Pison Data Scientist Technical Interview Challenge

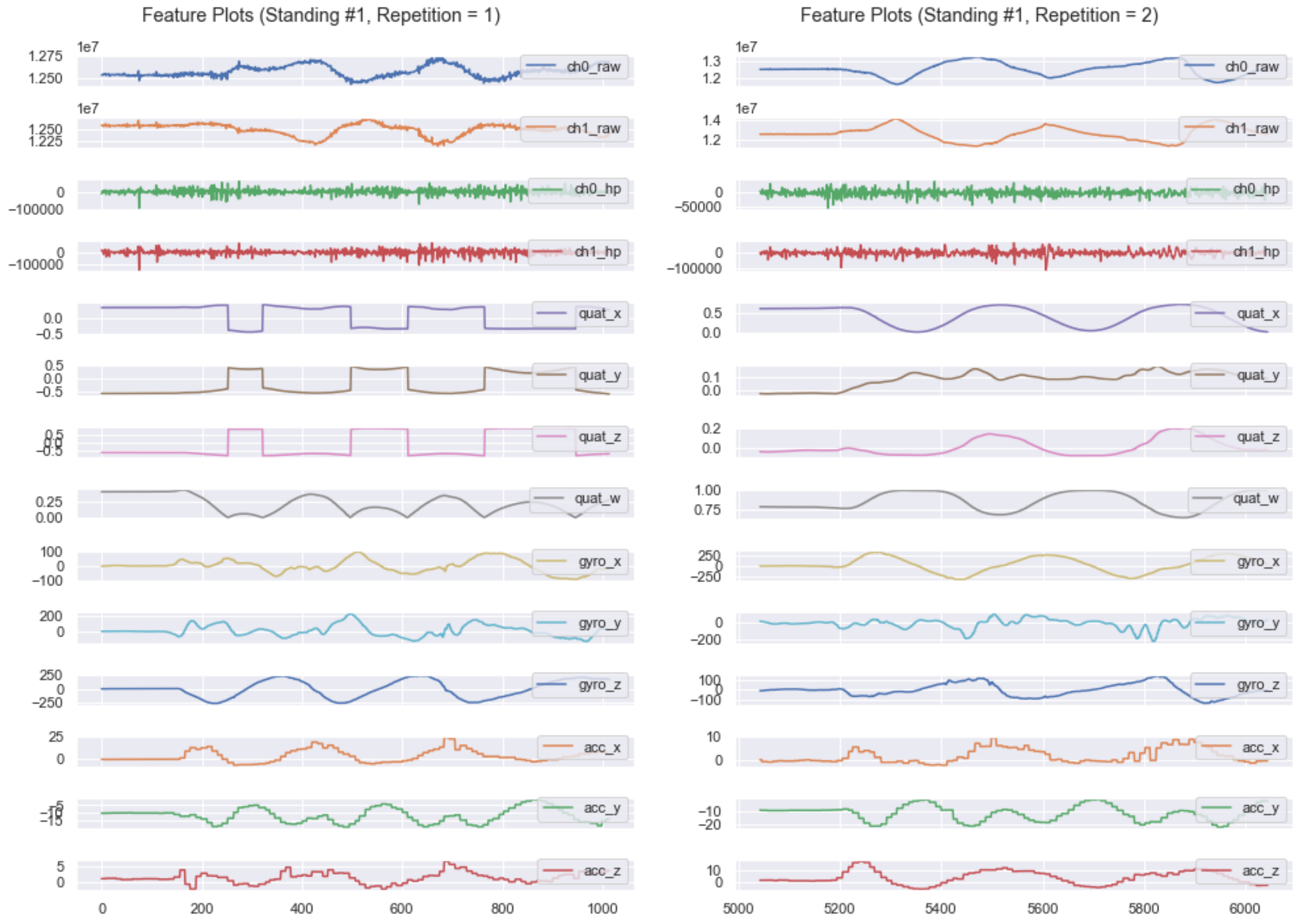
Vina Ro | Project link

1. Identifying the number of wrist motion classes

To identify wrist motion classes, it is easiest to observe from data recorded in static. Therefore, plots for standing #1 and standing #2 motion for all repetitions were plotted for further analysis (**Fig 1**).

From **Fig 1**, data for each repetition label share similarities. For example, quaternion data from all four axes follow the same pattern for repetition 1 across both standing motions. The same can be seen for all quaternion data for repetition 2. Similarly, motions in repetition 3 for quaternion data also follow a pattern of their own with a significantly high amplitude for gyroscope data in the y-axis.

