MSc project

Detecting Parkinson's Disease Tremors with Wearables in Free Living

Sam King

200244749

Submitted for the degree of Data Analytics Masters of Science
supervised by

Dr. Yordan Raykov

September 30, 2021

Abstract

There has been a lot of modern research into the detection of Parkinson's Disease tremors with various methods used to monitor patients' movements. The method focused on in this project is the highly effective method of using wearables within free living. In this project, the primary subject of discussion will be how best to evaluate a model's performance given the type of data that is being worked with. Methods such as different implementations of cross-validation will be utilised, and issues that arise such as lack of labels with certain parts of the data will also be addressed. The trained models with then be used on unlabelled smartphone data and the results will be assessed using domain knowledge to determine how likely the model is and whether there are any noticeable key trends.

Contents

Chapter 1

Introduction

A key characteristic and symptom of Parkinson's Disease is the tremors that the patients experience. The aim of this project is to detect Parkinson's Disease tremors by building a model that is able to use data that comes from wearables in free living. A model that is able to distinguish a Parkinson's patient from a non-Parkinson's patient through analysis of their gait is desirable for this. Currently the most effective way of diagnosing and spotting these trends is through an in-person clinic, but this is very cost and time inefficient so a personalised detection system would be most beneficial. Data from 8 Parkinson's patients will be analysed, and this data is collected whilst the patients are in their homes wearing sensors and additional camera footage to capture their various movements throughout the day. The patients are encouraged to go about their day as they normally would instead of carrying out set tasks. This makes it easier to create a model that is effective within the real world although this phase is still carried out within a controlled environment so that the data can be labelled. The patients' bradykinetic gait is characterised using the sensors. This is related to the patients' balance and can affect motions such as turning as well as their balance. The gait of Parkinson's patients is often exhibited by small shuffling steps and slow movements. [1] The sensors used would allow us to pick up on these movements and detect any abnormalities compared to the control subjects who do not have Parkinson's disease. The data used in this project will be purely from people who do have Parkinson's Disease; there will be no control patients who do not have the disease. If further research were to take place

after this project, then comparisons could be drawn between those with the disease and those control subjects who do not have the disease. Building a model that learns from those who display the tremors that are characteristic of the disease will be the focus of the training process.

1.1 Previous methods

There have been various studies into the classification of tremor and gait, all with varying methods of collecting data as well as different models used to analyse the data. They will be highlighted in this section, although it will become clear that some papers in this field will be more useful to this project than others due to the methods of data collection and the nature of the data that has been collected.

A recent similar study was carried out recently by Hammerla et al. [6] who too investigated the use of sensors for detecting tremors. The results are collected within a controlled environment in a laboratory. The models utilised in this paper are different from the ones used in this report, as Hammerla et al. makes use of more deep learning methods using models such as binary RBMs.

Another method commonly used in research when recording and measuring the gait of patients is using silhouette images. Photographs are taken of patients walking, and a black and white silhouette image is analysed to measure the differences between each step and analysis of the gait is conducted to notice any abnormalities or any tremors. Hong [7] makes use of this and models the images using a Bernoulli distribution $p(x \mid \mu, \pi) = \sum_{m=1}^{M} \pi_m p(x \mid \mu_m)$ where x is a silhouette image, M is the number of mixture components, and μ is the probability that a pixel will be white. This paper will not be as useful to the project due to the different approach that has been taken to model the gait. There may be some similarities with regards to sequence frequency and conclusions drawn on the characteristic of the gait itself through the probabilistic gait model, but the input source that is being modelled from means a different angle to approach things. Using a Fourier transformation is a common method on the input signals and is one used in our main two papers, so this will be a method that could also be applied but however will not be used in this project. This paper follows a very Bayesian approach and

is clear to define its priors and posteriors as it goes on. They do not use any regression methods; they use maximum a posteriori estimation in order to maximise the parameters using in the posterior probabilities for the images used.

In Parakkal's paper [11] he makes use of freezing gait which is identified when there is no forward motion of the stance leg. It analyses walk cycle which includes metrics such as heel strike phase. This paper mainly deals with the characterisation from gait that comes from stepping off so therefore is not as flexible to use on a real-world environment where a patient may carry out various tasks. The methods used for analysis in this paper are much more computational and look at the problem from more of a physics point of view looking at the force at which a patient may step off with. This means that this paper is not going to be that influential to this project, but some interesting conclusions could be utilised at some point.

Di Biase's [4] paper is a well-rounded overview on the various methods that can be used within probabilistic gait modelling. It cites supervised learning algorithms to be the most effective when building a model, and the other papers that have been found support this claim so it is more than likely that some form of supervised learning algorithm will be used for classification. The methods in which it is suggested that data can be collected varies, and it does touch upon using wearables as an effective method for gait analysis.

The two main papers that this project will be following on from are by Evers et al. [5] [12] as it is part of the same experiment and the data used in this report will be the ones that were collected by them. They are both very recent bits of research conducted on the area and will prove rather useful in this project. The data is collected in the same way through sensors, video recordings, and the use of smartphones and smartwatches as described in the introduction. They both serve different functions in terms of their use to the project. The first [5] offers more of a descriptive nature surrounding how the sensors work and the means in which they are utilised. As well as this, graphical results from the sensors are shown displaying statistics on the effect of the various factors measured by the sensors such as the medicine. It is basically a description of how the data is collected. The methods for

classification are still described and the process is described. The second of the important papers [12] gives a more mathematical overview of the methods that were utilised to achieve the results, with different equations shown. The method for classification that is used is logistic LASSO regression that uses uniform prior class probabilities. This helps keep it computationally simple. The classification aim was to be able to distinguish whether a sample gait was from a patient with or without Parkinson's disease. One classifier was trained for each sensor location as well as all the sensors combined. This is a very efficient method and allows us to get a better overview of how the sensors perform and whether any individual sensor carries any significance. In this project the overall view of all sensors will be classified, not the individual sensors.

1.2 Aims and objectives

There are two main parts to what will be done in this project, and this will be on the two different data sets mentioned previously. There is the labelled data that will be used as training data and unlabelled data that will be used to test the trained models on. An overview of what kind of data there is and what kind of variables that are being worked with will be looked at, and any considerations that will need to be made after having seen the data will be considered. Things like if there is an imbalanced distribution for some of the causal factors or if the data would benefit at all from a Principal Component Analysis will be looked at.

After the exploration of the data, the training process can begin. The focus of that will be how can the accuracy of a model on the training data can be determined. What different styles of cross-validation work on the models and how biased they are in their figures will be investigated. 10-fold and Leave One Out Cross-Validation are the ones that will be used, as well as a new method of Nested Cross-Validation which aims to improve the reliability of the accuracy from a usual Naive Cross-Validation by calculating the bias.

Following on from this, the imbalances between certain classes will be investigated and whether performing a causal bootstrap will influence the figures that the cross-validation produces. Measures such as the sensitivity and

specificity will also be investigated, as well as the ROC curve for each style of cross-validation that is performed. The importance of these figures and what they could show regarding what the future results will be in the second half of the project will be investigated. This will be things like whether the model could be expected to underestimate or overestimate the frequency of the tremors.

