## Method

### Design

For this study forty eight subjects completed a series of activities while carrying a Samsung phone in their right pockets. During the study participants wore a Jaeger device, and we collected VO2, VO2kg, VCO2, VE, BF, RER, EqO2, MET, and EE.

The study starts with five minutes of lying down, followed by a five-minute sitting period. Then, participants walk on a treadmill for ten minutes at their self-paced speed. After that, they lie down for another five minutes before running at 3 METs for ten minutes. After that, they lie down for five minutes and run at 5 METs for ten minutes. The next step is sitting for five minutes, and lastly, they end with a run at 7 METs for ten minutes to finish the study. Data was collected between January 7th, 2019 and May 9th, 2019. We used a Samsung Galaxy S7 (SM-G930W8) and Ethica app to record the data.

Ethical approval was obtained by the Memorial University Interdisciplinary Committee on the Ethics in Human Research (ICEHR #20180188-EX). All participants provided signed informed consent.

### Measures

There were three main data categories in this study. Accelerometer data, metabolic data and participant demographics. The raw accelerometer data was collected by the Ethica mobile application for x, y, and z directions. Since this data did not have a constant frequency, we changed the frequency to 30Hz using a resampling method introduced by J. Smith, 2002 [ref number]. The metabolic data was collected by Jaeger had a frequency of 1Hz and contained VO2, VO2kg, VCO2, VE, BF, RER, EqO2, MET, and EE. We used the measured MET to categorize the data into seven different activity levels. Sitting, lying, walking self-pace, running 3METs, running 5METs, running 7METs, and transit. The ***transit*** label represents periods during which the subject was in-between tasks. Therefore, we did not use the transit period in the modelling process. Lastly, participant characteristics, were collected, which included participants' age, weight in kg, height in cm, and sex.

### Data analysis

We generated four different cases to train our model. Each case used a different combination of the available data, as mentioned in Table 3. For Case One, the dataset which was fed to the ML models was comprised of four columns. We used x, y and z accelerometer data as the predictive features to classify the dependent variable, activity type.

For Case Two, we used an R package, activityCounts, to calculate Activity Counts. In this case, the dataset which was used for modelling had two columns. The predictive feature is the vector magnitude of the computed Activity Counts, and the target variable is the activity type.

For Case Three, we chose the acceleration data in x, y, and z-direction as the base features. Then we used the suggested feature list by Lue,2012 [Ref Number] to generate features based on these features. We selected a window of one second to create the new variables; the generated features’ means and standard deviations are specified in Table 1. Case Three uses these features in addition to x, y, and z accelerometer data. Since the new features have a frequency of 1 Hz, and the raw data frequency was 30 Hz, we reduced the raw data frequency to 1 Hz. At the end, 58 features were used as predictive variables.

Case Four was similar to the third case; however, we used Activity Counts as the base features instead of accelerometer data. We selected a window of ten seconds to calculate the features based on Activity counts in z, y and z-direction; thus, we obtained a dataset with a frequency of 10Hz. As the sampling rate of the base features, activity Counts, is different from the frequency of the new features, we needed to reduce the frequency of the base features to 0.1 Hz. Overall, 58 attributes predicted the outcome.

All these features were scaled and centred before being used with ML algorithms. We used the caret package to create balanced splits of the data. The data was split with a ratio of 70% to 30% for training and testing sets, respectively.

To create classification models, we used Support Vector Machines (SVM), Naïve Bayes (NB) and Random Forests classifiers (RF) as they are common in physical activity research using research-grade accelerometers. Also, we implemented the ***k-nearest neighbours*** algorithm (KNN) and "C5.0" algorithm, an improved version of ***C4.5*** algorithm, which creates Decision Trees.

We used accuracy, the area under the receiver operating characteristic (ROC) curve (AUC), and the area under Precision-Recall curve (prAUC) to compare the models. Also, we plotted and compared confusion matrices for the aforementioned models. To implement the models in R, we used the caret package. This package is an interface for ML algorithms and calls other packages to generate a model.

The implement the RF algorithm we used the ranger package via caret. We tried different hyperparameters to achieve the best performance. The hyperparameters that we tuned are the number of variables to split at in each node (mtry), the minimum size of each node, and the split rule. For Case Two, mtry was set to one, since there is only one predictive feature. For other cases, this parameter was set to two. We selected a minimum node size of 50 for Case One; for others, we set it to 5. Finally, we chose the extra tree method as the split rule.

## Results

Our participants were comprised of 26 females and 18 males participated in the study. The average age was 30.2, with a minimum of 18 and a maximum of 56 years. The average height and weight were 169.5cm and 69.8Kg, respectively. In total, we had accelerometer data for 219138 seconds in x, y and z directions. Table 2 shows the data for each activity.

Table 2 – Activities Duration

Activity Time (s)

Lying 47158.8

Sitting 32927.63

Self\_Pace\_walk 28510.47

Running\_3\_METs 36738.9

Running\_5\_METs 36866

Running\_7\_METs 36936.07

We applied the mentioned ML algorithms to the first case. Since our dataset was considerably large, SVM was unable to converge, and KNN was also extremely computationally expensive. C5.0’s performance was slightly worse than RF. Overall, RF had the best performance among all the models. Therefore, we used RF to build models for the four mentioned cases and Table 3 contains the performance metrics and Figure 1 shows the confusion matrices for different cases.

Table 3 – Performance Metrics

Case Accuracy ROC PR-AUC Mean\_Sensitivity Mean\_Specificity Note

Case One 50 0.4 0.4 0.4 0.4 x, y, z raw

Case Two 50 0.4 0.4 0.4 0.4 vector mag of counts

Case Three 50 0.4 0.4 0.4 0.4 x, y, z and features based on x, y, z

Case Four 50 0.4 0.4 0.4 0.4 Counts in x, y, z direction and features based on those.

For Case One, RF achieved an accuracy of 72.9% while the areas under the ROC and P-R curves were 0.93 and 0.63 respectively. As Figure 1 shows, the model had some difficulties classifying self\_pace\_walk, Running\_3METs, Running\_5METs, and Running\_7METs classes. Case Two used only vector magnitude of Counts, performed very poorly as it cannot distinguish the Lying Class from the Sitting Class. Case Three had the highest performance, and the accuracy reached 92.2% and Area under ROC, and P-R curves were 0.99 and 0.97, respectively. Accuracy, AUC, and prAUC for the fourth case were 62.5%, 0.91 and 0.7, respectively.

RF can sort the feature based on their importance and their effect on classification. Table 4 shows the top ten important features for each case calculated by RF.

Table 4 - Top Ten Features

Table 1(Online) – Generate features and their mean and SD