In conclusion, the hypothesis that there are a total of 3 wrist motion classes for each repetition is drawn. Therefore, repetition numbers from the given dataset were used directly as ground truths.



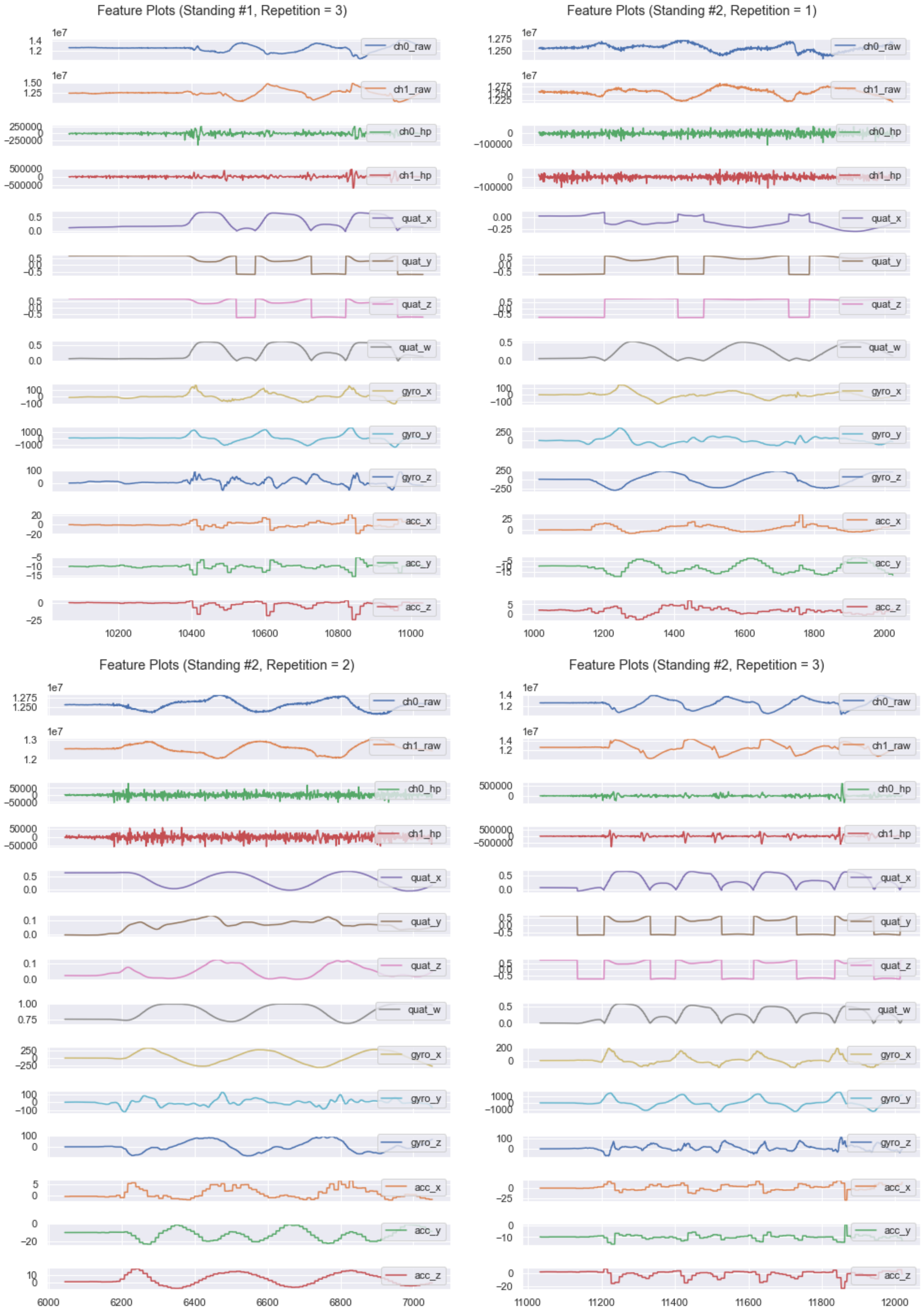


Fig 1. Plots for all sensor data for standing motion

2. Identifying the wrist motion for each class

For wrist motion in repetition 1, quaternions x, y, and z are discrete, indicating that there are pauses mid-air while doing the motion. Z-axis gyroscope data is most significant, indicating the wrist motion is moving horizontally (similar to a finger waggle "no" motion).

For wrist motion in repetition 2, x-axis quaternion data and x-axis gyroscope data are the most significant. Therefore, it seems that the finger motion for this class is the wrist moving horizontally left and right.

