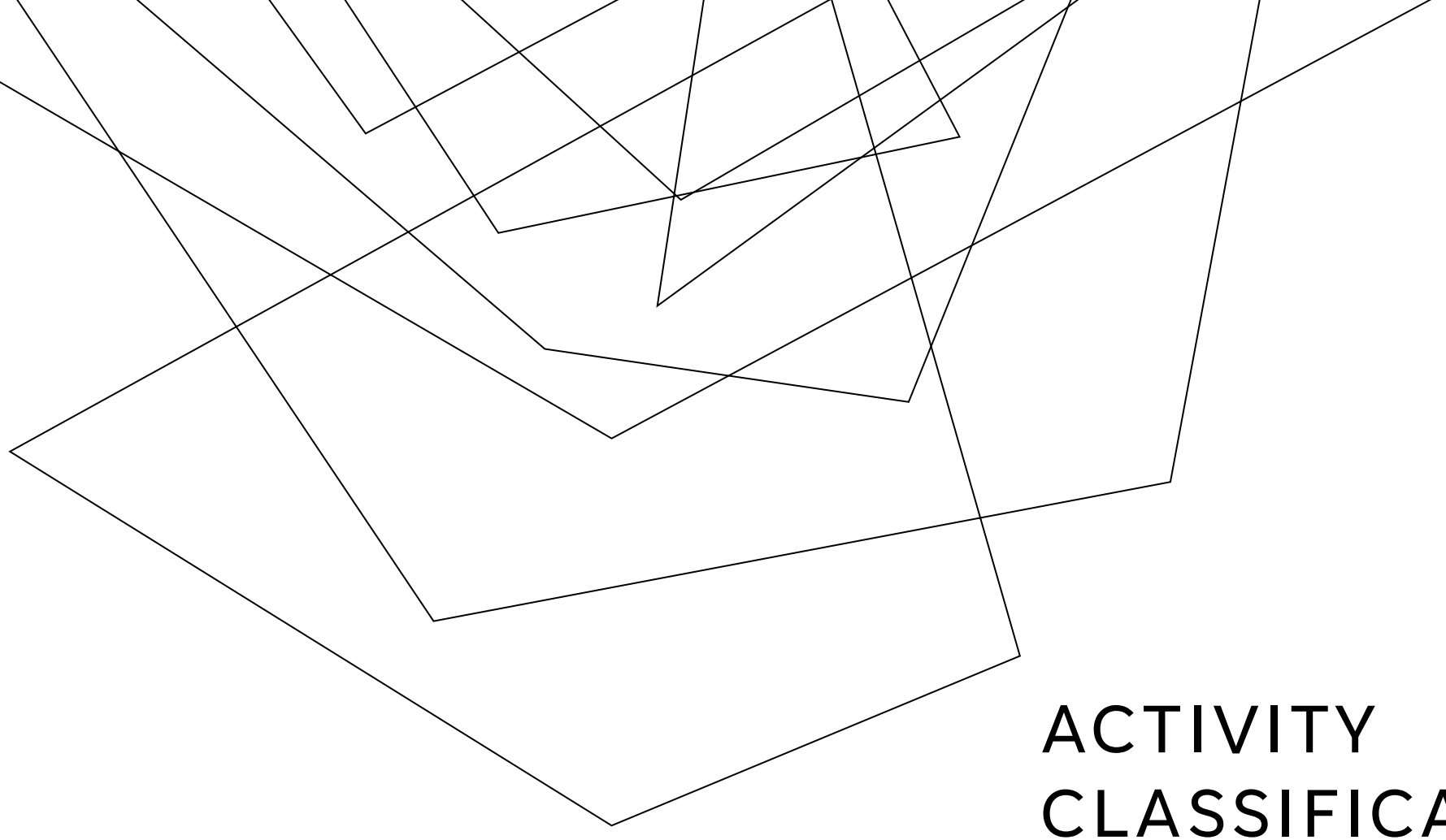


ACTIVITY CLASSIFICATION FROM SENSOR DATA

Allan Yeung

PROBLEM STATEMENT

Based on World Health Organization, the world's population with over 60 yrs old will double reaching 2.1B by 2050. Domestically, in US, older adults are projected to outnumber children by 2034. Beyond the complication with an aging population, over 50% of the population has at least 1 chronic condition such as diabetes, stroke, obesity. This will inherently drive the cost of health care higher over next decade.

OBJECTIVE

Over 95% of US population own a smartphone and over 25% of adults now owns a wearable like a smart watch. These communication devices are equipped with sensors such as accelerometer, gyroscope and magnetic sensor to determine location and movement.

The objective is to use sensor data to build a classification model to predict a person's activity such as standing, sitting, sleeping, to movement such as walking, running. These model can be used as monitoring application for healthcare and wellness purposes to better understand a person's lifestyle.

The data is collected with Samsung smartphone.

COMPANY

SAMSUNG

Samsung, is a South Korean multinational manufacturing conglomerate headquartered in Suwon, South Korea.

Samsung is one of the world's largest producers of electronic devices. Samsung specializes in the production of a wide variety of consumer and industry electronics, including appliances, digital media devices, semiconductors, memory chips, and integrated systems.

In Smartphone category, Samsung is the top brand in terms of volume as it ships about 54M phones in Q2 2023 follow by Apple with 45M.

In the wearable smartwatch market, Apple dominate the market 43% marketshare follow by Samsung.

