

Project and Dissertation for MSc in Data Science - CSC8639

**Machine Learning for Medicine:**

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

Author: Frestie Ngongo

1. 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.

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

• 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

2. Introduction

Mitochondrial dysfunction is identified as a particularly heterogenous pathological change (1). 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.

The use of Artificial Intelligence (AI) techniques such as machine learning (ML) may be able to improve the method of diagnosing mitochondrial dysfunction.

The dataset used in this study was obtained from a list of individuals within 3 controls and 9 patients. Imaging Mass Cytometry (IMC) was used to analyse the proteins in these samples. IMC works to analyse up to 40 protein markers simultaneously, using metal-labelled antibodies with laser ablation, followed by detection using mass cytometry by time-of-flight. (<https://www.jidonline.org/article/S0022-202X(20)32401-5/fulltext>) This was used to produce images that were then used to quantify the protein mean intensities. The proteins looked at were NDUFA13, NDUFB8, VDAC1, COX4+4L2, OSCP, MTCO1, SDHA, UqCRC2. The VDAC1 protein is known to be a mitochondrial mass marker (<https://www.nature.com/articles/s41598-020-70885-3>) and was used as a standard for creating a ratio with the other proteins.

The patients in this study had 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 8 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. The main focus was the proteins for this project because they were more realistic factors that could be playing a role in the diagnosis of whether an individual had the mitochondrial disease(patient) or they did not (control).

Mitochondria are present in every cell and responsible for respiration and producing energy. This occurs via the last step in respiration in a process called oxidative phosphorylation – where ATP (the molecule used for energy) is formed from the transfer of electrons from NADH or FADH2 to O2. However, the heterogenous expression of oxidative phosphorylation proteins, and resulting respiratory deficiency are characteristics found in fibres with a mitochondrial dysfunction.

The use of Machine Learning Algorithms may help to analyse these proteins and determine whether they are useful markers for diagnosis of mitochondrial diseases. 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 models are becoming increasingly involved in the clinical setting, including disease diagnosis according to an article by Kononenko 2001. 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 has the mitochondria disease or not and going further to identify what variation of the disease they may have, is important, but being able to do so accurately is even more essential. Therefore, running algorithms, including clustering algorithms, on the dataset, and analysing their accuracy and precision could potentially provide a means to make diagnosis of mitochondrial diseases easier and accurate.

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. SSL algorithms, 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. RL models, make use of a rewards system, in that the algorithms get rewarded for desired actions and punished for undesired actions. The clustering algorithms used in this project make use of unsupervised learning. This is ideal for this project as the model can predict what category to place the individual based on patterns identified by the model, hence aiding to see how accurately the model can be used in diagnosis of mitochondrial diseases. (<https://www.researchgate.net/publication/351428649_Machine_Learning_in_HealthCare>

**3. Aim**

The aim of this project is to design a model to accurately classify fibres into one of two groups: healthy or RC deficient. Then going further to classify those with mitochondrial disease into one of six groups.e. they don’t have the mitochondrial disease.

**4. 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>

**5. Methods**

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.

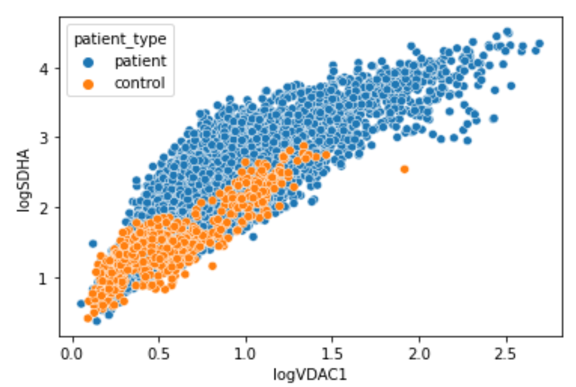
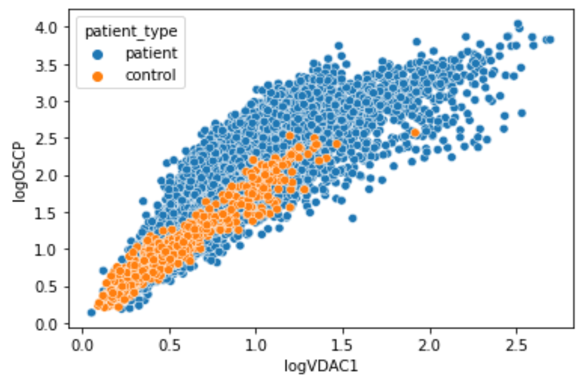
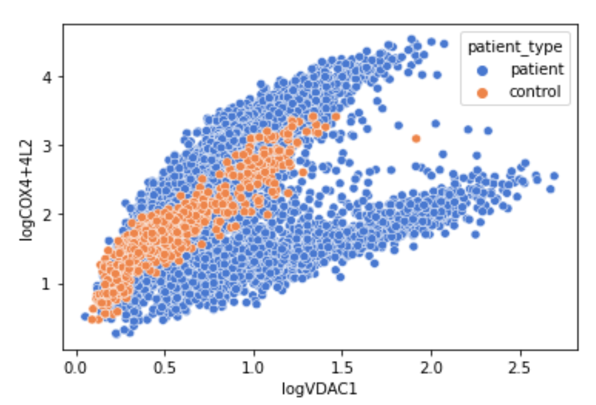
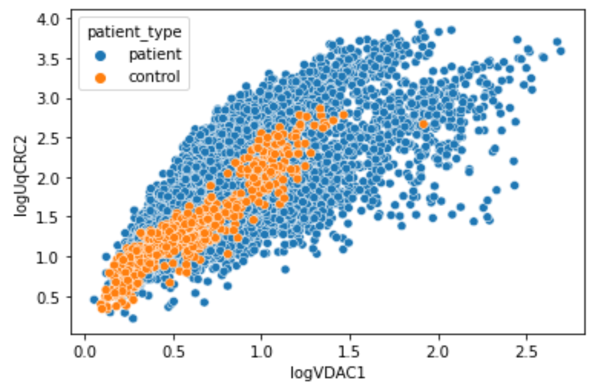
As a machine learning engineer machine agile methodology does not make decision but it give a much better way to make useful predictive ML models.

