2/14/2020

PROJECT REPORT

HUMAN ACTIVITY RECOGNITION

EED363
Applied Machine Learning

YATHARTH JAIN (1710110400)
YASHVARDHAN SINGH (1710110398)
ECE, SHIV NADAR UNIVERSITY

ACKNOWLEDGEMENT

We would like to express our deepest appreciation and gratitude to all those who provided us the opportunity to undertake this project. A special thanks is in order to our course instructor [Dr. Madan Gopal] and Teaching Assistant [Mr.Venugopala Kotha].

We would also like to thank Mr. Abhishek Aggarwal and Mr. Mubin Tirsaiwala for providing their invaluable guidance, comments and suggestions throughout the course of project

Contents

INTRODUCTION

Human Activity Recognition (HAR) refers to the problem of automatically identifying human activities by learning from data collected either from the individual or from the environment surrounding that individual. The form of the data might be inertial measurements, physiological signals from devices like wearable sensors or those such as image, audio and video data collected from environmental devices/sensors. HAR has been a well-studied problem for its potential capability to mine human behavioral patterns and for monitoring human activity for security and health purposes.

One real-life scenario where human activity recognition would help is monitoring heart rate in patients of chronic heart disease as they perform different activities on a daily basis. The heart rate of a patient might need to be monitored to observe how his heart rate changes with different activities. An automatic activity recognition solution would be of value in a scenario of this nature.

In this project, we focus on wearable sensor based single user activity recognition and the application of machine learning to this problem.

PROBLEM FORMULATION

The human activity recognition problem can be formalized as follows: Given a set $W = \{W_1, \ldots, W_m\}$ of m equally sized windows, a set of time series measured from k sensors, $S = \{S_i, \ldots, S_k\}$ where each S_i expands over severalwindows and a set of target activity class labels $A = \{A_1, \ldots, A_n\}$.

The task is to find a mapping function $f: W \rightarrow A$ such that $f(W_i)$ is as close as possible to the true activity performed during window W_i. Specifically, each window W_i consists of k time series chunks $S_{i,1}, \ldots, S_{i,k}$ that will be used as features to predict that window's activity class label.

The aim of our project is to use the dataset for classification of ambulatory activities (e.g., walking, sitting) using classifiers that have not so far been applied on this dataset and present a comparison across their performances.

In general, HAR is an extremely challenging problem to generalize across datasets due to the variations in data collection setting, properties of individuals and data collection environment. Due to this fact, each dataset poses a different HAR problem.

<u>**DATASET**</u>

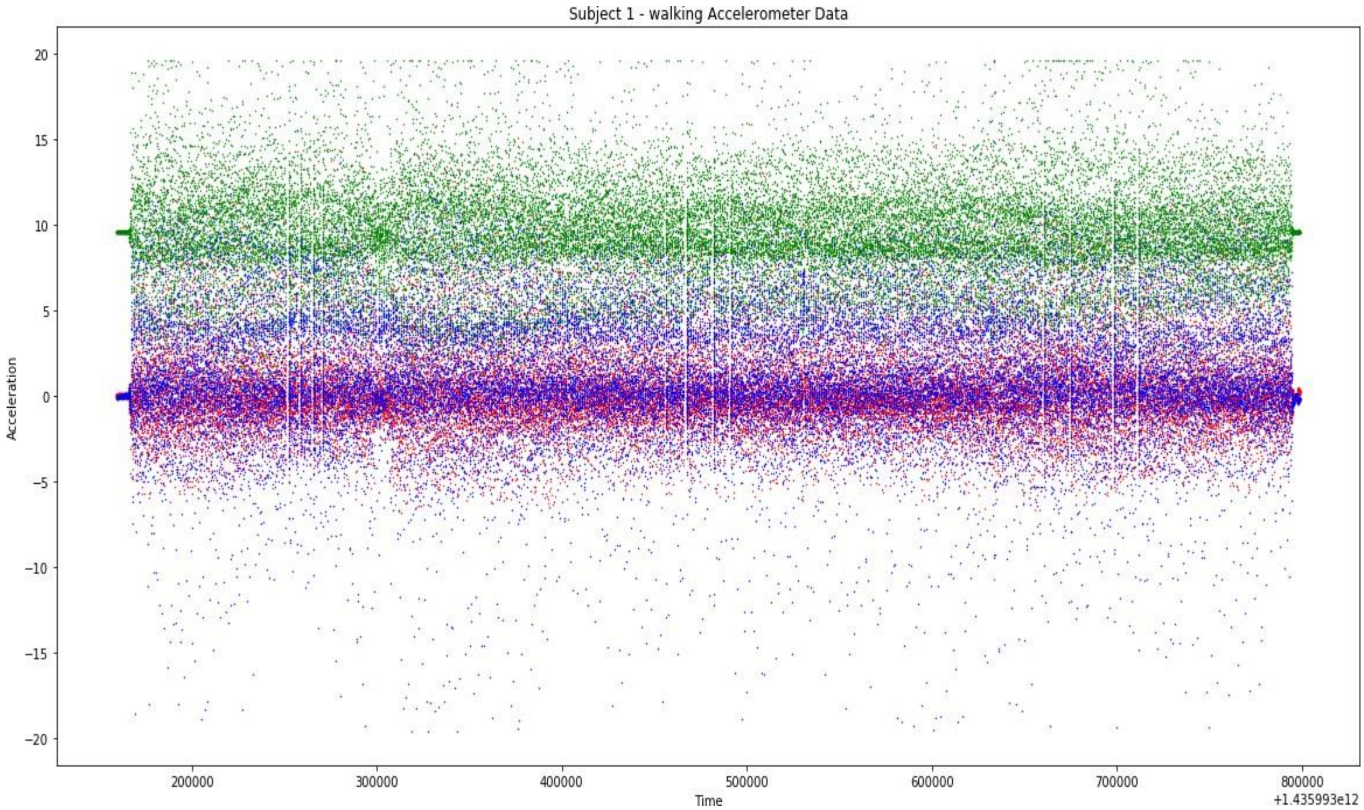
There are many publicly available datasets of human activity data that could be used for model development and testing. We used the *Real World Dataset* created by the *University of Mannheim — Research Group Data and WebScience.*

This dataset includes **8 activity classes** (*climbing down stairs, climbing upstairs, jumping, lying, standing, sitting, running/jogging, and walking*). The researchers outfitted <u>15 subjects</u> with smartphones or smartwatches in <u>7 body positions</u> (*chest, forearm, head, shin, thigh, upper arm, and waist*). We decided to use only the data from the **"thigh" (pocket)** position under the assumption that it is the <u>most commonly used</u> of those tested. The dataset includes readings from the following <u>sensors</u>: *acceleration, GPS, gyroscope, light, magnetic field, and sound level*. In our experiments we only use **accelerometer and gyroscope** data. Each of these two sensors contain <u>3</u> sets of readings along the *x, y and z axis* aligned to time. The sensors were recorded at 50 Hz (50 observations persecond).

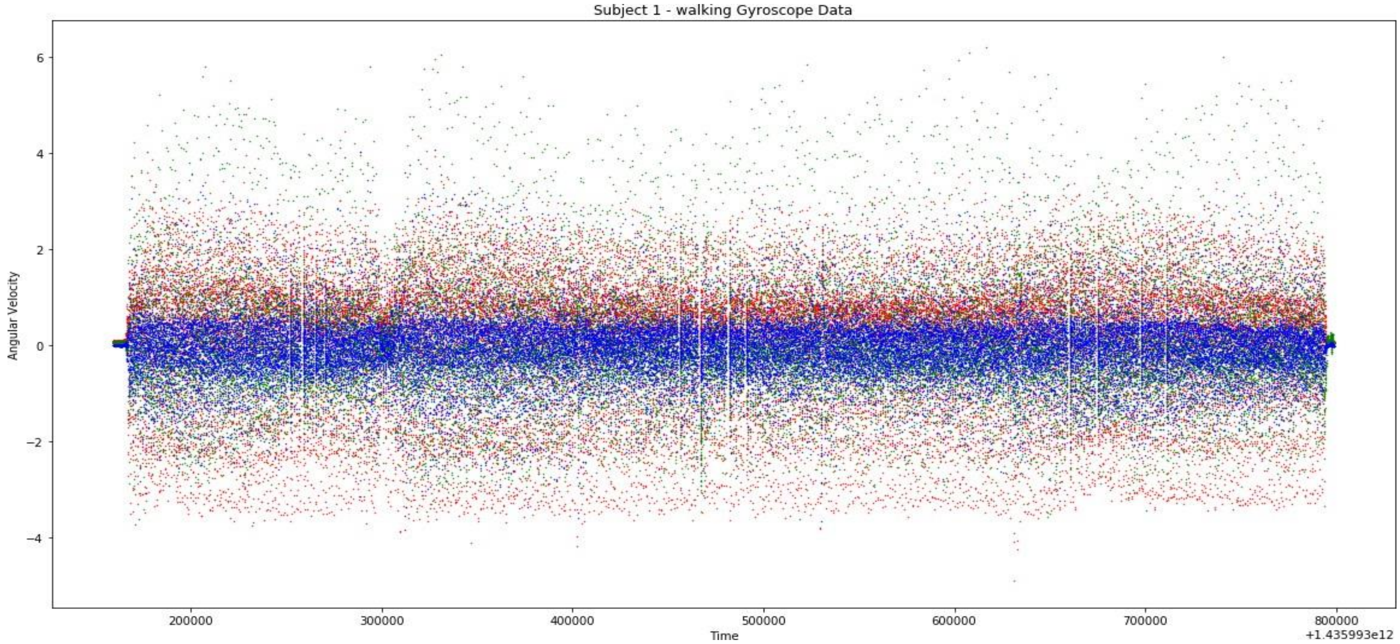
Out of the 15 subjects we are using the data for **6 subjects** only. Data for first **4 subjects (2 Males and 2 Females)** will be used for **training** of the model while the data of remaining **2 subjects (1 Male and 1 Female)** will be used for **testing** of the fitted models.