To diversify the models and explore different routes that can be taken with this, alternate methods to the standard linear models will be looked at. How the data fares with other models such as linear mixed effects models where the causal factors can be used as random effects in the model to judge how much they can affect the model's performance. An aim of this is to see whether an alternate model like this would be better than the standard models used in the previous cross-validation chapter, and whether any unique insights about the data can be used.

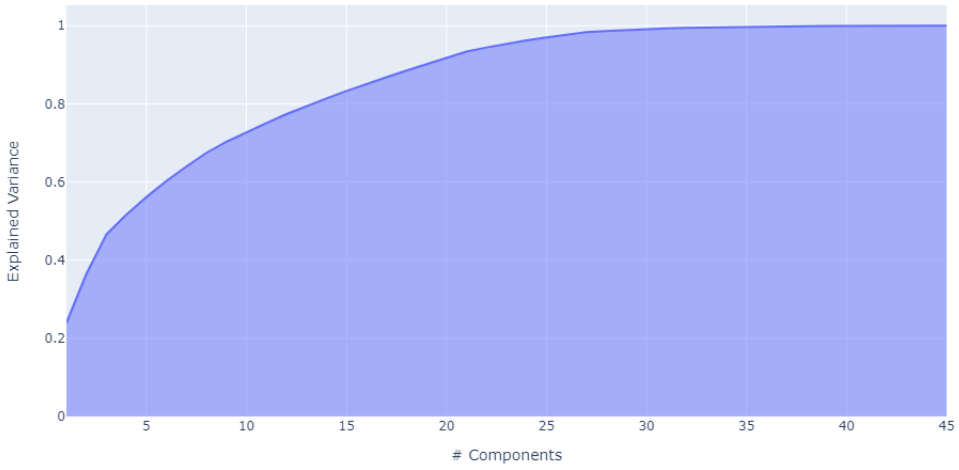
The aims of the final chapter are to see how the trained models perform on the unlabelled data, and if predictions made from figures calculated are close to what happens. Models that were assessed to be the most suitable in the previous stages will be used on the unlabelled accelerometer data and the aim is to study the graphs produced for each of the 8 new patients and to notice any trends that appear.

1.3 Initial data overview

The sensors used in the monitoring are worn on both wrists and ankles, on the lower back, and in the front pants pocket. [5] The sensors provide a much more convenient way to monitor movements and tremors as previously patients would have needed to fill out paper diaries. Not only can this be a lot of effort to carry out, but some patients are not able to even recognise when they are experiencing these tremors and the full extent of what is happening. As well as the sensors and camera, the patients are monitored using a smartphone and a smartwatch. This phase takes place over the course of two weeks. Unlike the first phase, this phase is not as controlled as the first so there will be no labels for this data so there are no causal factors recorded at all. The data picked up is from visits that are approximately two hours long

each. Furthermore, the effects dopaminergic medication has on the patient's gait is also monitored viewing how the medication works overtime as well as how withdrawal from the medication affects the results. Dopaminergic medication releases dopamine which is commonly known as the reward hormone, and this can help ease movement for patients.

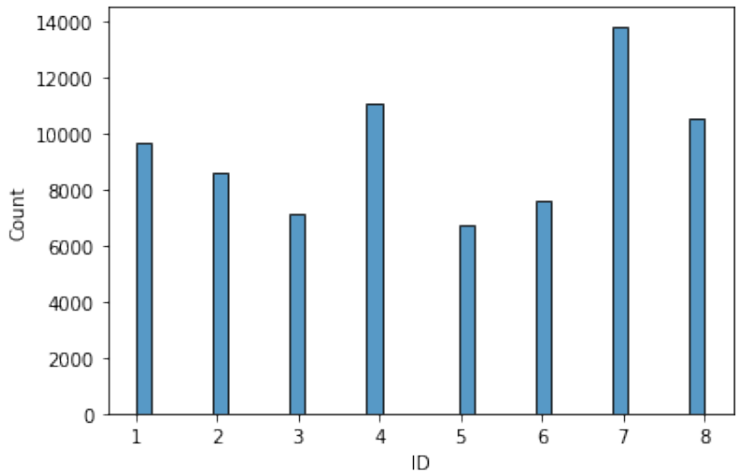
The initial data consists of 50 variables. These are the Patient ID, 45 sensor variables, Prototype ID, Medication Intake, Activity Type, and Non-tremor/Tremor. Aside from the target variable and the binary tremor variable, the non-sensor variables are called the causal factors. These are factors which have a proven link in causing or influencing whether a patient has a tremor. The activity label is not as significant for this project so for the most part it will be discarded when completing analysis on the data. The study consists of 8 patients with varying amounts of data for each one. Using the 45 sensor variables, a Principal Component Analysis was done to see if this could be reduced for faster computation. Later in the project when the datasets have millions of rows, it would be a good plan to reduce the number of variables to make it easier to work with. Below is a graph detailing the number of components compared with the cumulative explained variance.



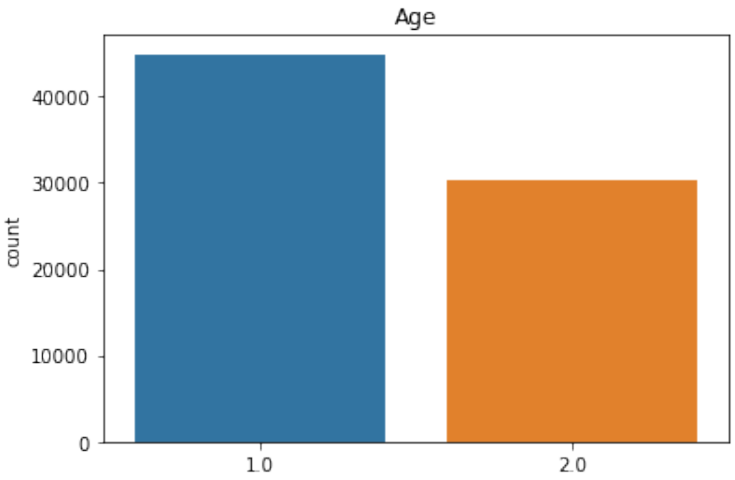
This shows that the data cannot be successfully visualised using 2 components as this only explains 30% of the variance. It is around 23 components

that a 95% explained variance is achieved meaning that half of the sensor variables can be discarded without losing a significant amount of information. The sensor data was transformed into the 23 components, before which it was whitened by using the mean. The use of PCA means later when discussing the impact of certain sensor variables, we are looking at significant components. Having assessed the sensor variables, we can now explore the causal classes that exist in the data.

Below is the graph for the distribution of the patient data.