DESIGN PROCESS

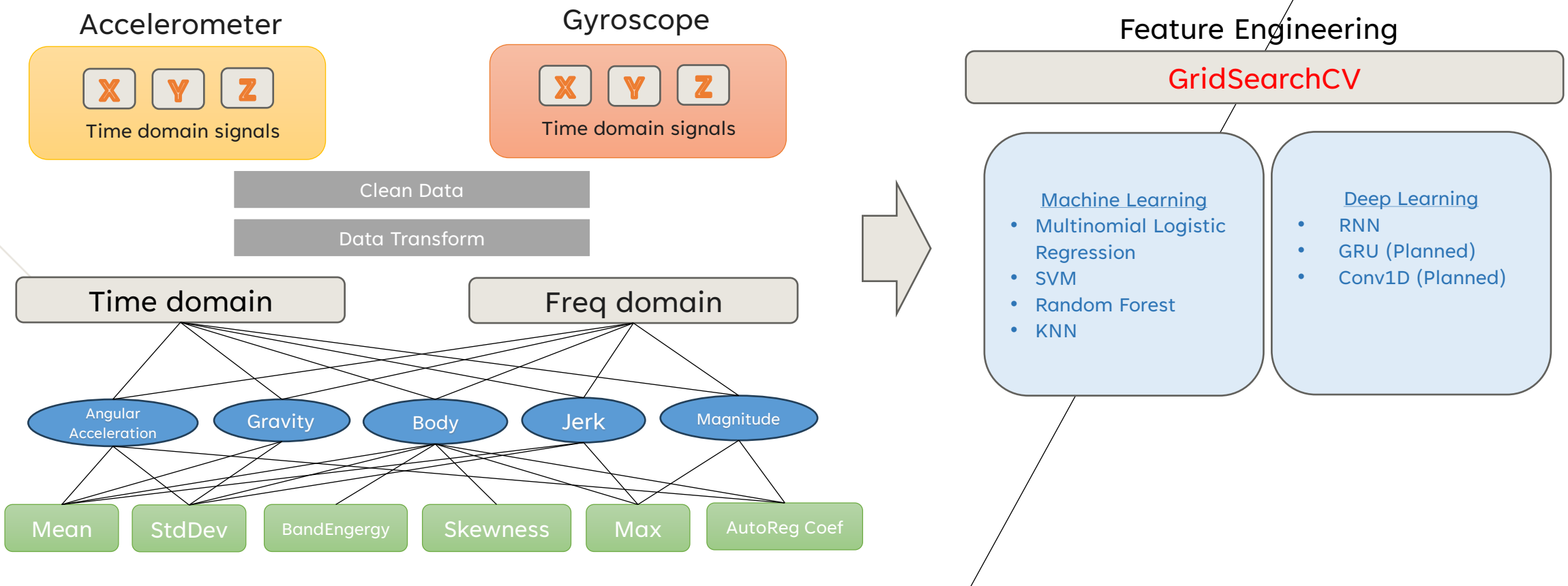
Empathize	Define	Ideate	Prototype	Test
For monitoring application, the wearable can be used to track an user daily activity to better understand their movement throughout the day.	What data should this activity classifier API serve?	Build an activity classification model	Dataset is modified to maximize the classification model's accuracy. Several models are used to evaluate the best model.	Recruit participant to perform all target 6 activities while wearing a smartphone. An application is preloaded on the phone to capture raw signals from sensors.
Activities track: 1. Laying Down 2. Standing 3. Sitting 4. Walking 5. Walking upstairs 6. Walking downstairs	The model would leverage sensor data such as accelerometer and gyroscope in the device. Developer would call the activity classifier API to identify what activities the user have done throughout the day.	The API should take raw signals from accelerometer and gyroscope where it generate time-series signals in X, Y and Z axis. Transform the data	Models include: 1. Logistic Reg (Multinomial) 2. SVM 3. Decision Tree 4. kNN	After model has train data (70% of the set), one can use a test set to compare with the prediction model. The model is assessed the based on training data compare with test data on its possible. One can evaluate based on error rate.

DATA UNDERSTANDING & EXPLORATION

Data Source

Data is captured by embedded sensor with accelerometer and gyroscope. Sensor data is captured for 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz . The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

Data Pipeline

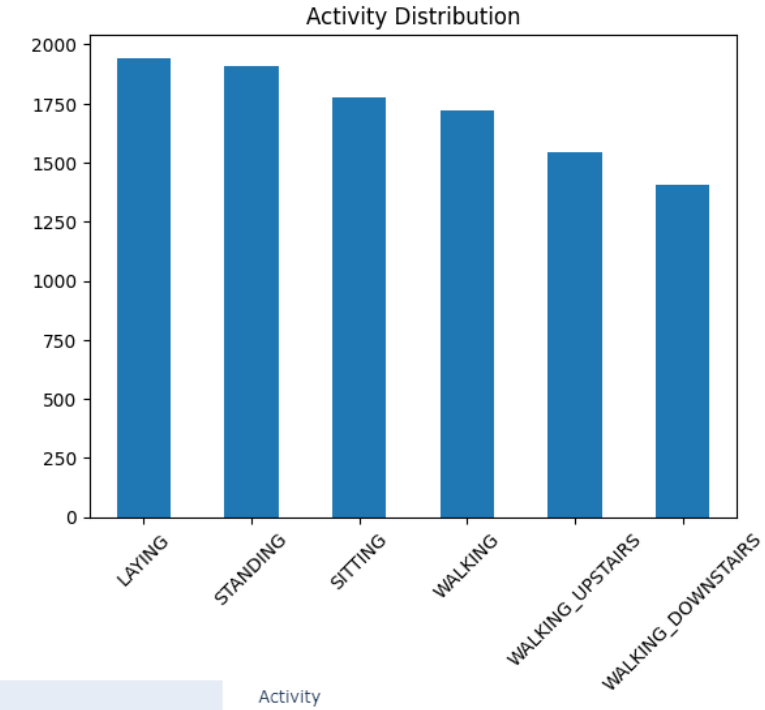


DATA UNDERSTANDING & EXPLORATION

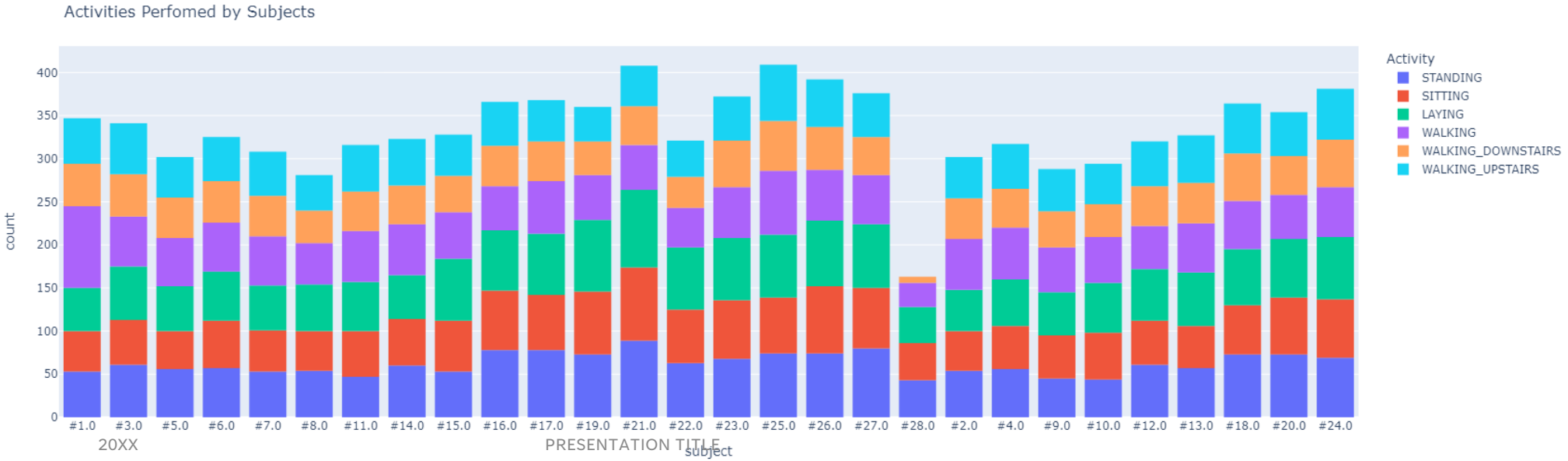
Activity Distribution

There are 6 activities performed by the research participants labeled as target:

1. Laying Down
2. Standing
3. Sitting
4. Walking
5. Walking Upstair
6. Walking Downstair



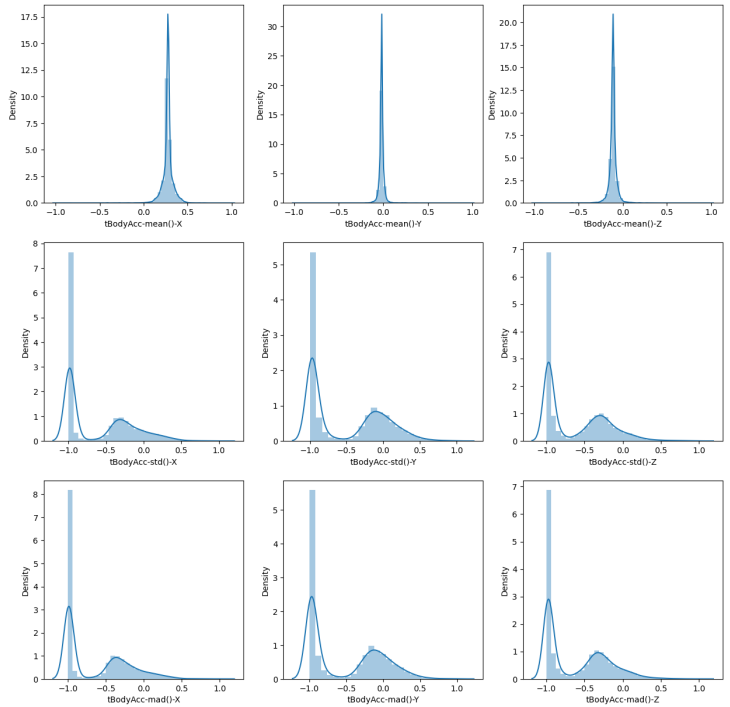
Activity Distribution by Subject



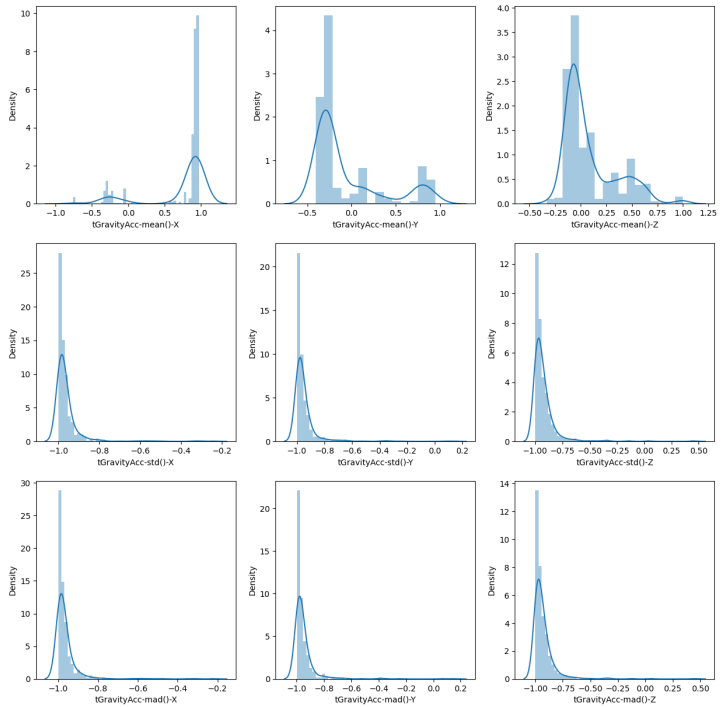
DATA EXPLORATION & FINDINGS

Distribution Plots on Time Domain

