mitoML: Machine Learning to Understand Mitochondrial Disease Pathology

Mir Atif Ali Khan

School of Computing

Newcastle University

In Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

September 2024

Acknowledgements

Firstly, I would like to say a massive thanks to my PhD supervisors Dr Stephen McGough, Dr Amy Vincent and Dr Conor Lawless for their invaluable guidance, feedback, patience and encouragement throughout my PhD research. Their constant support, expert knowledge and experience made the completion of this thesis possible. You three have been my guiding light throughout this exciting and sometimes tiring PhD journey. Thank you!

I want to express my gratitude to my sponsor EPSRC CDT for Big Data. The generous funding provided by them allowed me to complete my PhD research without much financial constraint. The CDT itself would not exist without its founding director Prof Paul Watson. Paul, I would like to thank you for giving me the opportunity to undertake this research. The CDT would not function without its amazing manager Ms Jennifer Woods. Jen, without your constant support I would not be able to publish papers or go to conferences, thank you for making this so easy for me. I would also like to thank all my colleagues at CDT and WCMR for the amazing last four years, specially Tiago and Valeria for providing expert feedback and testing software.

Now, I would like to acknowledge the contributions of my family. There are no words to describe my gratitude towards my parents Mr Mir Gazanfar Ali Khan and Mrs Monis Unnisa. Baba and Mommy, the reason for all the accomplishments in my life is your nurturing, you instill in me sound moral values and unshakable perseverance, for this I am forever thankful and love you both to bits. I would like to thank my immensely supportive wife Dr Mahwesh Naaz Khan. Mehwu you've been my greatest strength and my best friend. I really appreciate all the support throughout my career, your sacrifices enabled my progression, thank you Mehwu! I would also like to thanks my amazing siblings Eeliya, Adil, Imran, Irshad and Anum, you guys are the best! Thank you guys for believing in me.

Finally, I would like to thank all my teachers for imparting an education: the most valuable asset a human being can possess.

Abstract

Mitochondria are organelles that reside in virtually every cell of the human body and provide the energy for the cells to function. OXPHOS is the main metabolic pathway through which mitochondria generate energy, it is a machinery made up of five complexes each built with sub-units of multiple proteins and molecules. Defects in OXPHOS machinery manifest as results of genetic mutations and lead to mitochondrial disease. Mitochondrial diseases are currently untreatable due to our limited understanding of their pathology. The study of mitochondrial disease pathology involves discovery of OXPHOS protein expression patterns linked to various genetic mutations.

Mitochondrial disease affects high energy demanding cells like Skeletal Muscle (SM) cells (myofibres). The expression of various OXPHOS proteins in myofibres taken from SM biopsies is studied. These OXPHOS proteins in SM tissue are observed using various imaging techniques such as Imaging Mass Cytometry (IMC). IMC produces high dimensional (up to 40 channels) multiplexed pseudo-images representing spatial variation in the expression of a panel of OXPHOS proteins within a tissue, including sub-cellular variation. In previous methods good quality 'analysable' myofibres in these multichannel images are segmented and various statistical summaries, such as mean protein expression, are computed per myofibre. Statistical summaries of various groups of myofibres linked with different genetic mutations and a healthy control group are compared to analyse and understand the OXPHOS protein expression patterns of various mitochondrial diseases.

Theses methods have a number of limitations i) profiling OXPHOS protein patterns in high dimensionality data: Due to high dimensionality multiplex data, it is not possible to classify and discover the OXPHOS protein expression pattern for four out of five groups of genetic mutations affecting mitochondria that have been studied [1] i.e. except for one group of genetic mutation the classification accuracy for all other groups was below 90%. ii) Precise segmentation and curation of myofibres: It is not possible to precisely segment and curate myofibres with existing applications without heavy manual corrections. iii) The use of statistical summaries per myofibre ignores all intra-myofibre features. There are many hypotheses [2, 3] that theorise the existence of differential features within myofibre in various mitochondrial dysfunctions.

In this thesis I use Machine Learning (ML)-specifically logistic regression and XGboost, and various Deep Learning (DL) methods to address the three limitations mentioned above with the following contributions. I) Classify myofibres of mitochondrial patients affected by various genetic mutations, using explainable ML and myofibre statistical summaries. I show that using ML the classification accuracy for all five mutations exceeds 90% . I also demonstrate the use of explainable ML methods to discover the OXPHOS protein expression patterns associated with these high predictive accuracy ML models. II) Precise myofibre segmentation and curation pipeline: I developed 'myocytoML' a precise myofibre segmentation and curation pipeline that meets the quality of gold standard manual human annotations. This also led to the building of NCL-SM: A large dataset of more than 50k manually annotated myofibres, which is now available for public use. III) Classify myofibres of mitochondrial patients affected by various genetic mutations, using explainable DL and segmented multichannel raw images. I show that using DL the classification accuracy for all five mutations exceeds 98%. I also demonstrate the use of explainable DL methods to discover the OXPHOS protein expression patterns associated with these high predictive accuracy DL models.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms/Abbreviations