ANALYSIS OF DATA

Initially, we created scatter-plots of the accelerometer and gyroscope data for each of the 6 subjects performing each of the 8 activities with the x, y, and z dimensions shown in red, green and blue. This was done as part of data preprocessing. For example, this is a plot of the *accelerometer* data for subject 1 who was walking:

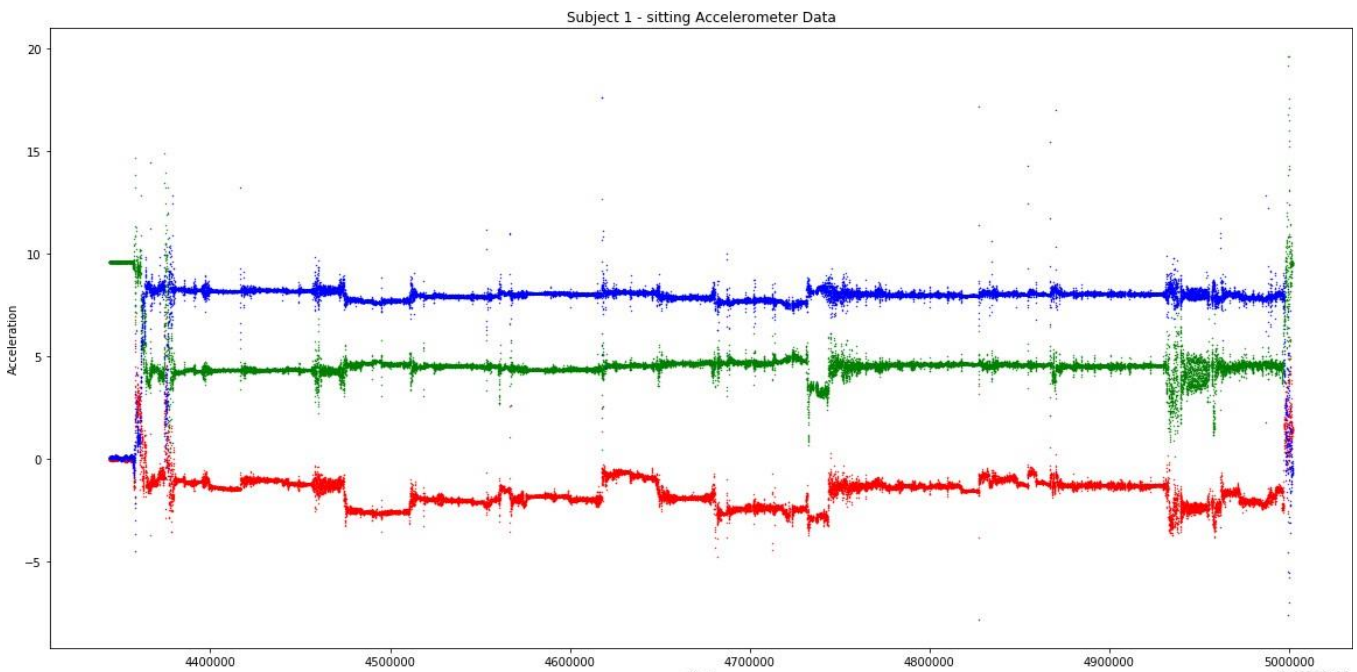


This is a plot of the *gyroscope* data for subject 1 who was walking:

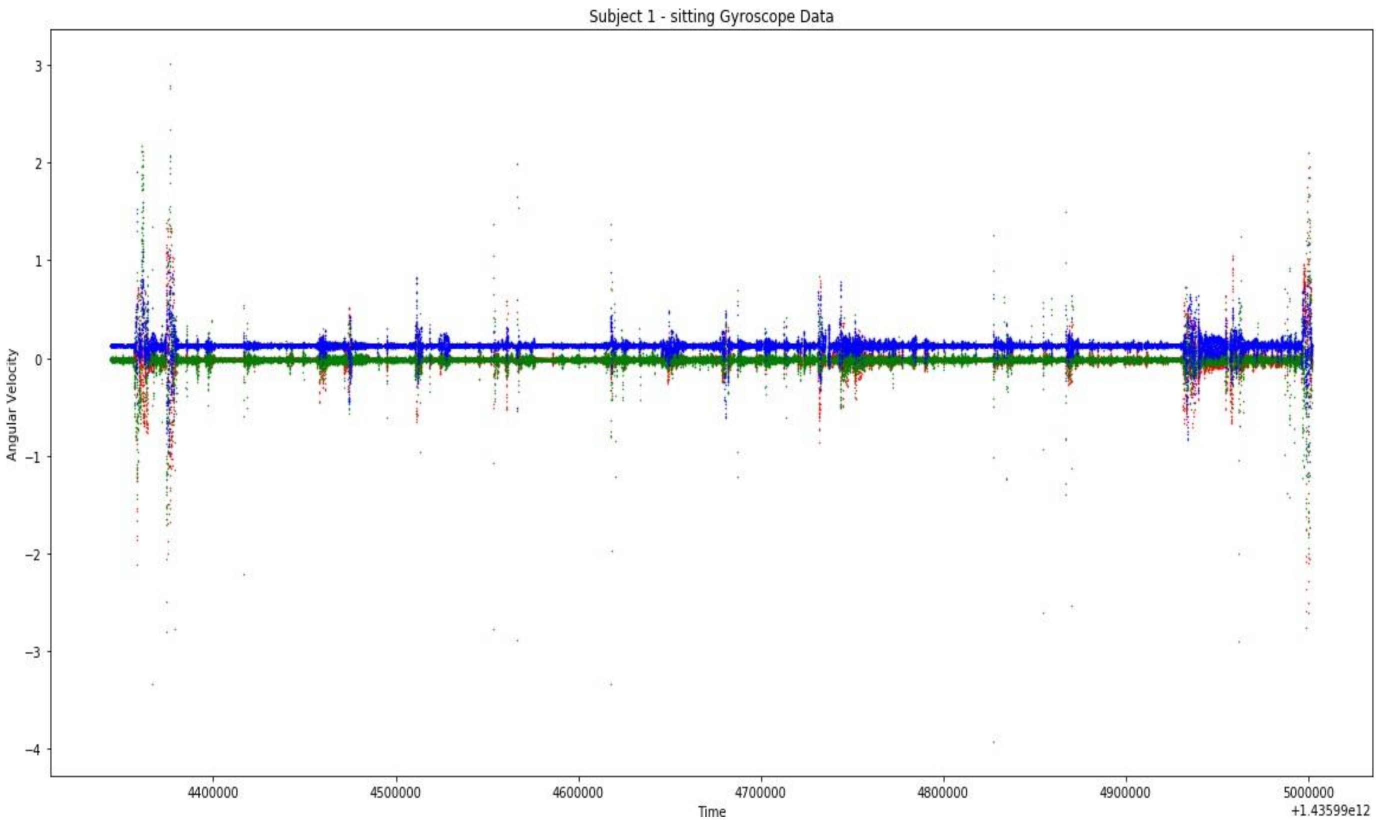


A plot of the same subject sitting looks very different:

Accelerometer:

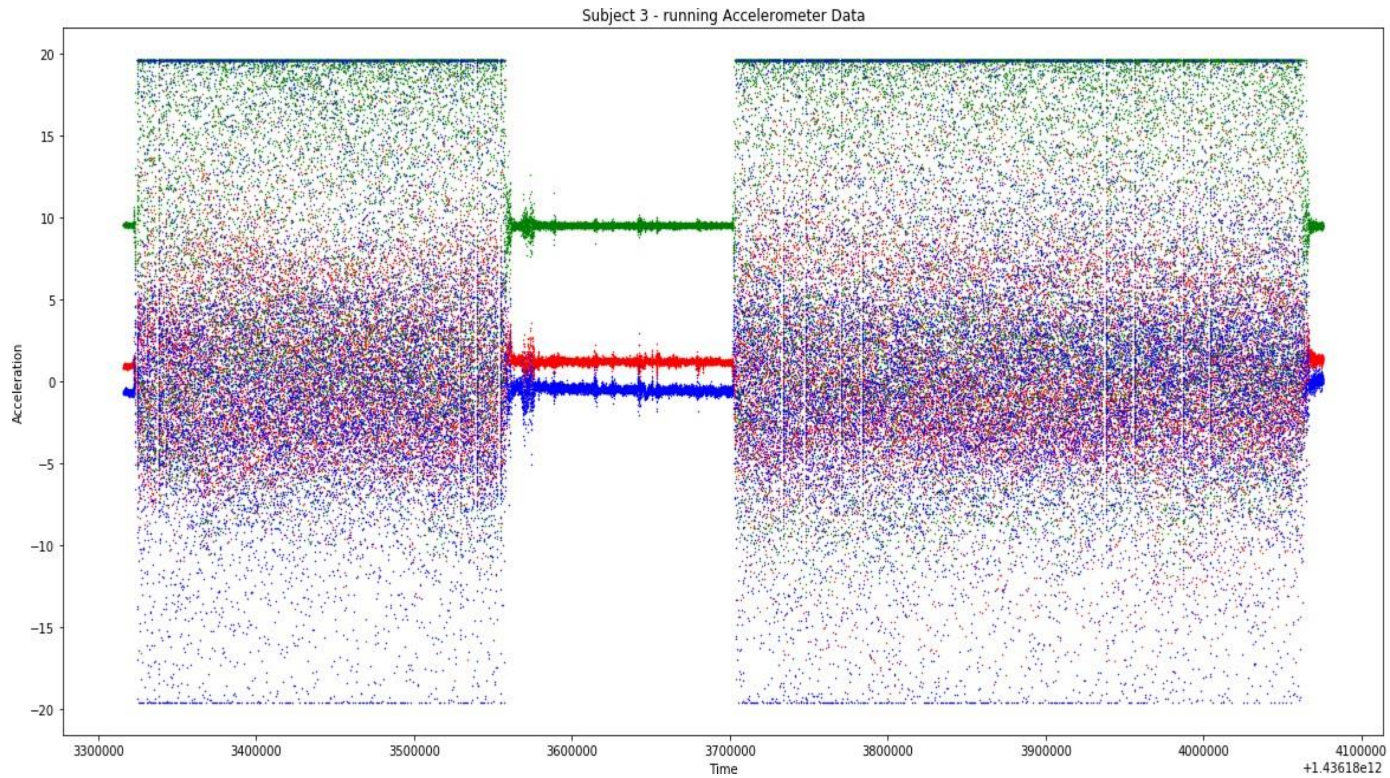


Gyroscope:

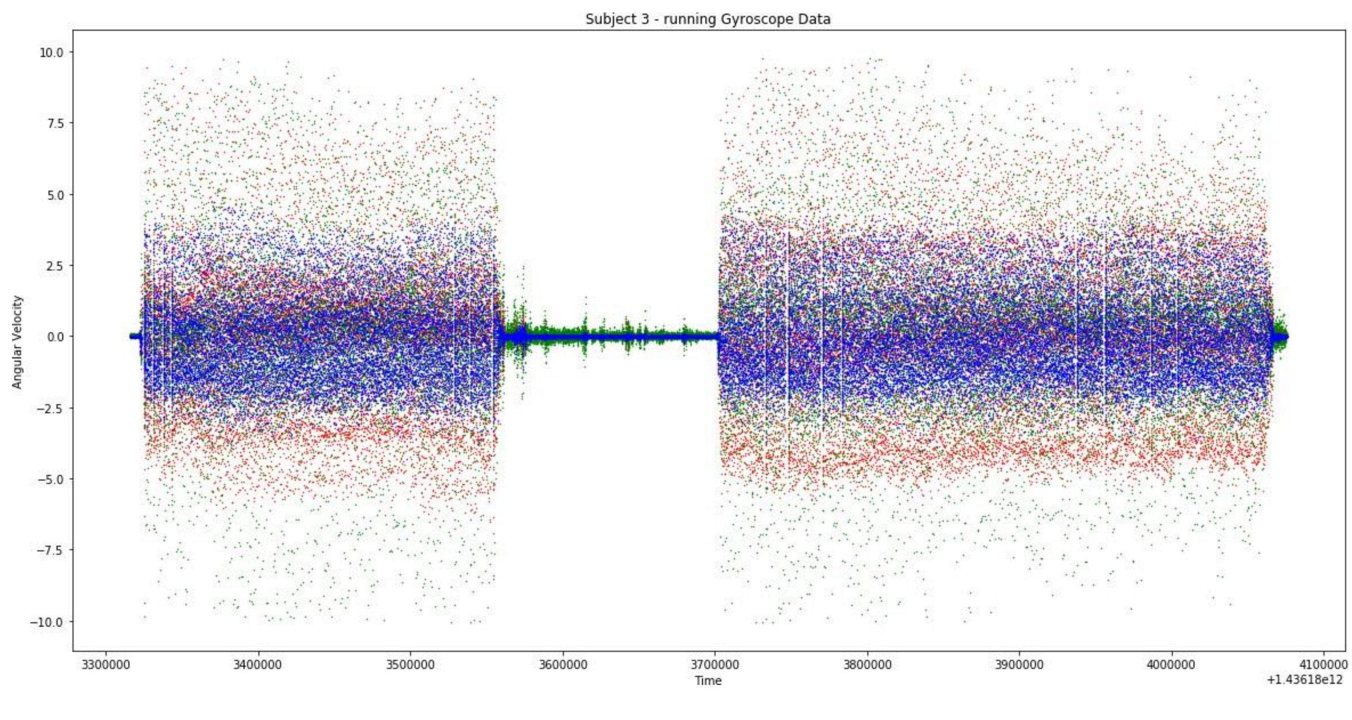


As can be seen from following plots, the subjects were asked to wait a short time before beginning the assigned activity. Thus, the labels for the beginning and end of each activity period are frequently incorrect. In addition, it is clear from inspecting the plots that not all subjects performed the assigned activity continuously. For example, *Subject 3* stood for short period between bouts of running, as this can be seen from the following:

Accelerometer:



Gyroscope:



DATA CONSOLIDATION

Till now we have the data in a very scattered form that is the data is distributed among the subjects, sensors and the activities as well. For fitting a model what we need is a combined dataset for training purpose and a similar dataset for testing purpose. Hence we decided to merge these separate datasets with features being attribute time, coordinates for x, y and z axes for both the sensors and for first 4 Subjects to form training dataset and similar merging of Subject_5 and Subject_6 resulted in the formation of testing dataset.

We have taken a window size of 100 (2 seconds) for sampling the data. Separate columns for describing the activity, label, Subject number and sample number were also formed.

ENCODING SCHEME:

Label 0: Climbing Down

Label 1: Climbing Up

Label 2: Jumping

Label 3: Lying

Label 4: Running

Label 5: Sitting

Label 6: Standing

Label 7: Walking

There are total **9029** samples ranging from **0 to 9028** in the *Training dataset* while there are total **4315** samples ranging from **9029-13343** in the *Testing dataset*. This makes **32.33%** of the total data as testing data.

The subject column determines the subject from whom a particular datapoint is taken from.

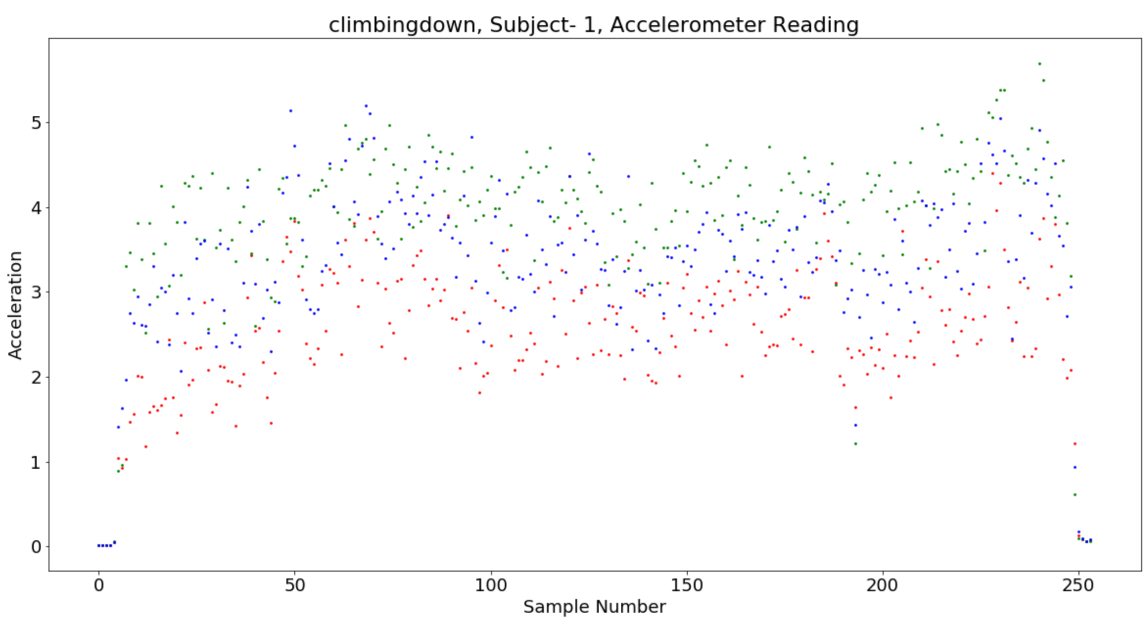
Hence, using all this 2 consolidated data files are formed one for training purpose and another one for testing purpose.

DATA CORRECTION

As we have seen from earlier graphs, the subjects did not perform the assigned activity continuously, rather it was performed in bouts with some rest period in between and also the activity was not started immediately neither it lasted for exact 10 minutes. Hence, it is necessary to filter the data to remove samples that would be otherwise mislabeled.

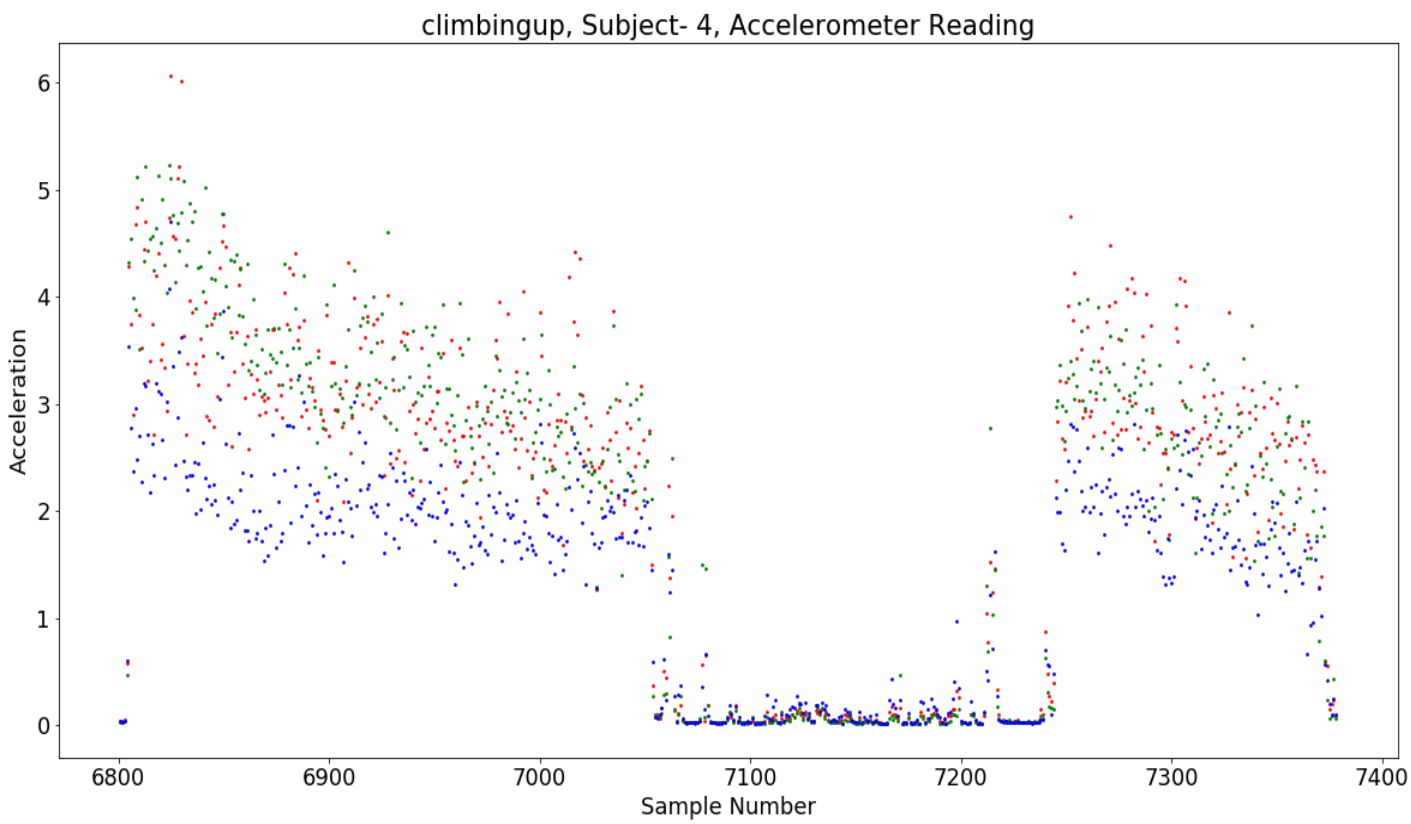
The data was sampled into 2 second (100 observations) windows for this very purpose. We calculated the standard deviation of each sample for accelerometer reading and plotted them against the samples numbers for each activity for each individual. Some example plots are shown:

- *Climbing down for subject 1:*



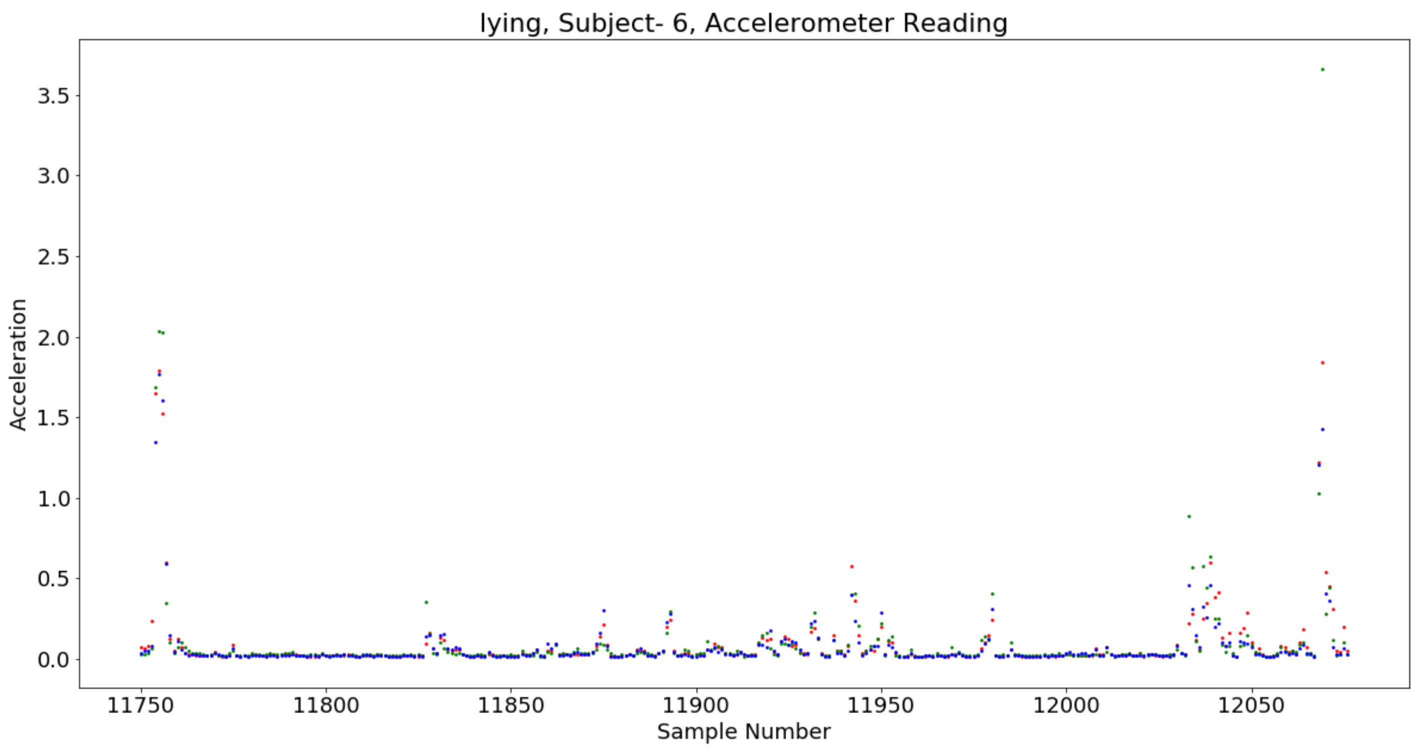
It can be clearly seen that though the subject continuously performed the climbing down activity once he started but there were mislabels in the starting and at the end as well. So, we needed to remove these samples to get correct data.

