dependant variable) appear to be important for both modelling strategies. A similar observation can be made for sensors L_1 – L_3, although to a lesser extent. For the most part, sensors L_4 – L_8 and R_1 – R_3 do not appear to be as important for both the resulting SR and XGBoost models. When referring to Figure 1, the sensor locations of R_4 – R_8 and L_1 – L_3 correspond to the front of the right foot and the back of the left foot respectfully. As noted in previous work studying a different set of ground force data [13], [14], given the physical manifestation of walking, it would make sense that these portions of the feet would be related to one another as they would be in contact with the ground at the same time. Since R_8 was the sensor being regressed to, all sensors that were activated at the same time as R_8 would be more likely to be included in the models and considered *important*.

Although the authors have no explanation for it, an interesting difference between the heatmaps of the models is the inclusion of R_1 in the XGBoost models, but not in the SR.

Figures 5 and 6 show similar information for the feature importance measures. Figure 5 presents the average percentage of times a given feature (sensor) appeared in all models generated by SR for the experiment set. Figure 6 is similar, but is the average F-scores for the XGBoost models. Figures 5 and 6 also include p-values obtained with a Mann-Whitney U test comparing the distributions of feature importance measures between different experiment sets. For the p-value matrices (right side of both figures), the first column compares *Control* x *Normal* subjects to *Control* x *Cog* subjects. The second column is similar to the first, but for PD patients. The third compares the features for *Control* x *Normal* with *PD* x *Normal*. The fourth compares *Control* x *Cog* and *PD* x *Cog*.

Again, the authors emphasize the small sample size of Control subjects. Although we do an analysis using the small

