Is Human Walking a Network Medicine Problem? An Analysis Using Symbolic Regression Models with Genetic Programming

Pritika Dasgupta[1], James Alexander Hughes[2], Mark Daley[3], Ervin Sejdić[1,4,5]

[1]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15261, USA

[2]Department of Computer Science, St. Francis Xavier University, Antigonish, Nova Scotia, B2G 2W5, Canada

[3]Department of Computer Science, Middlesex College, University of Western Ontario, London, Ontario, N6A 3K7, Canada

[4]Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA

[5]Department of Bioengineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA, 15261, USA

Corresponding Author:

Pritika Dasgupta, MPH MHI MS

Department of Biomedical Informatics, School of Medicine, University of Pittsburgh,

5607 Baum Blvd,

Pittsburgh, PA, 15206, USA

Tel: 763-742-7036

Email: prd17@pitt.edu

ABSTRACT

Background and Objective:

Human walking is typically assessed using a sensor placed on the lower back or the hip. Such analyses often ignore that the arms, legs, and body trunk movements all have significant roles during walking; in other words, these body nodes with accelerometers form a body sensor network (BSN). BSN refers to a network of wearable sensors or devices on the human body that collects physiological signals. Our study proposes that human locomotion could be considered as a network of well-connected nodes.

Methods:

While hypothesizing that accelerometer data can model this BSN, we collected accelerometer signals from six body areas from ten healthy participants performing a cognitive task. Machine learning based on genetic programming was used to produce a collection of non-linear symbolic models of human locomotion.

Results:

With implications in precision medicine, our primary finding was that our BSN models fit the data from the lower back's accelerometer and describe subject-specific data the best compared to all other models. Across subjects, models were less effective due to the diversity of human sizes.

Conclusions:

A BSN relationship between all six body nodes has been shown to describe the subject-specific data, which indicates that the network-medicine relationship between these nodes is essential in adequately describing human walking. Our gait analyses can be used for several clinical applications such as medical diagnostics as well as creating a baseline for healthy walking with and without a cognitive load.

Keywords: walking, genetic programming, mathematical model, symbolic regression, wearables, acceleration gait measures

1. INTRODUCTION

The human body is a network of moving parts, and the concept of “network medicine” can investigate how these moving parts interact. Network medicine is the biometric concept of modeling and identifying a person by using their body attributes [1-9]. This branch of clinical analysis overlaps with the goals of the fields such as “personalized medicine” or “systems biology.” Specifically, it refers to the concept of considering the relationship between the unique traits or features of a specific individual to make diagnostic decisions.

In particular, the topics of network medicine and human gait analysis have been investigated by multiple medicine-adjacent fields, such as biomechanics, computer science, and robotics [2-3]. Using model-based approaches, researchers have modeled the human body's motion through the individual's body structures [1,10]. Walking requires nodes of the body, such as the arms, legs, chest, and lower back, to move the body through space while preserving stability and balance [11]. While walking, the body's center of gravity vacillates between the right and left sides; the coordination of all these nodes makes walking a complex task, where all the nodes of the body must work in concert [2,11-14]. Viewing walking as a network can help create individual-specific gait models based on walking patterns [15].

In walking-related network medicine studies, how walking can be modeled using the current statistical and analytical approaches is the open research question. In conjunction with signal processing algorithms and wearable devices, machine learning algorithms may provide a way to monitor walking as a network [16]. Gait data collection is often done through non-invasive, inexpensive accelerometers, which can be placed in various locations on the body as a body sensor network (BSN), allow for continuous monitoring, and are used in clinical settings [17-19]. The resulting datasets often result in large time-series datasets per subject, which has been shown to give favorable results in studies with low sample sizes [19-22]. While these datasets can sometimes be "noisy," where some of the data are artifacts of the accelerometer, signal processing methodology analysis of these datasets can provide reliable measurements and the creation of clinical gait variables.

Symbolic regression (SR) networks are machine learning models that may be more adaptable and robust than other statistical approaches in modeling walking behavior [23]. SR is similar to other regression techniques. SR searches for parameters for a mathematical model to fit data; however, unlike linear regression, SR also frames the structure and operators within the model. SR has been widely used in medicine and physiology to study various topics, such as transcriptomics, metabolomics, the dynamics of the human gut microbiome, and the cardiovascular signals in sleep apnea patients [24-26]. SR has also been used for signal modeling in modeling bipedal locomotion and physiological signals [27-28]. In gait studies, genetic programming (GP) can be used to search through clinical gait variables for the best predictors to put into SR models of human locomotion [29-32].

We hypothesize that human locomotion can be considered as a network of well-connected nodes, or areas of the body. The objective of this study is to use GP to perform SR on accelerometer data to generate human BSN models for each individual in the study. By doing so, a “network of

walking" will emerge where the models are a representation of distributed nodes, proxied by the placement of accelerometer on different parts of the body, and their interactions. Explicitly, this study records and uses signal processing methods to transform gait accelerations from six separate accelerometers on the body, which are referred to as the nodes in the human network. The novelty of this approach is that these SR models can be used for gaining insight into the underlying kinematics and can also be used as a predictive model, without the use of other machine learning algorithms, such as artificial neural networks or support vector machine, since the models should represent the system's metastable state.

2. MATERIALS AND METHODS

2.1 Participant Demographics

This study included ten human volunteers (five male and five female) from the University of Pittsburgh in Pittsburgh, Pennsylvania, USA. The ages of the participants ranged from 18-35 years, with a mean of 21.40 and a standard deviation of 4.38. The mean height was 1.72 m (sd = 0.09) and the mean weight was 66.36 kg (sd = 8.41). Based on these basic measurements, the participants were considered "healthy" individuals. Further information about the participants is found in Table 1 of Dasgupta et al. [19].

2.2 Materials

Six wGT3X-BT triaxial accelerometer sensors (produced by ActiGraph LLC, Fort Walton Beach, Florida, USA) were placed on participants' chest (C), bilateral ankles (left ankle (LA) and right ankle (RA), wrists (left wrist (LW) and right wrist (RW), and lower back (LB) (a figure of these locations can be found in [19]). These sensors are known to be accepted monitoring sensors, and they present minimal risks to participants. Each sensor collected linear accelerations (meters per second squared) at a frequency of 80 Hz over from the mediolateral (ML), vertical (V), and

anteroposterior (AP) directions. Treadmills were set at a constant speed of 2.2 miles per hour (or 0.98 m/s).

2.3 Data Collection

Data were collected from ten participants from two sessions. These two sessions were wholly identical and were scheduled at least 48 hours or more apart.

For each session, there were five phases: 1) the six sensors were fastened on the participants, 2) the first walking trial, consisting of walking for 10 minutes, 3) the participants were asked to rest for a period of 10 to 20 minutes, 4) the second walking trial, consisting of walking for 10 minutes with a counting activity, and 5) then the sensors were unfastened from the participant.

The first trial consisted of walking on the treadmill, and the second trial involved walking while under an arithmetic cognitive load – counting backward from 10,000 in increments of 7. During each trial, each of the six sensors transmitted 48,000 time points for each of the mediolateral (ML), vertical (V), and anteroposterior (AP) directions, resulting in 18 streams of data for each participant for each trial.

2.4 Experimental Design

First, PCA was performed on each of the six sensors for all recording session. Each sensor recorded three dimensions or axes (ML, V, and AP); however, variations in the physical orientation of the sensor may have had negative consequences as the dimensions may be askew. Thus, for each of the sensors, the ML, V, and AP values were transformed into three principal components. Even if the ML, V, and AP axes were misaligned, the PCA will linearly transform and order the axes based on the amount of variation in the recorded data. For these principal components (PCs), the first dimension on all sensors had the most variation, the second orthogonal dimension had the second most variation, and the third orthogonal dimension had the least variation.

