

Project and Dissertation for MSc in Data Science - CSC8639

Interim Report

**Machine Learning for Medicine:**

**Predicting clinical outcomes for patients with Leigh syndrome**

Author: Frestie Ngongo

Abstract

• A ‘Structured Abstract’ for your project as a whole.

Context: In one to two sentences summarise the background context to your work; clearly state why it is an important problem to study.

Objective: Describe the purpose of your project, clearly state the problem you set out to investigate in the work you have done.

Method: Summarise the technical approach or apparatus you have used or developed in the project.

Results: Describe the results that you have obtained; be as specific as possible and if appropriate quantify the results.

Novelty: Please clearly state the novel contribution of your work in comparison to previous research and policy in the field.

• An introduction to the problem addressed by your research project, including the motivation for the project and its wider significance.

• A clear statement of the aim and scientific objectives of your work.

• A discussion of background and related work relevant to your problem.

• A description of the research undertaken and its conclusions.

• An appraisal of the contribution that your project makes to the state of the art.

• An assessment of the scope and limits of your work and relevant future work

Introduction

The use of machine learning models is becoming increasingly involved in the clinical setting, including disease diagnosis according to an article by Kononenko 2001. Mitochondrial dysfunction is identified as a particularly heterogenous pathological change (1). The mitochondria are present in every cell and responsible for respiration and producing energy. The mitochondria are made up of phospholipid bilayers and proteins, which form five distinct parts of the mitochondrion: the outer membrane; the intermembrane space; inner membrane; cristae space; the fluid matrix (2).

Mitochondrial diseases have been difficult to diagnose due to the different ways it presents in various individuals, including the many different organs. (3) In addition, there is no single lab or diagnostic test that can be carried out to confirm whether an individual has the disease. (3) Currently, gene analysis is the most reliable method used to confirm mitochondrial disease states in individuals, alongside taking a family history, blood and urine tests, and physical examinations. (4) However, studies have found that 55% of patients with mitochondrial disease were initially misdiagnosed on first admission and of these, 32% were misdiagnosed twice. (5) These figures are alarming and highlight the impact of human error. Introduction of an alternative method of diagnosis could therefore help to solve this issue, i.e., the use of ML.

The dataset used in this study was obtained from a list of individuals with 3 controls and 9 patients. The patients have clinically and genetically characterised mitochondria diseases and supplied muscle samples for this study (6). Control samples were obtained from the distal part of the hamstring of those undergoing anterior cruciate ligament surgery (6). Information regarding the 9 different proteins within the sample, the myofibers locations, the area of the myofibers, cell circularity and perimeter were obtained, and results presented in a table. These factors could be considered and analysed to identify a way to better understand the changes in protein levels observed in patients with a specific mitochondrial disease. The use of Machine Learning Algorithms may help to achieve this. For example, a study looked at developing three ML predictive models for cancer diagnosis and managed to achieve a maximum accuracy of 96% using the support vector machines algorithm (7). This was used to separate the data into two groups - those with cancer and those without. (7)

Machine Learning as a branch of Artificial Intelligence and Computer Science uses algorithms to imitate the way humans learn, while gradually improving its accuracy (8). Being able to predict whether an individual is a patient or a control and what type of disease they may have, is important, but being able to do so accurately is even more essential. Therefore, running algorithms, including support vector machines, multi-classification, artificial and/or deep neural networks and random forest classification on the dataset, and analysing their accuracy and precision could potentially provide a means to make diagnosis of mitochondrial diseases easier and more accurate, regardless of the site affected by the disease.

There are four types of machine learnings: supervised learning (SL); unsupervised learning (UL), semi-supervised learning (SSL), reinforcement learning (RL). For SL models, input and output data are fed into the algorithm, whilst for UL models, they independently identify patterns in the input data and use this to predict an output. In SSL algorithms, they can use a mix of classified and unclassified data to build problem-solving models

For RL models, they make use of a rewards system, in that the algorithms get rewarded for desired actions and punished for undesired actions.

Various steps involved in building Machine Learning models: (<https://www.researchgate.net/publication/351428649_Machine_Learning_in_HealthCare> )

1. Collect Data
2. Pre-process data
3. Transform Data
4. Train the algorithm
5. Test the algorithm
6. Execute

Aim and Objectives

The aim of this project is to design a model to accurately classify fibres into one of two groups; healthy or RC deficient. with mitochondrial disease into one of seven groups including healthy patients i.e. they don’t have the mitochondrial disease.

