This is a report WISDM data which can be found at UCI repository here. The WISDM data contains biometric information from phone and smart watch sensor based on the activity of the user. In other studies only smart watch accelerometer is utilized for monitoring limited physical activities. Here the data contains not only accelerometer but also gyroscope data from both smart watch and phone leading to four times the data available. Included in the data are a list of 18 activities that corresponds to 51 users over a span of 50 minute for each activity. In this report analysis and modeling prediction is done to predict and classify the user activity and determine if such data is feasible for biometric identification.

Data Inspection

	id	activity	timestamp	x	y	z
0	1600	A	252207666810782	-0.364761	8.793503	1.055084
1	1600	A	252207717164786	-0.879730	9.768784	1.016998
2	1600	A	252207767518790	2.001495	11.109070	2.619156
3	1600	A	252207817872794	0.450623	12.651642	0.184555
4	1600	A	252207868226798	-2.164352	13.928436	-4.422485

The data set contains four repositories: phone_accel, phone_gryo, watch_accel, watch_gyro. Each of these repository have 51 files corresponding to the 51 users, and in each file there are over 73,000 sample points each corresponding to 50ms. In total there are over 15 million sample points.

Each file contain the following features: user id, timestamp, x, y, z. The activity code is listed below corresponding the 18 activities. The timestamp interval for each sample point is 50ms as stated from before. The x y z corresponds to either the sensor (m/s2) or gyro sensor (rad/s).

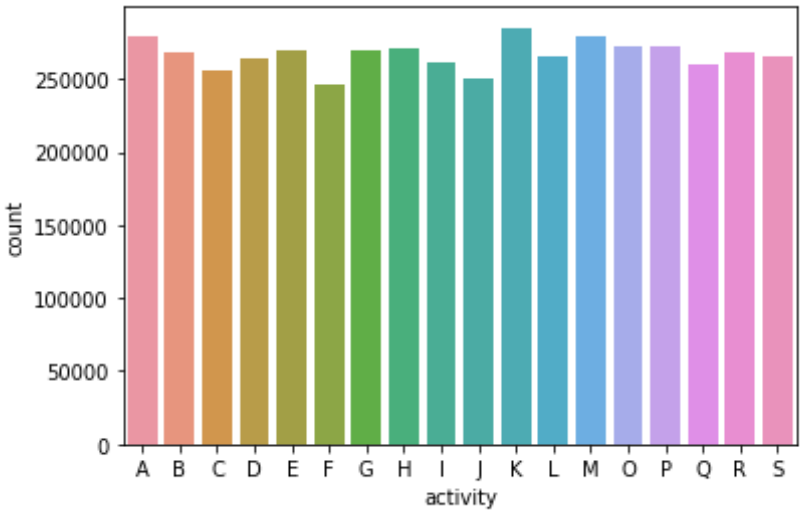
Activity	Code
Walking	A
Jogging	B
Stairs	C
Sitting	D
Standing	E
Typing	F
Brushing Teeth	G
Eating Soup	H
Eating Chips	I
Eating Pasta	J
Drinking from Cup	K
Eating Sandwich	L
Kicking (Soccer Ball)	M
Playing Catch w/Tennis Ball	O
Dribblinlg (Basketball)	P
Writing	Q
Clapping	R
Folding Clothes	S