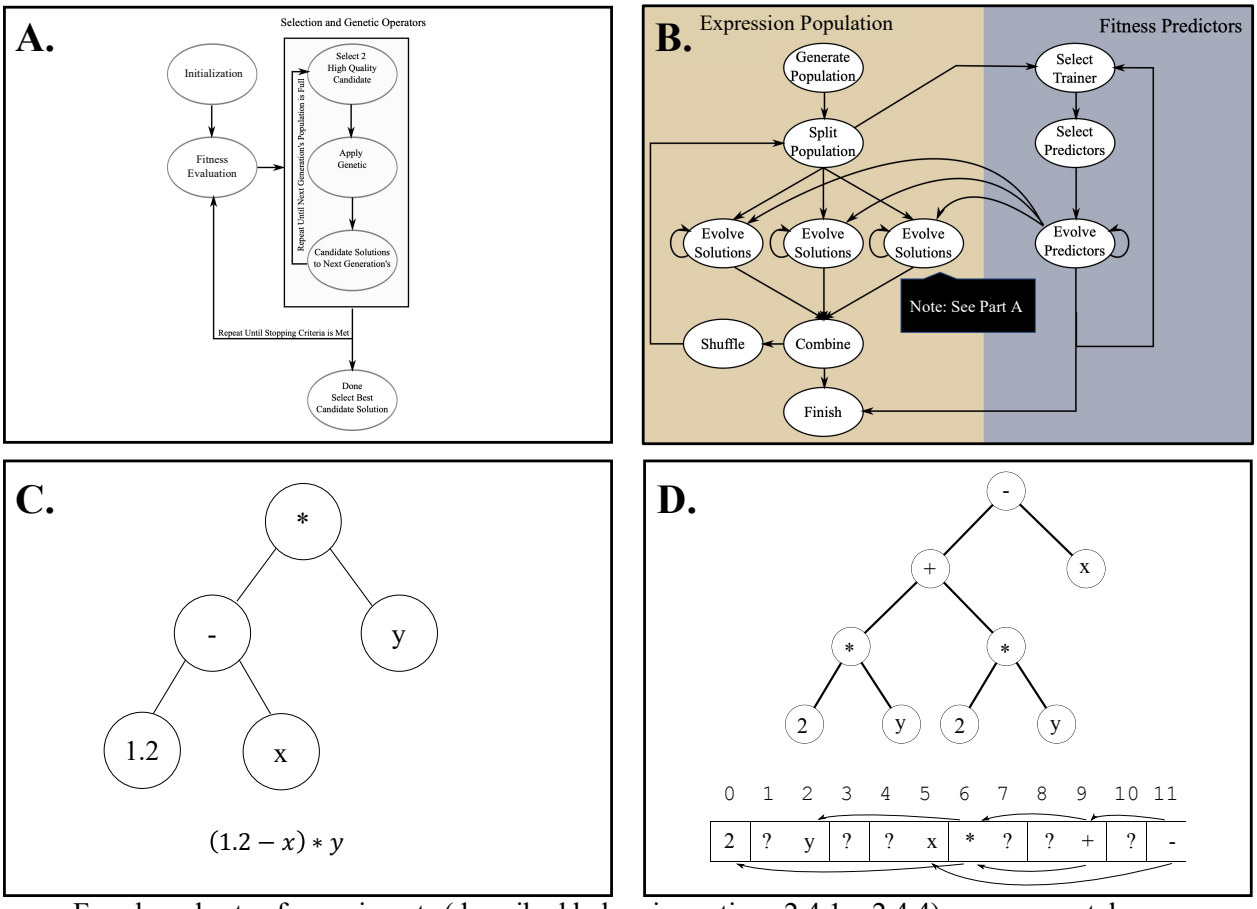
Evolutionary searches were used to find a candidate solution with a high-quality fitness value (Figure 1A). Typically, the evolutionary algorithm (EA) will run for a high number of generations as it will take some time before the algorithm can converge on a high-quality solution. EAs typically follow the same structure: 1) the initialization phase, 2) the fitness evaluation phase, 3) the selection of candidate solutions phase, and 4) the application of genetic operators phase (Figure 1A). In this study, symbolic regression is performed, and thus the fitness value is the mean squared error of the regressed model to the actual recorded signal (the typical fitness measure when performing symbolic regression). The parameters of the GP were optimized using this fitness metric through the evolutionary search.

The GP implementation used for this work is heavily inspired by our previous work, Hughes et al., and Schmidt et al.'s work and shown in Figure 1B [23, 31]. The symbolic regression models are $\hat{y} = f(X_1, X_2, \ldots, X_m)$, where the function f is some combination of the independent variables and linear and nonlinear basis functions defined by the GP system's language (Figure 1B) [17]. Unlike traditional GP systems where the chromosomes are tree-based S-expressions, such as the one shown in Figure 1C, this work uses an array-based acyclic graph representation of the underlying mathematical expressions. Figure 1D, which is derived from Hughes et. al, shows a comparison between the tree-based encoding and an array encoding of the acyclic graph; observe that only a subset of the genes encode information that make up the expression to be evaluated [32]. In Figure 1D, some nodes are also non-coding genes, denoted by question marks [32]. These vestigial genes will not negatively impact the fitness of a candidate solution and may even show up in later generations with positive effects [32].

Figure 1. This four-part figure describes the basics of this paper's implementation. **Part A** is a high level overview of the EA process. This process will repeat until a stopping criteria is met. Once complete, the candidate solution with the best fitness will be taken. The final population will

contain a collection of high-quality solutions. **Part B** describes the GP implementation, where the concepts of subpopulations, acyclic graph representation, and fitness predictors are used. Note that the step denoting "Evolve solutions" refers to the process in Part A. **Part C** is an example of a tree-based S-expression, which is how the encoding or representation of the GP process is done. **Part D** is an example array encoding of the acyclic graph representation. Both the tree and the array encode

$$(2 * y) + (2 * y) - x.$$



Four broad sets of experiments (described below in sections 2.4.1 – 2.4.4) were separately performed on the same data, which includes both walking regularly and walking under cognitive load. For every individual, session, and trial, each sensor's data (48,000 data points) were broken into ten collections (or batches) of 4,800 time points. For statistical significance and, given the

stochastic nature of GP, to increase the chance of generating a high-quality model, 100 models were generated for each of the batches of 4,800 time points. A total of 40,000 models were generated for each of the four-broad experiment sets described above (4 experiments *10 subjects *2 sessions * 2 trials * 10 batches * 100 = 160,000 in total).

For generating these models, one of the directions/axes from a single accelerometer was selected to be fit to the dependent variable ($\hat{y}$), and all other axis/features were used as independent variables of the equation (X). The goal is to develop a model of $\hat{y}$ in terms of X which accurately approximates $\hat{y}; y \approx \hat{y} = f(X)$.

A summary of the GP system settings for these runs are presented in Table 1. Throughout the evolutionary process, the mean squared error was used as the fitness metric. The number of mating events are low and the mutation rate of the GP is relatively high compared to traditional GP settings. Thus, preliminary results with longer run times showed significantly better results but provided minimal absolute improvements in the means. The higher mutation rate is a consequence of the acyclic graph representation as not all genes will be represented in the final expression, resulting in some mutations having no impact.

Table 1: Parameter settings for GP System. The values for migrations and generations per migration to 100 each when performing the analysis on models fit to a subset of data. The fitness metric and language for the GP system are listed below. Mutation refers to a single point mutation (change a value within a randomly selected index). Crossover refers to a single point crossover (select two candidate solutions, randomly select an index, then swap the elements between the two candidate solutions around that selected point).

Parameter	Values
Elitism	1
Population	101
Subpopulations	7
Migrations	5
Generations	100 per migration (500 total)
Crossover	80%
Mutation	10% (x2 chances)
Trainers	15
Predictors	20
Predictor Pop. Size	10% of Dataset
Max # Graph Nodes	16
Fitness Metric	Mean Squared Error: $\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2$
Language	$+, -, *, /, exp, abs, sin, cos, tan$

2.4.1 Experiment 1

The first experiment set was generated using only data collected from an accelerometer located on the subjects lower back due to it being the most typical accelerometer placement in the existing literature; both lower back and hip accelerometer placements have been touted to be the "best" location for providing data to detect an array of human activities [35-37]. The dependent variable ($\hat{y}$) was the axis with the most variance.

2.4.2 Experiment 2

The second experiment set included data recorded from the subjects' LA and RA, LW and RW, and C in addition to the lower back data. The dependent variable was the axis with the most variance from the LB.

2.4.3 Experiment 3

The third experiment set was the same as the second (included all axis from all six accelerometers), except the dependent variable was the axis with the most variance from the RA.

2.4.4 Experiment 4

The fourth experiment set was the same as the second, except the models were fit to only a tiny subset (100) of time points per data set. Since we only fit to a 100 timepoints, testing was two-fold: first, we tested the models against only 100 time points, and then we applied the models to all the data from each batch. These types of models are desirable as fitting to fewer data points will speed up run times, require less computing resources, and may be less susceptible to over fitting.

2.5 Computational Resources

Models were generated on a desktop computer with two quad-core (8 cores) i7-4770 CPUs at 3.4GHz with 8GB of RAM. Runtimes for models fit to 4,800 data points took between 7-11 s each, and the models fit only 100 data points (experiment set four) took between 0.4-0.5 s. Evolutionary searches have an element of stochasticity, and even though these runtimes are slow for linear regression, they are fast for a typical EA.