For wrist motion in repetition 3, recorded data from the y-axis of the gyroscope has a significantly larger amplitude compared to all the other data. This indicates that there is motion occurring exactly on that axis. Since it's fluctuating between positive and negative values, my guess for the wrist motion for this class is motion rotating clockwise and then counterclockwise around the y-axis.

3. Feature Engineering

Before applying standard classification algorithms, a sliding windowing technique is applied to the raw time-series data. Data is first divided into windows of 0.9 seconds with 66% overlap, and new features were calculated by aggregating the 300 raw samples within the window. Generated new features consist of the mean, standard deviation, maximum, and minimum from both the time and frequency domains. Other features were calculated as well such as skewness and energy, which decreased the performance of the machine learning model however in this case. The 0.9-second window and overlap percentage were selected empirically from a range of 0.5 to 1.5 seconds and 30% to 70% respectively. Overlapped windows are used to ensure that every subsequent row in the transformed data has some information from data in the previous window. For the labels, the most frequent activity within the window is taken.

4. Classifier construction

Data were scaled after splitting the previously calculated features into 80% train and 20% test to avoid data leakage. A Gaussian Naïve Bayes model was used to fit the training data. 5-fold cross-validation on the model was also performed to avoid overfitting.

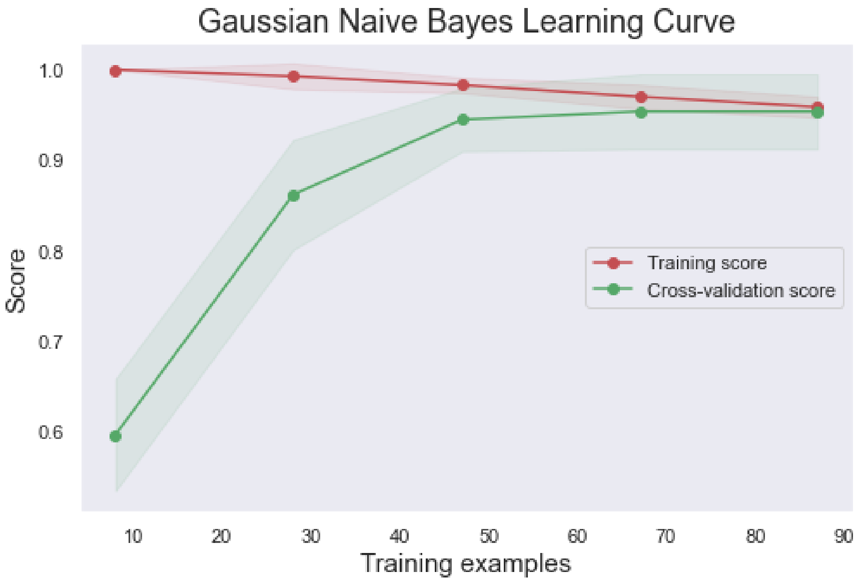


Fig 2. Learning curve

From the learning curve (**Fig 2**), we can see that the model has low variance and no signs of overfitting or underfitting.

After training the model, prediction on the test dataset was performed and resulted in a testing accuracy of 89.29 %. From the confusion matrix (**Fig 3**), it is shown that the model overall predicts correctly between the three motions with occasional mistakes.

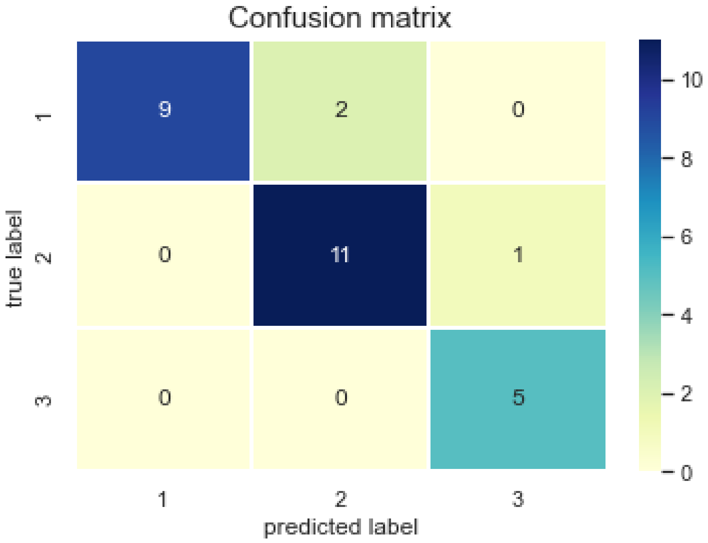


Fig 3. Confusion matrix of the model