ANN Artificial Neural Network 35

ASV Absolute SHAP Value 137

ATP adenosine triphosphate 9

CNN Convolutional Neural Network 36

COX cytochrome c oxidase 14

CV Computer Vision 31

DL Deep Learning 2

EM explainable method 3

FAM Freezing Artefact Myofibre 103

FR Folded tissue region 103

GUI Graphical User Interface 110

IAV Inter annotator variability 104

IMC Image Mass Cytometry 2

LR Logistic Regression 34

ML Machine Learning 2

mtDNA Mitochondrial DNA 1

nDNA Nuclear DNA 1

NTM Non-transverse sliced myofibre 103

OXPHOS Oxidative phosphorylation 9

PCA Principal Component Analysis 30

PI Predictive Interval 18

IF Quadruple Immunofluorescence 15

RC respiratory chain 9

ReLU rectified linear unit 35

SDH succinate dehydrogenase 14

SHAP Shapley Additive Explanation 44

SM Skeletal Muscle v

TS Tissue Section 114

WCMR Wellcome Centre for Mitochondrial Research 1

XGB XGBoost 34

Chapter 1

Introduction

Deep learning (DL) in medicine and healthcare is a rapidly advancing domain, making groundbreaking strides in diagnosis, prognosis and drug discovery [4–12]. This is particularly noticeable in bioimaging where DL is facilitating a paradigm shift in detecting disease [13–16]. The adaptation of general computer vision DL models such as Convolutional Neural network (CNN) [17], R-CNN (region based CNN) [18], vision transformers [19, 20] and DL models such as UNet [21] invented out of biomedical specific use cases are driving the invention of novel and groundbreaking DL pipelines for diagnosis and prognosis [14, 22, 23]. The use of explainable DL to discover underlying natural processes and phenomena in meteorology, astronomy, geology, biology is at the cutting-edge of AI research [24–33], however, to the best of my knowledge there is no prior literature describing such applications in mitochondrial disease i.e. the use of DL to aid in the discovery of clinical phenomena such as mitochondrial disease pathology.

Mitochondrial diseases are individually uncommon but are collectively the most common metabolic disorder affecting 1 in 5,000 people[34]. They can cause severe disabilities and adversely affect the life expectancy of patients [35]. They manifest either as a result of mutations in genes encoded by the mtDNA and/or in genes encoded by the nDNA whose products are imported into mitochondria [1]. Mitochondrial disease pathology is complex and highly heterogeneous. Diagnosis usually requires algorithmic analysis of clinical history and of results from multiple laboratory investigations [36]. Some of the latest approaches to classify mitochondrial diseases, quantify disease severity and understand the disease pathology are based on the analysis of single-cell protein expression in multiplex images of skeletal muscle (SM) tissue collected from patients and control groups.

The Wellcome Centre for Mitochondrial Research (WCMR) based within the Newcastle University is one of the leading institutes conducting research into mitochondrial diseases

worldwide, and WCMR have an unparalleled repository of clinical data and tissue from controls and patients with mitochondrial disease [37]. The data includes images of tissue sections that capture spatial variation in protein expression within tissue (including within cells) and from which single-cell level protein expression can be observed and measured. At WCMR advanced protein expression measurement techniques such as Image Mass Cytometry (IMC) are used to observe the spatial variation in the expression of up to 40 proteins in tissue simultaneously. Established statistical approaches to analyse this multiplexed high dimensional data can in some cases, i.e. one in the five cases that are studied [1], successfully identify defective myofibres and the proportion of defective myofibres seems to usefully segregate patients from control subjects [1, 38]. However, these approaches are based around statistical summaries of intensity per myofibre that were i) imprecisely segmented and ii) are fairly crude quantitative measures of myofibre morphology ,and iii) ignores intra-myofibre features that are theorised to differential as per the latest studies [2]. In contrast, in order to derive as much information as possible from these rare and valuable patient data to discover the possible protein expression patterns that can reveal mitochondrial disease pathology, the use of Machine Learning (ML) and Deep Learning (DL) can be explored for the analysis of these raw multiplex (IMC) images.

To profile protein expression patterns associated with various mitochondrial diseases using ML & DL is a significant challenge. This involves first building models that can accurately predict different classes of mitochondrial disease – especially difficult due to difficulties involved in precise myofibre segmentation; the subtle differences between the classes and second, building pipelines to interrogate these models to understand the basis for their predictions – through approaches of explainability. The work undertaken in this thesis is to research and develop various explainable ML & DL pipelines in order to derive as much information as possible from the raw multiplex (IMC) protein expression data with an aim to discover the possible protein expression patterns that can reveal mitochondrial disease pathology. This thesis details a system for automatic segmentation, classification and presentation of possible pathology of various mitochondrial dysfunctions based on unprocessed (raw) IMC data. This is achieved through three pipelines of trained computer vision ML, DL models and robust post-processing techniques capable of removing all non-ideal regions/myofibres in the tissue image, classify myofibres in these tissues from different subject classes and give possible pathological reasoning via explainable methods, a visualisation of which can be seen in Figure 1.1.

The proposed system led to development of the following dataset and pipelines.

- NCL-SM: a large dataset of >50k precise manually segmented myofibres that are made publicly available for training and evaluation of SM tissue image segmentation

pipelines. A consequence of this led to development of a) protocols for segmenting multiplex (IMC) SM tissue images that address the issues of precise segmentation; detecting frozen damaged and non-transverse sliced myofibres, and folded tissue; b) evaluation metrics that inform all aspects of annotation quality.

- myocytoML: an automatic SM tissue image segmentation pipeline that can precisely segment and curate myofibres that are of analysable quality.
- An explainable ML pipeline combines ML models with explainable methods applied to statistical summaries of myofibres to classify myofibres linked to five classes of genetic mutations and profile these myofibres based on associations discovered by ML Explainable Methods (EMs).
- An explainable DL pipeline that combine DL models with EMs to profile segmented multiplex images of myofibres linked to five classes of genetic mutations. This is further extended to classify and profile unsegmented SM tissue section multiplex images.