3. RESULTS

The results are organized in the following order: results on training and testing errors on the models generated for each subject (for the SR model and a comparison with an ordinary least squares regression model), examples of the model fits from Subject 10, feature counts in each specific subject, and the mean absolute percentage error matrices for each subject. For each experiment, from the 100 models generated for each subject, the one with the lowest training error was selected as the top model (no rigorous model selection strategy was done). Additionally, in our initial data analysis between the trials, clinical variables such as stride length, stride times and cognitive behavior had no noticeable difference between the trials (t-test p-value < 0.05). This is with respect to the idea that there would be a different manifestation of the action depending on if the subject's brain was under a load.

Table 2 presents the median training errors over all top models, for each subject's recording, (400 total) along with median testing error calculated by applying all top models to every batch from the same subject, excluding the batch the model was fit to (3,600 total); for each model (400 total), there were 9 unseen batches. For these, the mean absolute error (MAE), the measure of errors between observations, and mean absolute percentage error (MAPE), the measure of how accurate the predicted outcome is, was calculated for all cases, and then the median of the MAEs and MAPEs were presented here. Table 2 shows both the MAE and MAPE. Note that this percentage error is related to the mean relative error, which is the mean of the relative errors between observations, presented in the time series figures below (Figure 2). Although the MAE is reported, the MAPE is a more balanced view of how the models performed relative to one another because not all datasets were within the same range. Observe that although no rigorous model selection was performed and some overfitting is likely present, the models still generalized very well; the difference in errors from training and testing is small.

Table 3 shows a similar analysis to Table 2 but with ordinary least squares regression, as a comparison to Table 2. The p-values for the comparison between the training of the nonlinear (SR) and linear (OLS) for the four experimental sets (Set 1, Set 2, Set 3, Set 4 (Lower Back only), and Set 4 (Lower Back fit to all)) are 0.015, <<0.0001, <<0.0001, <<0.0001, and 0.07 respectively. The p-values for the comparison between the testing of the nonlinear (SR) and linear (OLS) for the four experimental sets are <<0.0001, <<0.0001, <<0.0001, <<0.0001, and <<0.0001 respectively.

Table 2: An analysis with SR models. Median errors and interquartile range (IQR) over all subject, session, and trial on all of the broad experiment sets. The median mean absolute error (MAE) and mean absolute percentage error (MAPE) are presented on both the data the models were fit to (training) and all batches from the same subject (from each trial), excluding the batch the model was fit to (testing). For the training, the models were averaged over all 400 instances, and for the testing, the models were averaged over all batches excluding the batch the model was fit to.

	Training		**Testing**	
	Median MAE (IQR)	**Median MAPE (IQR)**	**Median MAE (IQR)**	**Median MAPE (IQR)**
Set 1 (Lower Back Only)	0.099 (0.019)	150.40 (48.50)	0.10 (0.020)	151.97 (54.27)
Set 2 (Lower Back)	0.078 (0.017)	114.34 (38.32)	0.082 (0.023)	117.98 (42.96)
Set 3 (Right Ankle)	0.23 (0.071)	205.38 (87.95)	0.24 (0.081)	208.22 (96.84)
Set 4 (Lower Back 100tp)	0.074 (0.023)	103.60 (46.99)	0.096 (0.037)	123.04 (73.99)
Set 4 (Lower Back 100tp on All)	0.094 (0.032)	130.37 (56.12)	0.098 (0.035)	132.11 (60.64)

Table 3: An analysis with ordinary least squares (OLS) regression. Errors and interquartile range (IQR) over all subject, session, and trial on all of the broad experiment sets. The median mean

absolute error (MAE) and mean absolute percentage error (MAPE) are presented on both the data the models were fit to (training) and all batches from the same subject (from each trial), excluding the batch the model was fit to (testing). For the training, the models were averaged over all 400 instances, and for the testing, the models were averaged over all batches excluding the batch the model was fit to.

	Training (OLS)		**Testing (OLS)**	
	Median MAE (IQR)	**Median MAPE (IQR)**	**Median MAE (IQR)**	**Median MAPE (IQR)**
Set 1 (Lower Back Only)	0.11 (0.017)	155.70 (57.70)	0.10 (0.019)	156.99 (60.65)
Set 2 (Lower Back)	0.07 (0.014)	96.80 (39.09)	0.075 (0.020)	104.68 (41.95)
Set 3 (Right Ankle)	0.23 (0.080)	250.25 (100.38)	0.25 (0.081)	268.42 (120.45)
Set 4 (Lower Back 100tp)	0.050 (0.016)	65.53 (44.17)	0.10 (0.047)	132.62 (109.10)
Set 4 (Lower Back 100tp on All)	0.098 (0.036)	138.26 (69.94)	0.11 (0.042)	143.96 (82.54)

Table 4 presents a symmetrical matrix of probability values from a Mann-Whitney U test comparing the MAPE values obtained in testing each of the experimental sets (described in sections 2.4.1-2.4.4). The diagonal represents the case where the experimental set was compared to itself. The Mann-Whitney U test's p-values, for the four experimental sets (Set 1, Set 2, Set 3, Set 4 (Lower Back only), and Set 4 (Lower Back fit to all)), comparing the training and testing errors on the median MAPEs are 0.24, 0.006, 0.025, $<<0.0001$, and 0.085 respectively.

Table 4: Probability value obtained with a Mann-Whitney U test comparing the mean absolute percentage error (MAPE) testing results from each of the experimental sets to each other. Note that

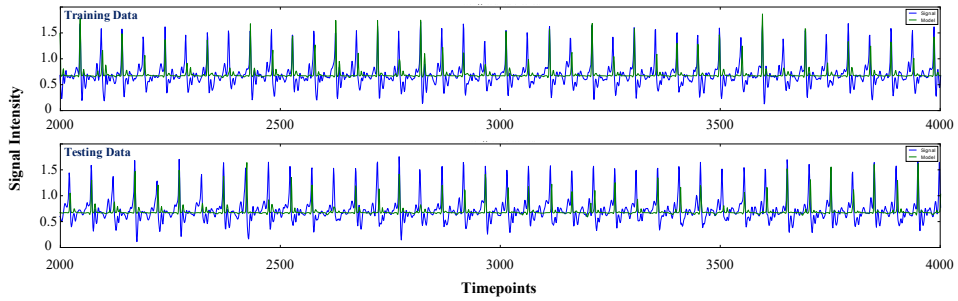
this matrix of information is symmetric and the diagonal compares each set to itself. Be aware that all p-values for set 3 are 0. This is due to precision errors and should be interpreted as very small, but not exactly 0.

	Set 1	**Set 2**	**Set 3**	**Set 4**	**Set 4 (on all)**
Set 1 (Lower Back Only)	0.5	<<0.0001	0.000	<<0.0001	<<0.0001
Set 2 (Lower Back)	<<0.0001	0.5	0.000	<<0.0001	<<0.0001
Set 3 (Right Ankle)	0.000	0.000	0.5	0.000	0.000
Set 4 (Lower Back 100tp)	<<0.0001	<<0.0001	0.000	0.5	<<0.0001
Set 4 (Lower Back 100tp on All)	<<0.0001	<<0.0001	<<0.0001	<<0.0001	0.5

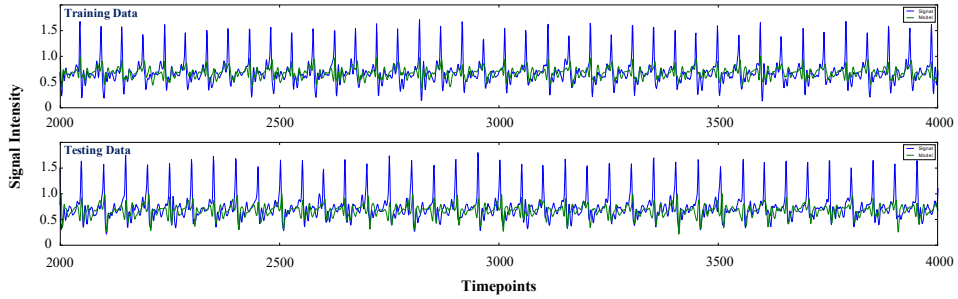
Figure 2: Example models from each of the 4 experimental sets (respectively). In these models, the variable corresponds to the component from the accelerometer the data came from, 0 being the component with the highest variance. Additionally, these models correspond to the models presented in the time-series. The blue lines and green lines represent the signal and model, respectively. **A.** Example times series of a model fit with lower back only. Top graph plots the model against the data it was fit to (mean absolute error of 0.14 and mean relative error of 0.25), and the bottom plots the model against unseen data from the same subject (mean absolute error of 0.14 and mean relative error of 0.20). This particular example came from the second and third segment of 4800 time points of Subject 10's first trial of session 1. **B.** Example time series of a model fit to lower back but fit with all features. Top graph plots the model against the data it was fit to (mean absolute error of 0.13 and mean relative error of 0.21), and the bottom plots the model against unseen data from the same subject (mean absolute error of 0.15 and mean relative error of 0.19). This particular example came from the second and ninth segment of 4800 time points of