Specific objectives to help achieve this aim:

1. Identify the best model(s) that can help to classify individuals into those with the disease and those without.
2. Identify key patterns in the characteristics of those with and those without the disease.
3. Identify patient fibres with a pattern different from the pattern in control subjects.
4. Identify the best model that can be used to classify the patients in the dataset into one of the six mitochondrial diseases based on the key patterns.
5. Identify the best algorithm to use to classify new individuals into one of the seven classes (6 diseases and 1 healthy group).

Background and Relevant Work

This project aims to apply machine learning to a study conducted by Warren. C et al., looking at the decoding of mitochondrial heterogeneity in single muscle fibres using imaging mass cytometry (IMC). (<https://www.nature.com/articles/s41598-020-70885-3> ) IMC makes use of the antibody-conjugated isotopes of rare earth metals with laser ablation, and detection using mass cytometry by time-of-flight. (<https://www.jidonline.org/article/S0022-202X(20)32401-5/fulltext> ) It can analyse up to 40 protein markers simultaneously to create images of high definition from a single tissue section. (<https://www.jidonline.org/article/S0022-202X(20)32401-5/fulltext> ) This produced the dataset used for the study by Warren. C et al., and the same dataset used in this project.

The result of the study by Warren. C et al., was that they were able to demonstrate the accurate quantification of protein levels using IMC. From this they accurately measured the deficiency of oxidative phosphorylation for common mitochondrial DNA variants and witnessed a compensatory upregulation in the number of unaffected oxidative phosphorylation components. (<https://www.nature.com/articles/s41598-020-70885-3> )

Taking this research further, machine learning algorithms were applied to the data set, to identify whether the fibres could be classified into two groups – healthy or RC deficient. Machine learning algorithms can process and find patterns in large data sets to enable decision making. Applying this to the dataset, machine learning has the potential to decide to classify the patient fibres into one of the 6 groups.

A recent trial had made use of machine learning algorithms (namely the supervised learning models) to predict those that fall into the at-risk category for COVID-19 in a timely manner, hence reducing death rates. In this study, they identified 20 features they deemed as significant for predicting survival chance of an individual and ran this against SL models (logistic regression, random forest, and extreme gradient boosting). The outcome of the study was that the random forest model outperformed the others. This study was similar to the one carried out in this research, in that specific features were used to predict one thing or the other. <https://www.hindawi.com/journals/sp/2021/5587188/> However, in this project, unsupervised learning algorithms were conducted as clustering has been proven to be powerful for discovering patterns. <https://www.frontiersin.org/articles/10.3389/fncom.2019.00031/full> This method of modelling has also been advantageous as the data passed through the model is unlabelled, so the model is required to identify its own pattern and predict an outcome.

In a study by Hany Alashwal et al., they used multiple clustering algorithms for partitioning patients of Alzheimer’s disease based on their similarity. They used K-Means clustering to identify whether it could classify individuals into the correct bio-profile. They observed that for those with Alzheimer’s disease, more than 69% of them and about half of those with mild cognitive impairment were always assigned to the pathological bio-bioprofile. This led to believe that the K-means algorithm could predict datasets with clinical features into specific labels.

<https://www.frontiersin.org/articles/10.3389/fncom.2019.00031/full>

Methods

Obtaining the Dataset

The data used in this research was obtained from the IMC carried out by Warren. C et al. The contents of the data consisted of features of the different patient’ fibres. This included the area, aspect ratio, perimeter, circularity, and amount of 8 different proteins, all within the fibres. Each fibre’s proteins were further numerically analysed by calculating logs and medians of the amount of each protein in the fibres.

Exploratory Data Analysis

To conduct this research several machine learning (ML) algorithms were used. To begin with, exploratory data analysis was conducted to find out if any of the features in the data showed a trend.

Classification Algorithms

Other ML algorithms used included Random Forest, Decision Tree, K Nearest Neighbour, K Means classifications were also carried out on the data.

<https://www.researchgate.net/publication/351428649_Machine_Learning_in_HealthCare> :

K-Means clustering

The k-Means clustering algorithm (Forgy, 1965) is a classical unsupervised learning method. This algorithm takes n observations and an integer k. The output is a partition of the *n* observations into k sets such that each observation belongs to the cluster with the nearest mean.