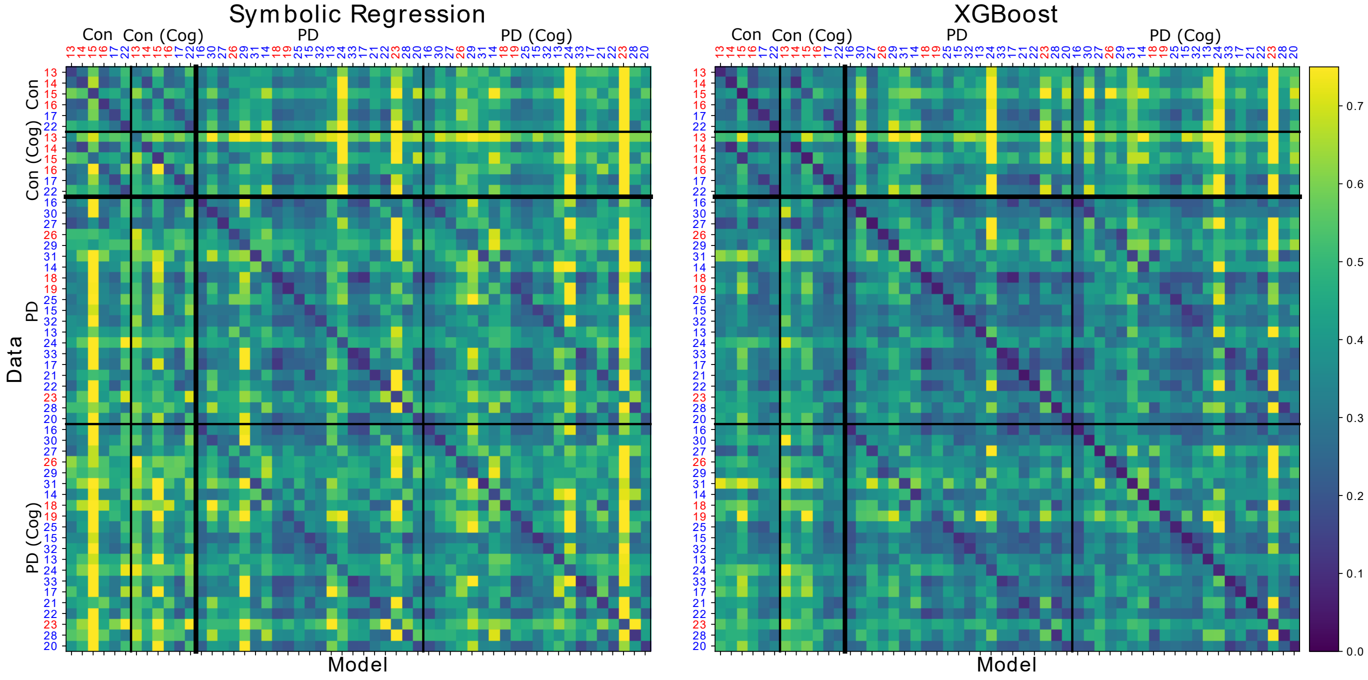

Fig. 3. Matrices showing the median mean absolute error values when a given model (column) is applied to data from a specific subject (row) for both the models generated with SR and XGBoost. The median was calculated over the five models and data segments for each subject. The colour of the labels represents subject sex (blue for male and red or female). Each row and column are divided into four parts: the first (left most) is the case when Control subjects walking normally, the second is for Control subjects walking while under a cognitive load, the third is for PD patients walking normally, and the fourth (right most) is for PD patients walking while under a cognitive load. Control subjects are ordered by subject number and PD patients are ordered based on their UPDRS rating in ascending order. Error values were capped at 0.75 for viewing purposes; any value of 0.75 (bright yellow) should be interpreted as a poor fit and not necessarily 0.75.

TABLE IV

Comparison of cohort errors to the errors obtained when applying data from the same subject to models fit to the different experiment set (diagonal in adjacent segments in Figure 3). Probability value was obtained with a Mann-Whitney U test.

		SR			XGBoost		
		Cohort Median	Other Diagonal Median	p-Value	Cohort Median	Other Diagonal Median	p-Value
Control	Normal	0.344	0.182	$4.052 * 10^{-25}$	0.293	0.153	$1.014 * 10^{-12}$
	Cognitive Load	0.407	0.187	$9.678 * 10^{-29}$	0.360	0.126	$4.889 * 10^{-28}$
PD	Normal	0.357	0.183	$2.988 * 10^{-153}$	0.324	0.166	$3.950 * 10^{-171}$
	Cognitive Load	0.392	0.183	$5.433 * 10^{-180}$	0.345	0.146	$1.467 * 10^{-212}$

sample size, more subjects are needed in order to make better conclusions.

The average matrices reinforce what was observed above when discussing Figure 4, including the unexplained appearance of R_1 within the XGBoost models, but they also highlight important differences between the different experiment sets. When focusing on the SR models (Figure 5), immediately it becomes obvious that L_1 becomes less important when Control subjects change from walking normally to walking while under a cognitive load. Similarly, in the PD models, regardless of experiment set (normal walking or cognitive load) L_1 is less important when compared to Control subjects. In general, it seems that the very back and outside back of the left foot becomes less important in the PD models versus Control, but the inside of the back part of the left foot becomes more important. Perhaps the PD patients are less likely to put much pressure on their left heel compared to Control subjects. PD subjects also seem to significantly change how the pressure is distributed through the front of the right foot (i.e. more importance at the very front of the foot) when they walk while under a cognitive load versus normal walking.

Many differences can also be seen in the XGBoost models (Figure 6), however most of these changes are between Control and PD. Although there are a few significant changes in the feature importance while the subjects transitioned from normal to cognitive load walking, mostly the less important features changed (L_4 – L_8 and R_1 – R_3). Perhaps the most interesting change is that the front of the right foot became less important when subjects changed from normal to cognitive load walking; this was also seen in the SR models but the changes were not significant. Unlike the SR models, there was no real change to the back of the left foot. Many more significant differences can