Subject 10's first trial of session 1. **C.** Example time series of a model fit to RA but fit with all features. Top graph plots the model against the data it was fit to (mean absolute error of 0.19 and mean relative error of 1.04), and the bottom plots the model against unseen data from the same subject (mean absolute error of 0.26 and mean relative error of 0.78). This particular example came from the tenth and seventh segment of 4800 time points of Subject 3's first trial of session 2. **D1 and D2.** Example timeseries of a model fit to lower back but fit with all features and only on 100 time points. Top graph plots the model against the data it was fit to (mean absolute error of 0.129 and mean relative error of 0.17), middle graph plots the model against all data points from the same data segment the 100 time points came from (mean absolute error of 0.16 and mean relative error of 0.22), and the bottom plots the model against unseen data from the same subject (mean absolute error of 0.16 and mean relative error of 0.23). This particular example came from the seventh and eighth segment of 4800 time points of Subject 10's first trial of session 1.

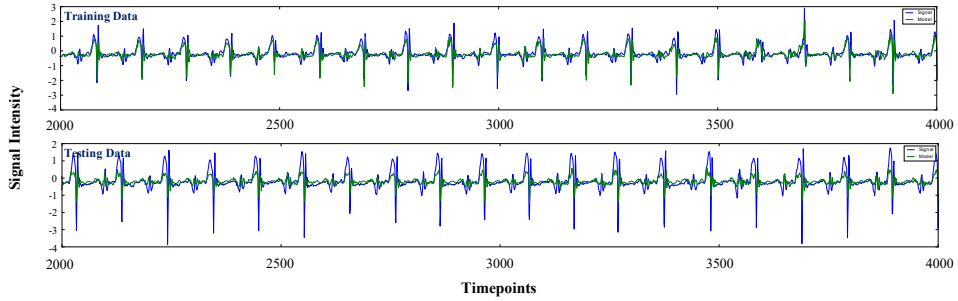
A. Models fit to Lower Back Only



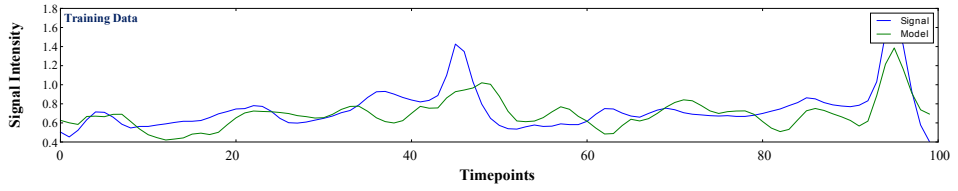
B. Models fit to Lower Back



C. Models fit to Right Leg/Ankle



D1. Models fit to Lower Back (100 timepoints)



D2. Models fit to Lower Back (100 timepoints on all data)

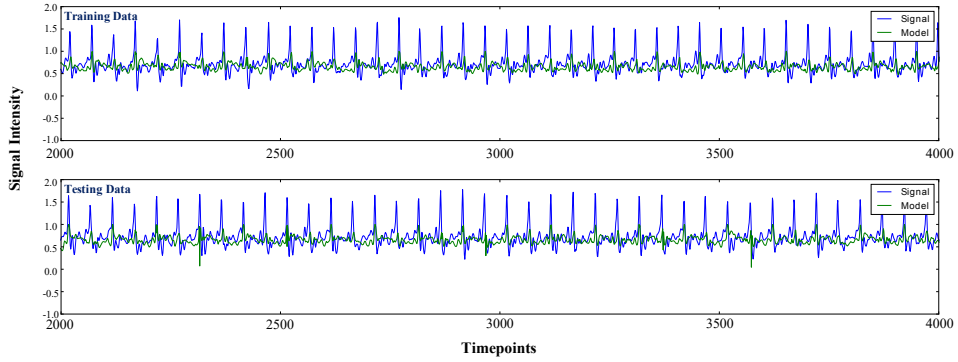


Figure 2 show excerpts of the models plotted against the data they were fit to (training data), and the models alongside unseen data from the same subject, session, and trial they were applied to (testing data). These particular models were selected as they were the ones with the best (lowest) error when applied to unseen data from their own experimental set (no rigorous model selection was performed). In other words, these models were not necessarily the ones with the lowest testing error. The errors reported here are MAE and mean relative error.

The equations from Figure 2's example models from each of the 4 experimental sets, respectively, are shown below (Equations 1-4). In these models, the variable corresponds to the component (one of the three PCs) from the accelerometer, where the subscript "0" is the component with the highest variance.

$$LB_0 = LB_2 - \cos(LB_2 * 5.0612 + 13.870) * 0.198 \quad \textbf{(Equation 1)}$$

$$LB_0 = \cos(RW_2 * \cos(RW_0 + LW_2)) \quad \textbf{(Equation 2)}$$

$$RA_0 = \frac{RW_2}{-3.688} - RA_1 * \tan(\cos(LB_2)) \quad \textbf{(Equation 3)}$$

$$LB_0 = RW_2 + \left(e^{C_0} * (LW_1 + LB_2)\right) \quad \textbf{(Equation 4)}$$

In Figure 3, four matrices are shown representing the feature counts in each specific subject, for each of the experimental sets respectively. There were 100 runs per instance for statistical significance; thus the feature count is within the range of 0 (dark blue) to 100 (dark red). The bottom row in each matrix is dark red, as it was set to the left hand side of each model/equation

from the experimental set (i.e., it was in 100% of the models). With the exception of the first matrix, which contains only the three PCs of the lower back accelerometer only, all other matrices contain 18 features: the three PCs from each accelerometer.

Figure 3: Number of times each feature appeared in the models generated, for each subject (denoted by the numbers on top). Please note that the y-axis labels refer to the accelerometer location, PC, and the "0", "1", and "2" represent the PC with the highest to lowest variance. Each row corresponds to the following experiments: A corresponds to experiment 1, B corresponds to experiment 2, C corresponds to experiment 3, and D corresponds to experiment 4. Rows correspond the feature and columns specific subject, session, trial, and batch instances. The percentage of models each feature appeared in each collection of data is represented by a color, as laid out in the color bar scale legend on the right. Note that the last row in each matrix is solid red as it was forced to be in each model since it was the left-hand side of the equation.