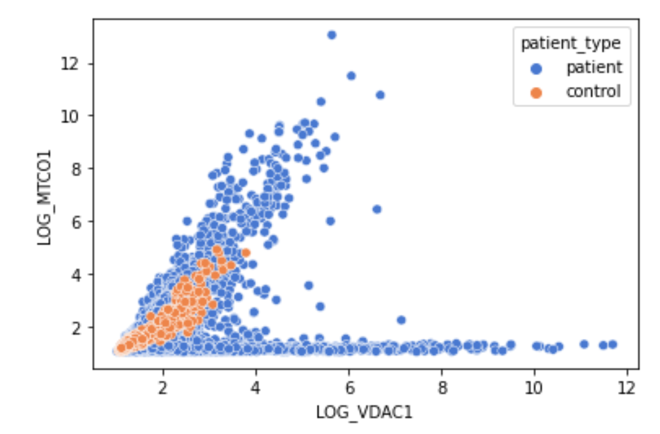
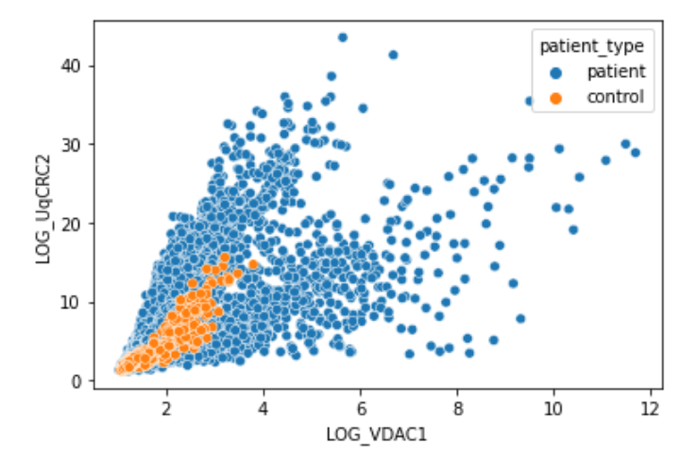
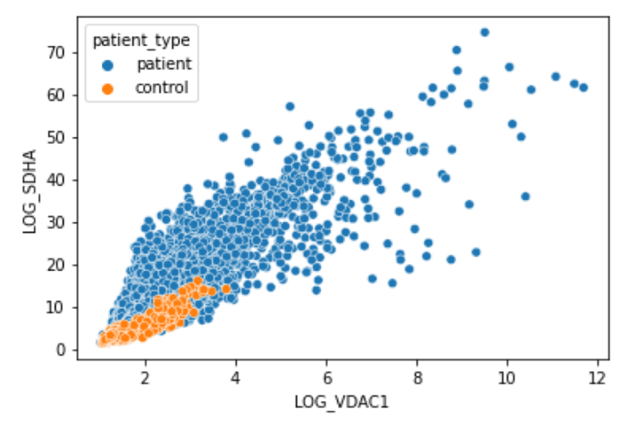
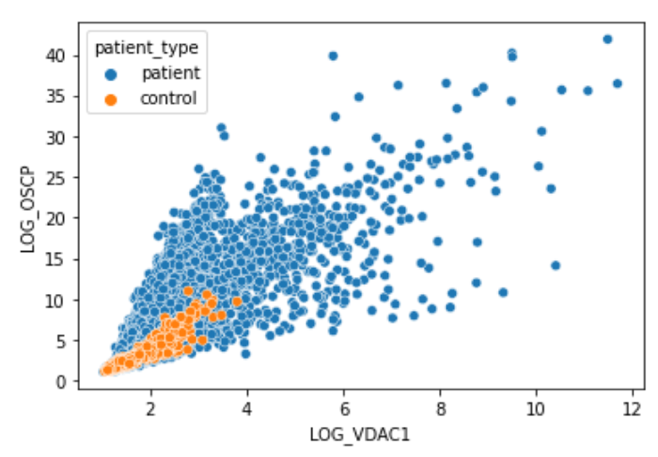
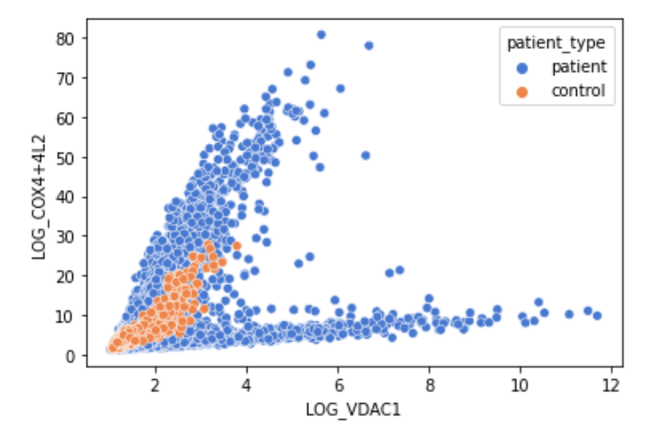
Agile method is a set of principles used for disciplines to avoid particular methods. Agile helps to achieve better results. The agile method helps to adapt and adjust the way of working in a team or individually. Agile procedure like stand-up meetings, product backlog, retrospective these methodologies really help planning and improving task and hand constantly.