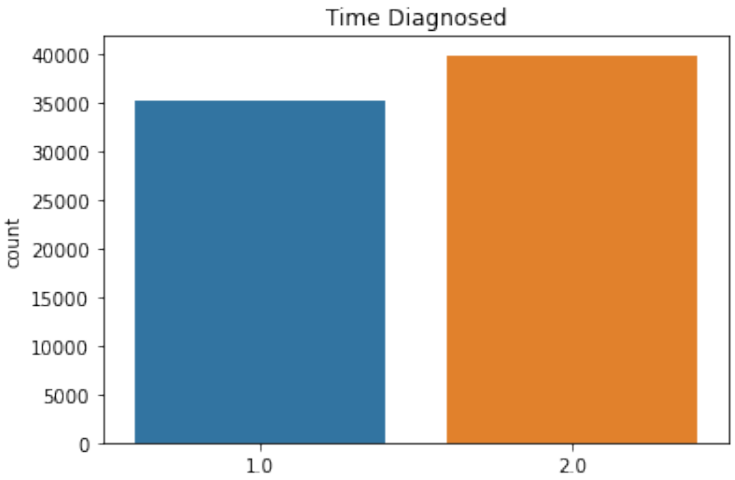


In some cases, certain patients have almost double the amount that other patients do. This will have to be taken into consideration later in the project when the evaluation of the models that are being used takes place. There are other variables within the patients in which can be separated out into. The first of these is age, and there are 2 age groups in which they can be separated out into. They are those that are above the age of 65 and those who are under it. This is another causal variable as the age of someone can have a huge impact.



Above is the distribution of the two age groups within the patient IDs. Like the patient IDs, this is also unbalanced so that could also be something that could be taken into consideration when performing some kind of sampling later in the project.

The other category that the patient IDs can be separated out into is the time since they were diagnosed. Like the age, this can be a binary variable split between those who diagnosed more or less than 7 years ago. This indicates the progression of the disease, so again is a causal factor that impacts the tremors.



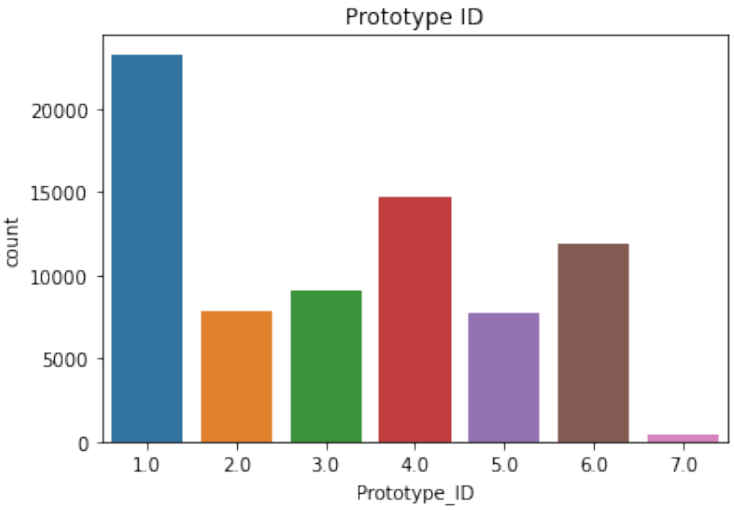
Above is the distribution for the class that the patients belong to with regards

to the occurances within the data. This distribution is the most equal so when considering sampling later it will not be a significant causal factor to keep in mind as age is clearly more significant in terms of needing to be more balanced.

The Prototype ID signifies what type of tremor has occurred if there was one. There are 8 unique classes for the Prototype ID detailed in the below table.

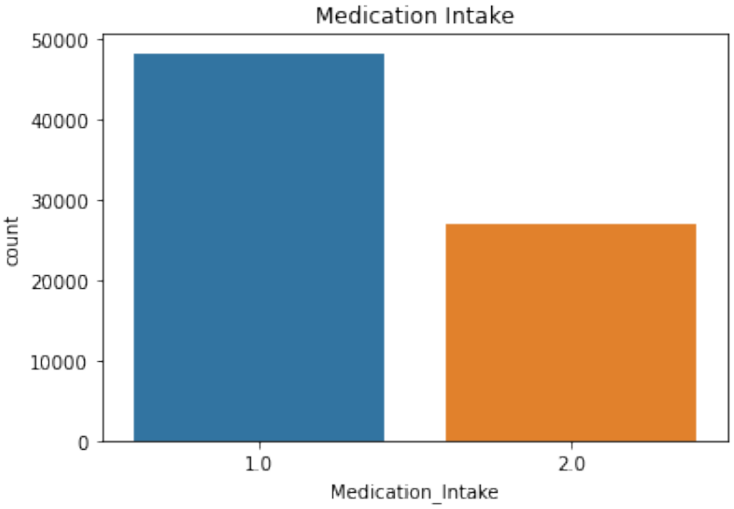
Prototype	**Prototype meaning**
0	Tremor has not been labelled
1	mostly wrist and/or fingers flexion, arm resting on surface
2	mostly wrist and/or fingers flexion, arm free
3	mostly elbow flexion, arm resting on surface
4	mostly elbow flexion, arm free
5	tremor during gait
6	mostly pronation supination, arm resting on surface
7	mostly pronation supination, arm free

This dataset has a lot of unlabelled prototypes, so most will have the value of 0. Using a Random Forest classifier, we are able to replace the missing values with a prediction of what they could be training on those instances in the data where the prototype is labelled.



Above is what the distribution of the prototype IDs looks like after having made these predictions. From the graph, it can be seen that most of the tremors that occur are wrist and/or fingers flexion while the arm is resting. This matches with clinical knowledge as this site is where tremors occur the most. There are very few for the pronation and supination compared to the other readings, this is probably not as significant to this research.

As mentioned previously, the medication intake is a binary variable which indicates whether the medication has been taken with 1 being the patient hasn't had their medication and 2 being that they have recently taken it.



This shows that most of the readings will be when the patient has not recently had their medication. This imbalance is another thing that will need to be taken into consideration since the medication intake is an important causal factor which would determine whether a tremor occurs, so it is significant to increase the representation of the readings following a medication intake in the training set.

Chapter 2

Cross validation

2.1 Issues with naive CV

Given the nature of what the model is predicting, it is important that the accuracy of the model is correct. Typically, a standard cross validation is used to determine the model's accuracy. There are several ways of doing cross validation, and the ones that will be used here are 10-fold cross validation and Leave-One-Out cross-validation being nested on a certain variable. The variable that will be nested on in this case is the patient ID. The typical Naive CV however has some issues with its performance and what it is estimating. Bates et al. [3] tackle this problem and investigate the drawbacks of a typical Naive CV. It shows that the estimate produced by Naive CV is not the accuracy of the fitted model on the data, but rather the average accuracy trained using many hypothetical data sets. They also show that the naive CV estimate of error produces a larger mean squared error (MSE) when it is estimating the prediction error of the final model than when it is estimating the average prediction error of models performing on many unseen data sets.