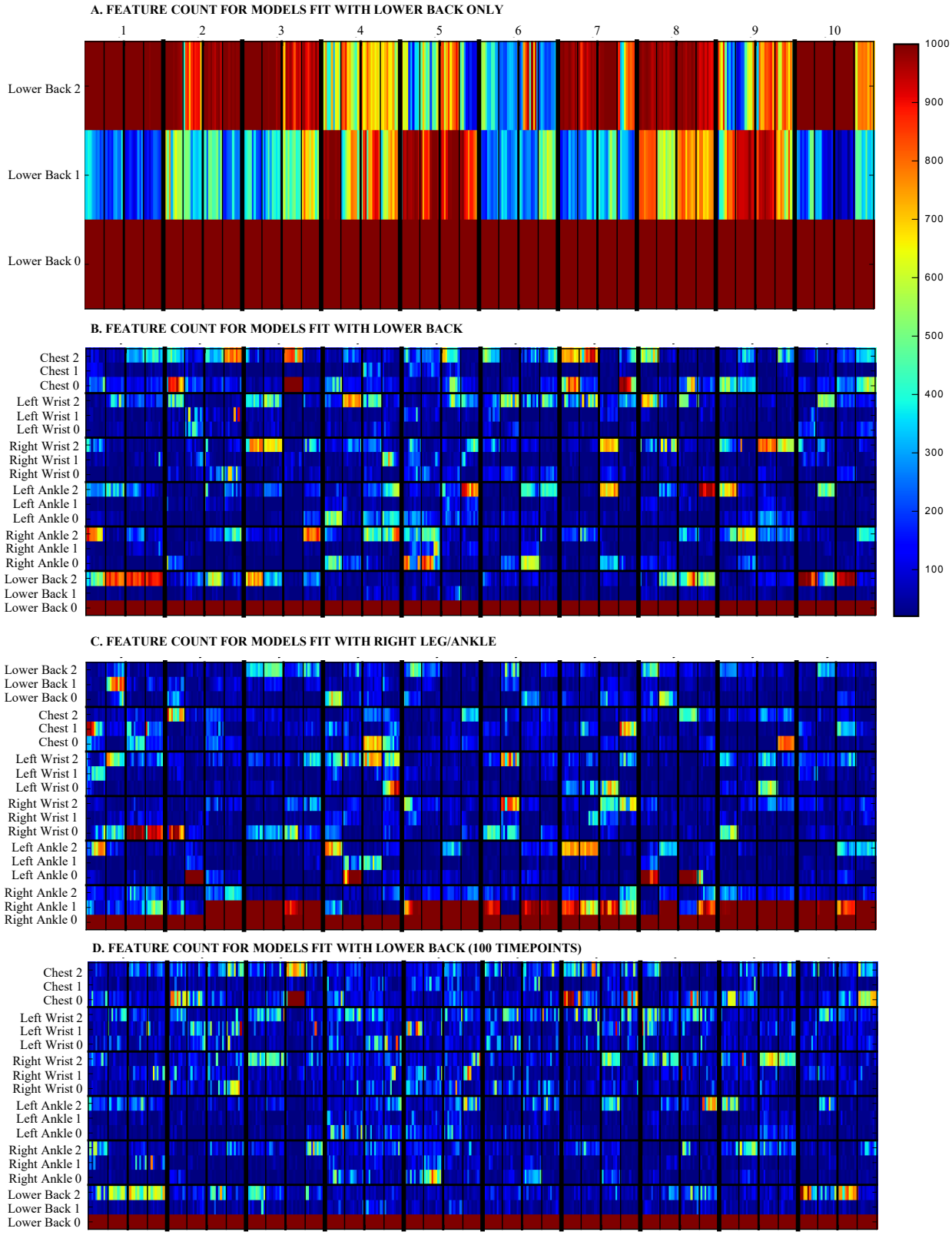
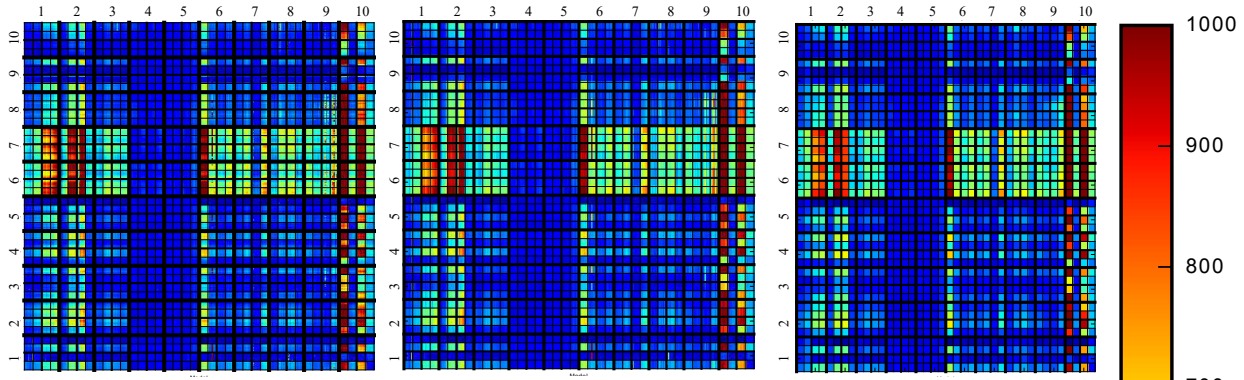

A. FEATURE COUNT FOR MODELS FIT WITH LOWER BACK ONLY
1
2
3
4
5
6
7
8
9
10
Lower Back 2
Lower Back 1
Lower Back 0
1000
900
800
700
600
500
400
300
200
100
B. FEATURE COUNT FOR MODELS FIT WITH LOWER BACK
Chest 2
Chest 1
Chest 0
Left Wrist 2
Left Wrist 1
Left Wrist 0
Right Wrist 2
Right Wrist 1
Right Wrist 0
Left Ankle 2
Left Ankle 1
Left Ankle 0
Right Ankle 2
Right Ankle 1
Right Ankle 0
Lower Back 2
Lower Back 1
Lower Back 0
C. FEATURE COUNT FOR MODELS FIT WITH RIGHT LEG/ANKLE
Lower Back 2
Lower Back 1
Lower Back 0
Chest 2
Chest 1
Chest 0
Left Wrist 2
Left Wrist 1
Left Wrist 0
Right Wrist 2
Right Wrist 1
Right Wrist 0
Left Ankle 2
Left Ankle 1
Left Ankle 0
Right Ankle 2
Right Ankle 1
Right Ankle 0
D. FEATURE COUNT FOR MODELS FIT WITH LOWER BACK (100 TIMEPOINTS)
Chest 2
Chest 1
Chest 0
Left Wrist 2
Left Wrist 1
Left Wrist 0
Right Wrist 2
Right Wrist 1
Right Wrist 0
Left Ankle 2
Left Ankle 1
Left Ankle 0
Right Ankle 2
Right Ankle 1
Right Ankle 0
Lower Back 2
Lower Back 1
Lower Back 0

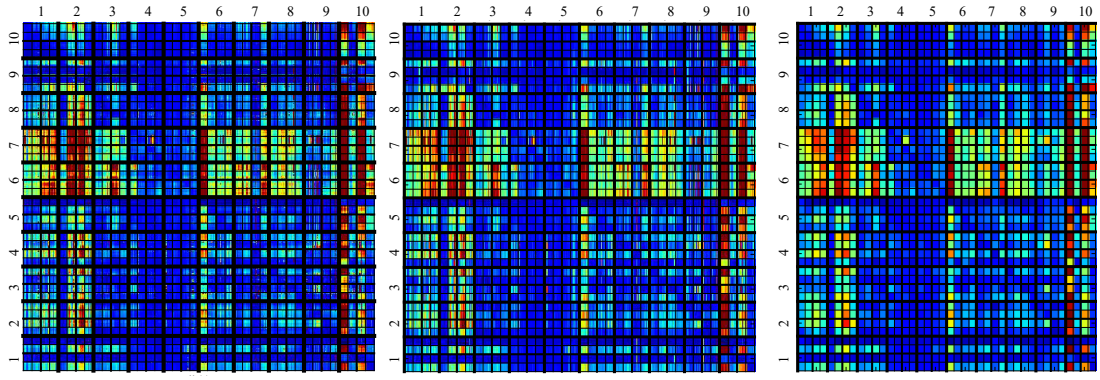
Figure 4: Three error matrices generated by applying all generated models to every collection of data for the first set of experiments. Each row corresponds to the following experiments: A corresponds to experiment 1, B corresponds to experiment 2, and C corresponds to experiment 3.The color bar scale legend on the right, represents feature count via color.

Each Dataset Applied to All Models | **Averaged Error for Each Model Over Subset Data** | **Averaged Error for Each Model Over Model and Subset Data**

