Fall Detection Model

Samuel Matsiko

12/29/2019

Introduction

The dataset was obtained from Kaggle(https://www.kaggle.com/pitasr/falldata). As detailed on the website, this dataset was generated by wearable motion sensor units fit to the subjects' body at six different positions. Each unit comprises of three tri-axial devices (accelerometer, gyroscope, and magnetometer/compass). Fourteen volunteers performed a standardized set of movements including 20 voluntary falls and 16 activities of daily living (ADLs), resulting in a large dataset with 16382 trials. The dataset comprises of 7 variables, namely; ACTIVITY,TIME, SL, EEG, BP, HR and CIRCULATION. Find details on each column.

ACTIVITY - activity classification TIME - monitoring time SL - sugar level EEG - EEG monitoring rate BP - Blood pressure HR - Heart beat rate CIRCLUATION - Blood circulation

The aim is to build a model that detects falls for people in the fall risk groups. With this dataset, i have built a model using 6 predictors to differentiate 6 human movements(captured under the target label variable, ACTIVITY) of Standing, Walking, Sitting, Falling, Cramps and Running that are represented by values of 0,1,2,3,4,5 repectively.

As indicated in the method section below, the data has first been explored using the different technicques that have guided on the machine learning approaches to deploy. Three machine learning algorithms have been considered and the final testing coducted with the best performing algorithm, random forest.

Method

Using different exploratory techniques detailed below, data is observed to be in tidy format with no null values. However, data has been observed to be of varying scales and has therfore been scaled. Further more, predictors are not correlated to the target label. It is also important to note that the variables are generally non-uniformly distributed. Given this nature of data, svm,knn and random forest machine learning approaches have been deployed

Data Overview

Dimensions of the dataset

```
## [1] 16382     7
```

Column names of the dataset

```
## [1] "ACTIVITY"    "TIME"        "SL"          "EEG"         "BP"
## [6] "HR"          "CIRCLUATION"
```

Data Types of the dataset

```
##    ACTIVITY        TIME          SL         EEG          BP          HR
##    "factor"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric"
## CIRCLUATION
##   "numeric"
```

Layout of the dataset

```
##   ACTIVITY    TIME       SL      EEG BP  HR CIRCLUATION
## 1        3 4722.92  4019.64 -1600.00 13  79         317
## 2        2 4059.12  2191.03 -1146.08 20  54         165
## 3        2 4773.56  2787.99 -1263.38 46  67         224
## 4        4 8271.27  9545.98 -2848.93 26 138         554
## 5        4 7102.16 14148.80 -2381.15 85 120         809
## 6        5 7015.24  7336.79 -1699.80 22  95         427
```

Checking for null values

```
any(is.na(FallData))
```

```
## [1] FALSE
```

Summary of the dataset

```
##  ACTIVITY      TIME             SL                   EEG
##  0:4608   Min.   : 1954   Min.   :     42.2   Min.   :-12626000
##  1: 502   1st Qu.: 7264   1st Qu.:   9941.2   1st Qu.:    -5630
##  2:2502   Median : 9769   Median :  31189.2   Median :    -3361
##  3:3588   Mean   :10937   Mean   :  75272.0   Mean   :    -5621
##  4:3494   3rd Qu.:13482   3rd Qu.:  80761.4   3rd Qu.:    -2150
##  5:1688   Max.   :50896   Max.   :2426140.0   Max.   :  1410000
##        BP               HR          CIRCLUATION
##  Min.   :  0.00   Min.   : 33.0   Min.   :    5
##  1st Qu.: 25.00   1st Qu.:119.0   1st Qu.:  587
##  Median : 44.00   Median :180.0   Median : 1581
##  Mean   : 58.25   Mean   :211.5   Mean   : 2894
##  3rd Qu.: 78.00   3rd Qu.:271.0   3rd Qu.: 3539
##  Max.   :533.00   Max.   :986.0   Max.   :52210
```

Seeing the different variables are of have varying scales(from above), all predictors have been scaled but not the target label variable, ACTIVITY. We also observe that the target label values are imbalanced as indicated below in percentage proportions.

