• 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

Abstract

Introduction

The use of machine learning models 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 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.

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 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).

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> )

* Research about ML/classification in health
* Research about ML/classification in mitochondria
* Research about classification in general

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> :

Random Forest Classification was initially carried out on the data. This method of classification makes use of randomly sorting the data into test and train sets. The vivid Decision trees of a Random Forest are trained using the various parts of the training dataset. For classifying a new model, the input data of that sample is essential to pass down with every Decision Tree. Every individual Decision Tree gives their own classification output. Then the one which is most classified by the different decision trees is selected (I.e., that types are voted the most) and that classification is given as the output of the Random Forest. Since the Random Forest algorithm takes the output from many every Decision tree, there is reduction in variance.

Decision Tree… - Decision trees which are extreme big may results in overfitting of the training data, which results a huge difference in classification output for a minor change in the input value. They are extreme reactive to their training data, which may lead to having more errors in the test dataset(value).

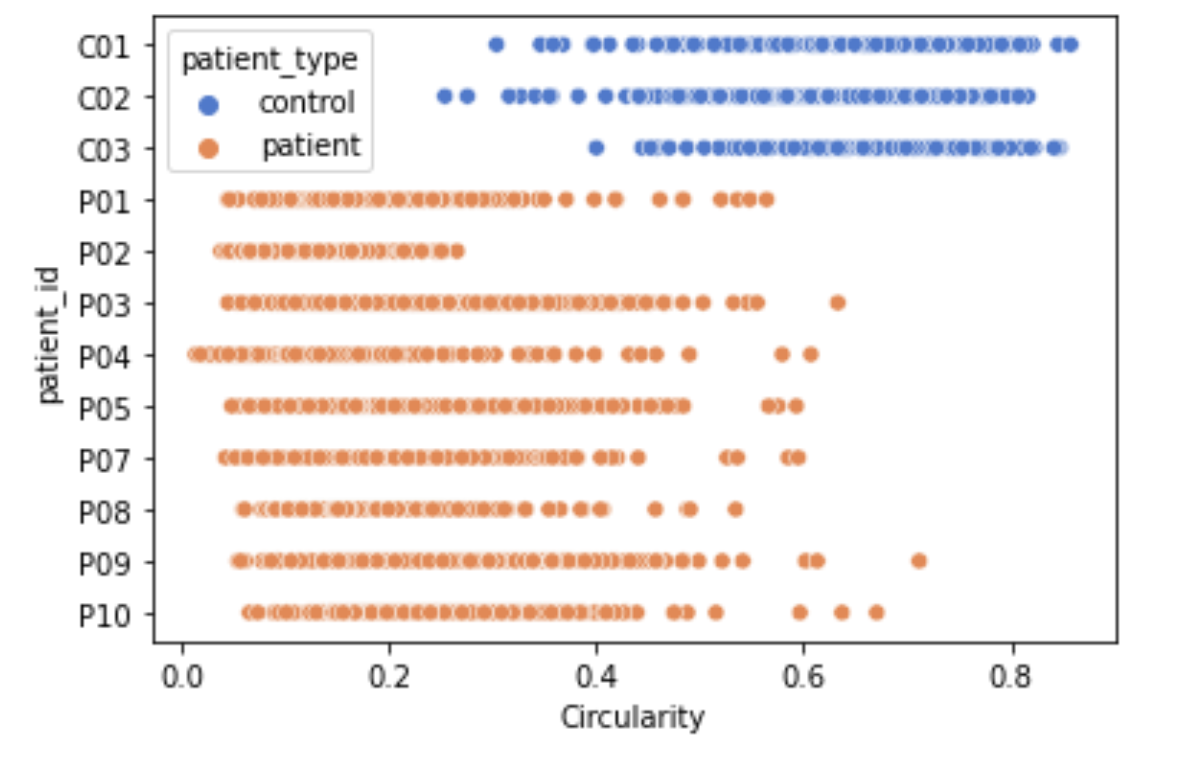
K-Nearest Neighbour Classification… The K-nearest neighbour (KNN) algorithm is one of the simplest and earliest classification algorithms. It is a simple machine learning algorithm which is helps in solving classification as well as regression problems. Although, this algorithm is not tough to implement, but it possesses some major downsides that is becomes extremely slow when the size of the data increases. K-nearest neighbour can be considered as the simple form of Naïve Bayes classification. In contrast Naïve Bayes process, the KNN algorithm does not need considering the possible values. The ‘K’ is the KNN algorithm is the total number of nearest neighbours who are eligible to take ‘vote’ from. The selection of vivid values for ‘K’ can give vivid results. Figure 12 illustrates how the K-nearest neighbour performs in classifying a new object. For k=3, the new element (big black star) is classified as ‘black’; whereas, it has been classified as ‘red’ when k=5