A. MEAN ABSOLUTE PERCENTAGE ERROR MATRIX FOR LOWER BACK ONLY



B. MEAN ABSOLUTE PERCENTAGE ERROR MATRIX FOR LOWER BACK



C. MEAN ABSOLUTE PERCENTAGE ERROR MATRIX FOR RIGHT LEG/ANKLE

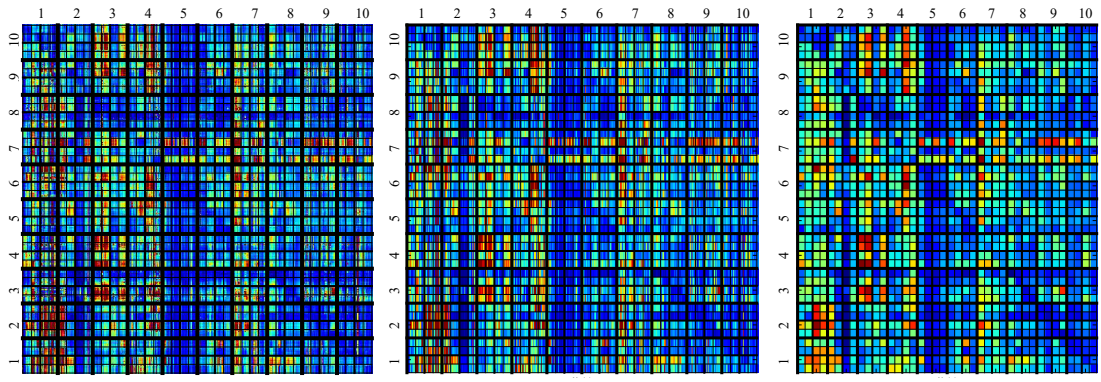
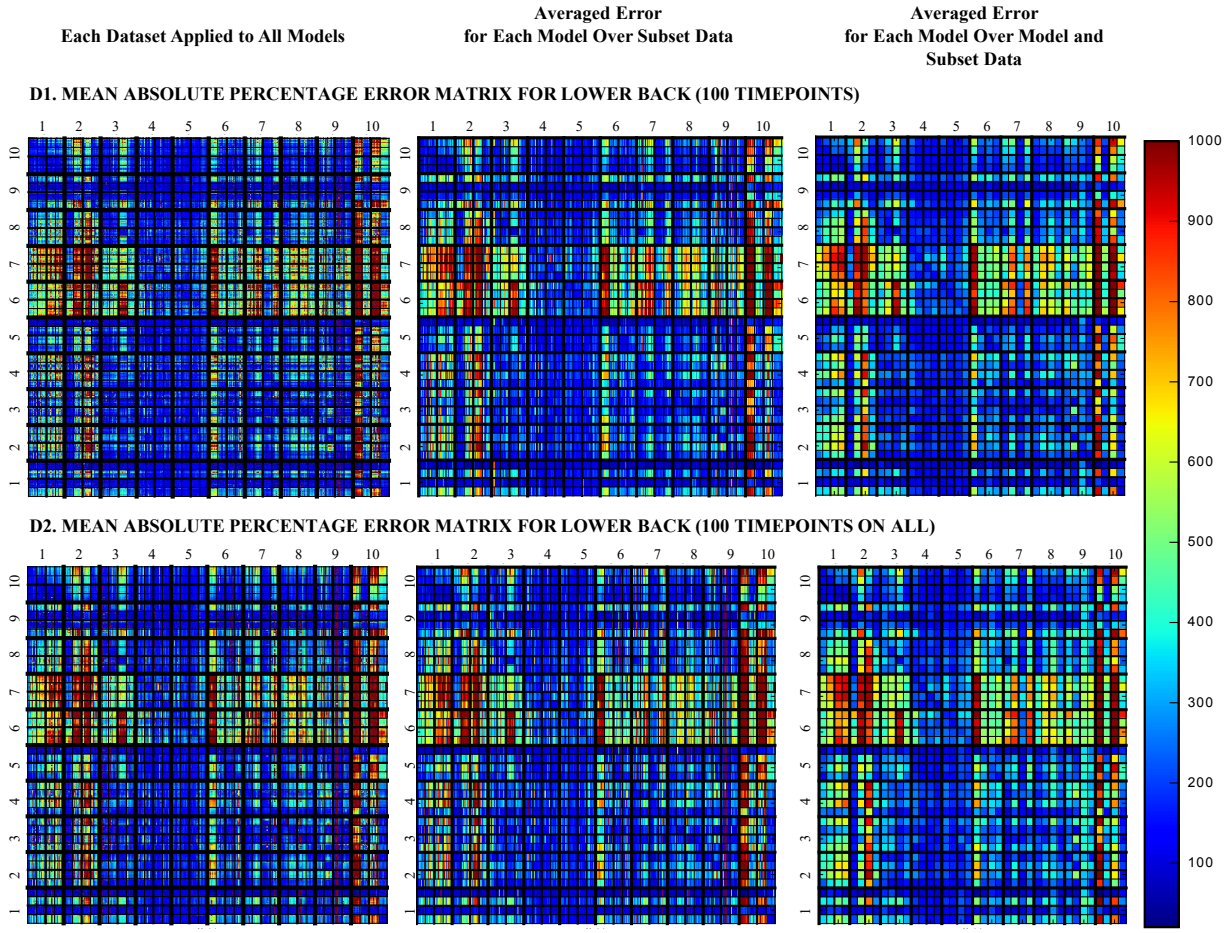


Figure 5: Three error matrices generated by applying all generated models to every collection of data for the first set of experiments. Each row corresponds to the following experiments: D1 and D2 both correspond to experiment 4. The color bar scale legend on the right, represents feature count via color.



In Figures 4-5, the left most matrix is all models applied to all data from each subject, session, trial, and batch instances. In the left matrix, rows correspond to the specific instance and columns to the top model created for that instance and diagonal represents the case when the top generate model for a given instance is being applied to the instance it was fit to. The center matrix contains the averaged errors each model obtained on all batches from the same subject, session, and

trial combination. The right most matrix averages the errors of all models generated for a given subject, session, and trial combination for all batches from the same subject, session, and trial combination. In all matrices, the mean absolute percentage error is represented by color.

In Figure 5, these error matrices depict how well these models generalize to other subjects. The best model in this case was the one with the lowest training error. The errors reported here are the MAPEs, not the MAEs. Although these matrices are of error values, they can occasionally be seen as a means of quantifying how similar subjects are. For example, if the subject X's model obtained a similar error when applied to itself and subject Y's data, then the two subjects' walking manifestations may be intrinsically similar. However, note that this will not be the case in general. For example, note subject 4 and 5 obtain similar error values on each other, but they also obtain similar error values on all other subject's data (Figure 4-5). Although there are similarities in the feature counts, it is by no means obvious that these subject's models produce these results (Figure 4-5).

4. DISCUSSION

This study created models that portray the network relationship between six different areas of the human body. By using accelerometers placed in various places of the body and treating the system as having a network relationship between these body parts, not only do the models fit their data better, but models are far more able to generalize to unseen data.

4. 1 Network of Walking using Sensors

One of the main interesting results of this study is that a network relationship was found between the different body parts. The BSN models fit to the lower back's accelerometer were able to describe subject-specific data the best when compared to all other models by a significant amount, suggesting the network relationship is essential in describing the system. This is as expected since there was more information to fit with, and the testing results show that overfitting was not a significant problem. The quality of these models depended on what feature the models were fit. The second best performing models were models fit to only 100 time points, as opposed to the 4,800 all others were fit to. Although these models did not perform as well as the lower back models, it shows that with high-quality data, very few data points are required to generate effective models.

Although the BSN models fit to the lower back were the most effective intrasubject, the models fit with only the lower back were the best at generalizing between subjects. However, subjects 4 and 5's models were still beneficial at generalizing to other subjects across all experimental sets. This is not surprising as these models were fit to data with fewer features, which would result in smaller errors.

No apparent trends can be seen in the feature counts in Figures 3-5 with respect to which features arose in the models. The largest trend that can be seen is the features arising in all batches from the same experimental set (from all the subjects), or sometimes between the same subject in the same experimental set. Unsurprisingly, the models' fit for experimental sets 2 and set 4 (both fit

to lower back with all features) have a lot in common. In general, the same features arise, but to a lesser extent in experimental set 4's models.

Because there is little consistency in features for each subject, it seems that human locomotion can be explained with a large variety of models, or perhaps the usefulness of a feature in a model is very dependent on the precise physical location of an accelerometer. Although PCA would align the axes, variations in the physical location of the accelerometer between subjects, or even between sessions or trials, cannot be corrected. More focus is concentrated on the latter as the commonality in counts between set 2 and 4's models show consistency. This does provide more evidence for the necessity of a network relationship for describing the system; however, some between-subject generality may be lost.

With the results presented here, little consistency arose in the features between subjects, and it seems that these models are very subject-specific. This could be a result of different physiology (weight, size, and shape) between individuals or our modeling strategy. Given this, it would be challenging to develop a general-purpose network relationship model with the presented strategy as is; however, alterations to the modeling strategy may be able to generate a more general model. Thus, this network of modeling using sensors can be an additional data perspective to existing biological databases that target cognitive or neuromotor disease.

4. 2 Clinical Application

Personalized or precision medicine, or the concept of identifying and aiding individuals by their data, can significantly benefit from these individualized models, perhaps in combination with other types of models, because they could be useful in situations where strict monitoring would be valuable. Being able to develop general models capable of describing human walking would have meaningful applications in the clinical setting and physical rehabilitation. Our SR models can be used as a means of quantifying how close a patient's manifestation of walking is to an idealized, personalized model. If patient data does not fit these models well, it could be a sign of poor posture, injury, or some other underlying health problem. Furthermore, poor model fits can be used as a possible predictor of cognitive impairment or falls. These models could also be used to track therapy effectiveness; if the patient's data becomes closer to the idealized model throughout treatment, then there is a good indication that the therapy is working.