1.1 Research aim and contributions

The high level aim of this thesis is to: **Design, implement, and evaluate novel explainable machine learning analysis pipelines that precisely segment the multiplex SM tissue images; collate multiplex myofibre images and genetic mutation data and make explainable predictions which allow users to discover underlying mitochondrial disease pathology.**

Implicit in this aim are a number of research questions:

1. Is it possible to precisely segment and classify 'analysable' myofibres in multiplex images using machine learning?
2. Is it possible using explainable machine learning to classify and profile mitochondrial disease using per myofibre statistical summaries?
3. ls it possible using explainable deep learning to classify and profile mitochondrial dysfunctional myofibres using raw segmented multiplex images of myofibres?
4. ls it possible using explainable deep learning to classify and profile mitochondrial diseases using raw unsegmented multiplex images of SM tissue?

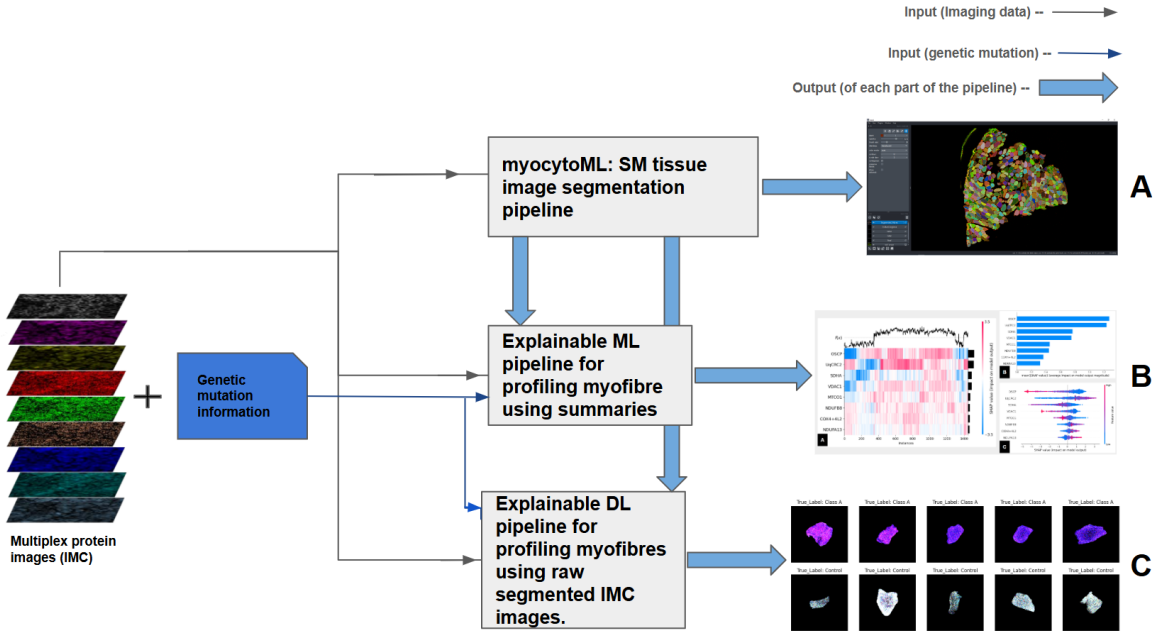

Figure 1.1 A high level overview of data flow through the proposed system. **A** Input multiplex IMC images (left) are passed through myocytoML: an automatic segmentation pipeline, producing an instance segmentation output mask of 'analysable' myofibres, remove all non-analysable (folded tissue) regions and myofibres (frozen artifact myofibres; non-transverse sliced myofibres). This also provides per myofibre morphological summaries and annotation quality metrics. **B** The 'analysable' myofibres mask, the multiplex input image and genetic mutation information are used as input to i) produce statistical pixel intensity summaries per myofibre for each input channel, ii) these are used to training ML classification models, iii) model/s with highest predictive accuracy is selected to apply EMs that produces SHAP plots which informs the ML model's prediction basis, this is used to discover associations between channels, pixels and genetic mutation class. **C** The 'analysable' myofibres mask, the multiplex input image and genetic mutation information are used as input to i) create a dataset of multiplex segmented images of 'analysable' myofibres, ii) these are used to train DL classification models, iii) model/s with highest predictive accuracy is/are selected to apply EMs that produces SHAP explanation masks which inform the DL model's prediction basis in terms of SHAP values per pixel, absolute SHAP value (ASV) per channel, this is used to profile myofibres linked to genetic mutation classes.

To answer these questions and thereby address its aim, this thesis presents the following contributions:

1. myocytoML: An automatic segmentation pipeline for precise segmentation of SM tissue images [39]. The pipeline is made up of three ML models and robust post processing techniques to 1) segment all myofibres, 2) classify and remove non-analysable myofibres, 3) detect and remove folded regions. This is presented in Chapter 5.

2. NCL-SM: A fully annotated dataset of images from human skeletal muscle biopsies. To enable the development of myocytoML, NCL-SM was built as a completely manually and precisely segmented dataset of more than 50,000 myofibres that captures diverse subjects (healthy controls and patients suffering from various muscle diseases) and imaging techniques (IMC, IF). In addition to this, NCL-SM has complete annotations of all folded regions and classification of frozen damaged myofibres, non-transverse sliced myofibres and 'analysable' myofibres. These annotations are accompanied with metrics that define the quality of these annotations, i.e. both comparative metrics derived from duplicate human annotations and objective metrics driven from domain specific factors i.e. inclusion of cell membrane and exclusion of cell mass. This is presented in Chapter 4.

3. Explainable ML pipeline to classify and profile mitochondrial disease using multiplex per myofibre statistical summaries: i) myofibre summaries are computed using 'analysable' myofibres mask, the multiplex input image, ii) these multiplex/multichannel myofibre summaries and genetic mutation information are used to train ML classification models, iii) models with the highest predictive accuracy are selected to apply EMs that produce explanation plots which informs the ML model's prediction basis, this is used to discover associations between channels, pixels and genetic mutation class. This is presented in Chapter 3.