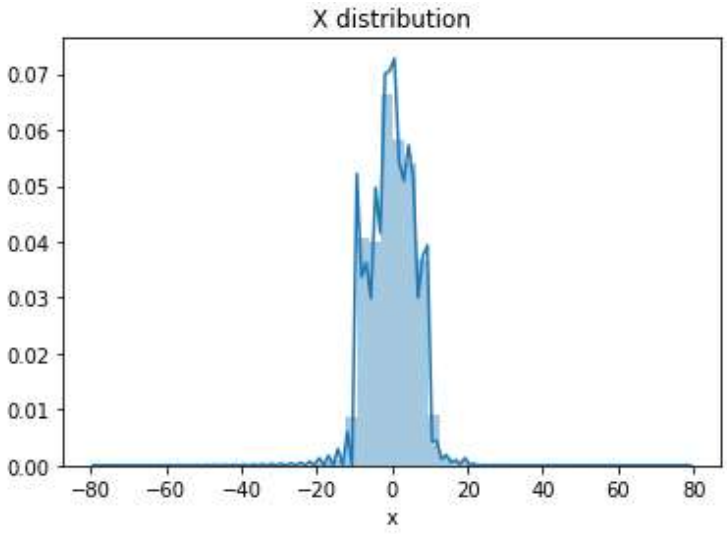
SUMMARY INFORMATION FOR THE DATASETS

Number of subjects	51
Number of activities	18
Minutes collected per activity	3
Sensor polling rate	20Hz
Smartphone used	Google Nexus 5/5x or Samsung Galaxy S5
Smartwatch used	LG G Watch
Number raw measurements	15,630,426

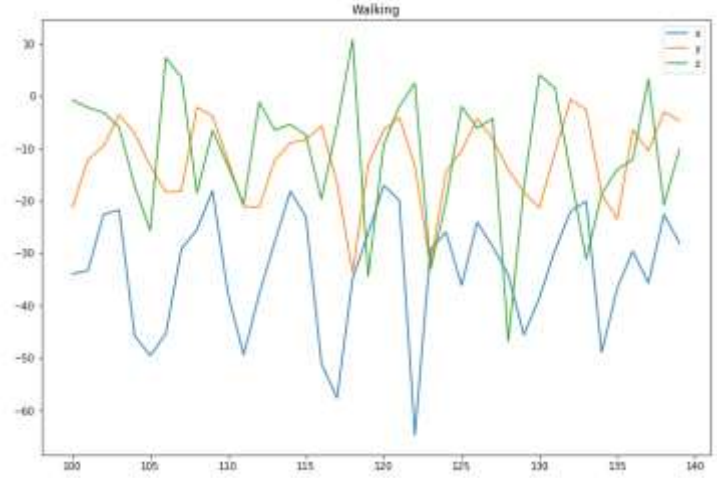
Data Exploration

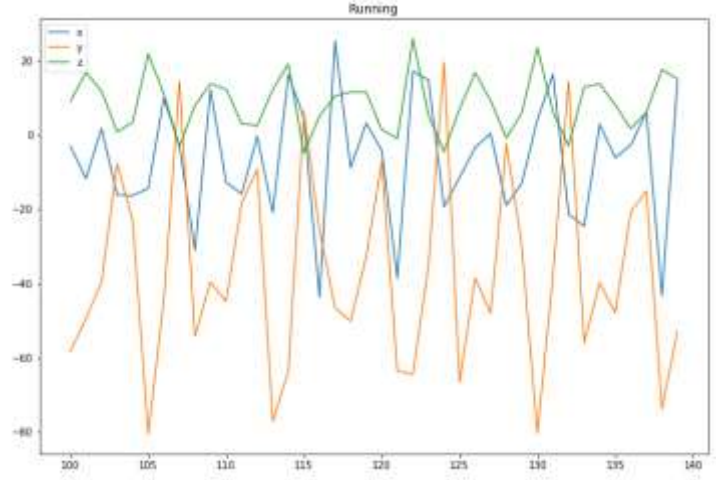
From initial examination the number of activities are evenly distributed as expected from a controlled environment. And the same goes for the X Y Z distribution.

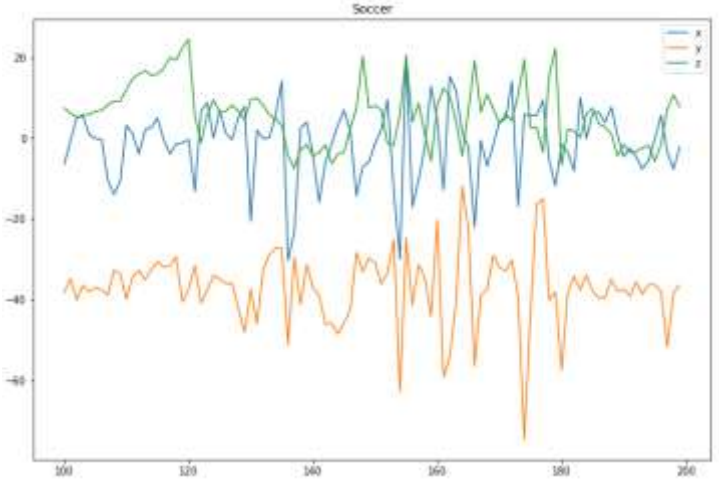




When looking at the time range of XYZ data it tells a different story. Below are three activities, walking, running, and playing soccer. Each point in the plot is 0.2 second over a range of 8 seconds. Under careful exanimation there are clear differences in the frequency and amplitude in these activities.