4. 3 Comparison of Experiments

As mentioned previously, lower back and hip accelerometer placements are ubiquitous in the literature. In nearly all of our results, the feature counts from the lower back accelerometer dominate the other nodes of the body; despite this, the other nodes of the body provide valuable information to the models in this paper. These findings are due to the lower back being a proxy for the center of mass of the body. After fitting the data collected from the lower back accelerometer only to the dependent variable (the axis from the lower back accelerometer with the most variance) (experimental set 1), the resulting models generalized to all other subjects far better than the other experiments. The error matrices for this experiment also indicate that these models could generalize the best, and the diagonal is not particularly well pronounced, indicating little intrasubject overfitting (Figures 4-5). This model has the second-best testing mean relative error at 0.203 (Figure 4).

Adding the data from the subjects' LA and RA, LW and RW, and C, as well as the lower back (experimental set 2), increased the performance of the models when considering the testing results (Table 2). Table 4 also indicates that these results outperformed other experiments by a significant amount. Figure 4 showing a model for experimental set 2 had the best testing mean absolute percentage error (0.188), which does align with the overall results seen in Table 2.

In our next experiment, we changed the dependent variable to the axis with the most variance from the RA, while keeping the data or features the same (experiment set 3). This change allows for an analysis of the model of human locomotion but from a different perspective. Despite having all the same data to fit with, these models had very distinct features when compared to the previous experiment (Figure 4). This finding also corresponds to the observations made when discussing Table 2; the network arising depends on the dependent variable. Unfortunately, these models appear not to generalize well; however, some of the observations made about subject outlier errors (subject's 1, 2, 6, and 10 discussed above) from the other matrices cannot be seen here, suggesting that the choice of dependent variable is crucial (Figure 4).

Next, we tried altering the second experiment by fitting the models to only 100 time points and then subsequently applied to all 4,800 time points (experimental set 4). These models were still performing very well and fit significantly better than those fit to only the lower back's accelerometer. It seems that very few data points are required to create a useful model of this complex system. This finding is advantageous in a real-world application as it would require much less data gathering, less pre-processing, and less computational power to generate the models.

Even though these models were able to fit its data the best (experimental set 4), it should be noted that this testing error was calculated on only 100 time points as opposed to the 4,800 time points all other sets used. For this reason, one should focus on the second version of experimental set 4, where the models fit to 100 time points were applied to the full 4,800 time points to calculate the errors. When comparing the second and fourth experimental sets, the error matrices all look

similar with a well pronounced diagonal (Figures 4-5). Both sets appear to have overfit nuances in the few data points they saw (Figures 4-5).

Regardless of the experimental set, one can see that the models track the signals well. The most considerable deviations seem to come from the spikes in the signals; some of the higher frequencies can be found near the middle of the signal. However, the example shown in time series corresponding to experimental set 4 did not fit the signal as well as those seen for sets 1 and 3 when considering the testing error (Figure 5).

In the error matrices in Figures 4 and 5, we can see some common trends between all four experiments, which may be due to the experimental setup, rather than any biological explanation. For example, subject 10's trial ones in both sessions appear to be markedly worse. A similar thing can be seen in session two for subjects 1 and 2 and on subject 6's trial one of session one. Perhaps these subjects' devices were in markedly different locations, or there were problems with accelerometer calibration. Additionally, the high mean absolute percentage errors are likely due to significant spiking events and other artifacts in the signal.

While the third experiment also treated the system as a network relationship, those models performed poorly and significantly worse than those fit to just the lower back's accelerometer. It seems that this network relationship is necessary for describing the system, but the choice of the dependent variable is critical. For example, the experiment fitting the lower back as the dependent variable substantially outperformed the experiment having the RA as the dependent variable.

4. 4 Methodological Considerations

This study produced symbolic regression models with GP. The major strengths of this particular implementation of producing symbolic regression models with GP are in the use of subpopulations, an acyclic graph representation, and the use of fitness predictors. Unlike traditional GP systems where the chromosomes are tree-based S-expressions, this study's application uses an array-based acyclic graph representation of the underlying mathematical expressions. The benefits

of such representation include faster runtimes, easy reuse of subexpressions, and the reduction of bloat (a GP phenomenon where the trees grow arbitrarily complex with no meaningful improvement in the quality of the effectiveness [23, 32]).

The human sample size of this study is ten subjects, while the actual sample size of this study consists of the 48,000 data points (well over 1000 strides per individual) from each accelerometer on each subject. Two trials of ten minutes of accelerometer data per person are sufficient enough to analyze how these subjects would walk. For example, in older adults, the six-minute walk test is a commonly used clinical tool to measure physical functional status (particularly for patients with heart and lung disease) [38-42]. Data consisting of at least four or more minutes is sufficient to gather enough data to generalize to a real-world walking scenario [38-42]. In addition to the six-minute walk test, other short treadmill and overground-walking tests have been developed and are in clinical use and gait research to assess a variety of physical functions [43].

The subjects were asked to walk on a treadmill with a set, constant walking speed. There is some evidence that, in healthy adults, treadmill walking may result in high gait stability [44]. Thus, the limitation of treadmill walking measurements is that these results may not generalize well to over-ground walking [44]. However, the controlled nature of the treadmill acts as a "dedicated pacer," which can be beneficial to control for speed and can make it easier to compare trials of data within each subject [19]. After treadmill familiarization, healthy adults have been shown to have imperceptible differences between overground and treadmill gait measurements [19].

4.5 Distinguishing Cognitive Load

While distinguishing cognitive load was the not the main aim of this study, we expected that cognitive load would affect these models. However, in this study, with our modelling approach, we found no noticeable differences between the trials (walking regularly vs. walking under cognitive load). This is in direct contrast to the study by Dasgupta et al., where they could distinguish between the features from the accelerometer data from the two trials, using standard

machine learning models [19]. While this is not a significant limitation of our study, this result could be due to a small human sample size in conjunction with a sample of solely healthy adults whose gait only changes slightly with an added cognitive load. Additionally, perhaps symbolic regression models are better at distilling noise; in this case, noise could be the inconsistent differences caused by cognitive load [46]. Hypothetically, symbolic regression models may only capture high level trends in gait, and other modelling techniques, such as those used in Dasgupta et al., may be better methods to classify cognitive load due to differences in gait [19].

4.6 Future Directions

In order to fully test if this approach is more generalizable to subject-specific data than node-specific model approaches is to do more experiments on healthy and unhealthy populations across a wide range of ages. For example, future studies could test a BSN model and compare it to a lower-back (or another body node) model; if the node-specific model as differs in errors from the whole BSN model, then there is more evidence of the utility of the BSN model.

5. CONCLUSIONS

Overall, we demonstrated that a collection of non-linear, SR models could represent a complex network system: human locomotion. These models were derived from sensor data from six body parts during locomotion; thus, the models analyzed any linear or non-linear relationships in the human body system's network. This study's main result was that the SR models fit from the data from the lower back's accelerometer were able to explain subject-specific data the most compared to all other models. SR was used to generate models that fit multiple training sets for different experiments and were tested against unseen test data. Our methodology and gait analyses, using GP and SR, may have an impact on personalized or precision medicine, because they could be valuable in situations where accelerometer monitoring is used.

STATEMENTS OF ETHICAL APPROVAL

The University of Pittsburgh Institutional Review Board approved the collection and use of demographic, acceleration, and treadmill data (No. PRO14060107). The collected data did not include any identifiers, and each participant was labeled with a subject number (e.g., Subject 1) for data analysis and reporting. Informed consent was obtained from volunteers.

COMPETING INTERESTS

The authors have no conflicts of interest to declare.