Summary after scaling

```
##  ACTIVITY      TIME               SL                  EEG
##  0:4608   Min.   :-1.7072   Min.   :-0.59003   Min.   :-116.61681
##  1: 502   1st Qu.:-0.6981   1st Qu.:-0.51239   1st Qu.:  -0.00008
##  2:2502   Median :-0.2219   Median :-0.34574   Median :   0.02088
##  3:3588   Mean   : 0.0000   Mean   : 0.00000   Mean   :   0.00000
##  4:3494   3rd Qu.: 0.4837   3rd Qu.: 0.04305   3rd Qu.:   0.03207
##  5:1688   Max.   : 7.5946   Max.   :18.43786   Max.   :  13.08084
##        BP                HR            CIRCLUATION
##  Min.   :-1.2062   Min.   :-1.3739   Min.   :-0.7552
##  1st Qu.:-0.6885   1st Qu.:-0.7121   1st Qu.:-0.6031
##  Median :-0.2951   Median :-0.2427   Median :-0.3433
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4089   3rd Qu.: 0.4576   3rd Qu.: 0.1685
##  Max.   : 9.8306   Max.   : 5.9597   Max.   :12.8899
```

Target label proportion

```
##   Count Percentage
## 0  4608      28.13
## 1   502       3.06
## 2  2502      15.27
## 3  3588      21.90
## 4  3494      21.33
## 5  1688      10.30
```

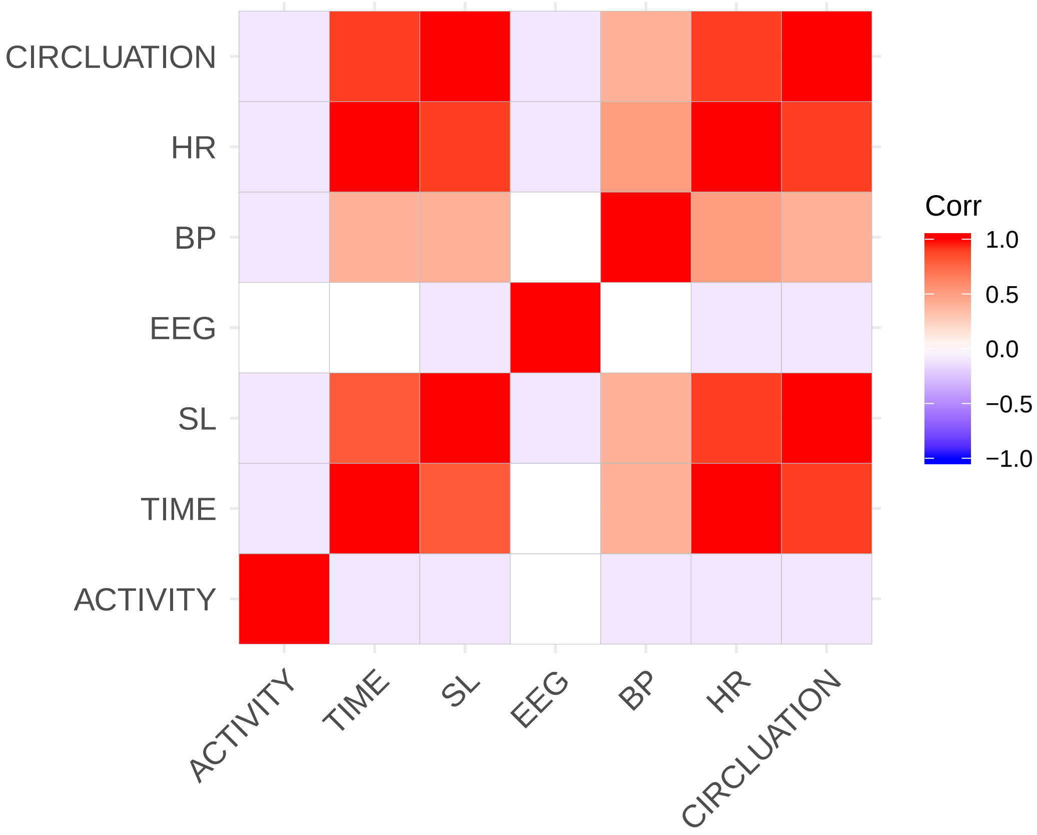
Understanding Correlation between Variables