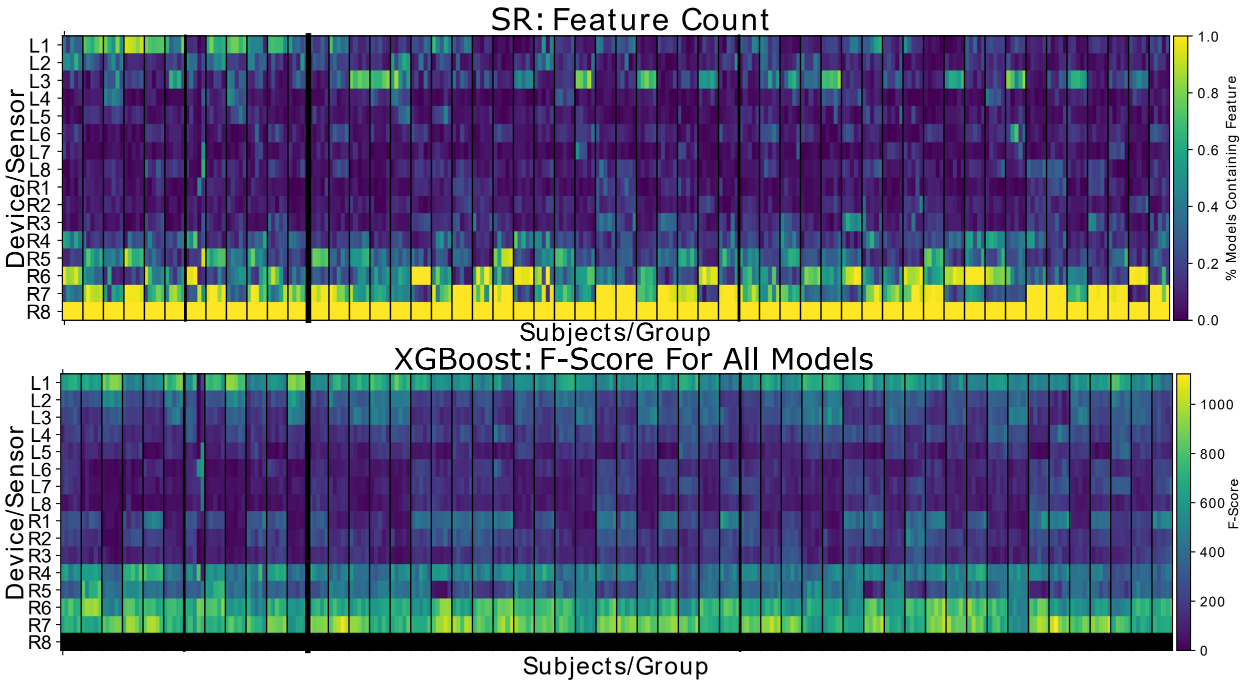

Fig. 4. Heatmaps of feature (row) importance proxies for each of the five sets of data for each subject. The heatmaps are broken into four segments corresponding to the four experiments. The left most segment corresponds to Control subjects walking normally, the second to Control subjects walking while under a cognitive load, the third is PD patients walking normally, and the last (right most) is PD patients walking while under a cog native load. The top heatmap shows the percentage of times a given feature appeared in all symbolic regression models generated for a given set of data. Note that R_8 is always included in these models are it is the dependant variable (left hand side of the equation). The bottom heatmap presents the F-scores for all models generated with XGBoost. Note that R_8 is not recorded, but is still the dependant variable (left hand side of the equation).

be seen when focusing on the differences between the Control and PD patient feature importance. Although some of these can be seen in the SR models (e.g. changes in importance of left heel), many more differences were found by the XGBoost models. This suggests that the XGBoost models are much better at distilling the differences in the physical manifestation of walking between Control subjects and PD patients, and SR emphasizes the differences between normal and cognitive load walking.

V. Conclusions and Future Work

This study has demonstrated that both SR and XGBoost are successful in identifying features that are important to discover gait changes for PD diagnoses. The models produced by SR are more explainable than those produced by XGBoost, however XGBoost has higher overall accuracy and are still interpretable. In general, these two methodologies are in agreement with each other in terms of the most important features, i.e. identifying the front and back of the foot as being of high importance. There was also an indication that that there was a difference between the pressure placed on outside of the foot vs the inside of the foot, for PD patients vs Control subjects.

Two different means of measuring feature importance were used, namely average feature count for SR and average F-score for XGBoost. Despite this, the two methodologies largely support each others' conclusions. Feature importance significantly differed between the *Control* x *Normal*, *Control* x *Cog*, *PD* x *Normal*, and *PD* x *Cog* experiment sets.

Currently the largest limitation of this project is the minimal data to which researchers have access. More data is required to increase the statistical significance of the results and to confirm conclusions.

It would also be interesting to analyze other types of data, such as wearable technology, to expand upon the results of the current study.

A deeper analysis into the models should be performed to gain the full benefit of the closed form mathematical expressions generated by SR. For example, the nonlinear models generated (SR, XGBoost, and ANNs) greatly outperform the linear models (OLS and LASSO) developed in previous work [13], [14]. The SR models may be analyzed to identify specific nonlinearities that arise that cannot be found in linear models.