During this research project, agile method was used for gathering information and building the prototype. This was helpful, for organisation and having a goal to meet in the end of every fortnight and discuss about the finding (the machine learning models) use to see if it is making sense to mitochondrial disease teams and the supervisor.

As a machine learning engineer machine agile methodology does not make decision but it give a much better way to make useful predictive ML models.

This approach was the best for this project because there was constant change. E.g. using different models to see which ML technique is best for predicting patient and control, proteins and fibres. Agile methodology makes the project flow because in the end there was a presentation of the results found from the research using ml prototype to predict task at hand and improving when requested especially when the graph did not make sense to the mitochondrial disease researchers

**2D Mito Plot**

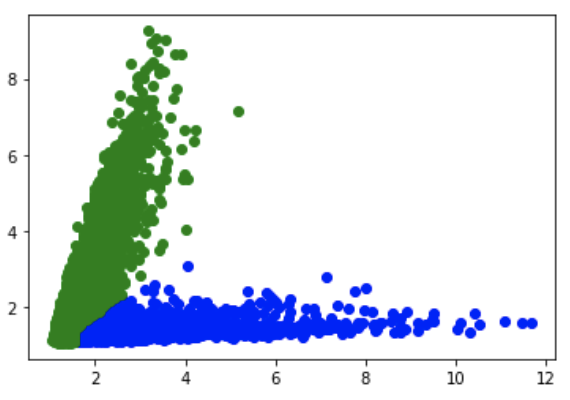
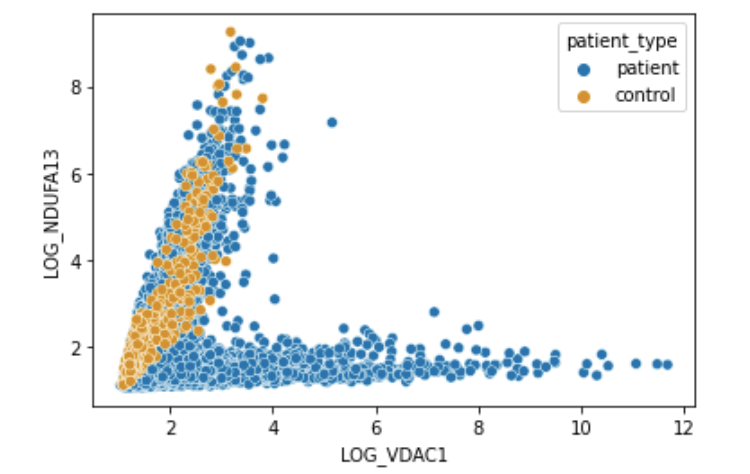
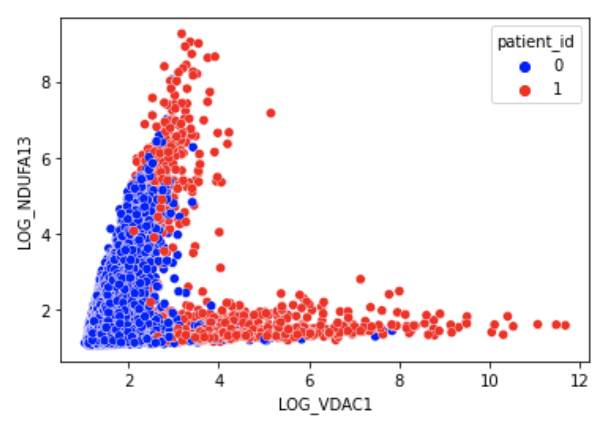