- *Climbing up for subject 4:*



It is visible from the graph that the activity was performed in bouts with periods of rest in between. Hence, these mislabeled samples were needed to be removed as well.

Lying for subject 6:



Lying, one of the easiest activities to perform also had some mislabels. So, this also needed to be corrected.

Running for Subject 3:

It is visible from the graph that the activity was performed in bouts with periods of rest in between. Hence, these mislabeled samples needed to be removed as well.

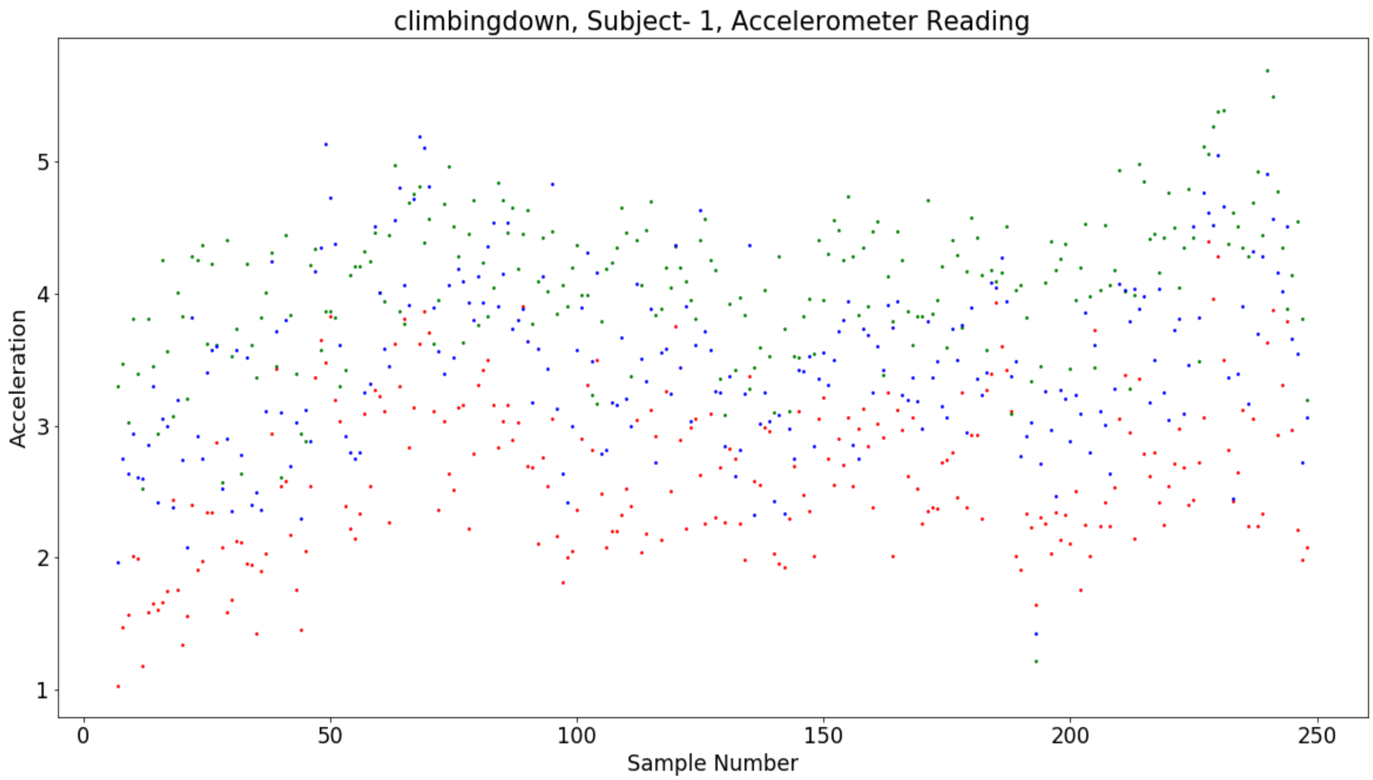
By inspecting similar plots for all subjects for each activity some thresholds for the standard deviation of ‘y’ accelerometer readings were decided as cutoff. Thresholds were established for all the activities: maxima for stationary activities (sitting, lying) and minima for moving activities. No filtering was done for standing, since the subjects did not transition to or from standing during the recording periods.

<u>List of thresholds:</u> Based on an inspection of the standard deviation plots, the following thresholds were picked for removing the samples based on the sample standard deviation of the y accelerometer readings:

- *Climbing Down*: **y<1.2**
- *Climbing up:***y<1.2**
- *Jumping:***y<3.5**
- *Lying:***y>0.5**
- *Running:* **y<5**
- *Sitting:***y>1**
- *Standing:***NONE**
- *Walking:***y<1.5**

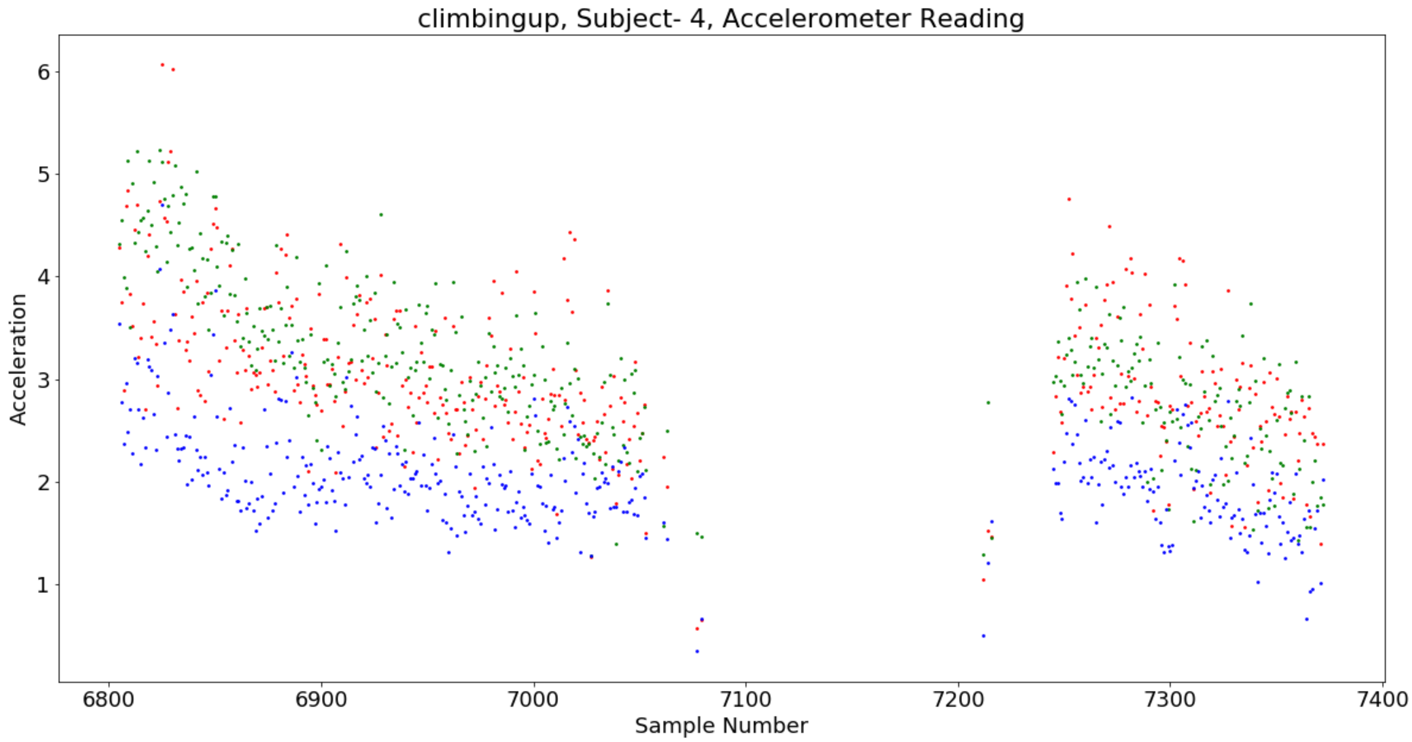
Same graphs after removing the samples using the above thresholds looked like this:

- *Climbing down for subject 1:*



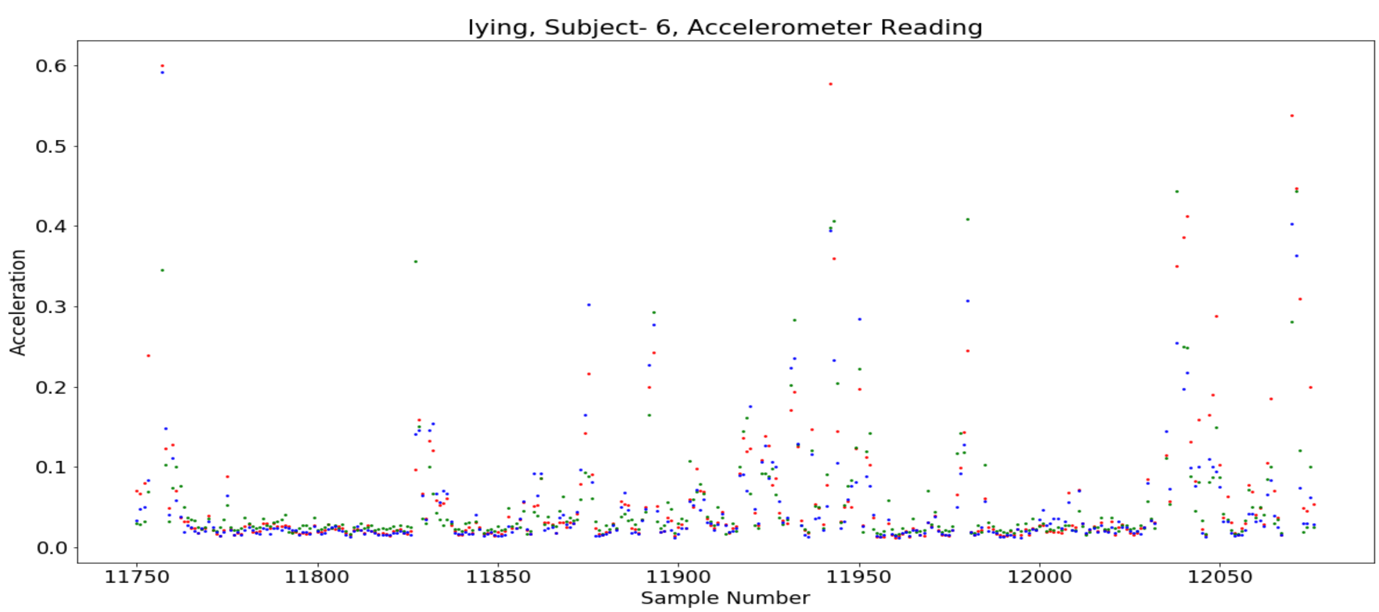
It can be clearly seen here that the mislabeled samples that were present at the beginning and end have been removed.