VI. Acknowledgements

This research was supported by ACEnet (ace-net.ca/), Compute Ontario (computeontario.ca), WestGrid (westgrid.ca), Compute Canada (computecanada.ca), and the Natural Sciences and Engineering Research Council of Canada (NSERC).

The authors thank Dr. Matea Drljepan for her artwork.

References

[1] T Chen and C Guestrin. Xgboost: A scalable tree boosting system. In *Proc. 22nd ACM sigkdd intl. conf. on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

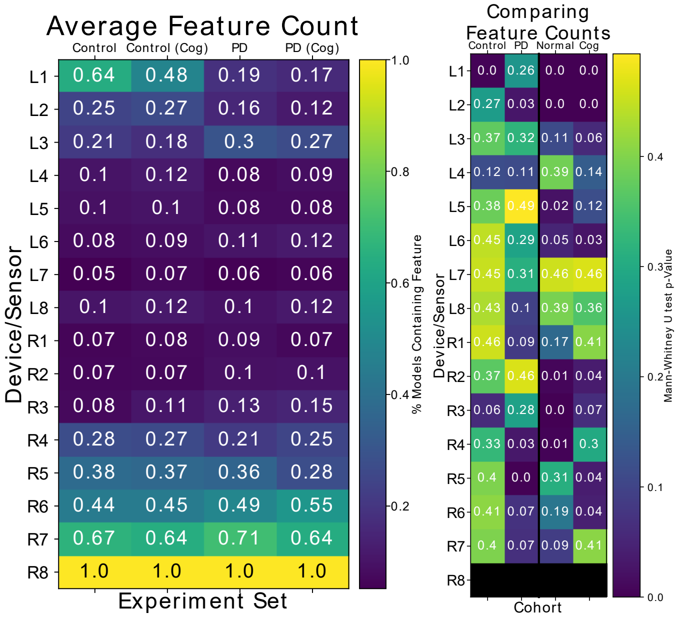

Fig. 5. Matrices showing summary statistics for the feature importance measure (percentage of times a feature appeared in all models for a set of data) used for the symbolic regression models. The left matrix shows the average number of times a given feature (row) appeared in all models generated for a given experiment set. The right matrix presents Mann-Whitney U tests comparing the distributions of average feature counts between experiment sets. In the right matrix, the first column (left most) compares the normal and cognitive load for the Control subjects, the second is the normal and cognitive load for PD patients, the third is the Control versus PD patient when walking normally, and the last (right most) compares Control and PD patents when walking under a cognitive load.

[2] A Entezari Heravi, K Tahmasebipour, and S Houghten. Evolutionary computation for disease gene association. In *2015 IEEE Conf. on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–8. IEEE, 2015.

[3] E. Dorsey et al. Global, regional, and national burden of parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet*, 17(11):939–953, 2018.

[4] S Frenkel-Toledo et al. Effect of gait speed on gait rhythmicity in Parkinson's disease: variability of stride time and swing time respond differently. *J. neuroengineering and rehabilitation*, 2(1):23, 2005.

[5] S Frenkel-Toledo et al. Treadmill walking as an external pacemaker to improve gait rhythm and stability in Parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, 20(9):1109–1114, 2005.

[6] T Gasser. Genetics of Parkinson's disease. *Current opinion in neurology*, 18(4):363–369, 2005.

[7] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[8] Elinor Harrison, Adam Horin, and Gammon Earhart;. Mental singing reduces gait variability more than music listening for healthy older adults and people with parkinson disease. *Journal of Neurologic Physical Therapy*, 43(4):204—211, 2019.

[9] J M Hausdorff. Gait in Parkinson's Disease database v1.0.0. https://physionet.org/content/gaitpdb/1.0.0/. Accessed: Sept. 10, 2019.

[10] J M Hausdorff et al. Rhythmic auditory stimulation modulates gait variability in Parkinson's disease. *European Journal of Neuroscience*, 26(8):2369–2375, 2007.

[11] F B Horak and M Mancini. Objective biomarkers of balance and gait for Parkinson's disease using body-worn sensors. *Movement Disorders*, 28(11):1544–1551, 2013.