**Figure 1. Plots of 5 different proteins against the LOG\_VDAC1 protein.**

The plots of 5 different proteins against a sixth protein (LOG\_VDAC1), all found in the subjects’ fibres. The plots identify that the control fibres seem to be in the closer to the bottom left corner, highlighting that all these proteins seem to be in smaller amounts in healthy subjects in comparqison to those with RC deficient fibres.

**NDUFA13**

Raw Data K-Means



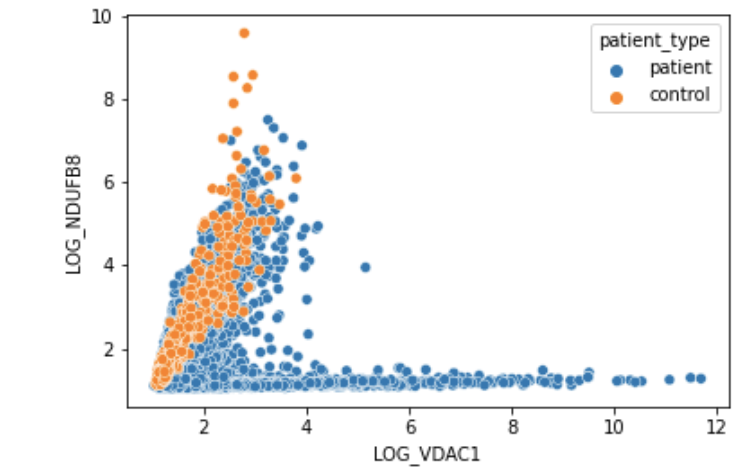
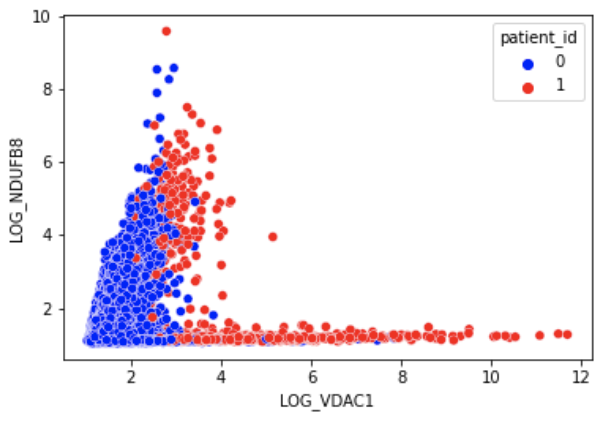
GMM

**Figure 2. Plots of the amounts of NDUFA13 vs the amounts of LOG\_VDAC1 proteins in the raw data, K-Means Model and Gaussian Mixture Model (GMM).**

All 8 proteins were used in both the K-Means and GMM algorithms to produce their respective plots shown above. The results of the K-Means algorithm are almost like the raw data and both plots appear to have a very similar structure. However, in comparison to the GMM, there is a clearer separation of the two clusters: green identifying the control, and blue identifying the patients. Additionally, the point on the GMM (circled) at which the clusters divide appears to be at the same point circled in the raw data.

**NDUFB8**

Raw data K-means



Chart, scatter chart

Description automatically generated

GMM

**Figure 3. Plots of the amounts of NDUFB8 vs the amounts of LOG\_VDAC1 proteins in the raw data, K-Means Model and Gaussian Mixture Model (GMM).**

All 8 proteins were used in both the K-Means and GMM algorithms to produce the plots above. The K-Means produces a plot like that of the raw data, however, there appears to be more control fibres in the 2nd fork of the K-Means than that of the raw data (circled). The GMM cluster seems to only produce one cluster instead of 2.