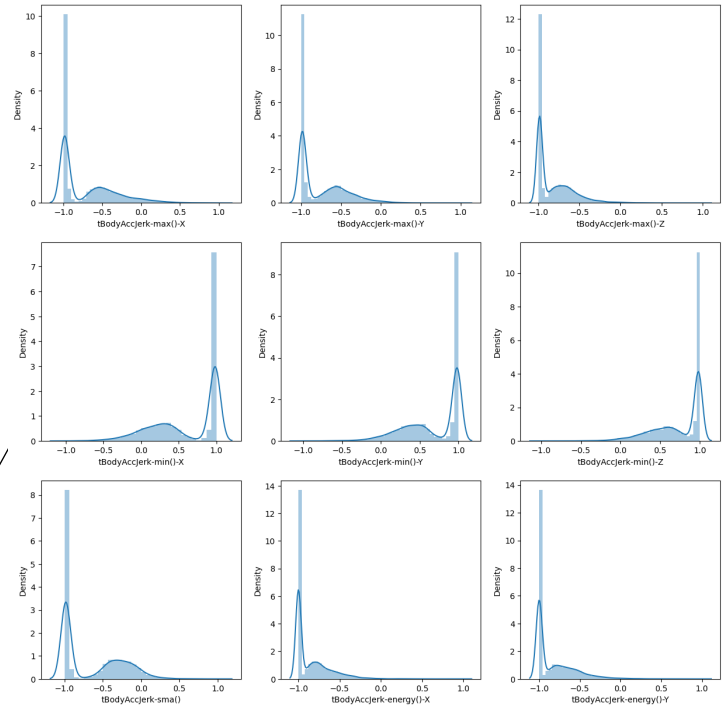
tBodyAcc



tGravityAcc



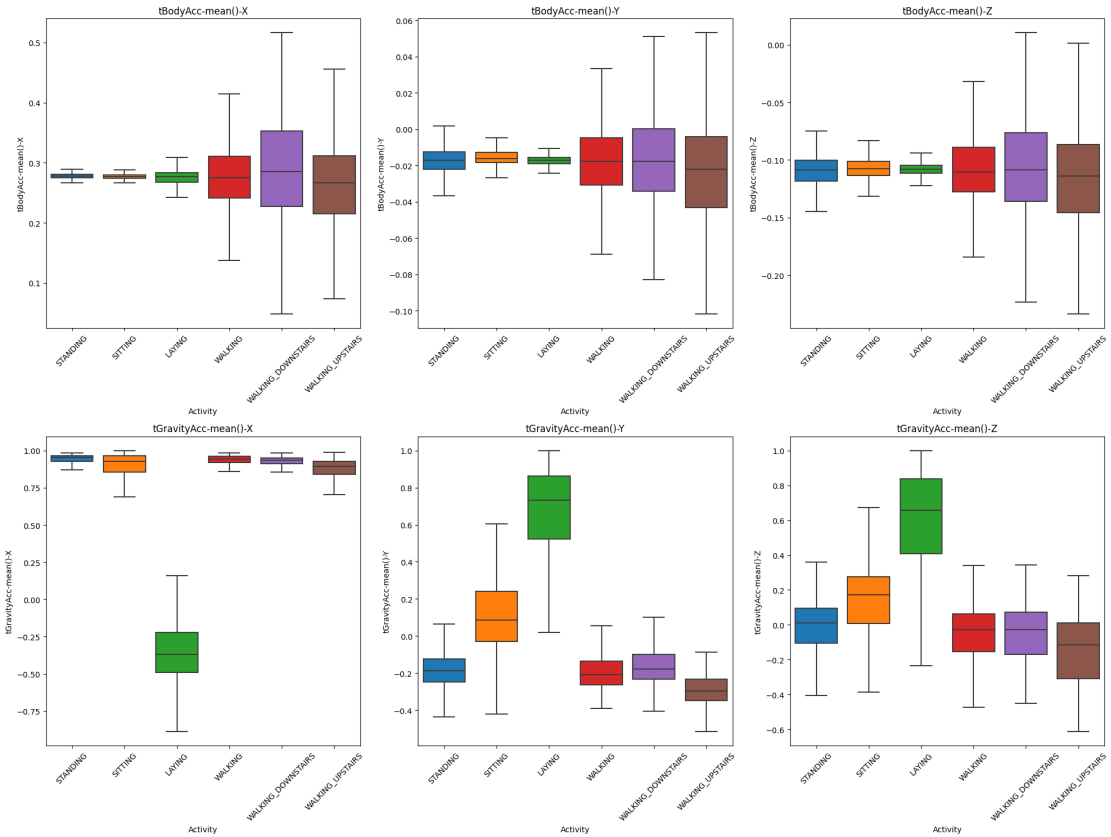
tBodyAccJerk



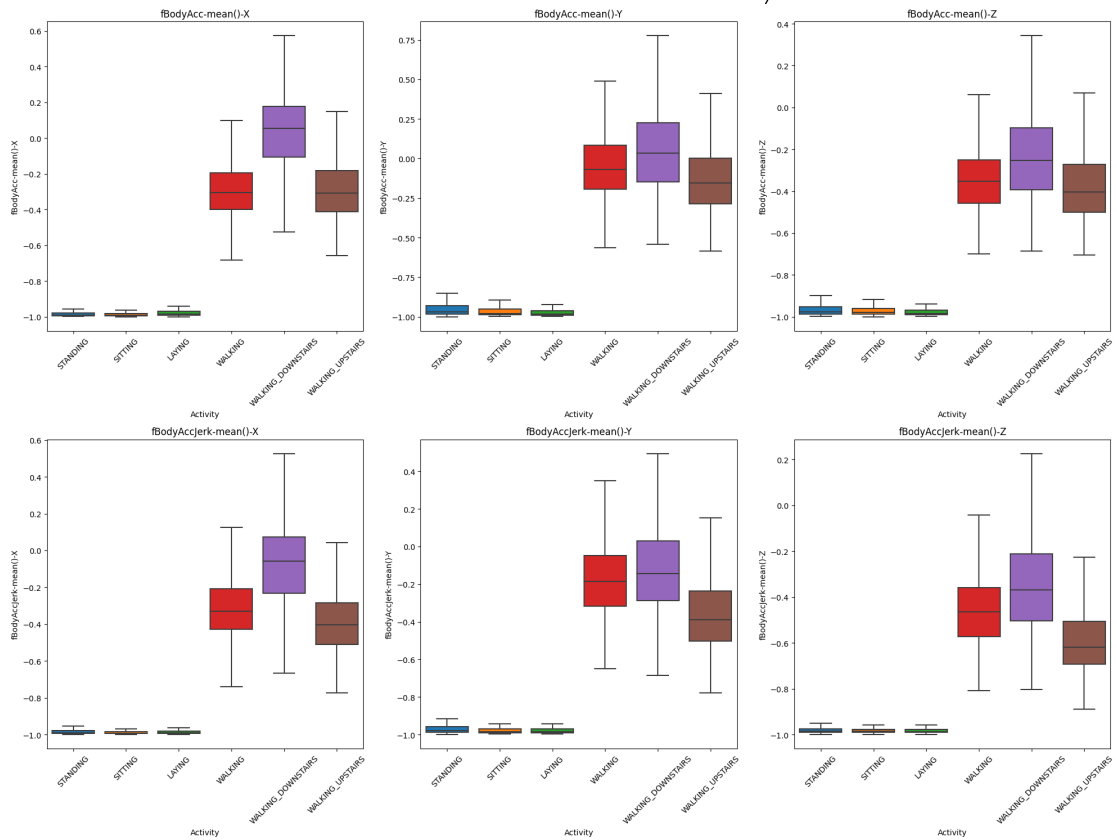
DATA EXPLORATION & FINDINGS

Time Domain vs Frequency Domain

Time Domain



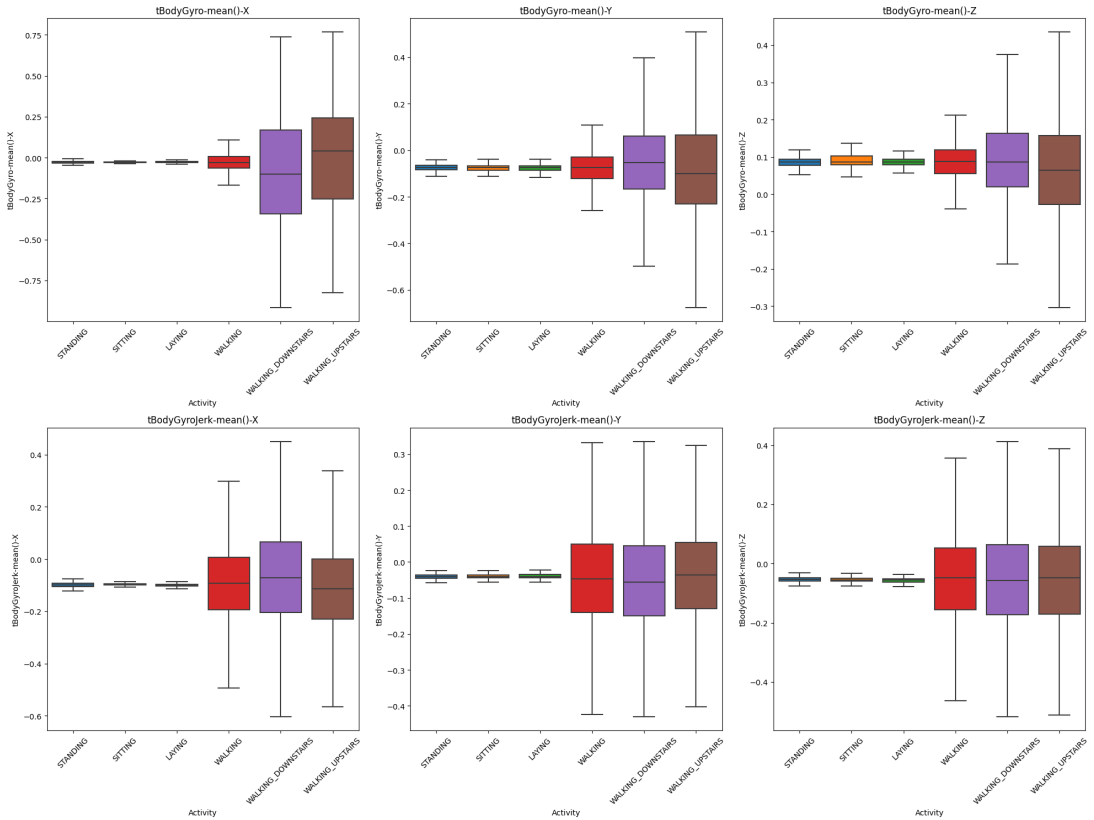
Frequency Domain