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

**6. Results**

To carry out this research, the logs of the mean intensities of the proteins were calculated. This was the values used in the analysis and modelling.

**6. 1 2D Mito Plot**

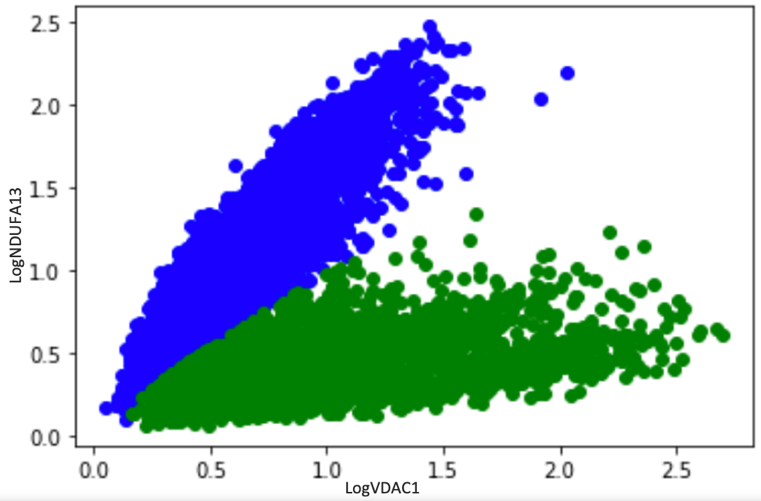
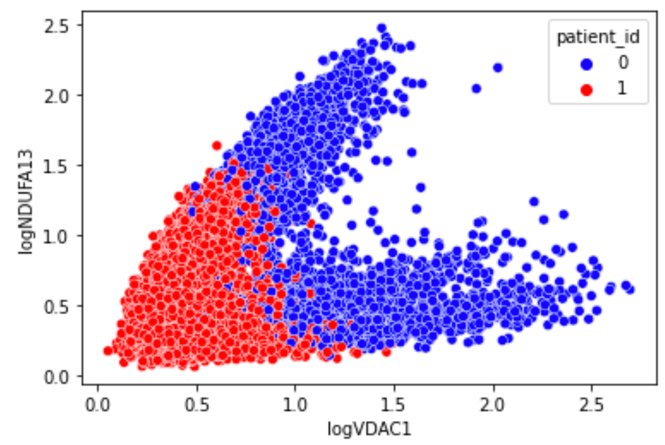
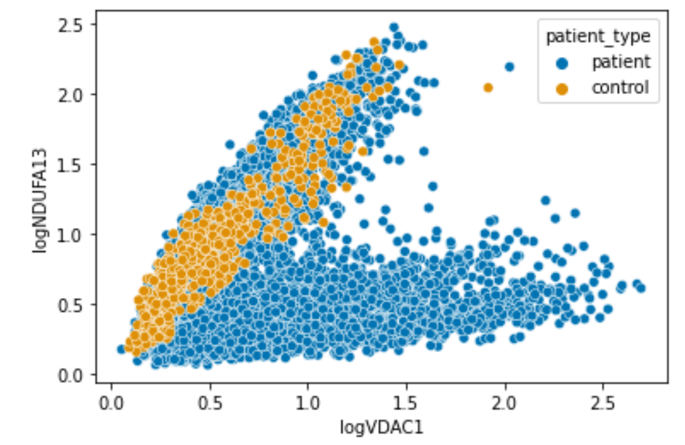


**Figure 1. Plots of 5 different proteins against the LOG\_VDAC1 protein.**

The plots of 5 different proteins against a sixth protein (LOG\_VDAC1), that are found in the individuals’ fibres. The plots identify that the control fibres seem to be less spread out than the patients’ fibres. Both patient and control fibres appear to have a positive correlation for all the proteins; as the logVDAC1 increases, so does the logX protein, where X is any of the other proteins.