K-Means clustering…

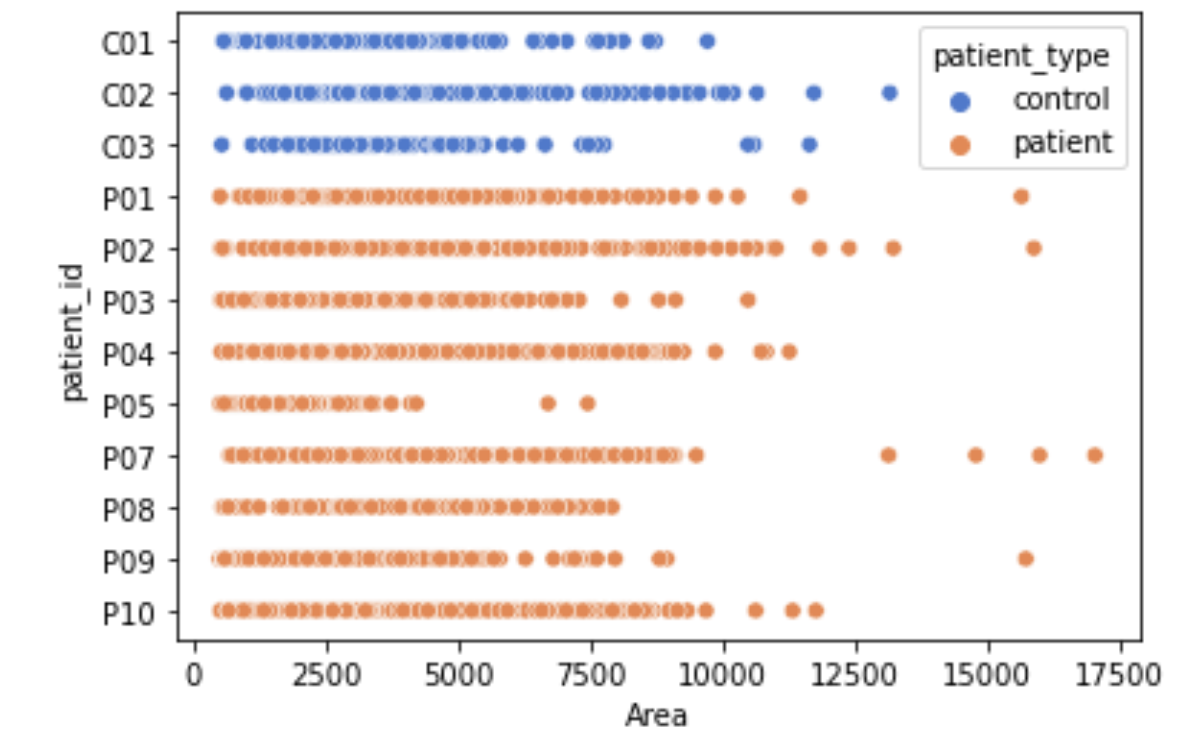
SVM…

Results

EDA of Morphological Features:



Circularity difference between patient and control groups.