4. Explainable deep learning pipeline to classify and profile mitochondrial dysfunctional myofibres using raw segmented multiplex images of myofibres : i) a dataset of multiplex segmented images of 'analysable' myofibres is created using an 'analysable' myofibres mask and the multiplex input image, ii) these multiplex/multichannel raw myofibre images and genetic mutation information are used to train DL classification models, iii) models with the highest predictive accuracy are selected to apply EMs that produce explanation masks which inform the DL model's prediction basis, this is used to profile myofibres linked to various genetic mutation classes. This pipeline is then extended to classify and profile mitochondrial diseases using raw unsegmented multiplex images of SM tissue. This is presented in Chapter 6.

1.2 Thesis Structure

The structure of this thesis is as follows:

Chapter 1 provides the motivation for the work undertaken in this thesis, and highlights its main contributions. An overview of the peer-reviewed publications produced as a result of work undertaken in fulfilment of this thesis is also presented.

Chapter 2 presents the required background knowledge for understanding the work presented in future chapters. An introduction to mitochondrial biology, IMC data analysis, machine learning and deep learning is provided, alongside a discussion of key computer vision concepts and their recent use in medicine. Existing analysis methods for multiplex (IMC) biomedical data is also examined.

Chapter 3 discusses the development of explainable ML pipeline to classify and profile mitochondrial disease using multiplex per myofibre statistical summaries. This includes selection of appropriate ML models and explainable methods; training experiments and building insights from predictive inference.

Chapter 4 outlines the creation of NCL-SML: a fully annotated dataset of images from SM biopsies. This is built to allow for work presented in later chapters of this thesis to be completed and NCL-SM is released for public use i.e. to train and evaluate ML models for SM tissue image segmentation. The NCL-SM consists of more than 50k manually segmented myofibres from diverse subjects and captured using two imaging techniques, i) a multichannel IMC dataset consists of ≈23k manual annotations of myofibre, this was collected at WCMR and consist of seven classes of mitochondrial diseases, ii) a multichannel IF dataset consist of ≈ 27k manual annotations of myofibre, this was collected at WCMR and consists of six classes of mitochondrial diseases. The chapter also introduces i) protocols to segment SM tissue image and curate 'analysable' myofibres within the image, ii) evaluation metrics for measuring the various aspects of annotation quality required for SM tissue image analysis.

Chapter 5 discusses the development of myocytoML: an automatic segmentation pipeline of SM tissue images. This discusses design decisions including selection of segmentation models, classification models, order of execution of tasks, graphical user inference and standard operating procedure. The chapter also discuses experimental design and training of ML models for the tasks involved in accomplishing precise SM tissue segmentation.

Chapter 6 discusses the development of an explainable DL pipeline to classify and profile dysfunctional myofibres using raw segmented multiplex images of myofibres. This includes preparation of data i.e. multichannel myofibre images of uniform size; selection of appropriate DL models and explainable methods; DL model training experiments and building myofibre profile images from predictive inference.The chapter also discusses an extension of this pipeline to be used in unsegmented multiplex (IMC) data.

Chapter 7 summarises the conclusions of the work presented in this thesis, highlights main contributions and explores avenues for future work in the area.

1.3 Related Publications

The work outlined in this thesis has led to the publication of the following peer-reviewed papers:

[40] - presented in Chapter 6 A. Khan, C. Lawless, A. E. Vincent, S. Pilla, S. Ramesh, and A. S. McGough, "Explainable Deep Learning to Profile Mitochondrial Disease Using High Dimensional Protein Expression Data", Proceedings – 2022 IEEE International Conference on Big Data, Big Data 2022, pp. 4375–4384, 2022.

[41] - presented in Chapter 4 A. Khan, C. Lawless, A. E. Vincent, C. Warren, V. D. Leo, T. Gomes, and A. S. McGough, "NCL-SM: A Fully Annotated Dataset of Images from Human Skeletal Muscle Biopsies", in 2023 IEEE International Conference on Big Data (BigData), pp. 3704–3710, 2023.

1.3.1 Publications reuse and contributions

Some of this published material is reproduced verbatim within this thesis, for which I have required copyright permissions. These publications are written by me as first author under the guidance of my PhD supervisors Dr Lawless, Dr Vincent and Dr McGough. The contributions of other coauthors are as follows- S. Pilla and S. Ramesh were MSc students who help with training classification models for unsegmented TS multiplex data; C. Warren, V. D. Leo and T. Gomes collected, processed, and imaged biopsies/tissue sections also performed expert annotations.

Chapter 2

Background

2.1 Mitochondrial biology

Mitochondria are organelles that produce $\approx$ 90 % of the energy consumed within each of the trillions of cells that make up a human body [42]. Dysfunction in mitochondria disproportionately affects cells with a high energy demand e.g., muscle cells and neurons. Mitochondria are unusual in that they have their own DNA (mtDNA). Genes in mtDNA code exclusively for mitochondrial proteins and their synthesis machinery, but most mitochondrial proteins are encoded in nuclear DNA (nDNA). Mutations affecting mitochondrial proteins, whether encoded in mtDNA or nDNA manifest as mitochondrial diseases [43]. Mitochondrial diseases are classified based on their genetic aetiology, i.e., the source and location (nDNA or mtDNA) of their mutation, as inherited or sporadic nDNA and/or mtDNA diseases. In mtDNA diseases, when genetic mutations in mitochondrially encoded genes reach high concentrations in an individual cell, this results in alterations in the concentration of mitochondrial proteins and associated subunits of the mitochondrial respiratory chain (RC) complexes which in turn results in mitochondrial dysfunction [44].

