CMPT 353 Report: Sensors, Noises and Walking

Naqsh Thind
naqsht@sfu.ca

Boxiao Wang
boxiaow@sfu.ca

1 Introduction

Walking data is a hot topic related to people's health. In this project, we collected accelerometer data from various participants, performed statistical tests and tested machine learning models on our dataset. The problems we addressed in this project are *a)* whether age affect how people walk, *b)* whether the sensor location (pocket, ankle or hand) has an effect on data collection, and *c)* whether machine learning models help us determine people's age based on sensor data.

2 Approach

2.1 Data Collection

45 walking data was collected from 4 participants, consisting the researcher, friends and family members. The participants are divided into 2 groups: younger group (age around 20) and elder group (age around 50).

All data was collected from iOS devices, using the same App *Sensor Data Recorder*. The data was collected in 3 ways: 15 of them are collected with the phone on the participant's right pocket, 15 are collected with the phone in hand, and 15 of them are collected with the phone tied with duct tape on the right ankle. Table 1 shows how the data can be divided into 6 groups.

		position of sensor		
		pocket	ankle	hand
age	$\sim$20	young_pocket	young_ankle	young_hand
	$\sim$50	adult_pocket	adult_ankle	adult_hand

Table 1: The 6 groups of data.

Each participant was instructed by a researcher and did the following steps for every walking session:

- Press record and stand still for 2-3 seconds
- Begin walking to the predetermined destination
- Stand still for 2-3 seconds, then press finish

Each walking session is 2-3 minutes long. The distance traveled for each session is also calculated using Google Maps as a reference.

2.2 Data Cleaning

The first and last 5 seconds of each walking data were trimmed because the participants were motionless when they start and finish recording.

For noise filtering, we experimented around LOESS filter, Kalman filter, Butterworth filter and Fourier Transform to filter the data. From the gyroscope and acceleometer data, we could not build a proper linear relationship to make good use of Kalman filter. A comparison of performance between LOESS filter, Butterworth filter and Fourier Transform are shown in the figures below.

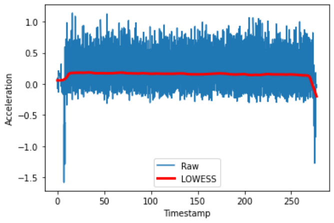


Figure 1: LOESS filter vs. Raw Data

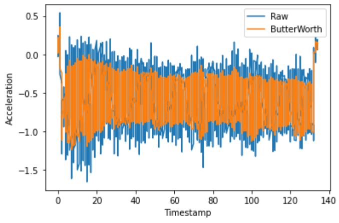


Figure 2: Butterworth filter vs. Raw Data

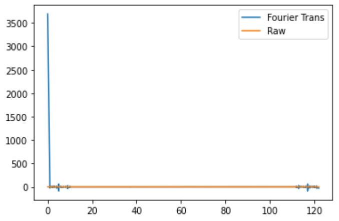


Figure 3: Fourier Transform vs. Raw Data

As we can see, LOESS did not perform well because the data was oscillating between positive and negative values, and LOESS did not take care well of that. Butterworth filter kept data in a good range and performed the best among the 3 filters. Fouried Transform did not produce a reasonable result.

2.3 Data Analysis

We performed both statistical tests and machine learning methods on the processed dataset to see if we can get interesting results.

For the machine learning part, we wanted to address these problems:

- Can we determine a person is younger (around 20) or older (around 50) from the walking data?
- Can we determine the position of sensor (hand, ankle or pocket) from the walking data?
- Can we determine the position of sensor and age at the same time?

These problems can be viewed as classification problems. First, we generate features that might be useful in training. While a single point of acceleration might not be helpful in determining any information about the person, we combined it with the moving average of adjacent acceleration data, much like what a LOESS filter did. We were also interested in the length of the acceleration vector, speed by intergration, as well as their difference versus previous data.

The Net Acceleration, which is the length of the acceleration vector can be calculated as:

$$a_{net} = \sqrt{a_x^2 + a_y^2 + a_z^2}$$

where a_x, a_y and a_z are accelerations on the respective axis.

After determining features, we tested various machine learning models including Naive Bayes, k Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting and Neural Network from the `sklearn` package. A total of 5232 rows of data was split into train and test set using `sklearn.model_selection.train_test_split`. Parameters are also tuned for the best performance of each individual model.

We will discuss how these models performed in the results section.

3 Results

3.1 Statistical Tests

Our approach for the statistical analysis was to perform normality tests on the data.

- If they pass the test, we will conduct ANOVA test, followed by Post Hoc Analysis using Tukey's HSD test.