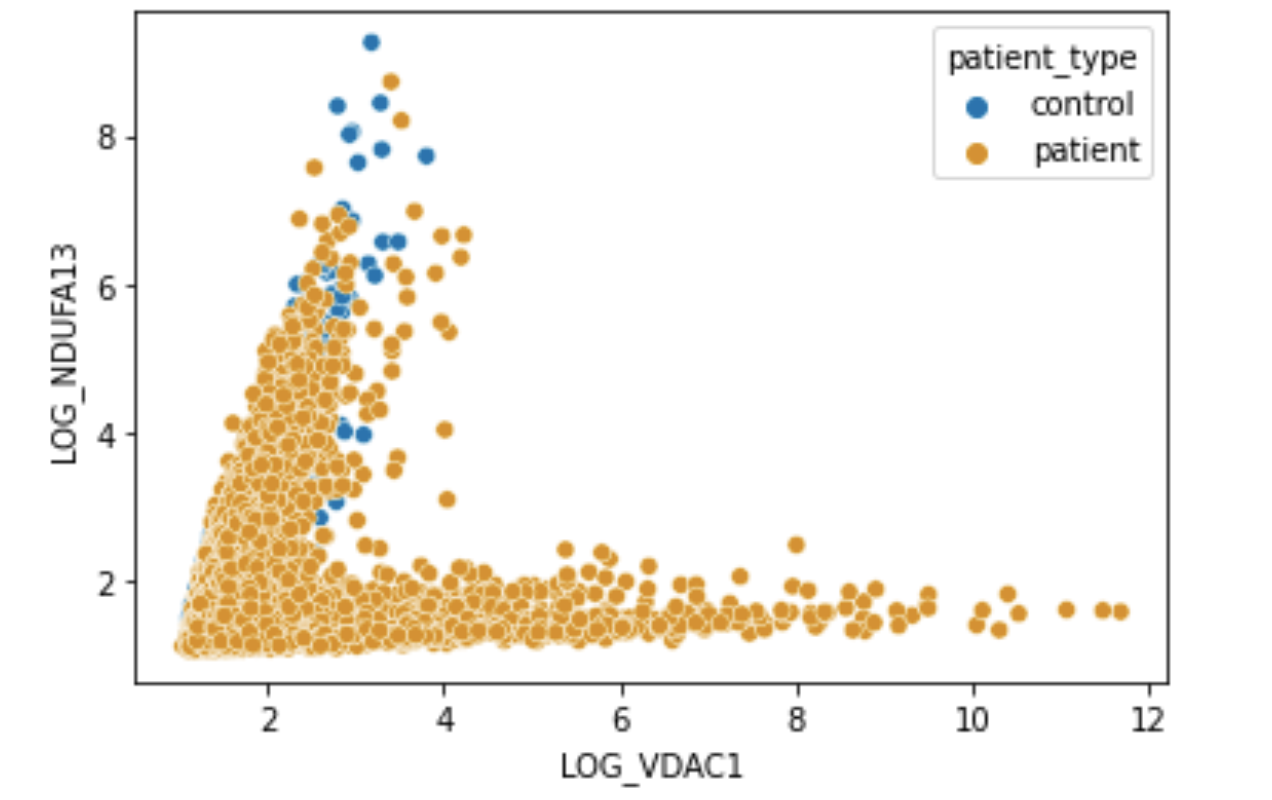


No huge difference in area for patients and control.