2.1.1 Oxidative phosphorylation

Oxidative phosphorylation (OXPHOS) is a vital metabolic process that occurs in the mitochondria of cells, where it generates ATP (adenosine triphosphate), the primary energy currency of the cell. The inner mitochondrial membrane hosts OXPHOS protein complexes that comprise the mitochondrial respiratory chain complexes (complexes I-IV) and ATP synthase (complex V). The whole electron transfer process through the OXPHOS complexes leading to creation of ATP is described in Figure 2.1.

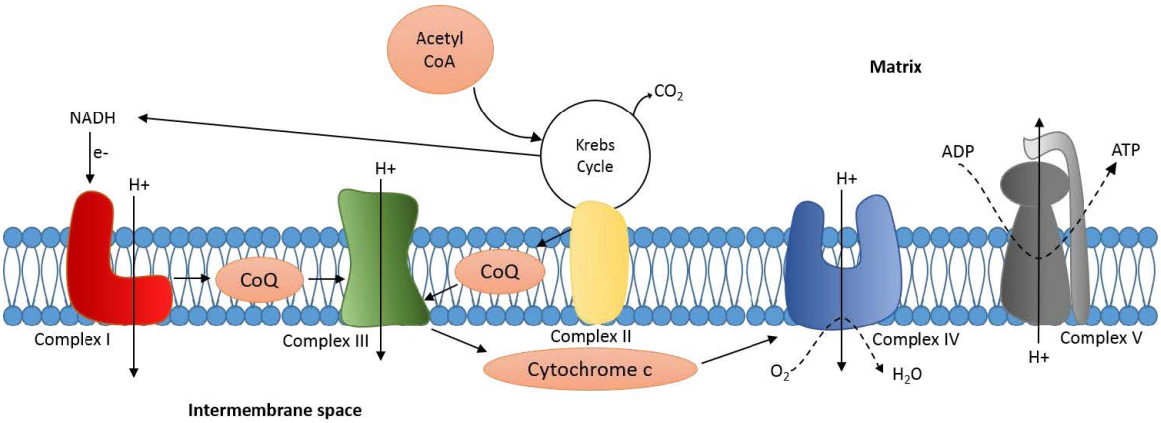

Figure 2.1 Oxidative phosphorylation: The inner mitochondrial membrane hosts OXPHOS protein complexes I-V. Complex I starts the first step of OXPHOS and accepts electrons derived from NADH, with complex II accepting electrons derived from $FADH_2$. These are transferred along the chain to complex III to complex IV through cytochrome c. The transfer of electrons through complexes I-IV is coupled with the pumping of protons from the matrix to the intermembrane space. This creates an electrochemical gradient which causes protons to flow back through complex V to convert ADP to ATP [45]. Figure taken from [46, 47]

2.1.2 Mitochondrial genetics

As mentioned earlier mitochondria have their own DNA (mtDNA) but the majority of mitochondrial proteins are encoded in nDNA.

mtDNA

The mtDNA located in the mitochondria of human cells is inherited almost exclusively from the mother, and is the small, circular double-stranded molecule – measuring 16569bp, which encodes 37 genes: 13 OXPHOS subunits, protein synthesis machinery(22 tRNAs and 2 rRNAs) as presented in Figure 2.2 [48].

mtDNA replication, hetroplasmy and the threshold effect

Replication of mtDNA occurs independently of the cell cycle and is reliant on its own replication machinery. This replication machinery when working correctly produces wild-type (normal) mtDNA; however sometimes defects (e.g. mutations in RRM2B and POLG) cause it to produce mutant (abnormal) mitochondrial mtDNA [49]. Due to the polyploid nature of mtDNA, it is possible for wild-type and mutant genomes to coexist within a cell. The cell is said to be homoplasmic if all copies of the mtDNA within the cell are identical, however a cell is heteroplasmic if more than one mtDNA species are present. Heteroplasmy

is measured as a percentage derived from the total number of mtDNA species and can vary in different cells. When the proportion of mutated mtDNA exceeds a certain threshold it leads to a phenotypic manifestation (disease symptoms) of the genetic defect see Figure 2.3. This threshold level varies across i) different tissues depending on their energy demand i.e. tissues which are highly dependent upon OXPHOS metabolism experience effects at a lower threshold, and ii) mtDNA mutations i.e. higher in point mutation than single deletion. Usually the threshold level falls between 60-90% mutant to wild-type mtDNA [50, 51].

mtDNA mutations

Genetic defects associated with the human mitochondrial genome were first demonstrated in 1988 [52, 53] and since then, a vast number of mutations have been identified and linked with mitochondrial disease. The main types of mtDNA mutations are as follows:

mtDNA point mutations

mtDNA point mutations are described as a single base pair substitution and are present in the adult population at a prevalence of 1 in 5000 [34] . These mutations, which can either be maternally inherited or sporadic, are hugely heterogeneous and cause a vast range of mitochondrial diseases. There is great clinical variability between carriers of the same mutation. Patients can present with other features such as ataxia, diabetes mellitus, optic atrophy, hearing loss and dementia [54].

Single, large-scale mtDNA deletions

Single, large-scale mtDNA deletions refer to the loss of a substantial segment of the mtDNA molecule, which can result in the removal of several genes essential for mitochondrial function, these deletions are thought to occur sporadically during early stages of development [55]. These deletions can lead to a variety of mitochondrial disorders; the most commonly reported deletion encompasses 4,977bp [55], but deletions can vary in size from 1.3 to 10 kb. A prevalence of 1.5/100,000 accounting for approximately 16% of adult mtDNA mutations, deletions are a common cause of mitochondrial disease [34]. There are three main clinical syndromes associated with deletions: CPEO, Kearns-Sayre syndrome (KSS) and Pearson syndrome [56].