2.2 Implementations of nested CV

The implementation that was used for nested CV (NCV) was the algorithm devised in the previously mentioned paper [3] and the ideas behind some of the key figures used will be briefly explained. One of the key figures calcu-

lated is the bias estimate for the algorithm. It yields a convenient estimate of the bias that exists for each NCV point estimate of error. On a smaller sample size of $n(K-2)/K$ there is no bias for the NCV estimate but when done on the full sample size the error typically will have some bias. To calculate this, both naive CV and nested CV were run, and the bias is a scaled difference between the two looking for differences between errors in a full sample versus a reduced sample

$$\hat{bias} := (1 + \frac{K-2}{K}))(\hat{Err}^{(NCV)} - \hat{Err}^{(CV)}) \tag{2.1}$$

As previously mentioned, the two types of cross validation that will be used are 10-fold and leave one out. These are both common methods, but the two implementations differ slightly in execution. For 10-fold as expected the data is divided into 10 roughly equal parts and each part is a fold meaning that causal factor distributions are kept the same. In the leave one out implementation it was decided to use one of the categorical variables. In this instance, the patient's ID was used to nest on. There are 8 patients, so the data is easily split into 8 parts. Another variable like prototype ID could also have been nested on.

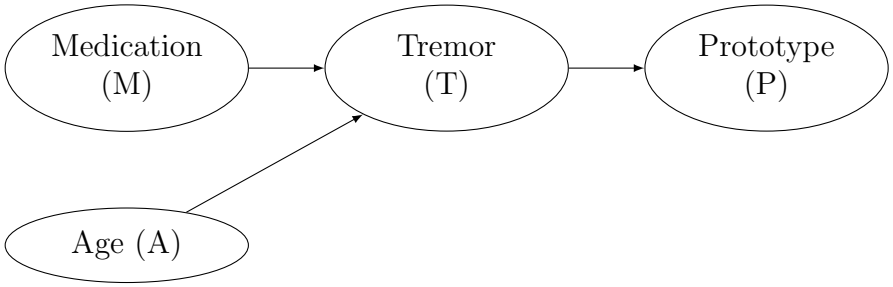
Model and CV type	**Lower CI**	**Higher CI**	**Error**	**Bias estimate**	**Standard Deviation**
Logistic Regression (LOOCV)	0.17497	0.188167	0.18157	0.00388	0.38866
Logistic Regression (10-fold)	0.1733	0.17787	0.17559	0.00044	0.38084
Random Forest (LOOCV)	0.09506	0.10527	0.10016	0.00046	0.30083
Random Forest (10-fold)	0.06542	0.06845	0.06694	0.00147	0.25246

From this table the bias estimates differ by a lot between the 10-fold cross validation and the leave-one-out cross validation. There are several reasons why this could be the case. As mentioned, the nested CV was being nested on the patient IDs. The problem with this is that there is not an equal amount of patient data so the model will be better at estimating those patients that have more data so it will skew it towards that. As well as this, patients can differ vastly so will not behave in the same way and experience tremors at different points and of different severity. It is also crucial to note that the 10-fold for a Random Forest and LOOCV for the Logistic Regression classifier carry higher biases than their counterparts, so this is something to keep in mind as they are more biased when doing nested CV on the hypothetical datasets produced rather than the actual data. This should be considered, as the Random Forest consistently outperforms the Logistic Regression classifier and when done with 10-fold cross validation it has the lowest standard deviation on the error.

A solution to balancing the data to make it more equally represented is to use causal bootstrapping

2.3 Causal Bootstrapping

As was seen earlier, the causal factors in the data are not distributed equally. To get a balanced training set with enough data from each causal factor causal bootstrapping is used. Below is the causal diagram with the causal factors that are most important for the bootstrapping.



The conditional relationship between the factors is used to perform the back-door causal bootstrapping as outlined by Little [10]. From the causal graph the following expression is used:

$$\frac{p(M)p(A)p(T \mid A)p(P \mid T)}{p(M, A, T, P)} \tag{2.2}$$

to calculate the conditional probability of the tremors given the causal factors.

Back-door bootstrapping is used here because the causal factors are measured confounders, which are labelled as $\mathcal{S}$ in the algorithm. In this case, both Y i.e. the response tremor variable and $\mathcal{S}$ are discrete so the method is parameter-free. This means that the only source of additional error is caused by the resampling variability which occurs during the bootstrapping process. Samples are selected from the data using the weight

$$w_i = \frac{K[y_i - y_n]}{N\hat{p}(y_n \mid \mathcal{S}_i)} \tag{2.3}$$

for each n in the dataset, where N is the length of the dataset and K is the Kronecker delta function. This method is useful because it helps any machine learning model that is used to learn the desired causal relationship that exists within the training data without having to change the algorithm in terms of the parameters used. It also eliminates the need to perform another highly controlled experiment to achieve a more balanced distribution of causal factors. The full implementation of the algorithm and the code can be found in the appendix.

2.3.1 ROC Analysis

Having performed the causal bootstrapping on the data, the effect on the results table for the performance of the nested cross validation on each of the models will now be looked at.

The sensitivity and specificity rate [8] as well as the ROC curve will be looked at to further analyse the suitability of certain models and to determine how well they perform to assist in achieving the aims and objectives set out in the introduction. Both figures make use of the balanced accuracy of the model,

which takes into consideration the false positive rate and the true positive rate. The formula for the balanced accuracy is simply

$$balanced\_accuracy = \frac{1}{2}(tpr + (1 - fpr)) \tag{2.4}$$

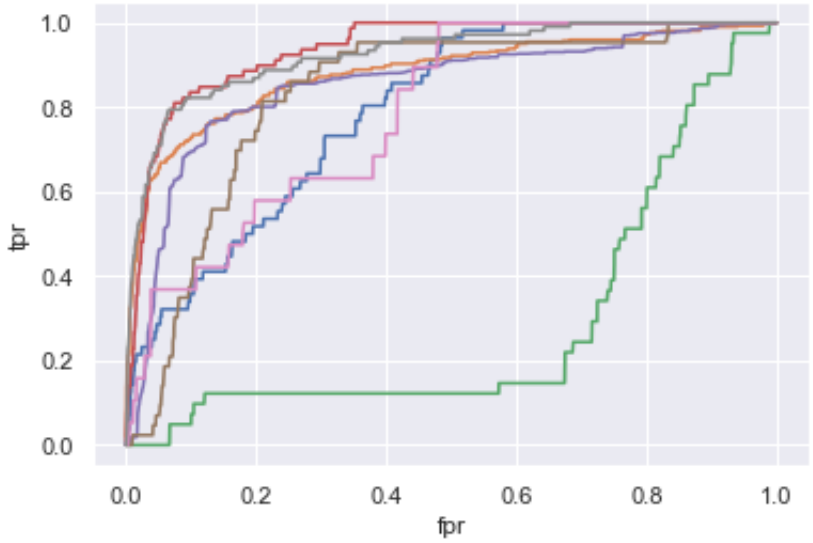
The sensitivity measures the ability of the model to correctly classify a tremor. The higher the sensitivity rate, the better the model is at identifying tremors. This figure is important as if it is low, it means that there are a lot of undetected tremors. It is given by the following formula:

$$sensitivity = \frac{tpr}{tpr + fnr} \tag{2.5}$$

where fnr is the false negative rate. In this case specificity is not as crucial, but it is still an important figure to look at as it can cause problems. The specificity of the model indicates the model's ability to correctly identify when a tremor is not occurring. The higher the specificity, the better the model is at detecting when there is no tremor. Classifying a non-tremor as a tremor can have serious impacts on treatment, especially if the specificity rate is rather low. Given the distribution of tremor against non-tremor data it would be expected that the model would have a reasonable specificity rate with a lower sensitivity rate. To get a view of these figures for the model, the average of the sensitivity and specificity of each patient fold was taken. The formula used for specificity was as follows:

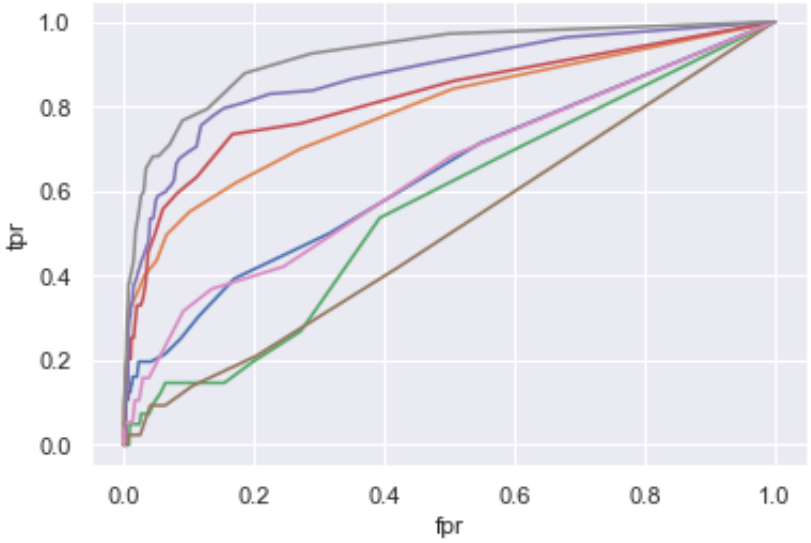
$$specificity = \frac{tnr}{tnr + fpr} \tag{2.6}$$

Below is the ROC graph for LOOCV of a Logistic Regression model, with each patient ID's false positive rate and true positive rate mapped against each other.



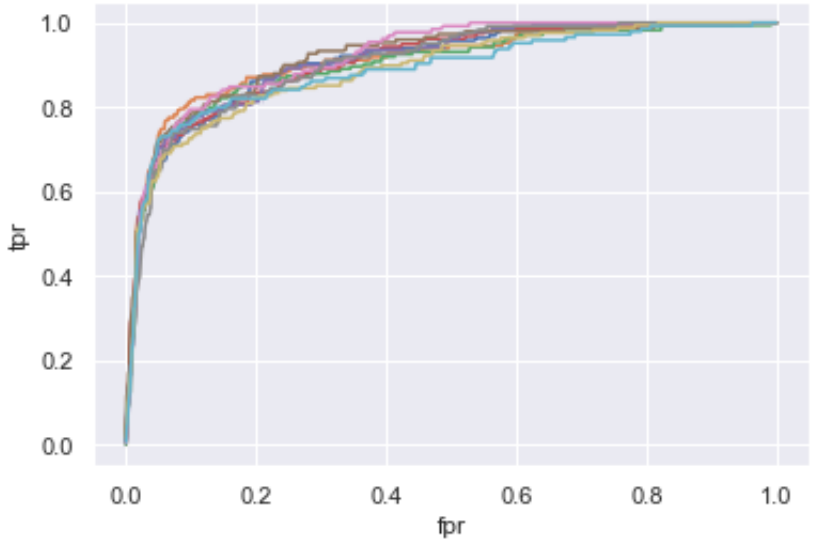
As can be seen, one of the patients results performs a lot more differently than the others. The true positive rate barely increases while the false positive rate does, and this indicates that the model performs very poorly on this one patient, as the curve dips below the fpr=tpr rate. This patient can skew the model's performance making it less accurate, meaning it is unreliable. The standard deviation of the error for this curve is 0.35 which could explain the high variance between the patient curves. It has a sensitivity rate of 0.84 and a specificity rate of 0.68 which indicates the model would incorrectly classify a non-tremor as a tremor meaning that the model will probably predict a lot of tremors over a certain period so this will have to be kept in mind. Remember that these figures are an average over all the patient IDs, so there is a possibility that the one patient who does not perform like the others could have greatly affected these figures.

Below is the ROC curve for LOOCV of a Random Forest model.



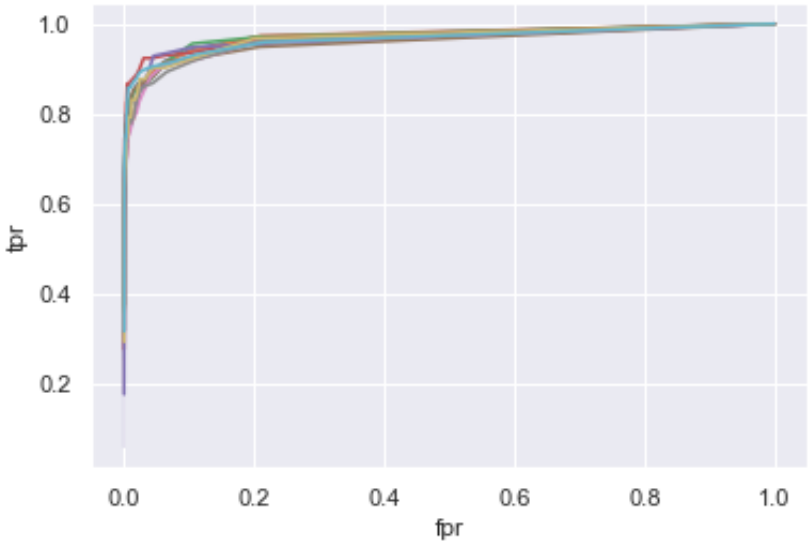
There is a big difference in the ROC curves between the LOOCV curves and the 10-fold curves. The standard deviation for the error greatly differs and it is shown in the graph as the curves are not as differently spread as in the Logistic Regression curve. There are no outliers with a few of the patients coming near the fpr=tpr line. The standard deviation of the error for this curve is 0.14 which is considerably lower than the figure achieved in the previous figure. Despite this, aside from the outlier the model performs reasonably well on the rest of the patients with them being grouped more tightly together. This differs from this curve as there are greater gaps between them. Despite the Random Forest model having a greater accuracy in the table shown earlier and having a lower standard deviation now, it has unexpected sensitivity and specificity figures. The specificity is much better than the Logistic Regression classifier being at 0.83 which indicates the Random Forest model does better at classifying when there is no tremor. The sensitivity figure on the other hand is rather interesting, as it is much lower than the Logistic Regression classifier being at 0.54. This shows that this model is not good at identifying when a tremor occurs, and in the predictions later it could result in the model having a low number of tremors predicted over a certain time period compared to the Logistic Regression classifier which may overestimate.