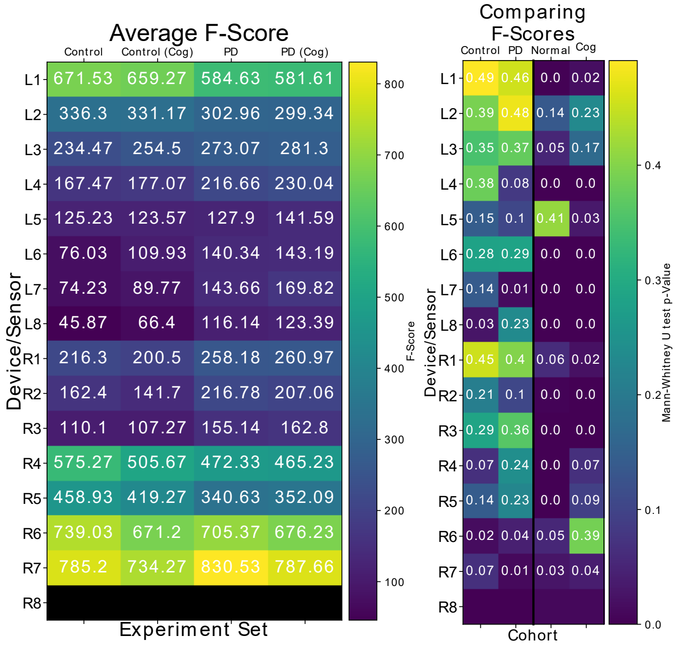

Fig. 6. Matrices showing summary statistics for the feature importance measure (F-score for a given model) used for the XGBoost models. The left matrix shows the average F-score of a given feature (row) for all models generated for a given experiment set. The right matrix presents Mann-Whitney U tests comparing the distributions of average feature counts between experiment sets. In the right matrix, the first column (left most) compares the normal and cognitive load for the Control subjects, the second is the normal and cognitive load for PD patients, the third is the Control versus PD patient when walking normally, and the last (right most) compares Control and PD patents when walking under a cognitive load.

[12] J A Hughes. jGP. https://github.com/jameshughes89/jGP, March 2015. Accessed: January 30, 2020.

[13] J A Hughes, S Houghten, and J A Brown. Descriptive symbolic models of gaits from Parkinson's disease patients. In *2019 IEEE Conf. on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE, 2019.

[14] James Hughes, Sheridan Houghten, and Joseph Alexander Brown. Models of parkinson's disease patient gait. *IEEE Journal of Biomedical and Health Informatics*, 2019.

[15] C Klein and A Westenberger. Genetics of Parkinson's disease. *Cold Spring Harbor persp. in medicine*, 2(1):a008888, 2012.

[16] D B Miller and J P O'Callaghan. Biomarkers of Parkinson's disease: present and future. *Metabolism*, 64(3):S40–S46, 2015.

[17] R L Nussbaum and C E Ellis. Alzheimer's disease and Parkinson's disease. *New england journal of medicine*, 348(14):1356–1364, 2003.

[18] M Plotnik, N Giladi, and J Hausdorff. A new measure for quantifying the bilateral coordination of human gait: effects of aging and Parkinson's disease. *Experimental brain research*, 181(4):561–570, 2007.

[19] M Schmidt and H Lipson. Comparison of tree and graph encodings as function of problem complexity. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1674–1679. ACM, 2007.

[20] M Schmidt and H Lipson. Coevolution of fitness predictors. *IEEE Transactions on Evolutionary Computation*, 12(6):736–749, 2008.

[21] Michael D Schmidt and Hod Lipson. Coevolving fitness models for accelerating evolution and reducing evaluations. In *Genetic Programming Theory and Practice IV*, pages 113–130. Springer, 2007.

[22] Michael D Schmidt, Ravishankar R Vallabhajosyula, Jerry W Jenkins, Jonathan E Hood, Abhishek S Soni, John P Wikswo, and Hod Lipson. Automated refinement and inference of analytical models for metabolic networks. *Physical biology*, 8(5):055011, 2011.

[23] G Yogev et al. Dual tasking, gait rhythmicity, and Parkinson's disease: which aspects of gait are attention demanding? *European journal of neuroscience*, 22(5):1248–1256, 2005.