From the correlation matrix below and the corresponding plot, data is largely not correlated. Hence, SVM,KNN and random forests approaches have been considered in modelling the data.

Correlation matrix

```
##             ACTIVITY TIME   SL  EEG   BP   HR CIRCLUATION
## ACTIVITY         1.0 -0.1 -0.1  0.0 -0.1 -0.1        -0.1
## TIME            -0.1  1.0  0.8  0.0  0.4  1.0         0.9
## SL              -0.1  0.8  1.0 -0.1  0.4  0.9         1.0
## EEG              0.0  0.0 -0.1  1.0  0.0 -0.1        -0.1
## BP              -0.1  0.4  0.4  0.0  1.0  0.5         0.4
## HR              -0.1  1.0  0.9 -0.1  0.5  1.0         0.9
## CIRCLUATION     -0.1  0.9  1.0 -0.1  0.4  0.9         1.0
```

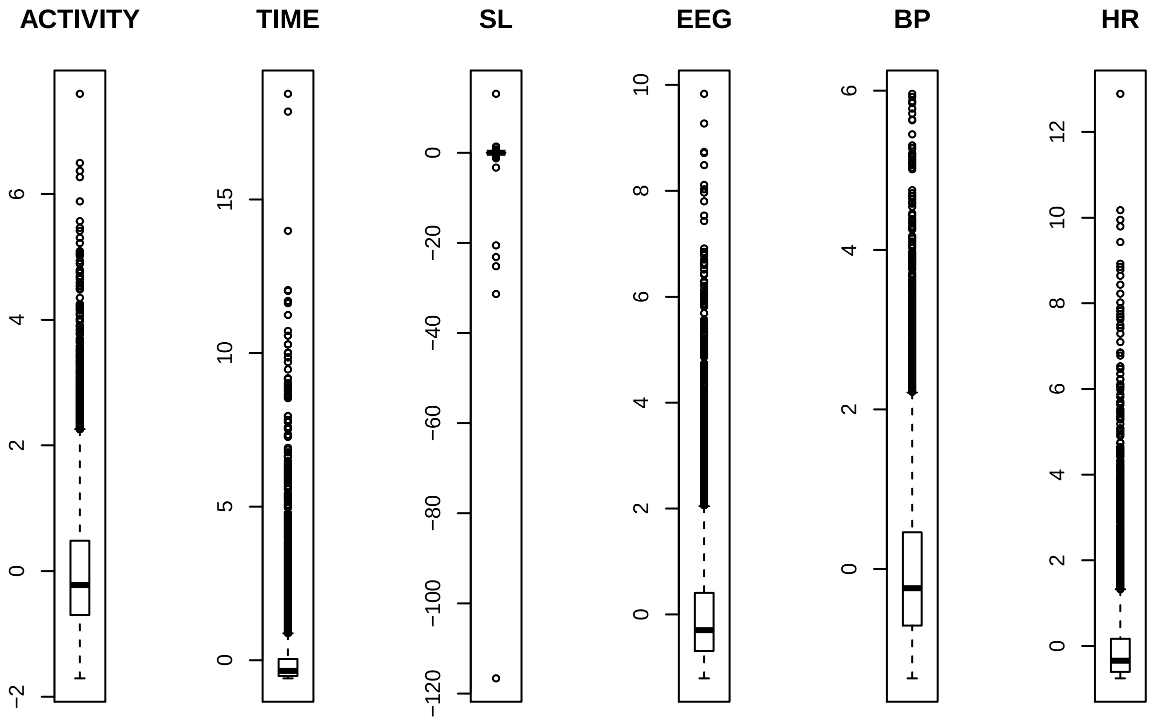