Below is the ROC curve for 10-fold CV of a Logistic Regression model. It is clear immediately from the graph that the ROC curve is much more accurate than the one produced for the LOOCV.



The reason for such a clear difference could be a result of the variance of the patients. As mentioned previously, no two patients will behave the same due to experiencing anxiety at different frequencies. It was seen previously that there was either one big outlier as a patient or varying levels of performance depending on the patient. Recall in the table that was produced that leave one out cross validation resulted in a higher bias than a 10-fold cross validation model by an order of magnitude of one for a Logistic Regression classifier. 10-fold cross validation changes this variance as the data is divided equally so it could be expected that portions of the data would behave more similarly than two patients would thus resulting in the fewer gaps between the folds as displayed in the graph.

Below is the ROC curve for 10-fold CV of a Random Forest model.



This curve shows that the Random Forest model performs remarkably well and that all the folds perform consistently well with them tending towards the top left corner. Recall that the 10-fold cross validation for a Random Forest model produced a higher bias than the LOOCV did so this could indicate why the model performs much higher than one would expect. There is a very low standard deviation in the error of 0.09 which further proves that the model has performed very well during this cross validation. We know that the 10-fold cross validation had a much lower bias than the LOOCV, so this evaluation of the models is more reliable. Despite the two styles of cross validation producing similar error estimates, the ROC curves show very different results with regards to how much the true positive rate and false positive rate is impacting the performance of the model. The sensitivity figure is 0.91 and the specificity figure is 0.95 which indicates the model should perform very well but the 10% misclassification on the tremors could prove to be an interesting point in the results. These sensitivity figures give a more reliable estimate to the error because as mentioned before this model with this style of cross validation produces a higher bias estimate.

2.4 Predictions

Having performed various forms of cross validation, the question must be asked about what is expected for the second phase of this project when the models trained here will be tested on the unlabelled accelerometer data. The plan is to use both a Logistic Regression classifier and a Random Forest classifier, so what can be expected and what is expected in terms of the model's performance will be investigated. Given the sensitivity and specificity, we would expect to see the Logistic Regression classifier over-estimate around 10-15% of the tremors and the Random Forest classifier to under-estimate and miss around 10-50% of tremors depending on the similarity to the patients. This means we could expect the models to find different trends depending on how it performs per hour, since they are expected to under and overestimate at certain points. There will likely be a gap between the two curves, but not a significant difference but the main prediction will be that the Logistic Regression classifier will predict a higher frequency than the Random Forest classifier.

Chapter 3

Bayesian analysis

3.1 Linear Mixed Effects Models

It was seen earlier that the imbalance of certain classes like patient ID can cause some issues with the training process as the model performs better on those with a greater portion. As well as this, patients can behave differently to each other. Sampling was used earlier, but another method that can be used to take the imbalance into account is a Linear Mixed Effects Model. Any of the descriptive variables can be used as the chosen mixed effects to model the data, but to get the best results it was decided to use all of them. Inspiration for the methods in this chapter were taken from Bates et al. [2], Lee et al. [9], and UCLA [13] with varying degrees of influence given the task at hand.

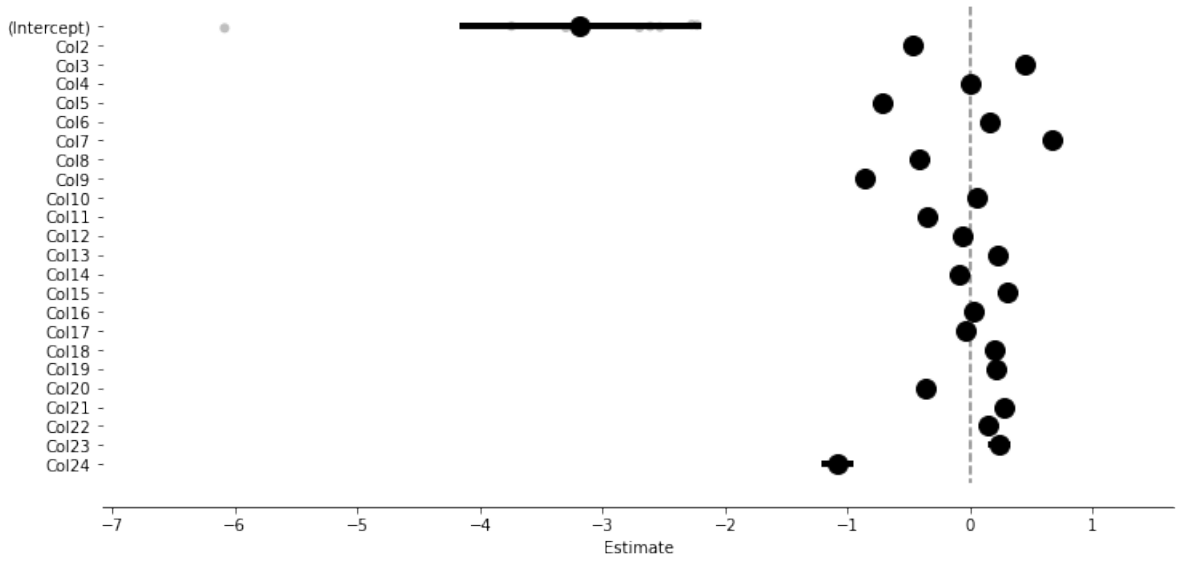
A usual model that is used would be based on variables called fixed effects. With a mixed effects model a new system is made in which certain variables can be used as random effects. In this case, the causal variables can be expressed as these random effects as they are the key separating factors in the data. It is useful to look for correlations within the random effects, for example the medication intake could correlate with the tremor rate. The sensor data remains as the fixed effects as it does not influence itself. In order to apply a mixed effects model to this classification problem, a generalised linear effects model is needed. A generalised linear mixed-effects model can

be expressed as the following:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon \tag{3.1}$$

where $\mathbf{y}$ is the response vector of the tremor values, $\mathbf{X}$ is the matrix fixed effects sensor data, β is the vector of fixed effects regression coefficients, $\mathbf{Z}$ is the design matrix of random effects i.e. the causal variables, $\mathbf{u}$ is the vector of the random effects, and ϵ is a vector of the residual noise that is not explained by the model.

When performing a 10-fold cross validation the model performs very similarly to a standard Logistic Regression classifier. It has a 87% accuracy with a standard deviation on the errors of 0.236 which is slightly lower than what it was with the Logistic Regression classifier which indicates some improvement and while it has a lower standard deviation in the error than the Random Forest classifier does, it does not have a good enough error estimate. It was decided that no prior correlations would be set for the random effects, so perhaps if further research were to be carried out then there could be some correlations set beforehand.