- If not, we will perform Mann Whitney U-Test, since that does not assume normality.

We performed Normality Tests on the dataset and found out that the data gives out very small p values, which means that it does not “look normal”. Hence, it does not pass the test. So we decided to go with Mann-Whitney U-Test. Plotting the histograms for the data suggests that the data isn’t too far from being normal, hence we can apply this test.

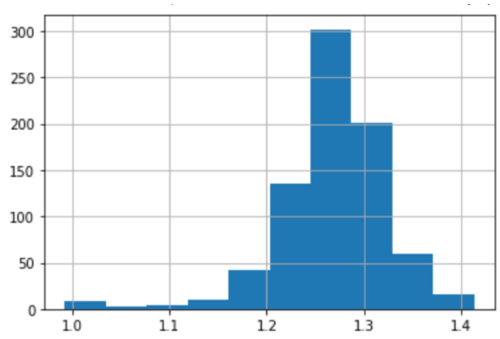


Figure 4: Histogram of young_ankle

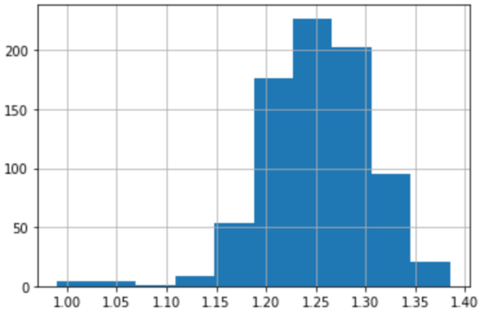
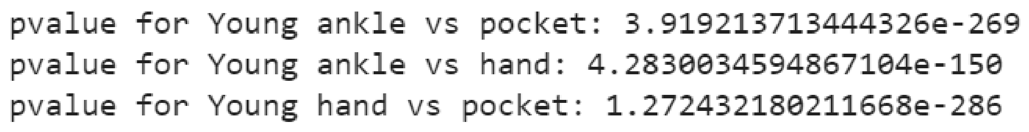


Figure 5: Histogram of adult_ankle

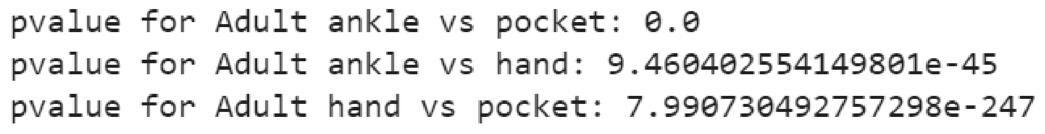
The pvalues we got from the test were very small, which suggests that values from one group tend to sort higher than the others. This was the case for all possible pairs of the groups. Hence, we can conclude that **we get different values when the location of the device is different.**

Given below are the pvalues for the various groups.



```
pvalue for Young ankle vs pocket: 3.919213713444326e-269
pvalue for Young ankle vs hand: 4.283034594867104e-150
pvalue for Young hand vs pocket: 1.272432180211668e-286
```

Figure 6: pvalues on young dataset



```
pvalue for Adult ankle vs pocket: 0.0
pvalue for Adult ankle vs hand: 9.460402554149801e-45
pvalue for Adult hand vs pocket: 7.990730492757298e-247
```

Figure 7: pvalues on adult dataset

Computed mean for the respective groups also differs (shown in Table 2 below), which further supports that the values tend to be different for different locations of the device.

Now we need to find out which location gives more accurate readings. We will discuss that in the next section.

Age and Position	Mean Value
Adult Ankle	1.252
Adult Hand	1.214
Adult Pocket	1.015
Young Ankle	1.268
Young Hand	1.411
Young Pocket	1.022

Table 2: Mean Net Acceleration among differnet age and positions

3.2 Other Numerical Analysis

Concluding which device location gives better readings is a bit tough to say, given the not so great accuracy of the device's sensors. But we decided to analyse this using two criteria.

- Comparing the Original data vs Filtered data for all the groups, to notice any obvious patterns for the respective device location.
- Computing the Velocity of the person, by doing Numerical Integration of Instantaneous Acceleration with respect to time using `scipy.integrate.trapz`.

