Final Report

Upon examining the dataset we were required to classify activities for, it was immediately apparent that there are 2 separate datasets included 'Protocol' and 'Optional' – the 'Optional' dataset does not complement the 'Protocol' dataset directly but rather adds lackluster data regarding different types of activities, as such our research focused on the more robust 'Protocol' dataset.

Exploratory Data Analysis and Preprocessing:

After further inspection we noticed that some data was either missing or flawed. To combat this issue we approached each scenario on a case-by-case basis:

- Transient activities (denoted as 0 in the activity ID) do not describe any specific activity, but rather unspecified or unknown activity or collection of activities. As such, we removed these samples from our dataset entirely.
- The 'orientation' columns were specified to be invalid in this data collection, as such, these were removed and have not been taken into consideration.
- The '3D-acceleration data (ms-2), scale: ±6g, resolution: 13-bit' was specified in the 'readme.pdf' to be less reliable than its counterpart with a scale of ±16g. Therefore, we decided to remove these measurements from our research.
- The sampling frequency of the heart rate monitor was specified to be ~9Hz, therefore, roughly every 0.1 seconds a value was measured, while all the other sensors yielded results every 0.01 seconds. We modified all missing samples that were tagged as NaN to the value of their previous real measurement.
- IMU sensory data had occasional missing values, in order to avoid losing potentially valuable data by dropping said rows, we decided to impute the missing data (i.e. infer them from the known part of the data).

Now that we have our dataset modified we are ready to perform some further analysis:

After inspecting the data given in 'PerformedActivitiesSummary.pdf' and after displaying the amount of times each activity was performed we can see that 'rope jumping' for example is significantly under-sampled as compared to most of the other activities. From this inspection alone we can expect our predictions on this activity and others that are not as well represented, to be less accurate.

We inspected the data more closely by plotting graphs from the available data and gathered the following insights:

- Activities that apply low physical strain on the body (such as lying, sitting, standing) have similar measurements in data such as heart rate or 3D-acceleration data. These types of activities are therefore expected to be more difficult to learn.
- Activities that contain different inclinations of the body parts measured have more apparent differences as part of the magnetometer measurements, this will allow the model to learn the difference between otherwise difficult activities to differentiate between, such as lying and sitting.

- We noticed that temperature and heart rate tend to change at a relatively low rate and in order to acquire a data sequence that is representative of the activity, a window size of ~5000 might even be required.
- We also noticed that using a sufficiently large sequence of the temperature and heart rate measurements gives us a pretty good idea of how to differentiate between more physically demanding activities to those that aren't.

Validation Strategy:

We've decided to validate our models using the Group K-fold validation strategy over the subjects.

Since our test set consists of new and unseen subjects the model won't be privy to during the training process, we want to assess its ability to generalize the input given.

Naïve Solution:

To establish our naïve model we decided to use various metrics gathered from the raw data:

Graph slopes – Chest Temperature, Heart Rate:

We've noticed that a more positive slope value usually indicates a more physically demanding activity (e.g. running or cycling).

Variance – Chest and Ankle Accelerometers:

We've noticed that these measurements provide a good way to differentiate between the different resting activities (e.g. lying or sitting).

Frequency – Ankle Accelerometer:

Different sport activities display varying frequencies in ankle movement, making it easier to differentiate between them using this measurement.

Average – Heart Rate:

A good measurement for differentiating between an activity that is less physically demanding (which will generally display a lower heart rate). This measurement can even allow to differentiate between different sport activities that are more demanding than others. For example: cycling almost always displays higher measurements than running.

We normalized the measurements specified above and took the relative portion of minimum and maximum values of the sensory data across all subjects.

Finally, upon receiving a window, in order to make the classification we checked which measurement he was closest to.

Our naïve solution yielded surprisingly relatively good results. Thanks to our exploratory data analysis we managed to extract relevant features from the raw data and process and iterate over them well.

Classical Machine Learning Algorithm:

The Decision Tree Classifier Algorithm we used yielded similar results to our naïve solution. We can deduce from this result that the features we decided to eventually use and impute were satisfactory but choosing a different algorithm may have yielded better results.

Moreover, according to the true positive/false negative matrix we created from the model's predictions we can see that it was unable to differentiate between activity archetypes (e.g. sports activities such as running or cycling, and resting or not physically demanding activities, such as sitting, walking, ironing and vacuum cleaning).`

Therefore, we can assume that a fully fledged neural network will produce more satisfactory results and will be able to generalize the data better and find the smaller nuances that differentiate between similar activities.

1st Model:

Our initial model implementation and training resulted in over fitting with moderately high validation accuracy sitting at roughly ~70%.

According to the matrix result displayed we can gather that the model failed to generalize mostly the activities of walking and ironing. This fits well with our analysis as the walking/ironing heart rate graphs are one of the potential measurements that are able to differentiate between the 2 activities. This means that providing the model with a larger window size will likely resolve this issue.

In order to improve upon this result we suggested various ways to tackle the problem and ended up implementing the 2nd version using, among others, a larger window size and stride which will allow us to take a more representative portion of the measurements into account when making a prediction.

2nd Model:

Due to the lower learning rate in this version we added more epochs to allow our model more time to converge. The result was that our 2nd model yielded generally better predictions across the folds and in average with a mean validation accuracy or ~78%.

In order to improve upon this result we may potentially use dropout to allow the model to generalize better. We can add additional extracted features and inject them to the classifier, seeing as the naïve solution generated good results for the walking activity, which this model failed to do.

Final Model:

Our final model's final performance was more similar to that of our 1st model, despite using the specs of the 2nd model. The reason for this could potentially be the fact that the tested subjects were different, and that the model did not generalize the features as well as we had been led to believe by the results of our 2nd model's validation performance.

Self-Supervised Pretraining:

We decided to pretrain our model on the classification of subjects. This allows us to initialize our model's weights to a good starting position from which we can start fine tuning by training the model for our initial objective.

The result from the pretraining wasn't as decisive as we had anticipated.

The most prominent reason for this outcome is probably the difference in classification intent. More specifically, when classifying subjects we fit the model to the subject's activity patterns, which is the opposite of what we want to do when fitting for the activity type – generalize the pattern so as to disregard the specific patterns laid out by the subject.

Another potential shortcoming that may have faulted our model is a sub-optimal learning rate which made our model converge on a sub-optimal minimum.