Above is the plot summary of the mixed effects model. It is a forestplot

overlaying estimated coefficients with the random effects that are shown as the intercept bar at the top. The 95% confidence interval bars are shown which is what the long bar on the intercept represents. There are three different random effects so this will explain by there is a long confidence interval bar. The other variables on the y-axis are the sensor variables. Recall that these were reduced using PCA earlier. This can be seen in the variance of the different intercepts, and the figures are not that surprising. As expected, the variance in the patient ID intercept is the highest at 1.49 due to the higher variability between different patients' habits and behaviours with the different causal factors applied. The prototype ID intercept has a lower variance of 0.296 and the medication intake intercept has by far the lowest at 0.04. This makes sense as to why medication intake is the lowest as it is a binary variable and is the most evenly distributed variable in the dataset. Most of the sensor variables have minimal effect on the model but there are certain variables that have a higher effect than the others. Column 24 seems to have the most effect, but also it produces the highest variability compared to the other sensor variables who seemingly have little to no variance in their confidence intervals. Due to the accuracy of the model and how it compares to a seemingly superior Random Forest model, it will be better to test the conventional linear models on the unlabelled data.

Chapter 4

Data without labels

4.1 Data overview

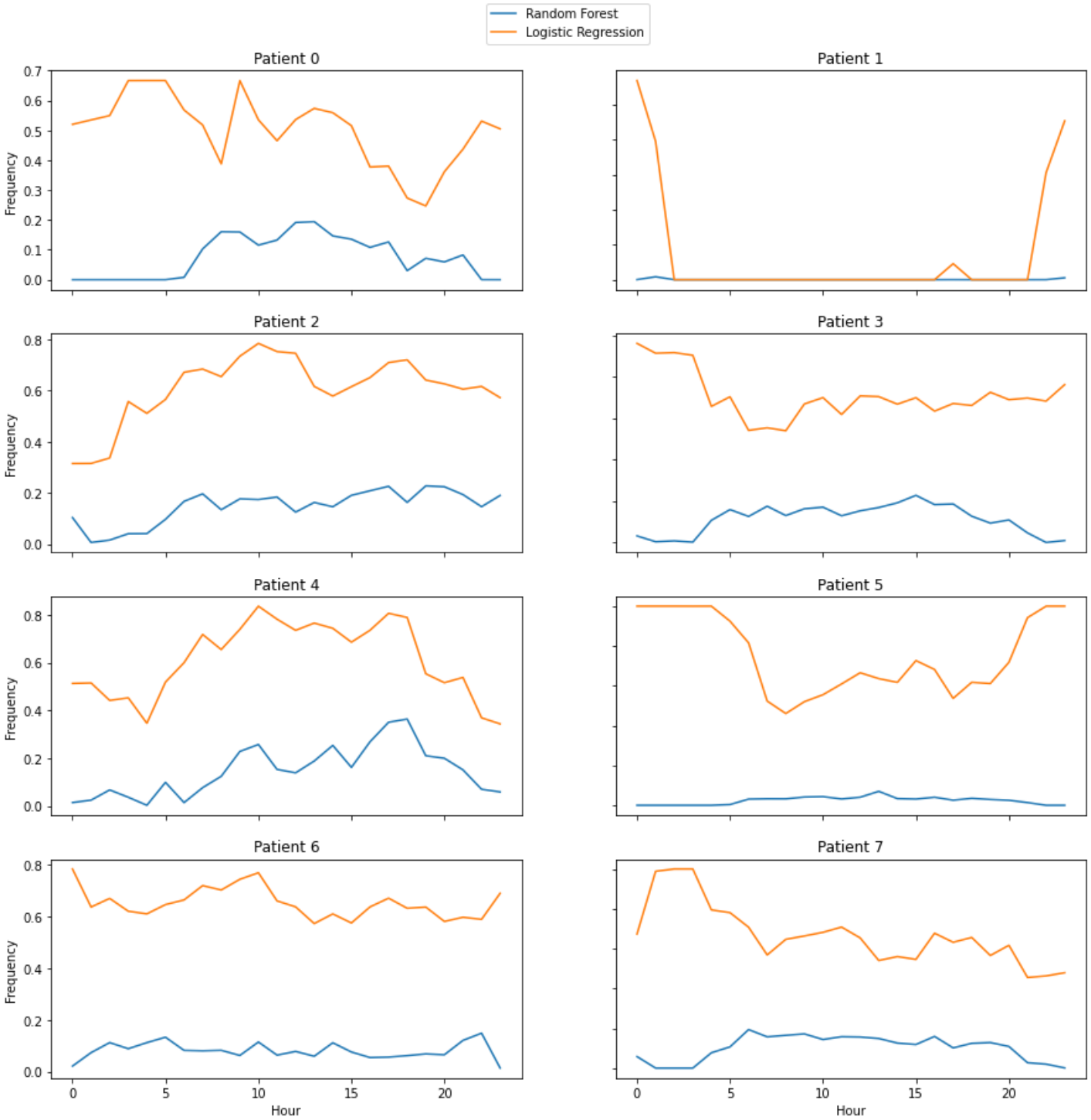
The accelerometer data from the patients will be used to test the models developed previously. The data is structed as before, with 45 features albeit this time as mentioned before there are no causal factors or prototypes included with this as there are no labels. There is a large increase in the number of readings that are being used for testing here. In the initial training set, there are around 75,000 readings, whereas for this stage there are millions of readings to be used for testing the models. Because of this huge difference, there may be some differences in how the model performs but from the training process it should be as general as it can be made. There are three additional time variables that indicate when the measurement was taken. These are a timestamp that indicate the proportion of the day that has passed, a date time object that gives the date and time of the reading, and an indication of which session visit the reading belongs to. Each session visit consists of roughly two hours' worth of readings. These time variables will be useful to the analysis in predicting which hours will have the most/least tremors depending on the patient.

4.2 Performance of trained models

As this data is not labelled, domain knowledge is required to see how likely the predictions the models make are. There is no labelled indication of whether a movement is a tremor or not as well as no causal factors. This highlights why earlier it was important to balance the causal factors in the initial set so that the models would effectively be able to predict over a range of causal factors. Unlike earlier, cross validation cannot be performed on the model with this data as there is no metric that can be used to show how accurate the predictions truly were. As with the previous data the data was whitened using the mean and the same fitted PCA was applied to reduce the data down to 23 components. The models were trained on all the training data that was seen earlier.

At first, it might seem like it makes sense to group these 8 patients together to find an aggregate average over all of them to spot any trends, however this is not a wise move. As was mentioned previously, patients can behave differently and can experience tremors at different points during the day. Anxiety is a leading cause in tremors, and naturally not every patient will experience anxiety at the same points. Some may have spikes in tremors at certain hours where others will not, and this would skew the aggregate average. One would expect most tremors to be during the day for some patients as that is when the most anxiety is experienced, however certain patients may have trouble sleeping due to anxiety so could have more tremors throughout the night. Therefore, it would be better of us to view the results of each patient separately and graph the performance of the models separately. This difference will be seen in the graphs that will be produced.

To show the contrast in the predictions and how different models can vary, a Logistic Regression model and a Random Forest classifier will be used to predict the frequency of tremors in the 8 new patients. It was seen earlier that the nested cross validation accuracy score was highest for the Random Forest, but it will be seen if that will translate well to the predicted frequencies and whether they make sense using domain knowledge.