**6.2 NDUFA13**

Raw Data K-Means



LogVDAC1

LogNDUFA13

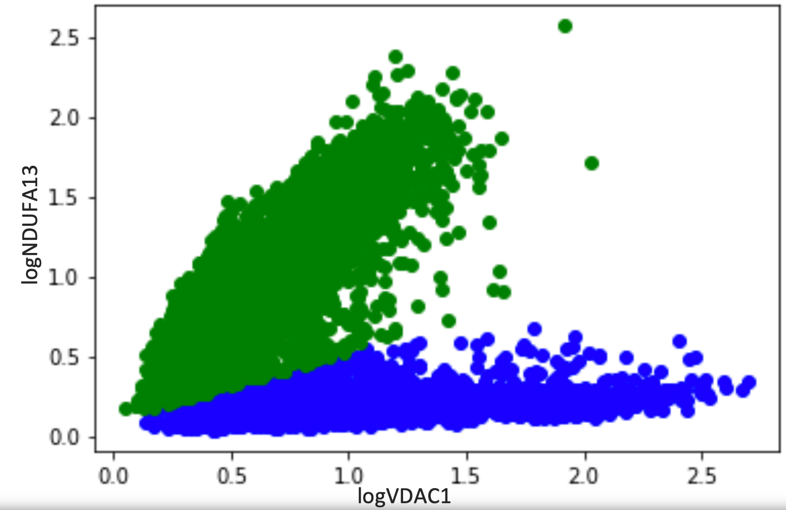
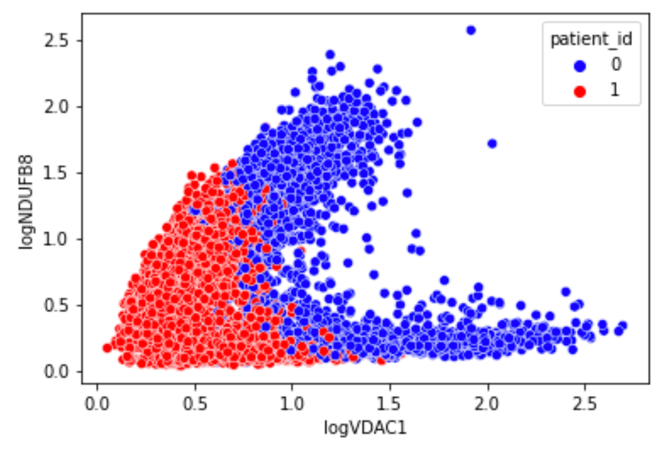
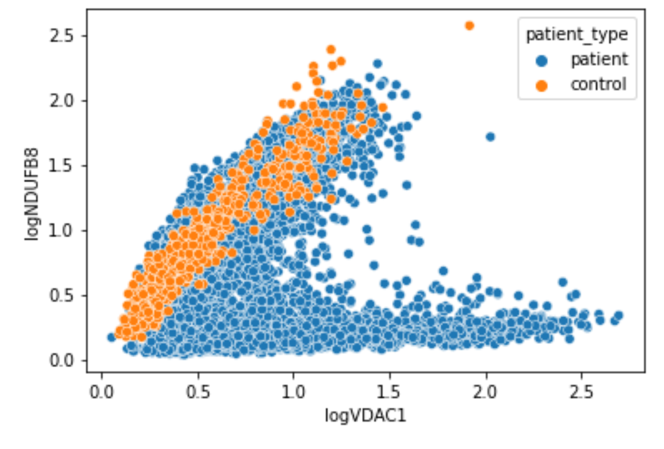
GMM

**Figure 2. Plots of the amounts of logNDUFA13 vs the amounts of logVDAC1 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 showed no distinct separation in the clusters in comparison to the GMM model. The control group appeared to be in the lower left corner of the graph of the KMeans, whilst the results for patients showed them in the higher end of the graph. For GMM, there was a clear separation with an observation of 2 distinct clusters forming the ‘V’ shape; control in blue, patients in green.

**6.3 NDUFB8**

Raw data K-means



logNDUFB8

GMM

logVDAC1

logNDUFA13

**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 produced a plot with no clear separation of the 2 clusters. The GMM cluster was observed to have produced 2 clear clusters, control in blue, patients in green.