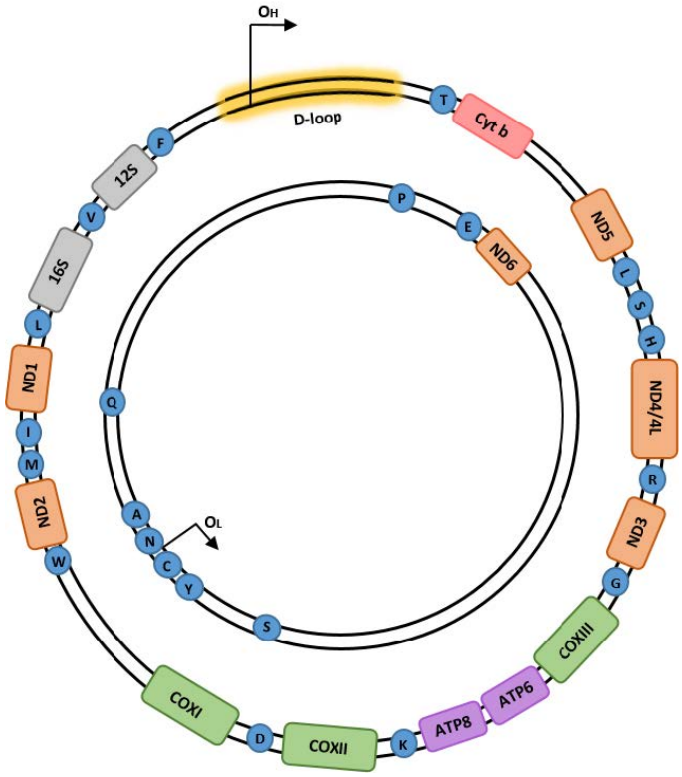

Figure 2.2 mtDNA: The mitochondrial genome encodes 37 genes – 13 OXPHOS subunits (complex I = orange, complex III = pink, complex IV = green, complex V = purple), protein synthesis machinery (22 mt-tRNAs (blue) and 2 mt-rRNAs (grey)). Figure taken from [46, 47]

2.1.3 nDNA mutations

As mentioned earlier mitochondria are under dual control of both mtDNA and nuclear DNA (nDNA). Although mtDNA encodes for 37 genes that are critical for protein synthesis, the remaining proteins that are required for RC complex assembly, mtDNA replication, repair, transcription and translation are encoded by nuclear DNA (nDNA). Currently there are approximately 1100 nuclear genes encoding mitochondrial proteins [57]. Therefore, some mitochondrial diseases are caused by defects in nuclear genes. Some of these nDNA mutations will have a secondary effect on mtDNA, resulting in deletions and mtDNA depletion causing various conditions [58]. The nuclear genes associated with mtDNA deletions encode proteins involved in 37 replications (TWNK [59], MFN2 [60], biogenesis (TFAM [61]), TK2 [62]), mitochondrial fusion (OPA1 [63]), POLG [64] and mitochondrial maintenance (RRM2B [65])).

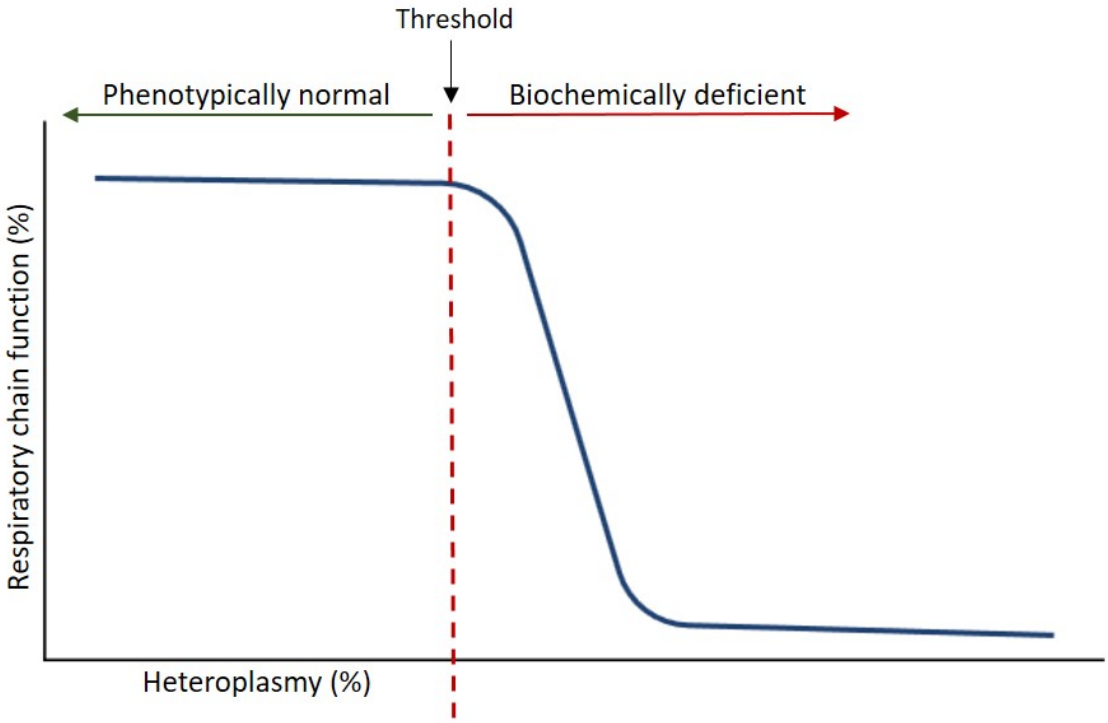

Figure 2.3 mtDNA heteroplasmy and the threshold effect. RC function (blue line) is maintained around 100% until the heteroplasmy level (proportion of mutant to wildtype mtDNA) reaches a threshold level (indicated by the red dotted line). After this threshold is reached, the percentage of respiratory chain function begins to decline. Figure taken from [46, 47]