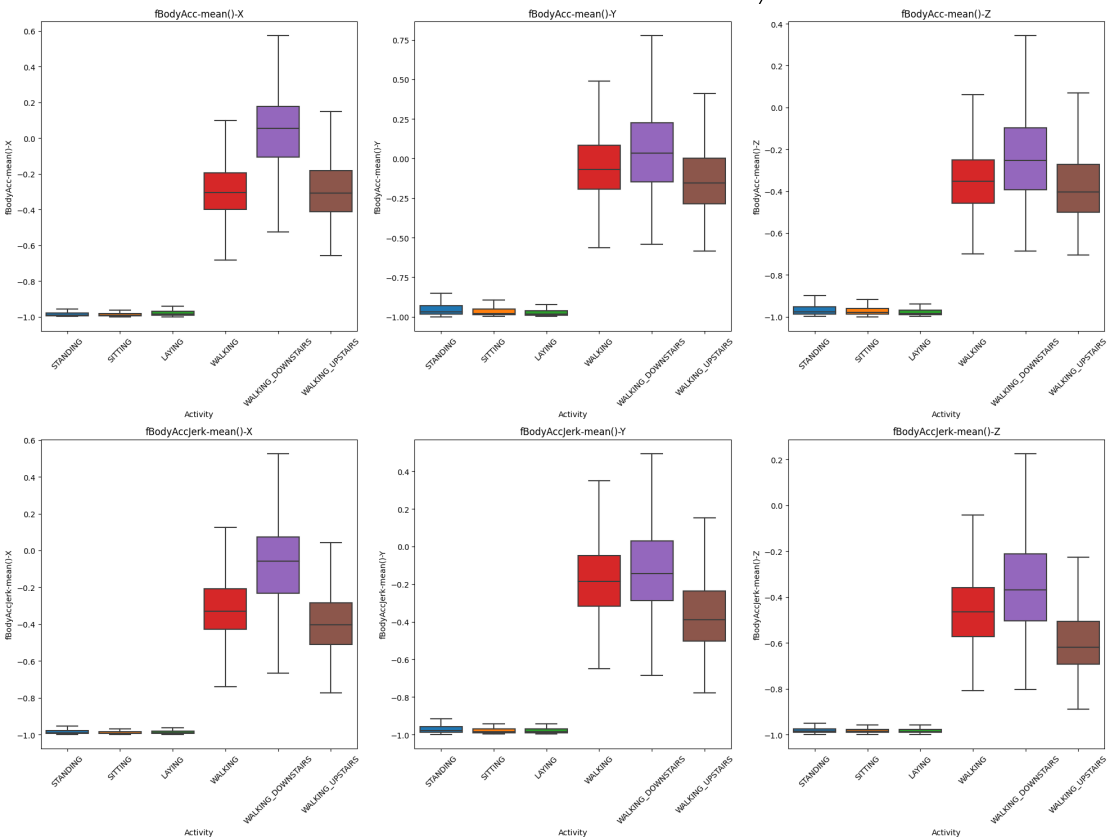
DATA EXPLORATION & FINDINGS

Accelerometer vs Gyroscope

Accelerometer Signals



Gyroscope Signals

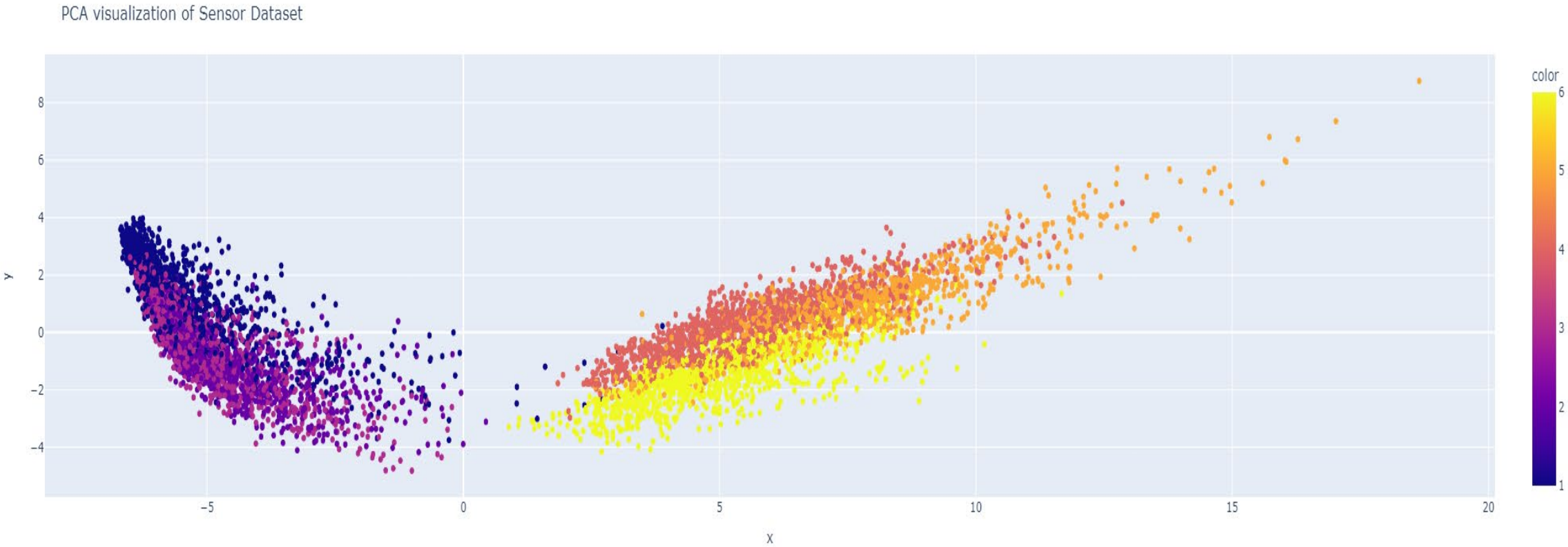


DATA EXPLORATION & FINDINGS

Clustering

PCA = 2

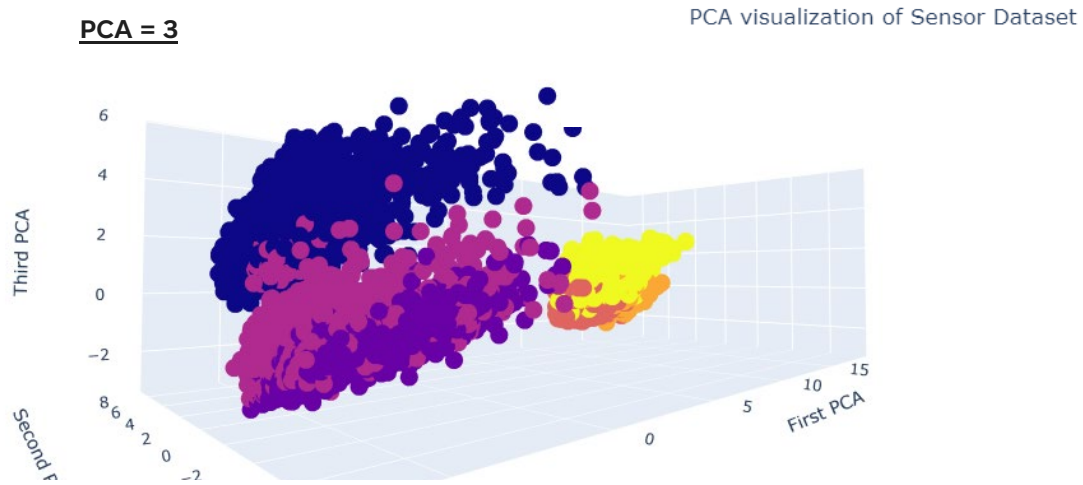
PCA = 3