K-Mean results

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual** | **Proportion of RC deficient Fibres (%)** | **Disease Type** | **Proportion of RC deficient Fibres (%)** |
| C01 | 0% |  |  |
| C02 | 0% | Control | 5.1% |
| C03 | 22.1% |  |  |
| P01 | 79.5% | CI | 83.7% |
| P02 | 89.7% |  |  |
| P03 | 0% | Deletion | 6.5% |
| P04 | 16.6% |  |  |
| P05 | 0.6% |  |  |
| P06 | 33.4% | MT-TL1 | 13.3% |
| P07 | 23.0% |  |  |
| P08 | 15.4% | MT-TG | 15.4% |
| P09 | 5.7% | MT-TE | 5.7% |
| P10 | 62.6% | MT-TW | 62.6% |

**Table 1. Proportion of Reactive Chain (RC) Deficient Fibres in each patient and in each disease from the K-Means Result.**

The algorithm predicted the control to all have mainly healthy fibres, whereas, for patients there is a mixture of healthy and disease fibres. This led us to investigate whether the type of disease had an impact on the number of the fibres.

The CI mutation disease presents with a much larger proportion of RC deficient fibres in comparison to the other fibres. The control and MT-TE disease were both predicted to have a rather similar proportion RC deficient fibres. MT-TW almost has equal amounts of RC deficient and healthy fibres.

**GMM Table of Results part 1: LOG\_VDAC1 & LOG\_NDUFA13**

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual** | **Proportion of RC deficient Fibres (%)** | **Disease Type** | **Proportion of RC deficient Fibres (%)** |
| C01 | 0% |  |  |
| C02 | 0% | Control | 0% |
| C03 | 0% |  |  |
| P01 | 95.5% | CI | 96.7% |
| P02 | 98.3% |  |  |
| P03 | 0.4% | Deletion | 1.6% |
| P04 | 3.4% |  |  |
| P05 | 18.5% |  |  |
| P06 | 9.9 % | MT-TL1 | 21.0% |
| P07 | 39.2% |  |  |
| P08 | 79.8% | MT-TG | 79.8% |
| P09 | 37.2% | MT-TE | 37.2% |
| P10 | 82.0% | MT-TW | 82.0% |

**Table 1. Proportion of Reactive Chain (RC) Deficient Fibres in each patient and in each disease from the GMM Result.**

It was observed that the CI disease type had almost 100% of fibres with RC deficiency. MT-TG and MT-TW both appear to have been predicted to have very similar proportions of RC deficient fibres. Finally, the deletion mutation was predicted to have very few RC deficient fibres based on the patterns identified by the GMM algorithm.

**GMM Table of Results part 2: LOG\_VDAC1 & LOG\_NDUFB8**

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual** | **Proportion of RC deficient Fibres (%)** | **Disease Type** | **Proportion of RC deficient Fibres (%)** |
| C01 | 0% |  |  |
| C02 | 0% | Control | 0% |
| C03 | 0% |  |  |
| P01 | 98.8% | CI | 98.9% |
| P02 | 99.1% |  |  |
| P03 | 2.2% | Deletion | 2.1% |
| P04 | 1.9% |  |  |
| P05 | 20.0% |  |  |
| P06 | 9.0% | MT-TL1 | 25.3% |
| P07 | 55.8% |  |  |
| P08 | 79.6% | MT-TG | 79.6% |
| P09 | 35.6% | MT-TE | 35.6% |
| P10 | 79.8% | MT-TW | 79.8% |

**Table 1. Proportion of Reactive Chain (RC) Deficient Fibres in each patient and in each disease from the GMM Result.**

It was predicted for the CI disease type that almost 100% of the fibres had RC deficiency. MT-TG and MT-TW both appear to have been predicted to have an almost identical proportion of RC deficient fibres, both of which are high in percentage. Finally, the deletion mutation was predicted to have very few RC deficient fibres, with its proportion being close to that of the control fibres. The MT-TL1 and MT-TE mutations both had less than half of the fibres predicted to be RC deficient.

Discussion

Conclusion

Contribution to the States of the Art

Scope and Limits of the Work and Future Work

For future work, might be good to see if there is a way to differentiate between MT-TG and MT-TW disease types considering the NDUFA13 and NDUFB8 proteins predicted them to have similar characteristics as shown in the GMM results table.

Week Plan:

Wednesday – code (overlay – I believe it’s the scaling of the graphs not that it is not using all the fibres. Table – done!!!) and results are done!!!!

Thursday – intro and aim

Sunday – code/discussion and conclusion

Monday – submit 1st draft