

**Project and Dissertation for MSc in Data Science CSC8639 Interim Report**

**Machine Learning for Medicine:**

**Predicting Clinical Diagnosis of Individuals with Mitochondrial Diseases**

**Author: Frestie Ngongo**

**1. Introduction**

The use of machine learning models on dataset is becoming increasingly involved in the clinical setting, including training several Machine Learning (ML) models on the dataset of patients with different mitochondrial diseases. Mitochondrial dysfunction is identified as a particularly heterogenous pathological change. (1) The mitochondria are a part of an organism's cell and is responsible for respiration and producing energy. It is 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.

Previously, identification of any patterns of the composition and/or characteristics of the mitochondria have been investigated to see how this could help with the diagnosis of individuals. The dataset used in this study was obtained from a list of individuals with 3 controls and 9 patients. Individuals with clinically and genetically characterised mitochondria diseases supplied samples for this study from skeletal muscle biopsies. (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 experimented on to identify a way to diagnose the individuals with a specific mitochondrial disease from the 6 classes. 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.

**2. Aim and Objectives**

The aim of this project is to design a model to accurately classify individuals 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).

**3. Overview of Progress**

Initial understanding of the data has been accomplished plus initial analysis of the data. There is an understanding of what the data is being used to predict, which is to be able to classify individuals into patients and controls based on patterns in their characteristics.

Following on, some exploratory data analysis on the data was carried out, trying to identify whether there is any difference in the characteristics of the control and patient cells, looking at the area, perimeter, proteins etc. From this, it was found that there was a difference in the circularity of those with the disease and those without. Additionally, it was observed that the log of proteins NDUFA13 and NDUFB8, showed some variation between those with mitochondrial disease and those without.

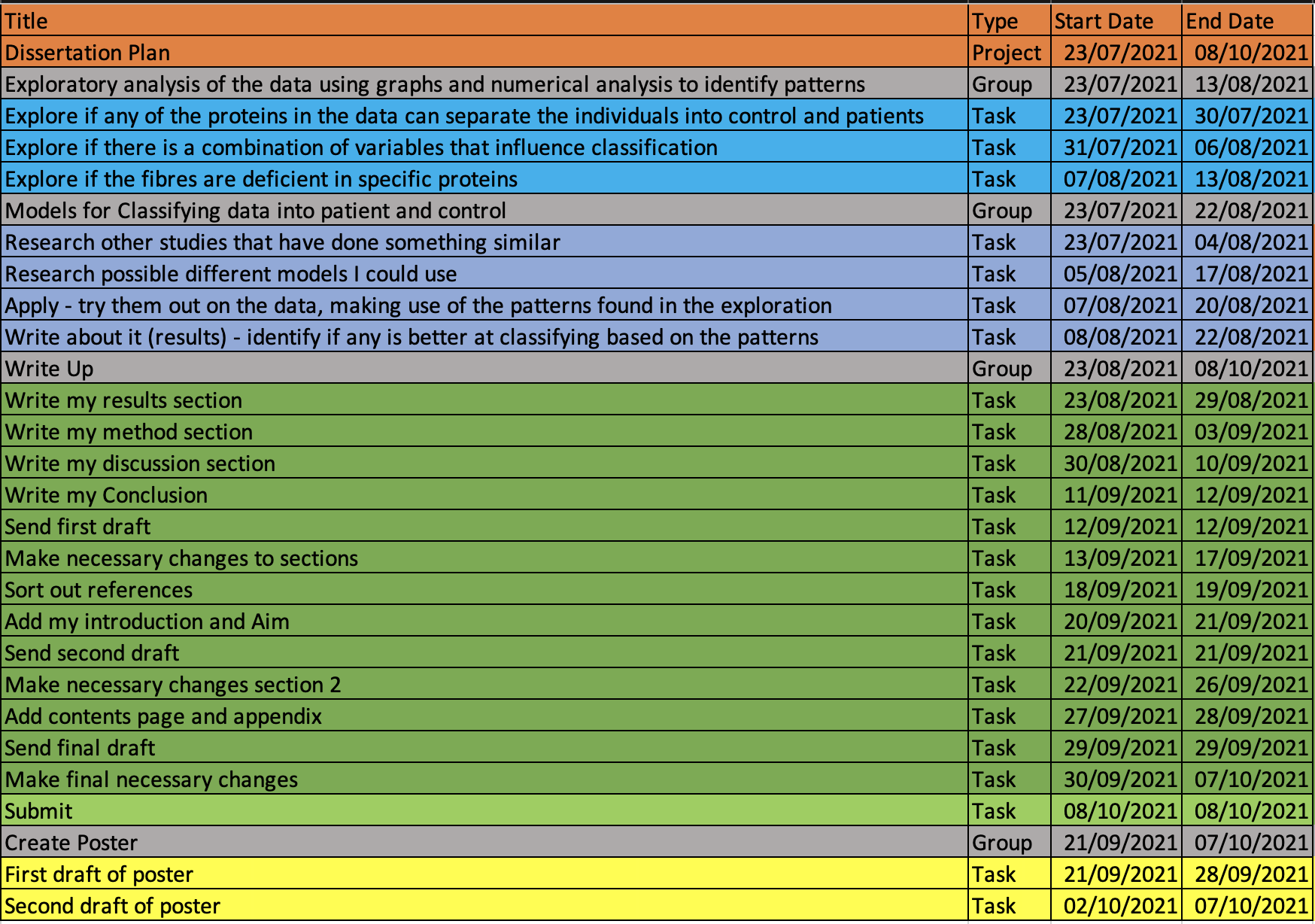
Additionally, research into the different machine learning models was conducted including Random Forest Classification; investigating its ability to predict a control from a patient from the data set. The aim of this was that it used the proteins, rather than the other factors, to predict whether the individual had the disease or not. When used, the model was found to make predictions mainly from the area, perimeter, and circularity. These factors tend to be different based on cell location and from patient to patient. Considering this, these variables were not favoured as accurate predictors of an individual's disease classification.