Climbing up for subject 4:



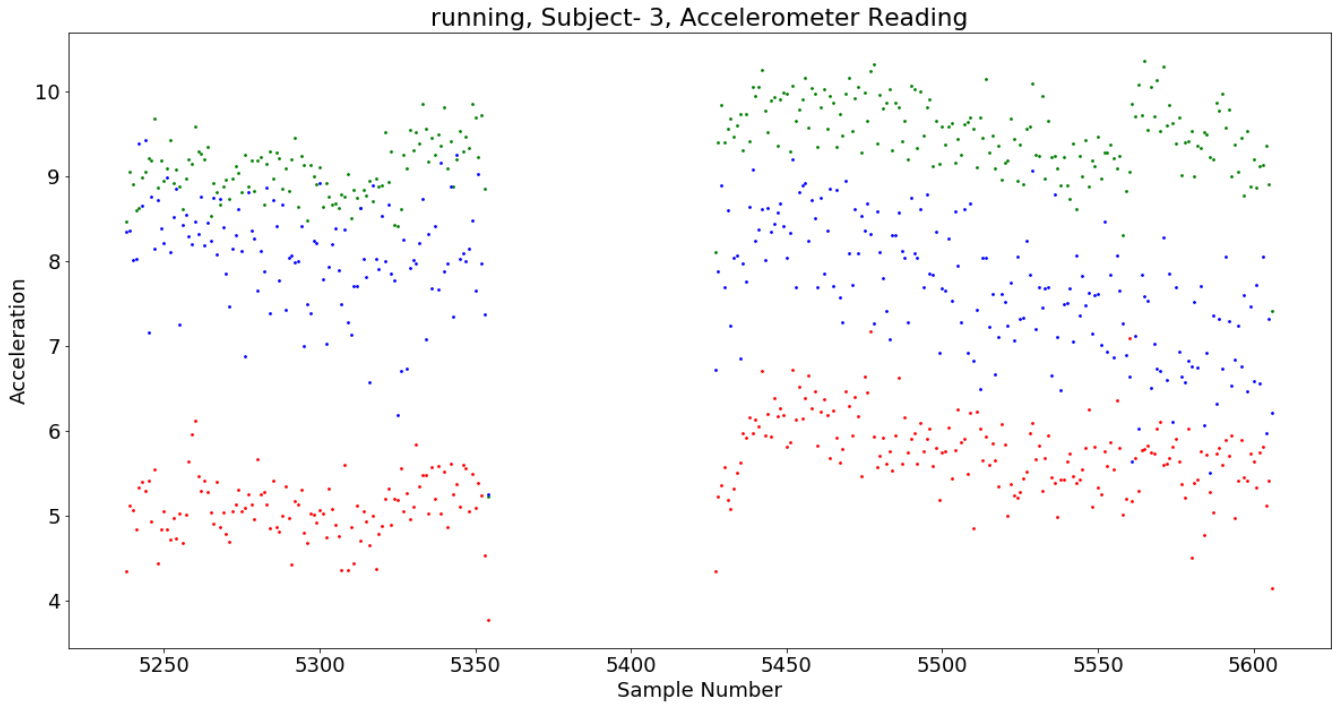
The samples at the beginning, in middle, and at the end as well which were mislabeled are removed.

Lying for subject 6:



This graph may look similar to the one that was plotted before removing the samples, but once we closely observe the y axis of the graph the difference which was made by removing the samples is clearly seen.

Running for Subject 3:



The samples at the beginning, in middle and at the end as well which were mislabeled are removed.

All this data for all the 6 subjects was combined into a single database & stored in a separate .csv file named clean_data.csv.

<u>MACHINE LEARNING ALGORITHMS</u>

<u>Extracted Features</u>: We created a dataset where samples of activities conducted by a subject were grouped into windows of 100 observations (representing 2 seconds of measurement). We calculated the mean, standard deviation, and range of the triaxial measurements (x, y, and z dimensions) for both the accelerometer and the gyroscope. These **18 (3 x 3 x 2) features** were used in various machine learning models to predict the activity performed.

<u>Training, Validation and Test sets:</u> Data from the first **4 subjects** was used to develop and **train** the model. Data from the remaining **2 subjects** was held out for **final testing**. The training data was divided into train and validation samples. The **training samples** had the data from **first 3 subjects** while the **validation sample** had the data from **4th subject**.We found that it was necessary to use a subject-based split of the training, validation and test data. A random split means that there are observations from all 6 subjects in the training data. Although the resulting test scores are high, the model performs poorly on data from subjects not seen before, since there are significant differences between subjects. Random sampling for time series data doesn't make much sense too. Since our use case involves the ability to detect performed activity from any

(previously unseen) user, it is essential to ensure that the model is generalizable to new subjects.

<u>Machine Learning algorithms used and their respective results:</u>

We decided to train and compare the performance of our problem using three machine learning algorithms namely **LOGISTIC REGRESSION**,

K- NEAREST NEIGHBOUR CLASSIFIER and **SUPPORT VECTOR MACHINE.**

TRAINING:

We trained our model using each one of these and compared their training results. The model which was performing best on training samples was selected for final testing.

Detailed analysis:

- <u>LOGISTIC REGRESSION:</u>

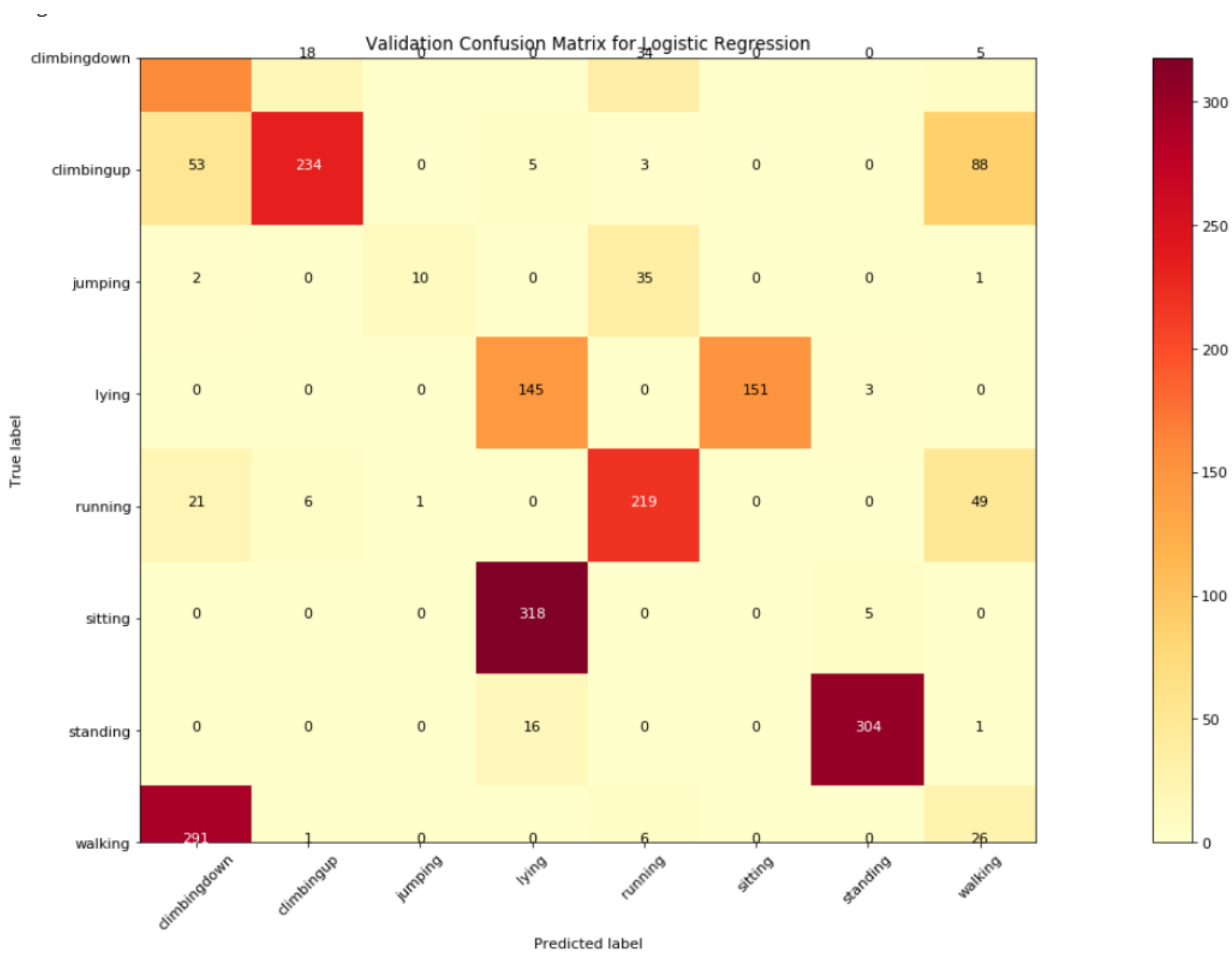
  Training F1 score is<u>: 0.955</u>

  Validation F1 score is: <u>0.459</u>

  Training Accuracy is: <u>0.953</u>

  Validation Accuracy is: <u>0.497</u>

  As we can see, the validation accuracy seem to be quite low as compared to what is required by this type of machine learning model. Confusion matrix for this is also shown:



- <u>K-NEAREST NEIGHBOR CLASSIFIER:</u>

  Training F1 score is: <u>0.981</u>

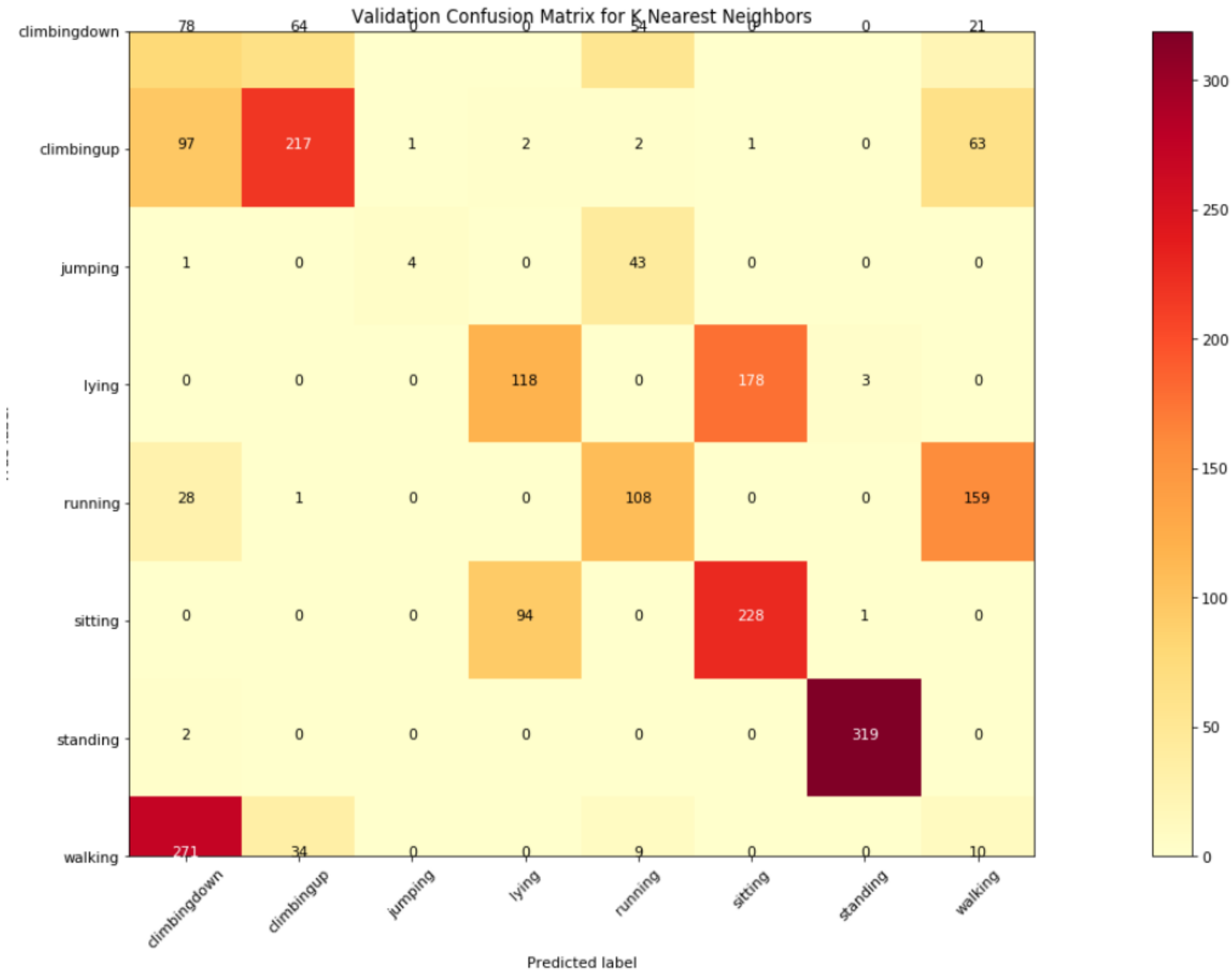
  Validation F1 score is: <u>0.441</u>

  Training Accuracy is: <u>0.981</u>

  Validation Accuracy is: <u>0.489</u>

We can see that though the model gave a high percentage of accuracy on training, even more than what logistic regression gave, but the validation accuracy seems to be quite low. The validation accuracy for KNN was less than that for Logistic Regression.

Confusion matrix for this is also shown:



- <u>SUPPORT VECTOR MACHINE:</u>

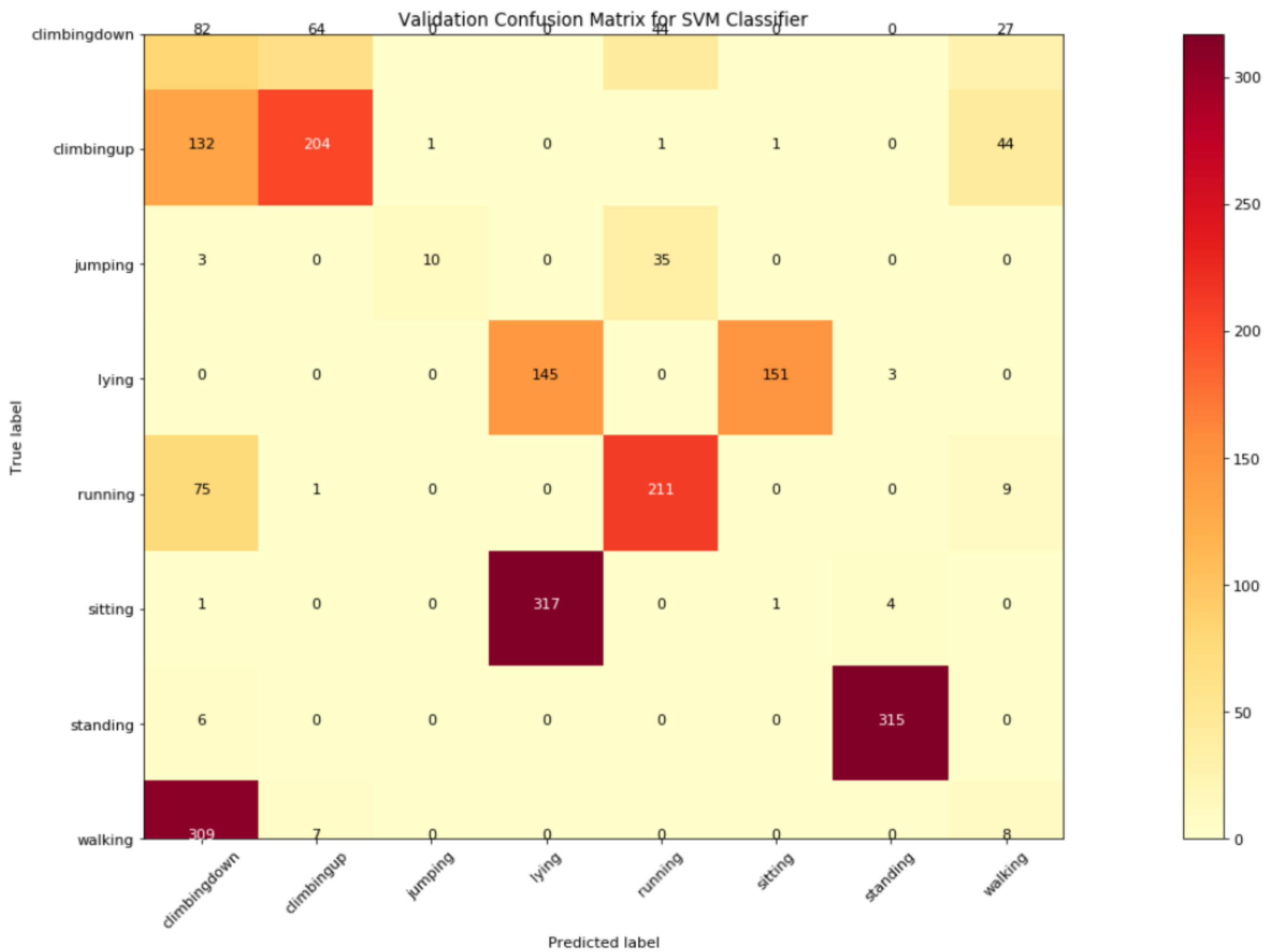
  Training F1 score is: <u>0.939</u>

  Validation F1 score is: <u>0.410</u>

  Training Accuracy is: <u>0.942</u>

  Validation Accuracy is: <u>0.441</u>

As seen in previous 2 models this one also had relatively high training accuracy with low validation accuracy. Confusion matrix for this model is shown:



After comparing the validation accuracy for all the three models logistic regression model was selected for final testing after training it for the data from all 4 training subjects simultaneously.

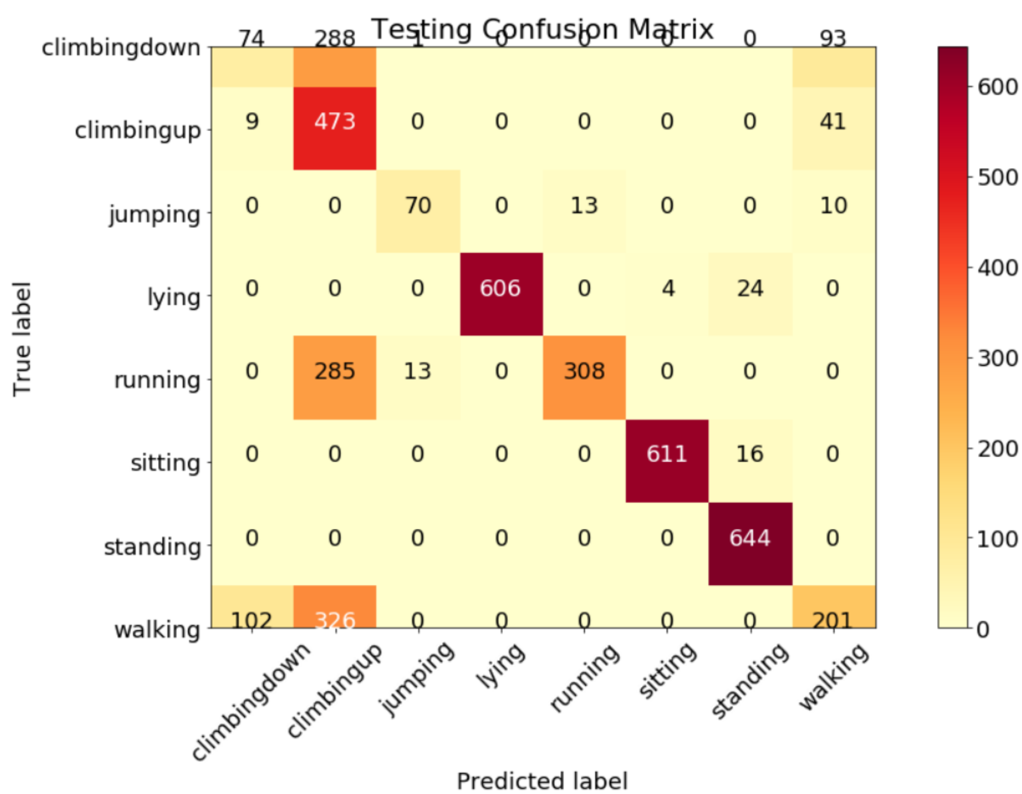
FINAL TESTING:

The results on testing data comprising data from subject 5 and subject 6 are:

```
Testing F1 score is: 0.691
Testing accuracy is: 0.709
```

	precision	recall	f1-score	support
climbingdown	0.40	0.16	0.23	456
climbingup	0.34	0.90	0.50	523
jumping	0.83	0.75	0.79	93
lying	1.00	0.96	0.98	634
running	0.96	0.51	0.66	606
sitting	0.99	0.97	0.98	627
standing	0.94	1.00	0.97	644
walking	0.58	0.32	0.41	629
accuracy			0.71	4212
macro avg	0.76	0.70	0.69	4212
weighted avg	0.77	0.71	0.70	4212

Confusion Matrix for final testing is as shown:



So, this model predicted 16% of climbing down, 90% climbing up, 75% jumping, 96% lying, 51% running, 97% sitting, 100% standing, 41% walking correctly with an **overall testing accuracy as 70.9%**. This is insufficient accuracy to be useful, since too many errors will cause users to mistrust and ignore such a deployed model. Moreover, for applications like physical activity promotion it is not necessary to know the exact type of activity. Thus, the activities were further sub-grouped into three physical activity classes and the model was then again trained using the same 3 machine learning algorithms and results were compared.