Mitochondrial diseases are rare metabolic disorders affecting 1 in 5,000 people [34]. They can cause severe disabilities and adversely affect the life expectancy of patients [35]. They manifest either as a result of mutations in (mtDNA) and/or (nDNA) [43]. Some of the latest approaches to understanding mitochondrial disease pathology involves analysis of single cell protein expression data in tissue that are highly dependent upon OXPHOS metabolism, i.e. need more energy e.g. skeletal muscle [43].

2.2 Skeletal muscle in mitochondrial diseases

Skeletal muscle (SM) is one of the primary tissues affected in mitochondrial diseases, given its high energy demands [66, 67]. SM is affected by defects in RC function, however theoretically RC dysfunction can give rise to any symptom, in other tissues[68]. Due to great clinical and genetic variability presented by patients suffering from mitochondrial SM disorders (myopathies), its diagnosis is challenging. However, genetic sequencing techniques such as next generation sequencing (NGS) has revolutionised the way in which these diseases are diagnosed making diagnosis quicker and more straightforward i.e. by detecting the genetic mutations [69]. Histochemical (microscopy) image analysis of SM biopsies is still widely-used for detecting mitochondrial abnormalities for diagnosis. One method, the Gomori trichrome stain, allows visualisation of ragged red fibres, a hallmark of mitochondrial myopathies, under the microscope. These myofibres have an abnormal appearance due to the accumulation of dysfunctional mitochondria around the edges of myofibres [70]. Other techniques that look for specific mitochondrial enzymes, e.g. cytochrome c oxidase/succinate dehydrogenase (COX/SDH) dual histochemistry [71], are more informative as they allow the observer to assess function of components of mitochondrial energy production. COX/SDH is a technique that in mitochondrial myopathy typically shows a mosaic pattern of complex IV enzyme COX deficiency in myofibres. This mosaic pattern is due to a mixture of COX-positive (brown precipitate) and COX-negative (lack of brown precipitate and therefore presence of blue precipitate) myofibres in a muscle biopsy. Although this allows the classification of myofibres into COX-positive, intermediate or COX negative, only complex IV deficiency can be evaluated with this method, and any deficiency in other OXPHOS complexes such as complex I, III and V cannot be detected.

This issue is addressed by an immunofluorescent (IF) technique that quantifies the levels of complex I and IV together with a marker for mitochondrial mass [72].

2.3 SM tissue imaging technique: Quadruple Immunofluorescence

IF is an antibody-based semi-quantitative imaging technique that visualises specific proteins or antigens within cells or tissues using antibodies linked to fluorescent dyes. It allows the capture of high resolution and high bit depth microscopic images and is less expensive and faster than other advanced techniques such as imaging mass cytometry (IMC), but with IF only up to five proteins can be observed.

2.4 SM tissue imaging technique: Image Mass Cytometry

IMC is a recently developed method allowing quantitative analysis of protein levels in a highly multiplexed way, at single cell and subcellular resolution as described in Figure 2.4. The output from the image mass cytometer is either text or proprietary .mcd file which can be converted into pseudo-images using MCD viewer [73]. These pseudo images have resolution of 1μm x 1μm per pixel and can be exported as a 16 or 32 bit grayscale image stack saved in OME-TIFF(.ome.tiff) format for downstream analysis [73].

2.5 Leveraging full potential of IMC data in the context of mitochondrial disease SM tissue analysis

IMC allows observation of up to 40 protein markers [74] with 1μm x 1μm resolution and can provide insight into the pathology of various genetic mutations that cause mitochondrial diseases, which require observation of proteins/enzymes/subunits in the myofibre (including sub-cellular areas to test various theories [2]) involved in the function of OXPHOS complexes (1-V). IMC allows us to select multiple target protein markers to observe i.e. multiple markers within each OXPHOS complex, mitochondrial mass markers, cell morphological markers (e.g. cell membranes, nuclei). The IMC data in context of mitochondrial disease pathology in SM tissue present us with the opportunity to observe and understand the associations between these various protein markers, intra-myofibre regions and genetic mutations. There are many analysis methods [75–81] for high dimensional multiplex (IMC) data, all of which employ i) segmentation of cells in the multiplex data, ii) followed by application of dimensionality reduction techniques, either spatially (e.g. protein marker's pixels mean intensity per cell) or channel-wise, iii) followed by classification or clustering methods to visualise various groups/classes/populations of cells. In the context of our use case this type of analysis