Another classification method researched was the multiclass classification. This type of classification could be useful for this dataset when grouping individuals into one of the six different mitochondrial diseases. It may help to remove ambiguity and provides definitive groups i.e., an individual has to fall into one group, they cannot belong to several groups. Research into the imbalanced dataset classification was also carried out. This refers to data where the classes are not represented equally. In relation to the sample, the control group only consisted of 3 individuals as found in the exploratory data analysis, however the patient group consisted of 9 individuals. Additionally, there were varying sample sizes for both

groups, therefore encouraging me to consider using this classification for the dataset. For both classifications, a sequential, random forest and linear regression model could be used.

Finally, looking at using just the LOG\_NDUFA13, the LOG\_NDUFB8 alone, the circularity alone and the three variables combined to see if the Random Forest classifier was able to accurately classify the individuals into patients and controls based on these variables. The best outcome was with the three variables combined with a precision of >0.95 for both control and patient groups. The two proteins played roles in this classification, however the protein with the highest ranking was LOG\_NDUFA13.

**4. Project Plan**



**5. References**

1. Alston, Charlotte L et al. “The genetics and pathology of mitochondrial disease.” The Journal of pathology vol. 241,2 (2017): 236-250. doi:10.1002/path.4809
2. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. The Mitochondrion.Available from: https://www.ncbi.nlm.nih.gov/books/NBK26894/
3. Koenig, Mary Kay. “Presentation and diagnosis of mitochondrial disorders in children.” Pediatric neurology vol. 38,5 (2008): 305-13. doi:10.1016/j.pediatrneurol.2007.12.001
4. Cleveland Clinic. 2021. Mitochondrial Diseases: Causes, Symptoms, Diagnosis & Treatment. [online] Available at: <https://my.clevelandclinic.org/health/diseases/15612-mitochondrial-diseases>
5. Columbia University's Mailman School of Public Health. "Mitochondrial disease patients face difficult road to diagnosis: On average, patients see more than eight physicians, undergo multiple tests, and receive misdiagnoses before finally being diagnosed with a mitochondrial disease." ScienceDaily. ScienceDaily, 26 March 2018. <www.sciencedaily.com/releases/2018/03/180326161004.htm>.
6. Warren, C., McDonald, D., Capaldi, R. et al. Decoding mitochondrial heterogeneity in single muscle fibres by imaging mass cytometry. Sci Rep 10, 15336 (2020). https://doi.org/10.1038/s41598-020-70885-3
7. Sidey-Gibbons, J., Sidey-Gibbons, C. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 19, 64 (2019). https://doi.org/10.1186/s12874-019-0681-4
8. Education, I., 2021. What is Machine Learning?. [online] Ibm.com. Available at: <https://www.ibm.com/cloud/learn/machine-learning>