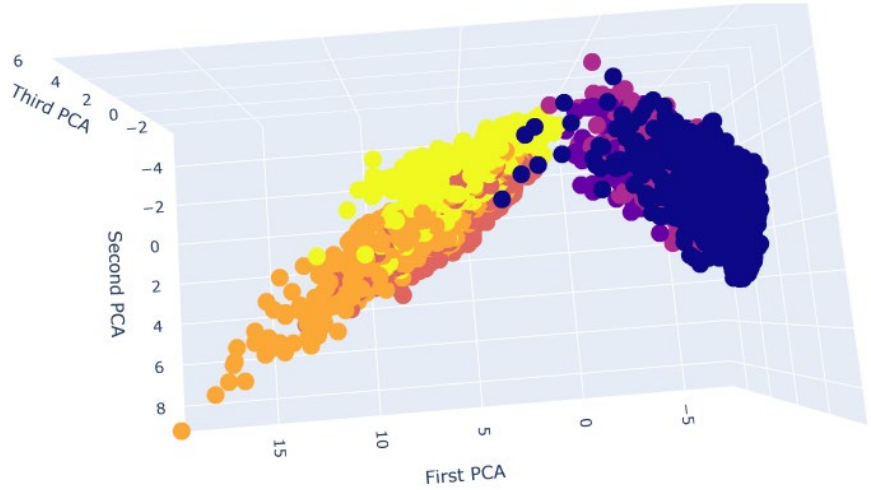


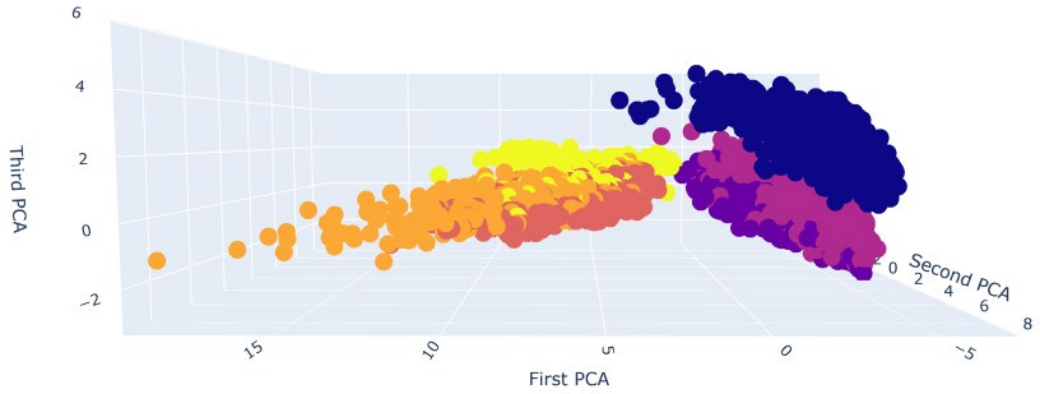
DATA EXPLORATION & FINDINGS

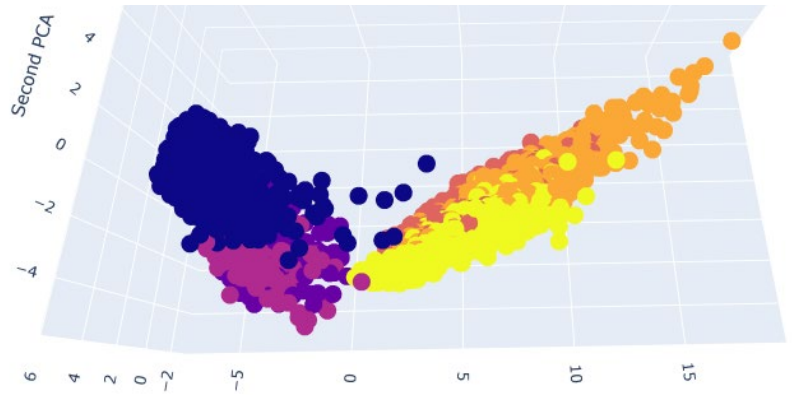
Clustering

PCA = 3









DATA EXPLORATION & FINDINGS

Visualization using t-SNE (T-distributed Stochastic Neighbor Embedding)