Random Forest
Logistic Regression
Patient 0
Patient 1
Patient 2
Patient 3
Patient 4
Patient 5
Patient 6
Patient 7
Frequency
Hour

The frequency is the average proportion of the hour's readings that were tremors over the course of two weeks. As was expected from the sensitivity and specificity figures, the Logistic Regression model overestimates a lot more than the Random Forest model does. Therefore, the model mostly predicts that more than half each hour has a tremor happening, and from a clinical point of view this is simply unrealistic. Taking patient 6 for example, over 60% of the day is spent having a tremor which is highly unlikely. As was mentioned earlier, there may be spikes during certain hours which could cause a 60% rate, however maintaining this over the whole day is unrealistic. Despite the high prediction rates that are produced by the Logistic Regression classifier, the model often predicts a similar trend or pattern that the Random Forest classifier predicts. Patients 1, 2, 4, and 6 are good examples are where the models predicted very similar trends. For Patient 1, both models agree that most of the tremors come during the night with almost little to no tremors during the day. This could signify that high anxiety levels are present throughout the night as the patient could have trouble sleeping. The predictions follow what was expected earlier from the sensitivity and specificity figures calculated previously. The Logistic Regression classifier is likely overestimating the number of tremors due to the lower specificity and the Random Forest classifier is likely underestimating the number of tremors due to the lower sensitivity. This matches the graph because, as discussed, the Logistic Regression classifier is predicting an unlikely number of tremors throughout the day. Despite this, it was not predicted that there would be this much difference between the levels of over and under estimating. The Logistic Regression classifier mostly classifies over 10 times the number of tremors per hour that the Random Forest classifier does.

Chapter 5

Summary

During this project there were two main stages carried out, each of which produced various conclusions on the aims and objectives that were set out at the beginning of this project. The completion of these aims and objectives was successful, albeit with some chapters there could be room for further more in depth research. The two main stages mentioned were the training process carried out on the labelled sensor data, and the testing process that was carried out on the unlabelled accelerometer data. Recall that the major aims and objectives of this project were to assess how well models perform on the labelled data using various methods of cross validation as well as investigating alternative models, namely a linear mixed effects model. Moreover, the other main aim and objective was to see how these models performed on the unlabelled data and to look at whether the graphs produced were plausible and if they made any medical sense.

To evaluate the initial models used, two types of cross-validation were utilised. 10-fold and leave one out cross-validation were used to evaluate a Logistic Regression model and a Random Forest model. As expected, the Random Forest model outperformed the Logistic Regression model in both the types of cross-validation used. Using naive cross-validation can have some issues with regards to what it is estimating, so the method of nested cross-validation was used. The key difference between the two types of cross-validation, 10-fold and leave one out, was the bias figures calculated. Due to what was being nested on, it made sense that the 10-fold resulted in the lowest bias figures. To deal with class imbalances among the causal factors, causal bootstrapping

was performed on the data to attempt to construct a more balanced sample. This difference between 10-fold and LOOCV was further seen in the ROC curves due to the differences in the standard deviation of the error.

A generalised linear mixed effects model was utilised to see how much it differed from the conventional models that were used in the initial training process. This differs from conventional models that use purely fixed effects, as we could now introduce the causal factors and random effects that had impact on the intercept of the model. It was concluded that it was not worth attempting to use the model for the final testing phase as it did not outperform a standard Logistic Regression model. If there were to be further research past this project, then more complicated mixed effects models could be utilised to experiment with different results to further investigate the correlation some of the random effects can have on the data.

With the models trained and evaluated using the cross-validation methods mentioned earlier, they were tested on the new batch of unlabelled accelerometer data to see whether the new results could be plausible taking into consideration the sensitivity and specificity levels of the models calculated previously. Some of the things that were predicted from the training process occurred, namely the Logistic Regression classifier overestimating the number of tremors, and the Random Forest classifier underestimating the number of tremors. This prediction was based on the sensitivity and specificity figures calculated during that process. What was not foreseen was the level of overestimation by the Logistic Regression classifier. For a lot of the patients, the model was predicting that they were having tremors for over half of the day which is simply unfeasible. One interesting point though was that for a lot of the patients, the two models agreed on the trends and patterns throughout the day albeit at different frequency levels. If the model were to be improved further following this, then the sensitivity and specificity rates could be factored into the model and the predictions could have taken this into consideration when using them as new parameters.

Appendix A

Appendix

All code and supporting files used for this project can be found in this GitHub page `https://github.com/kingy434/MScProject`. It is a mixture of my own written code as well as code from my supervisor Dr. Raykov. Note that the accelerometer files were too big to upload to my Github page.

Bibliography

[1] Parkinson's disease movement symptoms. `https://www.parkinson.org/Understanding-Parkinsons/Movement-Symptoms`. Accessed: 09-05-2021.

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

[3] Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv preprint arXiv:2104.00673*, 2021.

[4] Lazzaro di Biase, Alessandro Di Santo, Maria Letizia Caminiti, Alfredo De Liso, Syed Ahmar Shah, Lorenzo Ricci, and Vincenzo Di Lazzaro. Gait analysis in parkinson's disease: an overview of the most accurate markers for diagnosis and symptoms monitoring. *Sensors*, 20(12):3529, 2020.

[5] Luc JW Evers, Yordan P Raykov, Jesse H Krijthe, Ana Lígia Silva De Lima, Reham Badawy, Kasper Claes, Tom M Heskes, Max A Little, Marjan J Meinders, and Bastiaan R Bloem. Real-life gait performance as a digital biomarker for motor fluctuations: The parkinson@ home validation study. *Journal of medical Internet research*, 22(10):e19068, 2020.

[6] Nils Hammerla, James Fisher, Peter Andras, Lynn Rochester, Richard Walker, and Thomas Plötz. Pd disease state assessment in naturalistic environments using deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[7] Sungjun Hong, Heesung Lee, and Euntai Kim. Probabilistic gait modelling and recognition. *IET Computer Vision*, 7(1):56–70, 2013.

[8] Abdul Ghaaliq Lalkhen and Anthony McCluskey. Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia Critical Care & Pain*, 8(6):221–223, 12 2008.

[9] Wooyeol Lee and Kevin J Grimm. Generalized linear mixed-effects modeling programs in r for binary outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(5):824–828, 2018.

[10] Max A Little and Reham Badawy. Causal bootstrapping. *arXiv preprint arXiv:1910.09648*, 2019.

[11] Midhun Parakkal Unni, Prathyush P Menon, Mark R Wilson, and Krasimira Tsaneva-Atanasova. Ankle push-off based mathematical model for freezing of gait in parkinson's disease. *Frontiers in bioengineering and biotechnology*, 8:1197, 2020.

[12] Yordan P Raykov, Luc JW Evers, Reham Badawy, Bastiaan R Bloem, Tom M Heskes, Marjan J Meinders, Kasper Claes, and Max A Little. Probabilistic modelling of gait for robust passive monitoring in daily life. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[13] UCLA. Mixed effects logistic regression. `https://stats.idre.ucla.edu/r/dae/mixed-effects-logistic-regression/`. Accessed: 09-05-2021.