|  |  |  |  |
| --- | --- | --- | --- |
| **0. Project title, author, version and date** | | | |
| *Project: Machine Learning for Medicine: Predicting Clinical Diagnosis of individuals with Mitochondrial Diseases* | | | |
| *Author: Frestie Ngongo* | | *Version: 1* | *Date: 15/07/2021* |
| **1. Description of the data** | | | |
| **1.1 Type of study**  *The data was collected for a clinical study using machine learning algorithms. It will help to try to use artificial intelligence techniques to help with the diagnosis of individuals with mitochondrial disease and potentially diagnose what class of the disease they may have.*  **1.2 Assessment of existing data**  *The data currently being used consists of a list of variables including details about 9 different proteins, the cell circularity and area, myofiber locations and the classification of each individual. Currently I have not been able to identify any gaps in the data.*  **1.3 Types of data**  *The data in this study is being used by the different Machine Learning Algorithms and are all quantitative generated from analysis of musculoskeletal biopsies from individuals. Python will be used to help to process the data using several classifications such as binary class, multiclass and imbalanced data classification. Sequential modelling and Random Forest may be used to test these classifications and to see how they help to reach the desired aim. There is also qualitative data telling us which individual is a control, and who has what type of mitochondrial disease.*    **1.4 Format and scale of the data**  *The data will be in an online format through GitHub and the code to process it will be written in python. The data consists of 3 control individuals each with varying myofiber counts, plus 9 patients that have the mitochondrial disease. Being able to access the data online through GitHub as well as my python code makes it easy to share with my supervisors.* | | | |
| **2. Data collection / generation** | | | |
| *Focus on the good practice and standards for ensuring new data are of high quality and processing is well documented.*  **2.1 Methodologies for data collection / generation**  *The data was already collected in a previous study investigating whether mitochondrial diseases were heterogeneous using imaging mass cytometry. I will be using the data for analysis to create a model to successfully classify individuals as having the mitochondrial disease based on any found patterns.*  **2.2 Data quality and standards**  *To check that my processing of the data is accurate, I will be carrying out tests on the model to provide me with the precision scores of the models. I will also validate my work by use of peer reviews with the PhD students assigned to me as well. All of which should help me to ensure I use the best models and maintain high standards.* | | | |
| **3. Data management, documentation and curation** | | | |
| *Focus on principles, systems and major standards. Focus on the main kind(s) of study data. Give brief examples and avoid long lists.*  **3.1 Managing, storing and curating data.**  *The data in this project will be stored on my computer, on a USB stick, pushed to my GitHub account and also saved on my google drive. This way I would be able to retrieve data that is backed up if I lose my work in one place. Also, having multiple locations enables me to always have a means of accessing my work no matter where I am. For example, I could access GitHub on any device so long as I know my login details. Also, creating several folders within one dissertation folder, to organise different aspects of the project, would be ideal. This makes it easier to work on a specific part of the project.*  **3.2 Metadata standards and data documentation**  *Creating a readme.txt file of the work done would be useful to be able to share the data with not only my team but others outside my team. This will enable them to be able to load the data and understand how the code used to analyse the data would work. Also, documenting the method used to obtain the data, the code used to analyse the data and the outcome from this analysis could be produced.* | | | |
| **4. Data security and confidentiality of potentially disclosive information** | | | |
| *This section should be completed if your research data includes* ***personal data relating to human participants in research****. For other research, the safeguarding and security of data should also be considered. Information provided will be in line with your ethical review. Please note this section concerns protecting the data, not any potential patients.*  **4.1 Main risks to data security**  *The data used in this research contains patient information, however the patients are not identifiable as they have been anonymised to have numerical identities rather than their names. The patient’s data is also protected from unknown users by only sharing it with those that are working on the project with me. The data was obtained with patient consent in a previous study and will be protected from outside use as it is only saved on GitHub. Therefore, by having a secure GitHub password, this reduces the risk of my account being hacked. Also, by making my repository private, it would protect the data from being seen by others without permission.* | | | |
| **5. Data sharing and access** | | | |
| *Identify any data repository(-ies) that are, or will be, entrusted with storing, curating and/or sharing data from your study, where they exist for particular disciplinary domains or data types.* [*Information on repositories is available here*](https://www.ncl.ac.uk/library/academics-and-researchers/research/rdm/sharing/)*[.](https://www.ncl.ac.uk/library/academics-and-researchers/research/rdm/sharing/)*  **5.1 Suitability for sharing**  *I would say it is suitable for sharing to an extent because although it is obtained from patients, the data has been anonymised and may not make sense to outside users. Therefore, sharing the data would not breach any form of data protection.*  **5.2 Discovery by potential users of the research data**  *Potential new users may find out about my data by me making my GitHub repository public so they can access my report and read it. Here they will also be able to access the data for the project and all other necessary parts.*  **5.3 Data preservation strategy and standards**  *The data will be shared for my supervisors throughout the project. However, no part will be made available for the public unless I allow it and it is necessary to do so.*  **5.4 Restrictions or delays to sharing, with planned actions to limit such restrictions**  *The patient ID has been anonymised to reduce risks of identification and permission has been granted prior to this project for the data to be shared with those working on the project. The data should also be kept confidential due to copyright risks for those that are carrying out the study. To avoid unknown users from stealing our research ideas, this project will only be shared between myself and my supervisors and through the use of GitHub, the repository will be made private to outsiders and only available to those on the project with me.* | | | |
| **6. Responsibilities and Resources** | | | |
| *To help me fulfil the plan, I need to ensure I have regular meetings with my supervisors. 6Additionally, I need to make sure I learn to use several ML algorithms so I can implement them in my work. I will need to ensure I have enough storage in my google drive and on my laptop for all the work I am doing. One resource that I believe would help me a great deal is the Outlook calendar. This would help me to be reasonable with my time, planning my tasks daily/weekly so that I can stay on track.* | | | |
| **7. Relevant institutional, departmental or study policies on data sharing and data security** | | | |
| **Policy** | **URL or Reference** | | |
| Data Management Policy & Procedures | <https://www.ncl.ac.uk/media/wwwnclacuk/research/files/ResearchDataManagementPolicy.pdf> | | |
| Information Security | [*https://services.ncl.ac.uk/itservice/policies/InformationSecurityPolicy-v2\_1.pdf*](https://services.ncl.ac.uk/itservice/policies/InformationSecurityPolicy-v2_1.pdf) | | |
| Other |  | | |