3.2.1 Filtering Comparison

We compared acceleration graphs for all the groups and noticed some patterns. We will discuss with an example below, for each LOESS and Butter Worth:

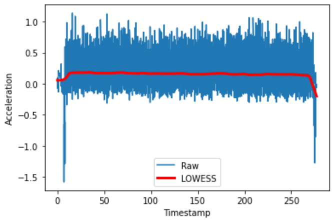


Figure 8: LOESS filter on pockets

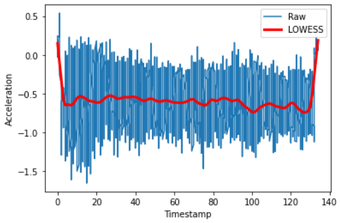


Figure 9: LOESS filter on hands

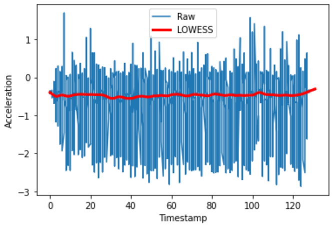


Figure 10: LOESS filter on ankles

Here, we can see that we can't obtain much from LOESS filter for the Pocket location (even after changing fractions), but it is not too bad for Ankle and Hand locations. But we still cannot draw any conclusions from this.

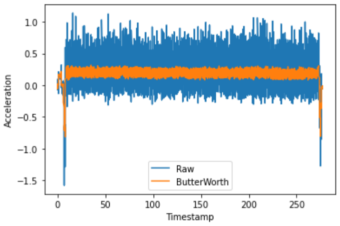


Figure 11: ButterWorth filter on pockets

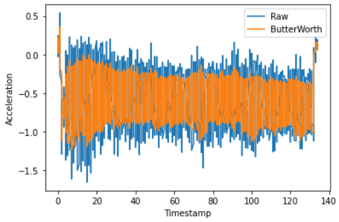


Figure 12: ButterWorth filter on hands

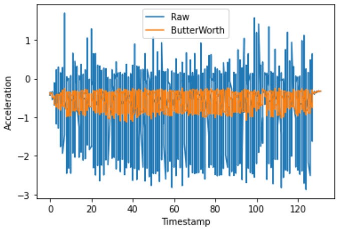


Figure 13: ButterWorth filter on ankles

Using the butterworth filter does a better job, as stated above.

Hand location: We can see that we are not able to filter out much noise and we're still getting values that seem very sensitive to the noise. Also, the pattern of the filtered data seems too much similar to the noisy data.

Pocket location: Here, we can see that the data is filtered a lot, and it looks like we might lose some important data while filtering out the noise.

Ankle location: This location seemed to lie in between the extremities of Hand and Pocket. Sure, we are filtering out noise, but it seems to be better than both Hand and Pocket. It is not too sensitive to the noise like Hand and also doesn't ignore the pattern as much as Pocket.

From the above information, we can say that Ankle location seems to give better results, at least till the filtering part.

3.2.2 Computing and Comparing Velocities

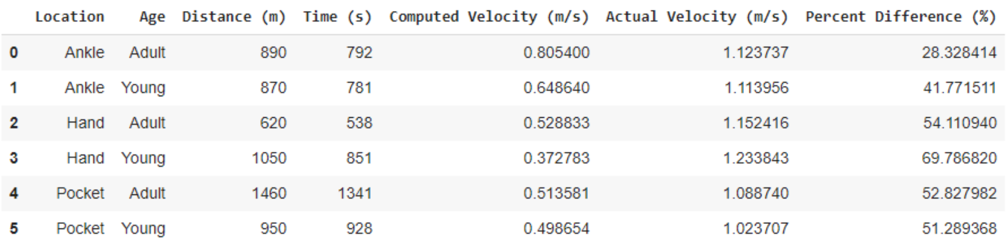
We computed the **instantaneous accelerations** for the dataset, for each timestamp and performed numerical integration of Instantaneous Acceleration with respect to Time, to obtain the average velocity of the age group.

$$\mathrm{d}v = a\mathrm{d}t$$

$$\int_{v_0}^{v} \mathrm{d}v = \int_{t_0}^{t} a\mathrm{d}t$$

Since, we already have the time intervals and the distances travelled from our dataset, we can compute average velocities (velocity = distance/time) of the age groups, for different Device Locations.

The table below shows the Computed Velocity, Actual Velocity, Percent Difference for the two age groups, for different Device Locations.



	Location	Age	Distance (m)	Time (s)	Computed Velocity (m/s)	Actual Velocity (m/s)	Percent Difference (%)
0	Ankle	Adult	890	792	0.805400	1.123737	28.328414
1	Ankle	Young	870	781	0.648640	1.113956	41.771511
2	Hand	Adult	620	538	0.528833	1.152416	54.110940
3	Hand	Young	1050	851	0.372783	1.233843	69.786820
4	Pocket	Adult	1460	1341	0.513581	1.088740	52.827982
5	Pocket	Young	950	928	0.498654	1.023707	51.289368

Figure 14: Computed Velocity, Actual Velocity and Percent Difference

From the above table , we can see that we are not able to obtain accurate velocity values from the data.