FUNDING

This research is funded by the National Library of Medicine (National Institutes of Health) (Grant Reference Number: 4T15LM007059), the National Institute on Aging through the Pittsburgh Claude D. Pepper Older Americans Independence Center under Grant NIA P30 AG 024827, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

1. Kathale SN, Solaskar S. A Method for Identifying Human by Using Gait Cycle. In Intelligent Communication Technologies and Virtual Mobile Networks 2019 Feb 14 (pp. 655-666). Springer, Cham.
2. Yoo JH, Nixon MS, Harris CJ. Extracting human gait signatures by body segment properties. In Proceedings Fifth IEEE Southwest Symposium on Image Analysis and Interpretation 2002 Apr 7 (pp. 35-39). IEEE.
3. Yoo JH, Hwang D, Moon KY, Nixon MS. Automated human recognition by gait using neural network. In2008 First Workshops on Image Processing Theory, Tools and Applications 2008 Nov 23 (pp. 1-6). IEEE.
4. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. 2011 Jan;12(1):56-68.
5. Ivanov PC, Liu KK, Bartsch RP. Focus on the emerging new fields of network physiology and network medicine. New journal of physics. 2016 Oct 13;18(10):100201.
6. Barabási AL. Network medicine—from obesity to the "diseasome". 2007 :404-407.
7. Moussa MN, Vechlekar CD, Burdette JH, Steen MR, Hugenschmidt CE, Laurienti PJ. Changes in cognitive state alter human functional brain networks. Frontiers in Human Neuroscience. 2011 Aug 22;5:83.
8. Chi YM, Cauwenberghs G. Wireless non-contact EEG/ECG electrodes for body sensor networks. In 2010 International Conference on Body Sensor Networks 2010 Jun 7 (pp. 297-301). IEEE.
9. Bartsch RP, Liu KK, Bashan A, Ivanov PC. Network physiology: how organ systems dynamically interact. PloS One. 2015 Nov 10;10(11):e0142143.
10. Rustagi L, Kumar L, Pillai GN. Human gait recognition based on dynamic and static features using generalized regression neural network. In 2009 Second International Conference on Machine Vision 2009 Dec 28 (pp. 64-68). IEEE.
11. VanSwearingen JM, Studenski SA. Aging, motor skill, and the energy cost of walking: implications for the prevention and treatment of mobility decline in older persons. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 2014 Nov;69(11):1429.
12. Rosenbaum DA. Human motor control. Academic press; 2009 Sep 11.
13. Rosenbaum DA, Chapman KM, Coelho CJ, Gong L, Studenka BE. Choosing actions. Frontiers in Psychology. 2013 Jun 3;4:273.
14. Inkol KA, Huntley AH, Vallis LA. Repeated exposure to forward support-surface perturbation during overground walking alters upper-body kinematics and step parameters. Journal of Motor Behavior. 2019 May 4;51(3):318-30.
15. Luu TP, Low KH, Qu X, Lim HB, Hoon KH. An individual-specific gait pattern prediction model based on generalized regression neural networks. Gait & Posture. 2014 Jan 1;39(1):443-8.
16. Gou H, Shi T, Yan L, Xiao J. Gait and Posture Analysis Method Based on Genetic Algorithm and Support Vector Machines with Acceleration Data. Journal of Robotics and Mechatronics. 2016 Jun 20;28(3):418-24.
17. Morris JR. Accelerometry—A technique for the measurement of human body movements. Journal of Biomechanics. 1973 Nov 1;6(6):729-36.
18. Del Din S, Hickey A, Hurwitz N, Mathers JC, Rochester L, Godfrey A. Measuring gait with an accelerometer-based wearable: influence of device location, testing protocol and age. Physiological Measurement. 2016 Sep 21;37(10):1785.

19. Dasgupta P, VanSwearingen J, Sejdic E. “You can tell by the way I use my walk.” Predicting the presence of cognitive load with gait measurements. BioMedical Engineering OnLine. 2018 Dec 1;17(1):122.
20. Cunado D, Nixon MS, Carter JN. Automatic extraction and description of human gait models for recognition purposes. Computer Vision and Image Understanding. 2003 Apr 1;90(1):1-41.
21. Wang L, Hu W, Tan T. A new attempt to gait-based human identification. In Object Recognition Supported by User Interaction for Service Robots 2002 Aug 11 (Vol. 1, pp. 115-118). IEEE.
22. BenAbdelkader C, Cutler R, Davis L. Stride and cadence as a biometric in automatic person identification and verification. In Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition 2002 May 21 (pp. 372-377). IEEE.
23. Hughes J, Houghten S, Brown JA. Models of Parkinson's Disease Patient Gait. IEEE Journal of Biomedical and Health Informatics. 2019 Dec 23.
24. Li S, Todor A, Luo R. Blood transcriptomics and metabolomics for personalized medicine. Computational and Structural Biotechnology Journal. 2016 Jan 1;14:1-7.
25. Chen Y, Angulo MT, Liu YY. Revealing complex ecological dynamics via symbolic regression. BioEssays. 2019 Dec;41(12):1900069.
26. Macbeth J, Sarrafzadeh M. Shrinking symbolic regression over medical and physiological signals. In 2010 2nd International Conference on Signal Processing Systems 2010 Jul 5 (Vol. 1, pp. V1-461). IEEE.
27. Ok S, Miyashita K, Hase K. Evolving bipedal locomotion with genetic programming-a preliminary report. In Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546) 2001 May 27 (Vol. 2, pp. 1025-1032). IEEE.
28. Macbeth J, Sarrafzadeh M. Shrinking symbolic regression over medical and physiological signals. In 2010 2nd International Conference on Signal Processing Systems 2010 Jul 5 (Vol. 1, pp. V1-461). IEEE.
29. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010 Jul;2(4):433-59.
30. Koza JR. Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems. Stanford, CA: Stanford University, Department of Computer Science; 1990 Jun 1.
31. Koza JR. Genetic programming as a means for programming computers by natural selection. Statistics and Computing. 1994 Jun 1;4(2):87-112.
32. Koza JR, Rice JP. Automatic programming of robots using genetic programming. In AAAI 1992 Jul 12 (Vol. 92, pp. 194-207).
33. Schmidt MD, Vallabhajosyula RR, Jenkins JW, Hood JE, Soni AS, Wikswo JP, Lipson H. Automated refinement and inference of analytical models for metabolic networks. Physical Biology. 2011 Aug 10;8(5):055011.
34. Hughes JA, Brown JA, Khan AM, Khattak AM, Daley M. Analysis of symbolic models of biometrie data and their use for action and user identification. In 2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) 2018 May 30 (pp. 1-8). IEEE.
35. McAdams ET, Gehin C, Noury N, Ramon C, Nocua R, Massot B, Oliveira A, Dittmar A, Nugent CD, McLaughlin J. Biomedical sensors for ambient assisted living. In Advances in Biomedical Sensing, Measurements, Instrumentation and Systems 2010 (pp. 240-262). Springer, Berlin, Heidelberg.

36. Bonomi AG, Goris AH, Yin B, Westerterp KR. Detection of type, duration, and intensity of physical activity using an accelerometer. Medicine & Science in Sports & Exercise. 2009 Sep 1;41(9):1770-7.
37. Cleland I, Kikhia B, Nugent C, Boytsov A, Hallberg J, Synnes K, McClean S, Finlay D. Optimal placement of accelerometers for the detection of everyday activities. Sensors. 2013 Jul;13(7):9183-200.
38. Enright PL. The six-minute walk test. Respiratory care. 2003 Aug 1;48(8):783-5.
39. Schenkman M, Cutson TM, Kuchibhatla M, Chandler J, Pieper C. Reliability of impairment and physical performance measures for persons with Parkinson's disease. Physical Therapy. 1997 Jan 1;77(1):19-27.
40. Steffen TM, Hacker TA, Mollinger L. Age-and gender-related test performance in community-dwelling elderly people: Six-Minute Walk Test, Berg Balance Scale, Timed Up & Go Test, and gait speeds. Physical Therapy. 2002 Feb 1;82(2):128-37.
41. Strijbos JH, Postma DS, Van Altena R, Gimeno F, Koeter GH. A comparison between an outpatient hospital-based pulmonary rehabilitation program and a home-care pulmonary rehabilitation program in patients with COPD: a follow-up of 18 months. Chest. 1996 Feb 1;109(2):366-72.
42. Du H, Newton PJ, Salamonson Y, Carrieri-Kohlman VL, Davidson PM. A review of the six-minute walk test: its implication as a self-administered assessment tool. European Journal of Cardiovascular Nursing. 2009 Mar;8(1):2-8.
43. Roerdink M, de Jonge CP, Smid LM, Daffertshofer A. Tightening up the control of treadmill walking: effects of maneuverability range and acoustic pacing on stride-to-stride fluctuations. Frontiers in Physiology. 2019;10:257.
44. Nazary-Moghadam S, Salavati M, Esteki A, Akhbari B, Keyhani S, Zeinalzadeh A. Gait speed is more challenging than cognitive load on the stride-to-stride variability in individuals with anterior cruciate ligament deficiency. The Knee. 2019 Jan 1;26(1):88-96.
45. Schmidt M, Lipson H. Comparison of tree and graph encodings as function of problem complexity. In Proceedings of the 9th annual conference on Genetic and evolutionary computation 2007 Jul 7 (pp. 1674-1679).
46. Hughes JA, Brown JA, Khan AM. Smartphone gait fingerprinting models via genetic programming. In 2016 IEEE Congress on Evolutionary Computation (CEC) 2016 Jul 24 (pp. 408-415). IEEE.