A plot of the correlation matrix

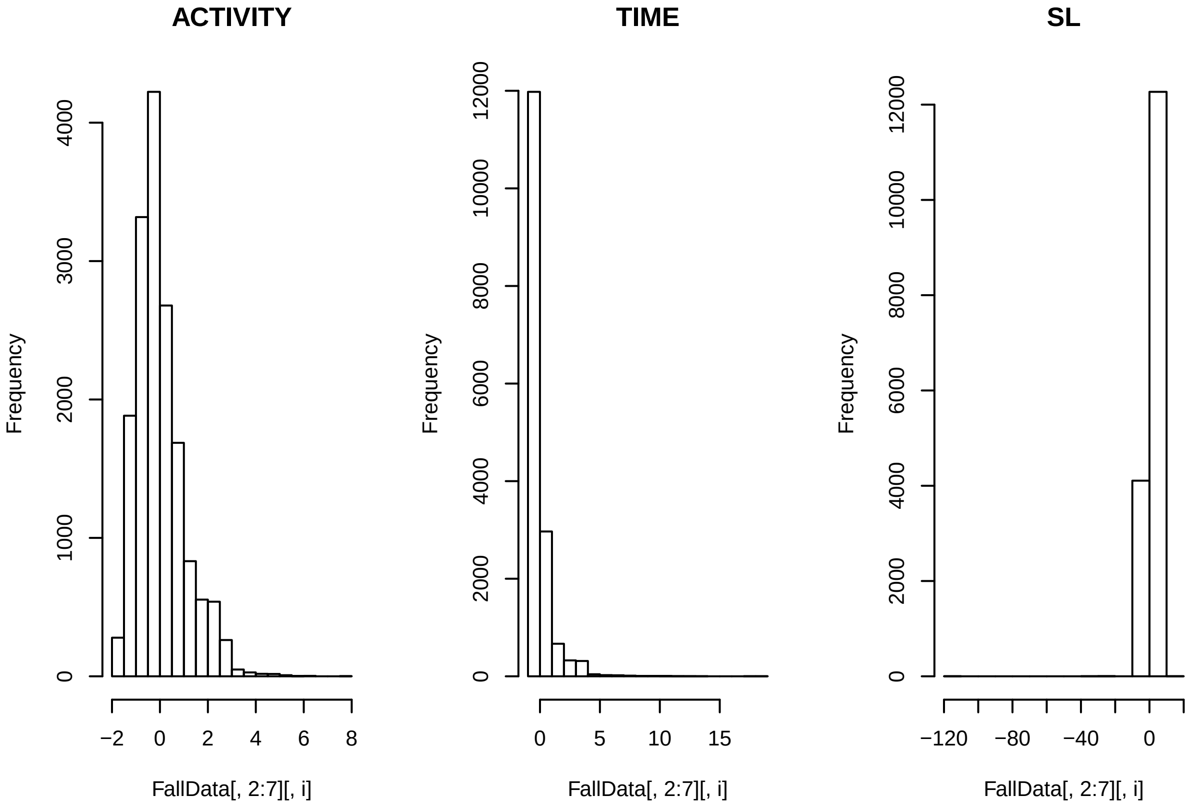


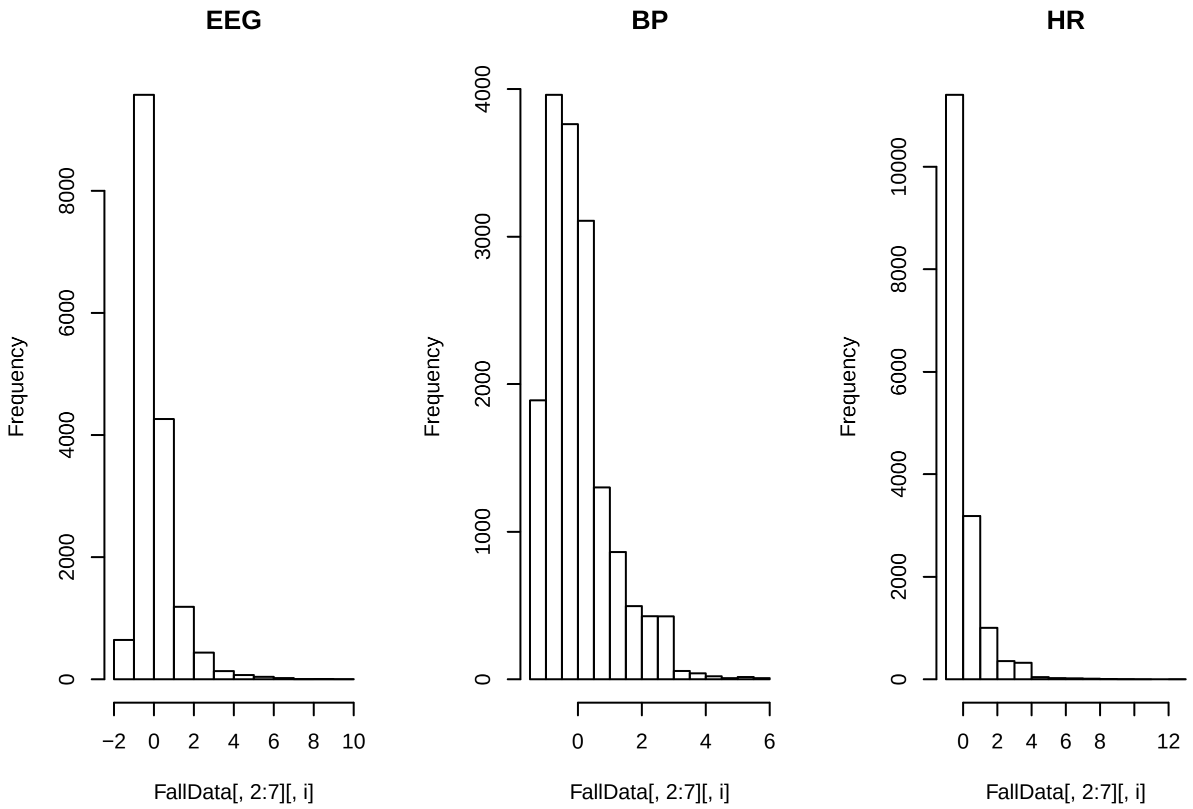
Distribution of Data

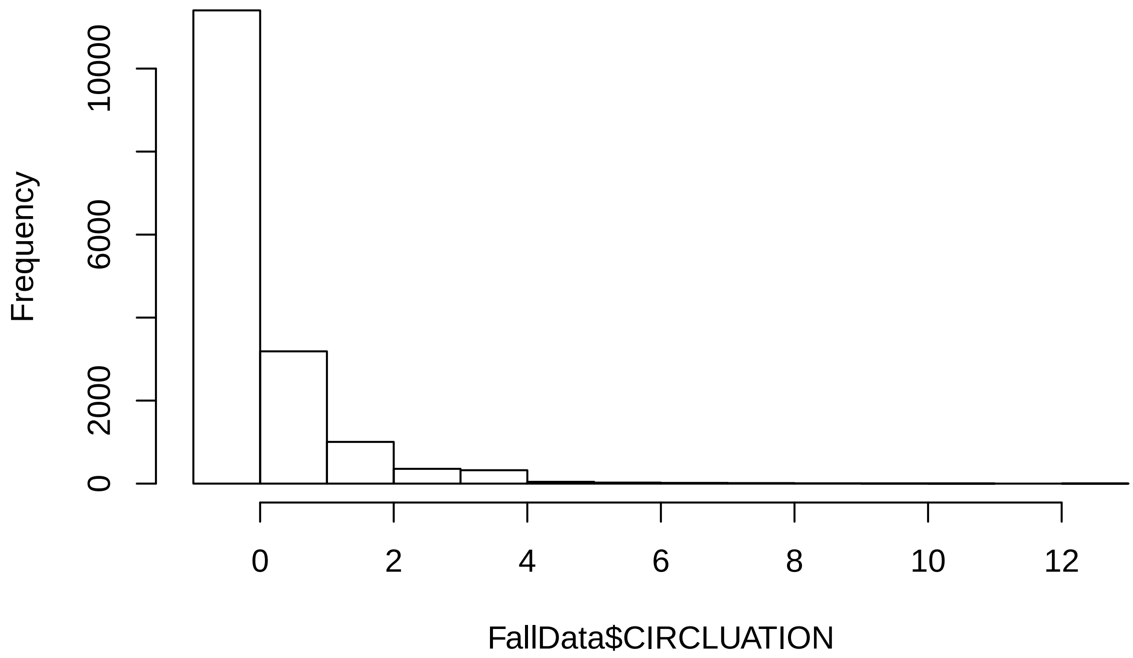
Exploring distribution of data by boxplot and histograms for each variable, and distribution of each variable for every target label, data is seen to be non-randomly distributed, even after scaling.