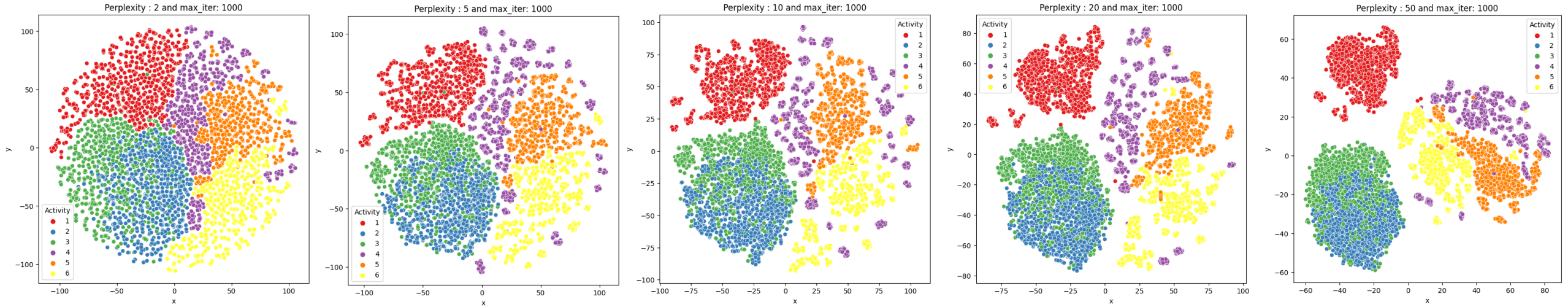
Perplexity = 2

Perplexity = 5

Perplexity = 10

Perplexity = 20

Perplexity = 50



T-SNE can help reduce the dimensionality of the dataset. Perplexity is a tuning parameter which balance the attention between local and global aspect of the data. The typical value is between 5 to 50.

Based on the visualization above of perplexity value of 2, 5, 10, 20 and 50 with a fixed iteration of 1000. You can see that some pattern in the cluster. Among the activity, 3 of the activities are stationary (#1, #2, #3) and other 3 are in motion when participant is walking (#4, #5, #6)

Coincidently, ‘stationary’ clusters are grouped together separated from the ‘motion’ clusters. One explanation is that gravitational and angular acceleration component is changing in periodic pattern when the participant is walking.

DATA MODELING & EVALUATION

6 Models are Evaluated to implement the Activity Classifier

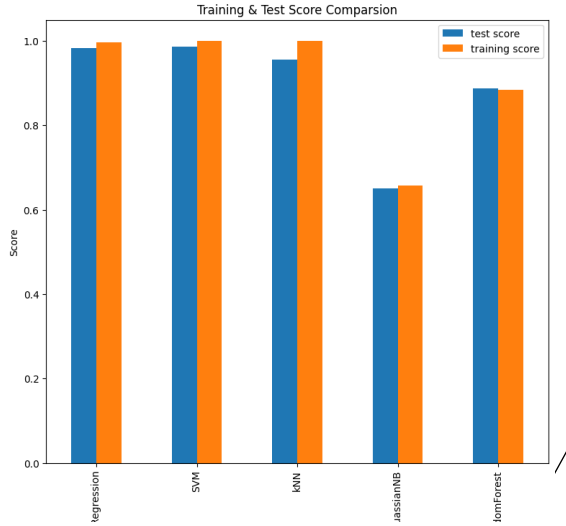
- Logistic Regression with Multinominal
- SVM
- kNN
- Gaussian Naïve Bayes
- Random Forest Classifier
- Recurrent Neural Network (RNN)

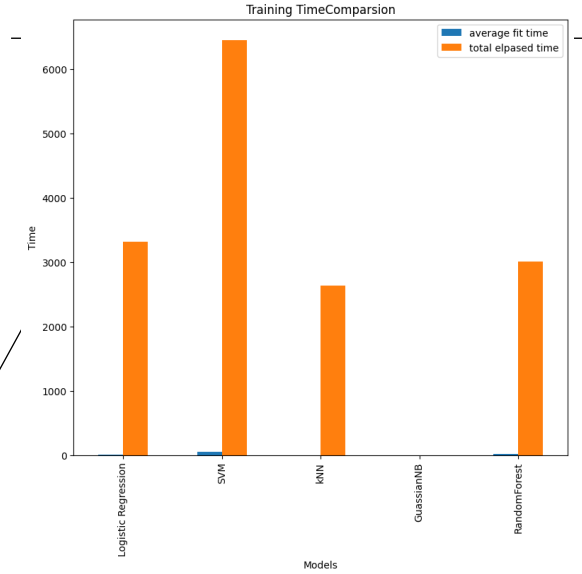
Based on Test Score generated by gridSearchCV, SVM model has the highest test score with its best parameters. Closely follow by Logistic Regression using multinomial. This is verified using accuracy metric generated through classification report. SVM yield the highest Precision, recall and F1 value. However the downside of SVM is reflected in training time as it takes at least 2x to 3x as long compare to other models.

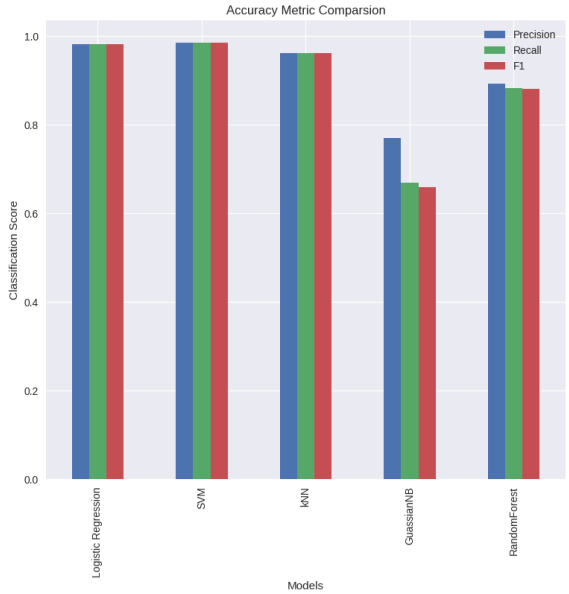
Regarding ROC Curve, SVM and Random Forest has the highest performance as a classification model. ROC curve plot TPR compare to FPR at different classification threshold. Interesting, Logistics Model has an invested curve. A potential remedy is remove the max iteration.