Walking
x
y
z
30
0
−10
−20
−30
−40
−50
−60
100
105
110
115
120
125
130
135
140


Running
x
y
z
20
0
−20
−40
−60
−80
100
105
110
115
120
125
130
135
140


Soccer
x
y
z
20
0
−20
−40
−60
100
120
140
160
180
200

Feature Engineering

Due to the short interval of each data point there is a need to transform the time series data to labeled examples. Since each data point is 50ms, a grouping of 200 sample points should be used leading to a partitioning of the whole data structure to into 10 second labeled samples.

With each 200 sample points the high level features are generated from XYZ data columns. The script for the transformation can be view on the github page. The following 46 features are:

- Avg: Average value per axis (3)
- Std: Standard deviation per axis (3)
- Var: Variance per axis (3)
- Abs Range: Range per axis (3)
- Resultant: Resultant per axis (1)
- XYZpeak: Period of fitted sine wave (3)
- Bin Distribution: From the range of each axis, 10 equal sized bins are formed and the fraction of the 200 values within each bin is recorded (30)

Classification Modelling

After engineering the high level features classification models are made to see if the sensors can predict the activity of the user. Three classification algorithm are used, these are Random Forest, XGBoost, and KNN. A 10 k stratified fold is applied for these classification.

Random Forest

To test the accuracy of the sensor the classification is split into 3 parts, accelerometer data, gyro data, and total data. For accelerometer the precision, recall, and f1-score is listed in the table below.

	precision	recall	f1-score	support
drinking from cup	0.99	1.00	0.99	993
basketball	0.98	0.99	0.99	919
stairs	0.94	0.99	0.96	914
jogging	0.94	0.97	0.96	944
soccer	0.91	0.95	0.93	970
eating chips	0.90	0.94	0.92	879
eathing soup	0.87	0.87	0.87	930
eating pasta	0.79	0.84	0.81	945
folding clothes	0.82	0.83	0.83	933
walking	0.86	0.80	0.83	881
eating sandwitch	0.85	0.82	0.84	996
tennis	0.80	0.77	0.79	915
writing	0.91	0.94	0.93	969
brushing teeth	0.88	0.91	0.89	929
clapping	0.92	0.88	0.90	969
sitting	0.92	0.86	0.89	903
typing	0.97	0.90	0.94	971
standing	0.97	0.97	0.97	921
accuracy			0.90	16881
macro avg	0.90	0.90	0.90	16881
weighted avg	0.90	0.90	0.90	16881

The accuracy and precision is very high at a predictive rate of 0.90.

For the gyro sensor the classification is listed below. The gyro sensor also have high predictability but somewhat lower than the accelerometer.

	precision	recall	f1-score	support
drinking from cup	1.00	1.00	1.00	993
basketball	0.99	0.99	0.99	919
stairs	0.95	0.98	0.97	914
jogging	0.92	0.93	0.93	944
soccer	0.86	0.91	0.88	970
eating chips	0.84	0.92	0.88	879
eathing soup	0.80	0.85	0.82	930
eating pasta	0.73	0.79	0.76	945
folding clothes	0.70	0.71	0.70	933
walking	0.74	0.68	0.71	881
eating sandwitch	0.75	0.73	0.74	996
tennis	0.74	0.65	0.69	915
writing	0.89	0.95	0.92	969
brushing teeth	0.90	0.90	0.90	929
clapping	0.92	0.88	0.90	969
sitting	0.90	0.83	0.86	903
typing	0.95	0.88	0.92	971
standing	0.98	0.96	0.97	921
accuracy			0.87	16881
macro avg	0.86	0.86	0.86	16881
weighted avg	0.87	0.87	0.86	16881