Boxplots of individuals variables

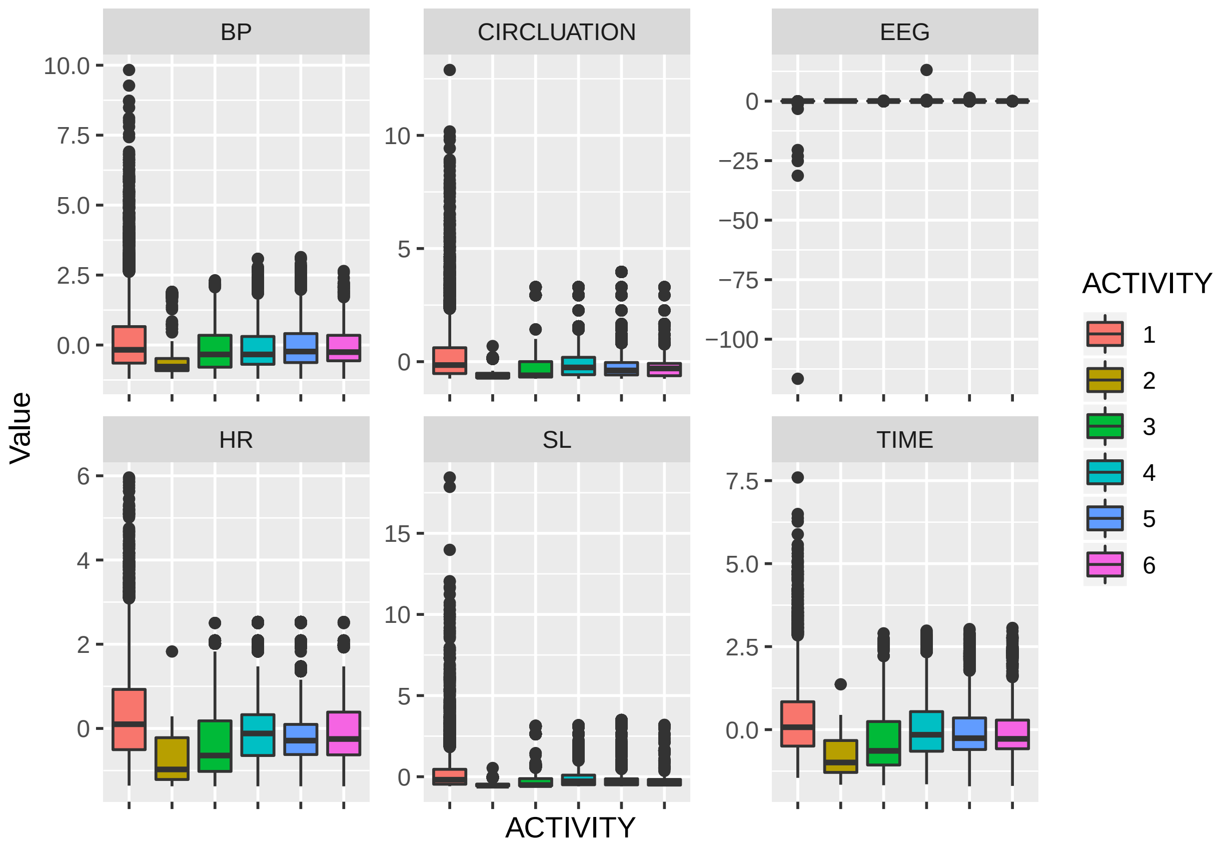



ACTIVITY
TIME
SL
Frequency
0
1000
2000
3000
4000
−2
0
2
4
6
8
FallData[, 2:7][, i]
Frequency
0
2000
4000
6000
8000
10000
12000
0
5
10
15
FallData[, 2:7][, i]
Frequency
0
2000
4000
6000
8000
10000
12000
−120
−80
−40
0
FallData[, 2:7][, i]


EEG
Frequency
0
2000
4000
6000
8000
−2
0
2
4
6
8
10
FallData[, 2:7][, i]
BP
Frequency
0
1000
2000
3000
4000
0
2
4
6
FallData[, 2:7][, i]
HR
Frequency
0
2000
4000
6000
8000
10000
0
2
4
6
8
12
FallData[, 2:7][, i]



Distribution of target variables against each predictor



Modelling methodology

Given the correlation and distribution of data, Support Vector Machines(SVM),knn and random forest machine learnining algorithms have been considered.The dataset is first split into main and validation sets in ratios of 8:2 respectively. While the validation set has been set aside for final tests, the main set has been split further into train and test sets in proportions of 50% each for intermidiate training and testing to determine the best performing approach.

The three machine learning algorithms have been deployed to train(with train set) and test(with test set) to determine the best performing algorithm. The best performing algorithm is the random forest approach that has been trained with the main set and tested with the validation set to obtain an accuracy of 77%. Accuracy is the preffered preformance metric as the model is expected to distinguish the different human movements.

Results

The table below shows results by the different machine learning approaches. Training and testing is first done on train set and test set respectively to determine the best performing algorithm. Lastly, since random forest is the best performing algorithm, it has been trained(with the main set) and tested(with the validation set) to obtain an accuracy of 77%.

```
##                                              Method  Accuracy
## 1                                  Training with Svm 0.4030215
## 2                                  Training with knn 0.5951473
## 3                      Training with random forest 0.7437815
## 4 Final test on Random forest with validation set 0.7792010
```

Determining imporatance of predictors

The importance of predictors is detailed in the table below.

```
##             MeanDecreaseGini
## TIME                1461.362
## SL                  2079.762
## EEG                 1873.739
## BP                  1465.218
## HR                  1567.558
## CIRCLUATION         1517.094
```

Conclusion

A fall detection system to detect falls from five other human movements has been built using a random forest machine learning approach with an overall accuracy of 77%. While the accuracy is still wanting, performance of the model can improved with more data as already observed in the difference of accuracy when training with less and much more data. Performance of the model might also be improved if trained with real-world scenarios data. Future works will look at feature engineering, ensembling and PCA in a bid to improve performance.