Ankle Location has the lowest percentage difference from the actual velocity values (28% and 42%). It is true for both Adult and Young data groups.

Hand Location on the other side, differs the most from the actual values (54% and 70%)

Pocket Location does a better job than Hand, but is still considerably less accurate than Ankle Location (53% and 51%)

So, overall, Ankle Location seems to do a better job in filtering, as compared to the other two locations. Alos, it did considerably well while computing the velocities using the Accelerometer values. Hence, considering the analysis done in 3.2.1 and 3.2.2, it would be fair to say that, out of the 3 locations, Ankle Location does the best job, according to our data.

Considering the analysis done in 3.2.1 and 3.2.2, it would be fair to say that, out of the 3 locations, **Ankle Location does the best job, according to our data.**

3.3 Machine Learning

As stated above, the first question we address is how machine learning models can classify between two age groups (around 20 vs. around 50) based on walking data. The results are shown in Table 3.

Model	Test Accuracy(%)
Naive Bayes	77.1
KNN	95.2
Decision Tree	92.5
Random Forest	94.2
Gradient Boosting	97.6
Neural Network	95.1

Table 3: Model Performance on determining age

We can see that every model except for Naive Bayes have more than 90% accuracy, with Gradient Boosting having the best performance. We then explored how machine learning models can identify the position of sensor (pocket, hand or ankle). The results are shown in Table 4.

Model	Test Accuracy(%)
Naive Bayes	99.2
KNN	99.6
Decision Tree	99.7
Random Forest	99.8
Gradient Boosting	99.8
Neural Network	99.5

Table 4: Model Performance on determining position of sensor

We can see that every model has suprisingly high accuracy. We conclude that the position of sensor has a significant effect on the collection of walking data, which is why they are easily differentiated. We then further explored if the models can determine the position of sensor and participant's age in the same time. The results are shown in Table 5.

Model	Test Accuracy(%)
Naive Bayes	88.6
KNN	94.0
Decision Tree	94.7
Random Forest	96.1
Gradient Boosting	99.0
Neural Network	93.2

Table 5: Model Performance on determining both position and age

We can see that gradient boosting took the lead again in performance, with a testing score of 99%. We also noticed two interesting facts.

The first thing is that neural network did not perform very well (rank 3 in determining age, rank 5 in determining both) as we expected, since neural network is usually considered a powerful tool in machine learning. The possible reason is that the amount of data we gathered is rather toy-sized for building powerful neural networks, and it can't really do much work based on 5000 rows of data.

The second thing we noticed is that some models (Naive Bayes, Decision Tree, Random Forest, Gradient Boosting) did a better job when determining both position and age than determining age only. It is possible that **the information about sensor position can be used as a prior in determining the user's age**, thus lead to a even better result. We believe that knowing the sensor location is very helpful in real-world health data collection.

4 Conclusion

We performed various statistical tests and machine learning methods to find out whether there is a difference in walking for different ages of people, whether sensor location affects data collection, and how machine learning models can help determine age or sensor position in walking data. We concluded that there is a significant difference in walking data when collected from different body locations, and established that ankle data was the best in collecting accurate results. On the machine learning part, we can see that Gradient Boosting did a really good job in differentiating ages from walking data, especially if it knows where the sensor is located.

5 Reflections on our work

Some improvement on our work would be gathering more data. Due to the pandemic, collecting data from a large group of people is particularly difficult, so we had to do it ourselves and with family members, resulting in only 4 participants and 45 pieces of data. We could also explore online datasets to get better results in the machine learning part. We also could explore more sensor options that might have additional features, which will help in our analysis. What's more, we learned to give a trial run before collecting a large amount of data, as the first data we gathered was not done in a scientific measure and was not usable.

6 Project Experience Summary

Naqsh Thind

- Worked in a group of two, to collect and analyze accelerometer data using statistical and ML techniques
- Compared device placement positions, hand, ankle and pocket to conclude which works best
- Used Butter Worth Noise Filtering, Statistical Tests and Velocity estimations to draw conclusions
- Used several ML models like Naive Bayes, KNN, Neural Networks to build and train a classifier that predicts the age of person and location of device

Boxiao Wang

- Organized, collected and processed a small accelerometer walking dataset from multiple participants
- Researched on various statistical methods as well as machine learning models to predict age from walking data
- Tuned the Gradient Boosting classifier and achieve 99% testing accuracy on the dataset
- Worked in group of 2 and finished a data analysis report