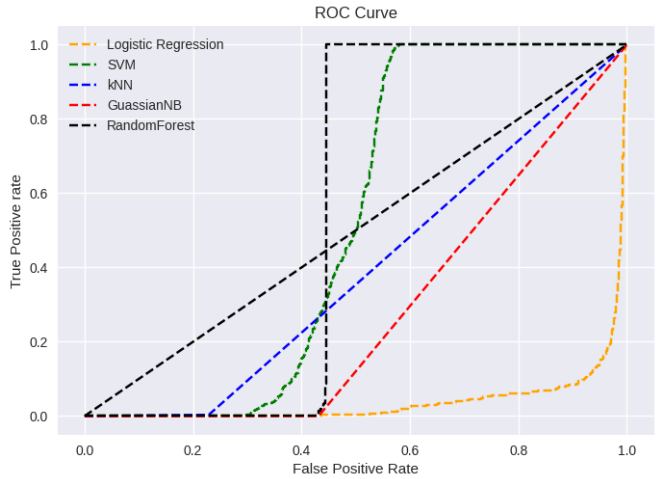
As for deep learning, a simple RNN is used with Sequential model. It produced a TP value off 3700 (True Positive) with no False Positive and vice versa with Negative scenarios. Also It produces a precision score of 1.0 which also mean no false positive. A recall score of 1.0 mean it has no false negative.

In conclusion, due to the fact that SVM require a lengthy modelling time, it will increase the cost of the model over the lifespan of the deployment. It is recommended to use RNN to develop the human activity classifier for its high precision and accuracy rate.





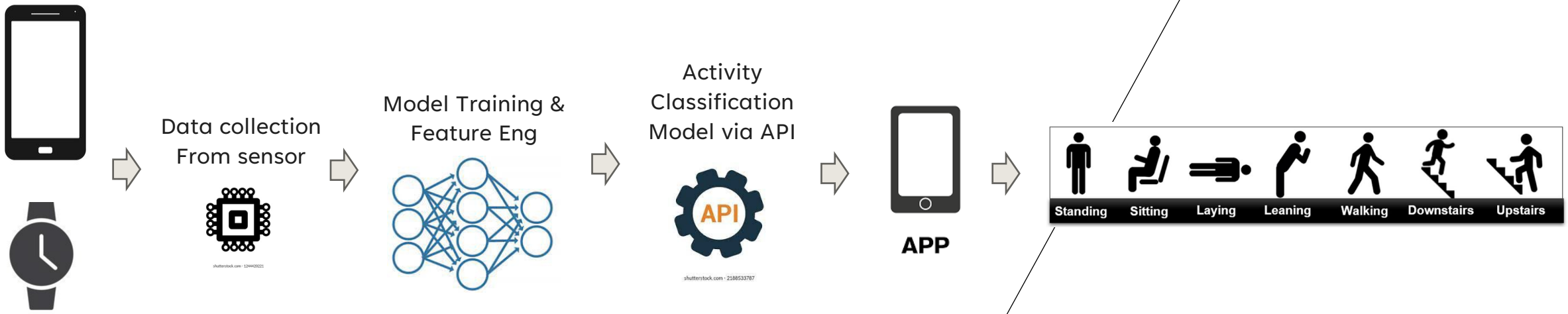


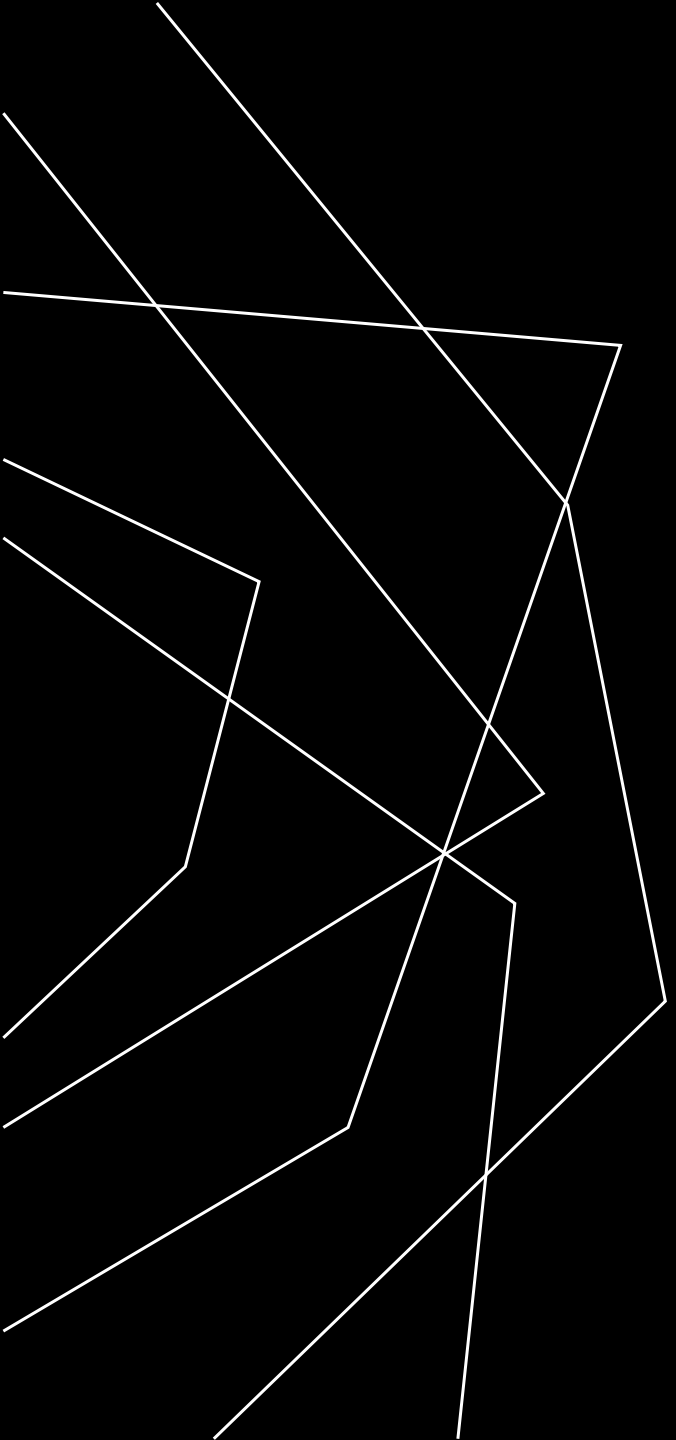


DATA MODELING & EVALUATION

Next Steps

Based on the activity classification developed by RNN, platform company such as Samsung can package the model as API for developer to develop application or Samsung's own usage based on its own sensor in its product linesup





THANK YOU