For the whole data set the classification is around the same where the score averages out.

	precision	recall	f1-score	support
drinking from cup	0.99	1.00	0.99	4350
basketball	0.99	0.99	0.99	4231
stairs	0.92	0.98	0.95	4124
jogging	0.91	0.93	0.92	4287
soccer	0.88	0.91	0.89	4327
eating chips	0.87	0.93	0.90	4084
eathing soup	0.85	0.87	0.86	4275
eating pasta	0.77	0.82	0.79	4265
folding clothes	0.74	0.76	0.75	4222
walking	0.79	0.76	0.77	4117
eating sandwitch	0.82	0.79	0.81	4424
tennis	0.80	0.71	0.76	4203
writing	0.89	0.94	0.91	4327
brushing teeth	0.89	0.90	0.89	4260
clapping	0.93	0.88	0.91	4327
sitting	0.91	0.86	0.89	4268
typing	0.96	0.89	0.93	4264
standing	0.98	0.95	0.96	4275
accuracy			0.88	76630
macro avg	0.88	0.88	0.88	76630
weighted avg	0.88	0.88	0.88	76630

XGB

For XGBoost classification there appears to be over fitting regardless of what data is passed through making it unreliable in general.

	precision	recall	f1-score	support
drinking from cup	1.00	1.00	1.00	993
basketball	1.00	1.00	1.00	919
stairs	1.00	1.00	1.00	914
jogging	1.00	1.00	1.00	944
soccer	1.00	1.00	1.00	970
eating chips	1.00	1.00	1.00	879
eathing soup	1.00	1.00	1.00	930
eating pasta	1.00	1.00	1.00	945
folding clothes	1.00	1.00	1.00	933
walking	1.00	1.00	1.00	881
eating sandwitch	1.00	1.00	1.00	996
tennis	1.00	1.00	1.00	915
writing	1.00	1.00	1.00	969
brushing teeth	1.00	1.00	1.00	929
clapping	1.00	1.00	1.00	969
sitting	1.00	1.00	1.00	903
typing	1.00	1.00	1.00	971
standing	1.00	1.00	1.00	921
accuracy			1.00	16881
macro avg	1.00	1.00	1.00	16881
weighted avg	1.00	1.00	1.00	16881

KNN

KNN classification performed the worse than RF classification. This could be due to the structure of the data where splits are a better classification tool than data grouping.

	precision	recall	f1-score	support
drinking from cup	0.70	0.74	0.72	1766
basketball	0.85	0.80	0.82	1693
stairs	0.55	0.52	0.53	1653
jogging	0.31	0.36	0.33	1719
soccer	0.36	0.44	0.40	1696
eating chips	0.36	0.46	0.41	1641
eathing soup	0.60	0.57	0.58	1663
eating pasta	0.40	0.44	0.42	1691
folding clothes	0.31	0.33	0.32	1644
walking	0.39	0.37	0.38	1614
eating sandwitch	0.38	0.39	0.38	1789
tennis	0.37	0.29	0.33	1705
writing	0.45	0.61	0.52	1729
brushing teeth	0.42	0.38	0.40	1694
clapping	0.52	0.45	0.48	1728
sitting	0.39	0.33	0.36	1747
typing	0.63	0.52	0.57	1720
standing	0.66	0.48	0.56	1760
accuracy			0.47	30652
macro avg	0.48	0.47	0.47	30652
weighted avg	0.48	0.47	0.47	30652

Conclusion

From these classification models the motion based biometric data can be used to identify activities of the user. These four simple features from either phone or watch sensor is suffice and feasible to determine the action of the user for commercial use. The best model out of the three was Random Forest, while XBG model is over fitting the data, and KNN failed in terms of grouping the data into separate classifier. After training the model a 10s sample is enough and able to be classify the activity. For the accuracy of the model the identification is quite high with the exception of eating and playing tennis where one have a more stationary monotonic data structure while the other have a high degree of variance. Given the eighteen activities that was evaluated they were all useful as biometric data for monitoring a user's daily live.