MODEL TRAINING AFTER DIVISION INTO CLASSES

The activities were sub-divided into groups as follows-

Light Activity – Sitting, Standing, Lying

Moderate Activity – Walking, Climbing upstairs, Going downstairs

Vigorous Activity – Running, Jumping

The previously mentioned models were then trained on the data after this class division had been done.

Detailed analysis:

- LOGISTIC REGRESSION:

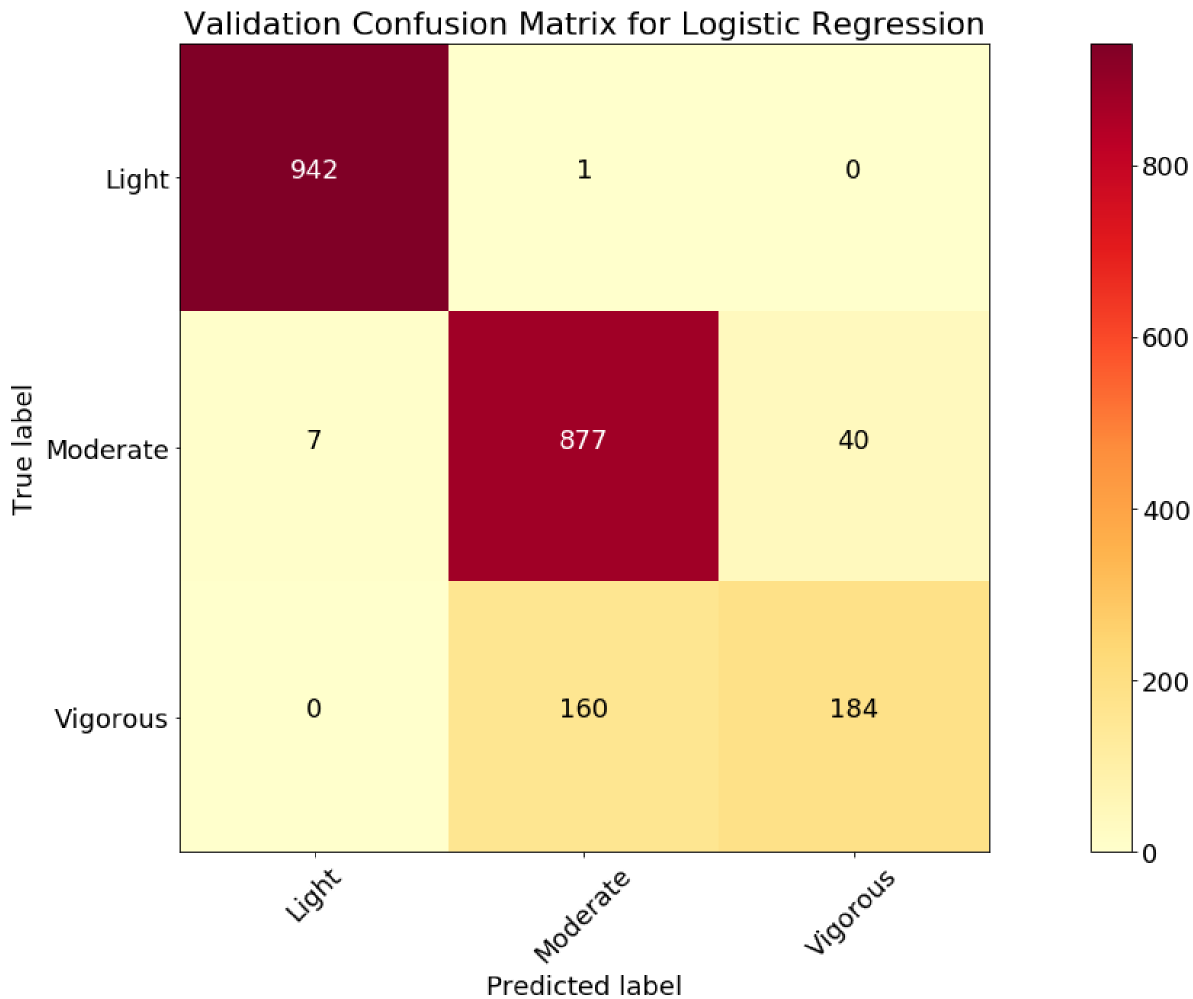
  Training F1 score is: 0.996

  Validation F1 score is: 0.846

  Training Accuracy is: 0.997

  Validation Accuracy is: 0.906

As can be seen, the validation accuracy has drastically improved, with far less misclassifications happening. All the activity classes are accurately predicted.



This trend is observed for the other models too, with validation accuracies having improved significantly across the board.

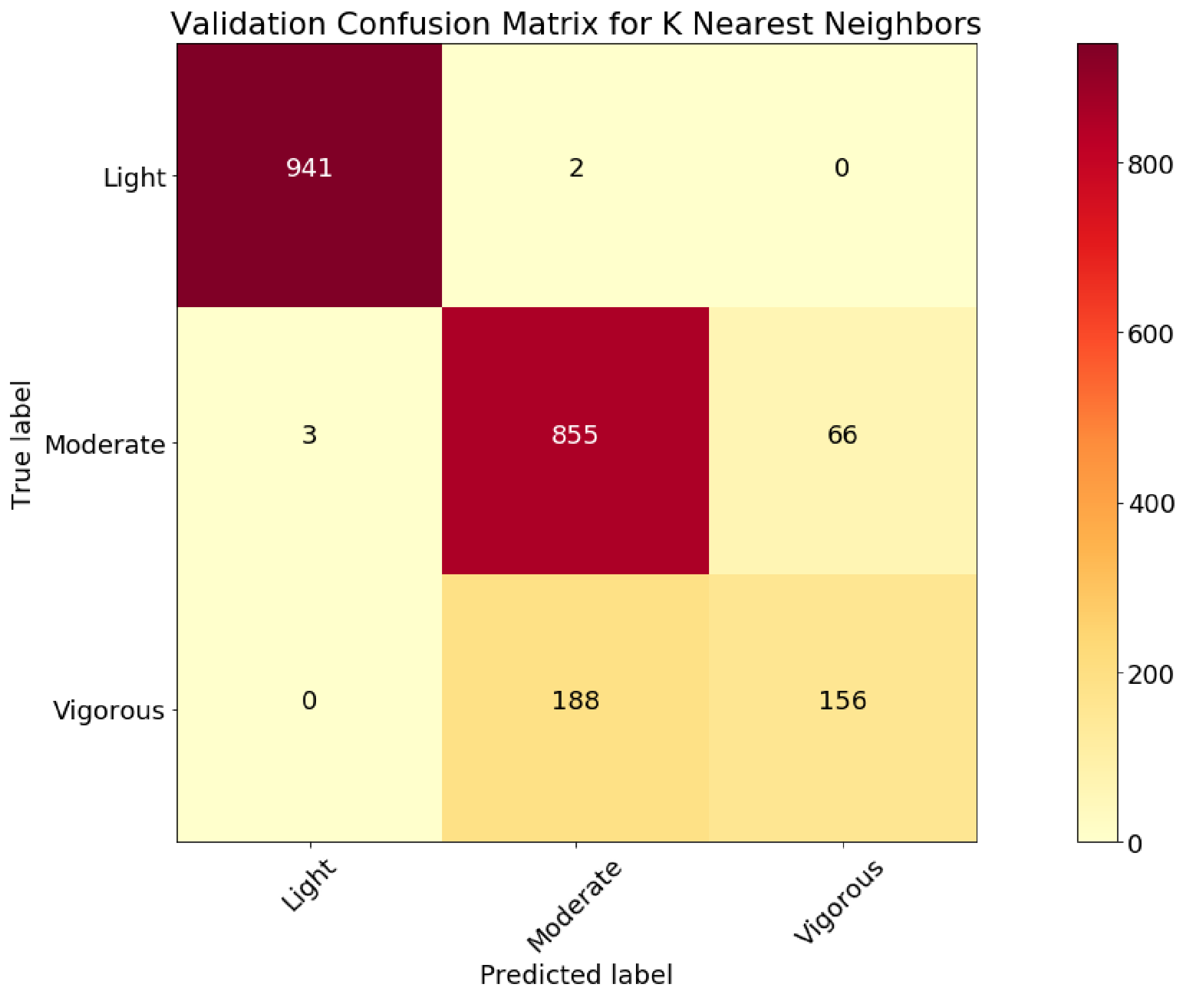
- K-NEAREST NEIGHBOR CLASSIFIER:

  Training F1 score is: 0.998

  Validation F1 score is: 0.806

  Training Accuracy is: 0.998

  Validation Accuracy is: 0.883



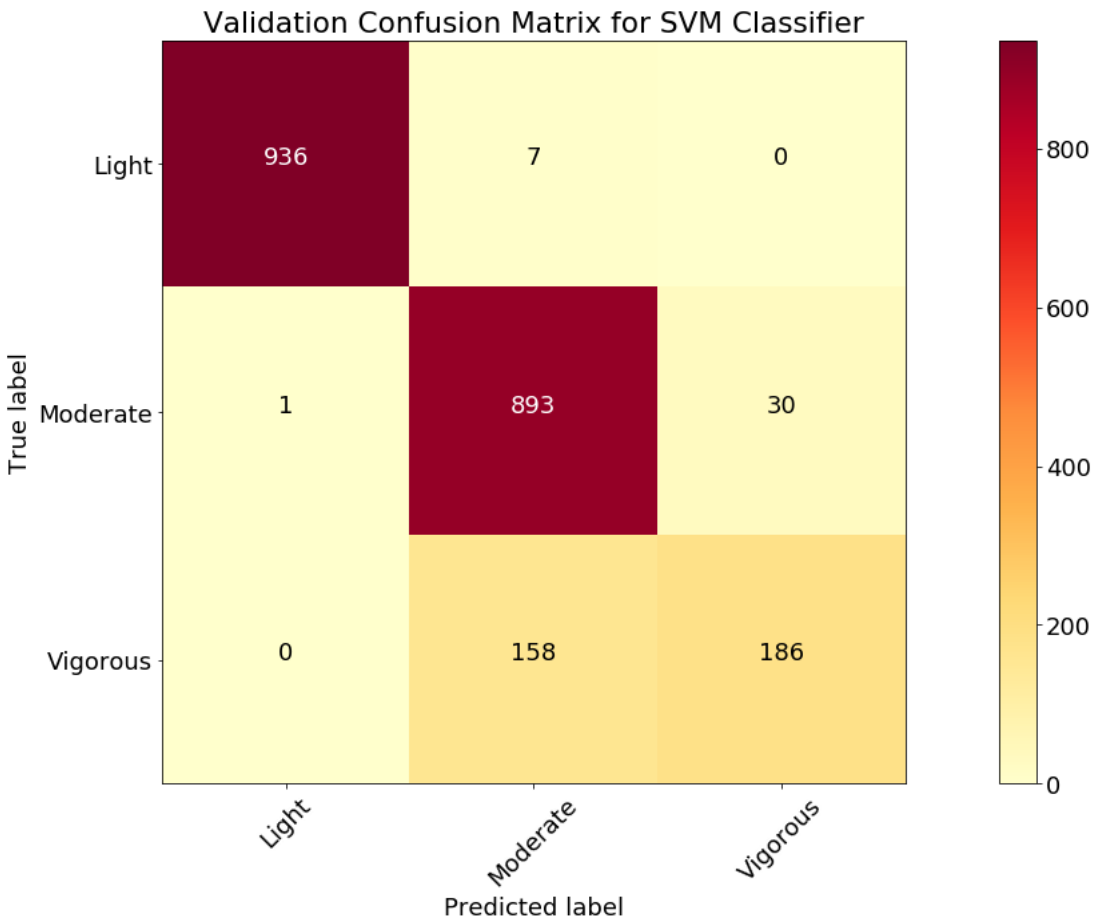
- SUPPORT VECTOR MACHINE:

  Training F1 score is: 0.989

  Validation F1 score is: 0.854

  Training Accuracy is: 0.992

  Validation Accuracy is: 0.911



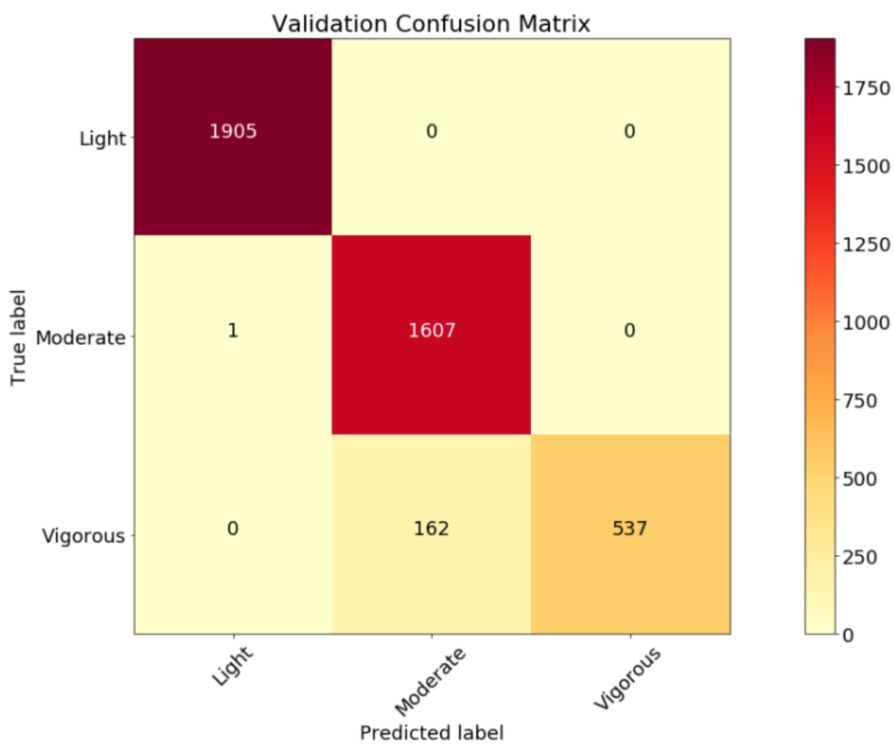
After checking validation accuracies for all 3 models, **SVM was chosen for final deployment** on unseen test data, as it gave an accuracy of 91.1% and an F1 score of 0.854. This is different from the case when training was done directly on the activities, as logistic regression was selected then.

FINAL TESTING:

The results on testing data comprising data from subject 5 and subject 6 are:

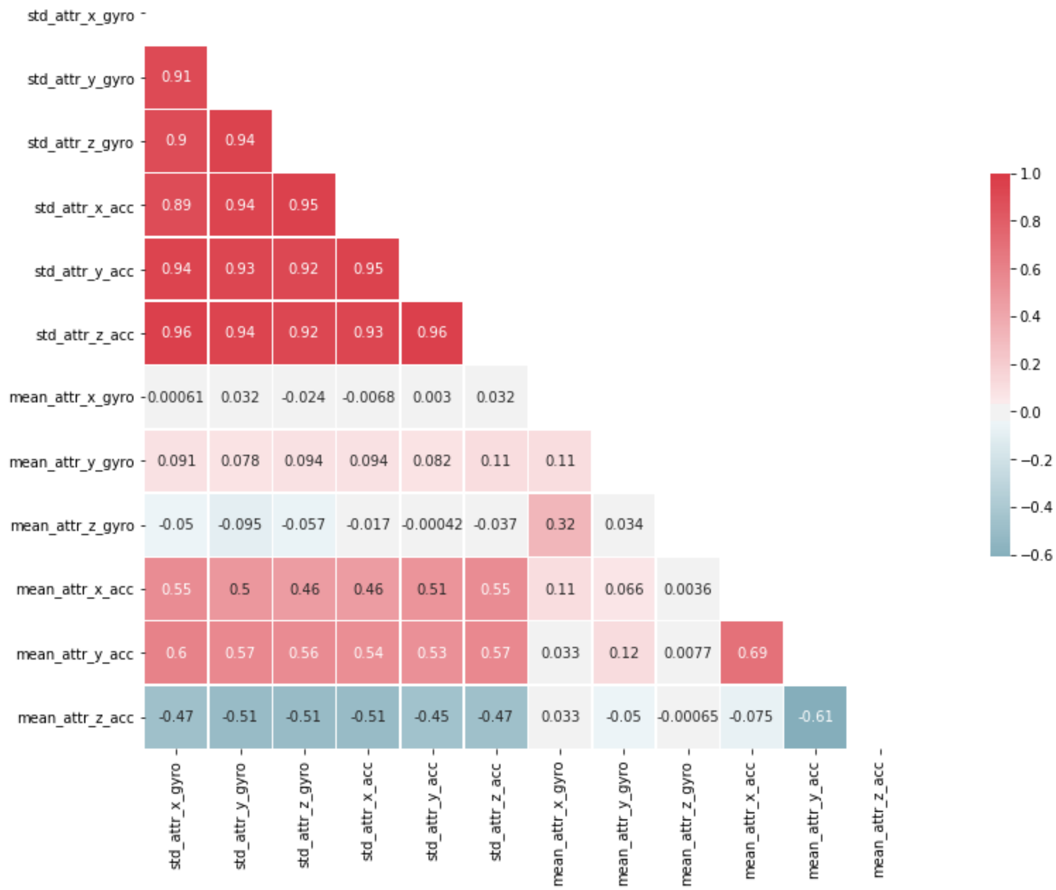
```
Testing F1 score is: 0.940
Testing accuracy is: 0.961
              precision    recall  f1-score   support

       Light       1.00      1.00      1.00      1905
    Moderate       0.91      1.00      0.95      1608
    Vigorous       1.00      0.77      0.87       699

    accuracy                           0.96      4212
   macro avg       0.97      0.92      0.94      4212
weighted avg       0.96      0.96      0.96      4212
```

The **final accuracy turns out to be 96.1%**, which is much higher than what was achieved by training directly on activities.



There are fewer misclassifications too. Overall, thus, SVM is chosen as our final model for Human Activity Recognition.

<u>Training on just accelerometer:</u> As a final check, we calculated the correlation coefficients between the various features used for training.



Upon inspection, we can see that many gyroscope features are highly correlated to the accelerometer features (r>0.9). Theoretically, therefore, elimination of the gyroscope features should have no appreciable effect on the final accuracy.

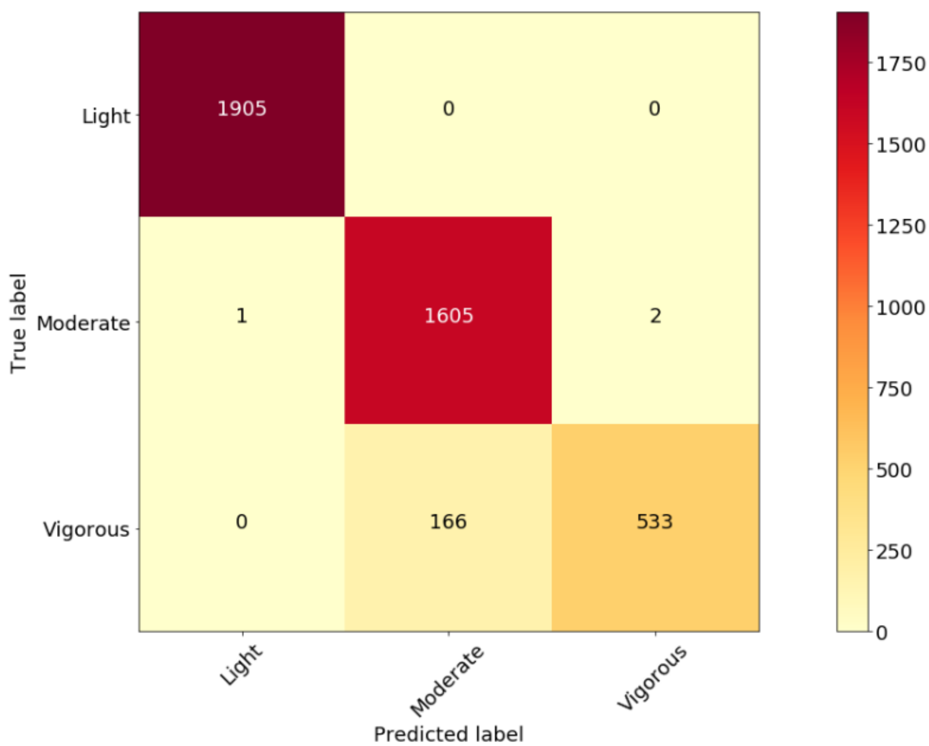
We thus trained and tested our model as before on the subjects, but only used the accelerometer data this time.

Testing Results:

```
Testing F1 score is: 0.938
Testing accuracy is: 0.960
```

	precision	recall	f1-score	support
Light	1.00	1.00	1.00	1905
Moderate	0.91	1.00	0.95	1608
Vigorous	1.00	0.76	0.86	699
accuracy			0.96	4212
macro avg	0.97	0.92	0.94	4212
weighted avg	0.96	0.96	0.96	4212

The actual results confirmed our expectations, with the final testing accuracy (96%) hardly lower than that earlier. It seems that due to high correlation between the two, gyroscope data is essentially redundant for purpose of training.



Misclassifications have increased slightly, but overall, this model is just as good, only with less complexity. This result is significant, and shows that physical activity can be predicted using just one sensor and with basic statistical features.

BIBLIOGRAPHY

- Applied Machine Learning by Mr. M. Gopal
- Link to Dataset
- https://en.wikipedia.org/wiki/Activity_recognition
- https://machinelearningmastery.com/deep-learning-models-for-human-activity-recognition/
- https://machinelearningmastery.com/how-to-load-and-explore-a-standard-human-activity-recognition-problem/
- https://towardsdatascience.com/physical-activity-monitoring-using-smartphone-sensors-and-machine-learning-93f51f4e744a

GITHUB LINKS

- Raw_Sensor_Data
- Jupyter Notebook for data analysis
- Jupyter notebook for Data Consolidation
- Data after consolidation
- Jupyter notebook for Data Correction
- Data after correction
- Initial models for predicting activities (Training)
- Testing activity model
- Model for predicting Classes
- Testing model trained for predicting classes
- Models using only accelerometer data