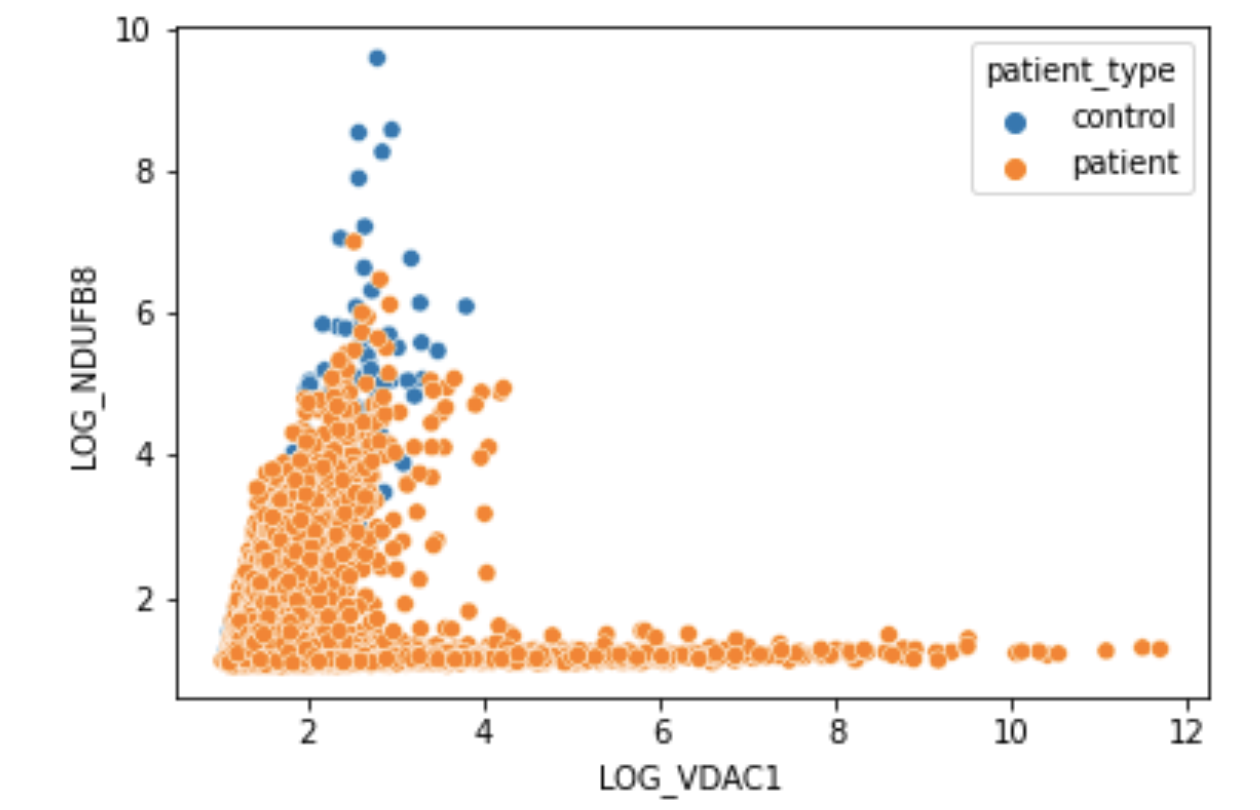
Chart, scatter chart

Description automatically generated

Control clearly has small perimeter in comparison to patient.



Clearly distinguished results between control and patients for NDUFA13



Clearly distinguished results between control and patients for NDUFB8

Chart, scatter chart

Description automatically generated

No huge difference

Chart, scatter chart

Description automatically generated

No huge difference but appears like there is a concentration of control fibres around a log\_vdac1 between 2-4 and a log UqCRC2 between 10-20 whilst for patients there is a widespread.

Table

Description automatically generated

Chart, bar chart

Description automatically generated Nice results from random forest. Any faults with using random forest e.g. imbalanced dataset?



Results from the Decision Tree classification. Again, accuracy seems high. Would be good to know what features were used to predict this.

Accuracy is at 95.36% which is rather high showing that it is likely for this algorithm to predict accurately the label an individual should receive.

A picture containing shape

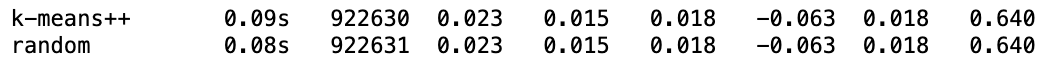
Description automatically generated

Accuracy is at 97.5% for K-Nearest Neighbour, which is also very good showing that it is likely for this algorithm to predict accurately the label an individual should receive. Its accuracy appears to be as good as the decision tree classification.

Graphical user interface, chart

Description automatically generated

The plot of the ground truth shows that the control seems to have a LOG\_VDAC1 nearer to 2.5 and a LOG\_NDUFA13 closer to 6-8, whilst the patients show a large variation across the two proteins. For the predictions, both control and patients are shown to be spread broadly across the two proteins.



The

* KMeans graph of all labels???
* Potentially try a support vector machine algorithm
* UMap

Discussion

Results from the analyses of the morphological features show in Figures X and Y that there is no significant difference in the areas of patients and healthy individuals, however the converse is true for the circularity of the individuals’ fibres. Patients appear to have a less circular fibre in comparison to the control group. This was further examined in the random forest classification algortithm…

Conclusion

Contribution to the States of the Art

Scope and Limits of the Work and Future Work