**6.4 K-Mean results**

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual** | **Proportion of RC deficient Fibres (%)** | **Disease Type** | **Proportion of RC deficient Fibres (%)** |
| C01 | 1.4% |  |  |
| C02 | 0% | Control | 19.0% |
| C03 | 80.9% |  |  |
| P01 | 97.6% | CI | 98.1% |
| P02 | 98.7% |  |  |
| P03 | 0.8% | Deletion | 33.8% |
| P04 | 84.8% |  |  |
| P05 | 20.1% |  |  |
| P06 | 76.4% | MT-TL1 | 39.8% |
| P07 | 49.5% |  |  |
| P08 | 54.8% | MT-TG | 54.8% |
| P09 | 50.2% | MT-TE | 50.2% |
| P10 | 80.9% | MT-TW | 80.9% |

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

The algorithm was ran using all 8 proteins as predictors. It predicted controls 1 and 2 to have a low to zero proportion of reactive chain (RC) deficient fibres, whilst for control 3, they were predicted to have a high percentage of RC deficient fibres.

Patients 1 and 2, who have the CI disease variation, were predicted to have almost 100% RC deficient fibres. This was followed by those with the MT-TW disease variant, who were predicted to have around 80% of their fibres as deficient.

Patients 3 and 4, although they had the same disease, the algorithm predicted them to have differing amounts of RC deficient fibres. Likewise, the same was true for patients 5, 6 and 7.

Those with MT-TG and MT-TE mutations had almost 50% RC deficient fibres and 50% healthy fibres.

**6.5 GMM Table of Results part 1: LogVDAC1 & LogNDUFA13**

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual** | **Proportion of RC deficient Fibres (%)** | **Disease Type** | **Proportion of RC deficient Fibres (%)** |
| C01 | 0% |  |  |
| C02 | 0.3% | Control | 0.2% |
| C03 | 0% |  |  |
| P01 | 100% | CI | 100% |
| P02 | 100% |  |  |
| P03 | 2.6% | Deletion | 1.5% |
| P04 | 3.4% |  |  |
| P05 | 21.4% |  |  |
| P06 | 9.5 % | MT-TL1 | 28.2% |
| P07 | 65.3% |  |  |
| P08 | 83.6% | MT-TG | 83.6% |
| P09 | 37.7% | MT-TE | 37.7% |
| P10 | 80.9% | MT-TW | 80.9% |

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

The GMM model was ran using only the logNDUFA13 and logVDAC1 proteins and produced almost 0% of the controls as having RC deficient fibres. It also predicted patients 1 and 2 to have 100% RC deficient fibres. Patient 8 (MT-TG) and patient 10 (MT-TW) were both predicted to have around 80% of their fibres as RC deficient. The deletion patients both have a very low amount of RC deficient fibres averaging 1.5%. The patients with the MT-TL1 disease were predicted to have varying proportions of deficient fibres.

**6.6 GMM Table of Results part 2: LogVDAC1 & LogNDUFB8**

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual** | **Proportion of RC deficient Fibres (%)** | **Disease Type** | **Proportion of RC deficient Fibres (%)** |
| C01 | 0.7% |  |  |
| C02 | 2.1% | Control | 1.4% |
| C03 | 0% |  |  |
| P01 | 99.7% | CI | 99.8% |
| P02 | 100% |  |  |
| P03 | 3.8% | Deletion | 4.0% |
| P04 | 4.3% |  |  |
| P05 | 23.7% |  |  |
| P06 | 11.3% | MT-TL1 | 30.7% |
| P07 | 68.9% |  |  |
| P08 | 87.3% | MT-TG | 87.3% |
| P09 | 42.5% | MT-TE | 42.5% |
| P10 | 83.6% | MT-TW | 83.6% |

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

The GMM model was ran using only the logNDUFB8 and logVDAC1 proteins and predicted controls 1 and 2 to have a low proportion of RC deficient fibres. It also predicted patients 1 and 2 to have close to 100% RC deficient fibres. Patient 8 (MT-TG) and patient 10 (MT-TW) were both predicted to have over 80% of their fibres as RC deficient. The deletion patients were both predicted to have a low amount of RC deficient fibres averaging 4.0%. The patients with the MT-TL1 disease were predicted to have a similar proportion of deficient fibres for patients 5 and 6, but patient 7 was significantly different from the rest.

7. Discussion

8. Conclusion

9. Contribution to the States of the Art

10. 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:

Monday – Abstract, Aim, method, discussion, conclusion and submit draft?

Tuesday – Try and do a line graph like they asked for?

**Draft:**

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).

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.

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).

(insert image of mitochondrion)

Table

Description automatically generated