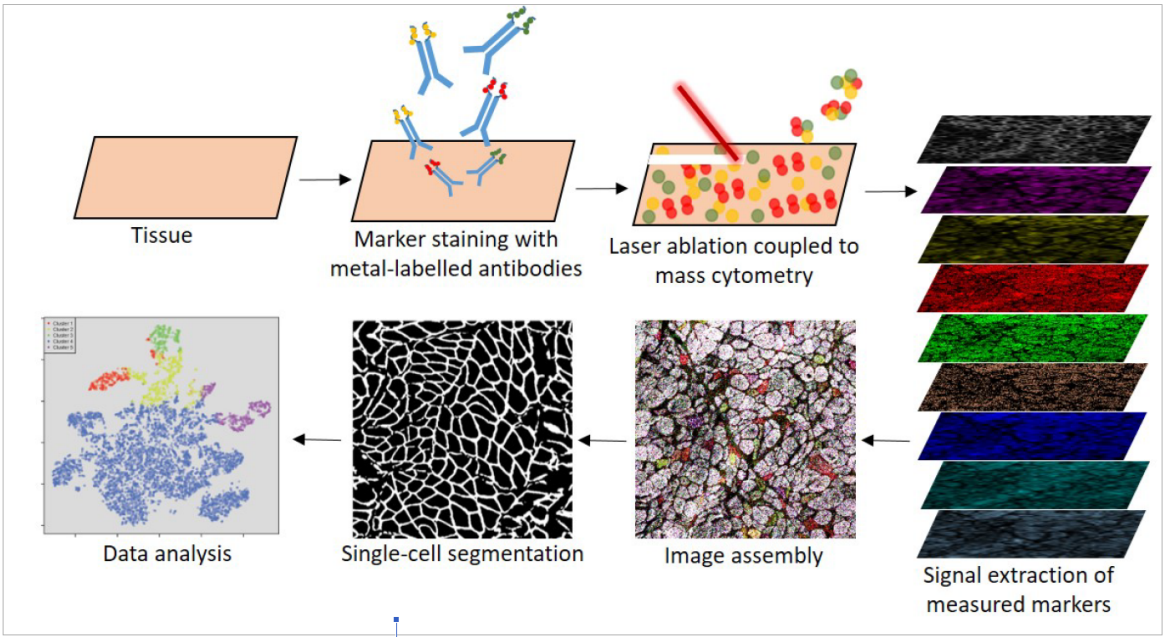

Figure 2.4 Imaging Mass Cytometry: experimental procedure in which IMC cells are stained with a panel of antibodies conjugated to heavy metals and tissue (cells) are scanned by a pulsed laser which ablates a spot of tissue section. The tissue is vaporised on each laser shot and enters the mass cytometer where the relative concentration of heavy metals can be quantified. These measurements are later combined as pseudo images where location and intensity of each pixel correspond to the amount of metal isotopes and location at each spot. Cells are segmented in these images for conducting single cell data analysis. Figure taken from [1, 74]

which requires either ignoring intra-myofibre features (when the data is reduced in spatial dimension) or ignore channels (when data is reduced in channel-wise dimension) means compromising with leveraging of full potential of multiplex IMC data.

2.6 Existing IMC analysis pipeline for mitochondrial disease SM tissue

For downstream analysis of skeletal muscle tissue sections, currently individual skeletal myofibers (SMs, skeletal muscle cells) are identified using semi-automatic image segmentation and then average protein expressions per cell estimated. This step is carried out using a Python image analysis package called mitocyto [1]. Average protein expression levels in single SMs are analysed using relevant statistical techniques and the R shiny tool plotIMC[82][75]. Comparisons are drawn between matched patient and control groups using relevant statistical models (e.g. linear regression, Gaussian Mixture Models (GMM)).

2.6.1 Myofibre segmentation using mitocyto

Before performing the analysis on the multiplex IMC data, individual myofibres need to be segmented. For this the mitocyto[1] application built in-house by Dr Conor Lawless is used. Segmentation software that were previously used by other groups for IMC multiplex data such as MiCAT [77], cell profiler [83] and Ilastik [84], were difficult to adapt in order to accommodate for the type of analysis that was required for IMC in skeletal muscle i.e. precise segmentation of myofibre, removal of unfit for analysis regions (folded tissue) and myofibres (freezing damaged and non-transverse sliced (NTM). This necessitated the development of mitocyto in-house at WCMR that can segment myofibres in IMC multiplex images. Mitocyto works by building an edge map of the cells which is used to segment the area of each individual cell. The edge map is constructed by applying computer vision algorithms from OpenCV [85] like watershed, threshold, Canny edge detection and finding contours. The edge map can be constructed automatically, either directly from an image representing cell membranes, or from a gradient map constructed from a channel (or from an average of all available channels). The edge map can also be drawn by manual tracing over individual channel images (or over an average of all available channels), using the mouse. It is suggested that using membrane channel images produces the best results compared to all other channels or combinations [1].
This is followed by a post processing step to retain myofibres that are of 'analysable' quality

[1] www.github.com/CnrLwlss/mitocyto

and remove the rest from the segmented myofibres in the edge map. The criteria of size (areamin=500, areamax=17500), convexity (convexmin=0.75, convexmax=1.0), aspect ratio (asp-ratiomin=0.0, asp-ratiomax=10.0), and circularity (circmin=0.0, circmax=100.0) were used to detect and remove non-transverse sliced myofibres.
But it was observed that both automatic steps of constructing edge maps and filtering 'analysable' myofibres do not usually yield a segmentation which is analysis ready. This might be due to weak signal or noise in the image, or factors such as tissue folding or freezing damaged myofibres that are not handled by mitocyto. For this reason the common workflow to use mitocyto is where a first draft of the edge map is constructed automatically, followed by manual updates to arrive at a segmentation of analysable quality [1]. An example of a final 'analysable' myofibres mask produced using mitocyto is presented in Figure 2.5.
Using analysis ready edge map mitocyto can produce a comma separated (CSV) file consisting of statistical summaries for each individual myofibre. This includes area, perimeter, aspect ratio, coordinates and mean, log mean and median intensity for individual myofibres.

2.6.2 SM tissue multiplex (IMC) data analysis using plotIMC

plotIMC[2] [1] is a R shiny web-tool that is built in-house at WCMR by Dr Conor Lawless to analyse the multiplex statistical summaries produced by mitocyto. plotIMC has three views, i) 2Dmito plot is the default view which produces a scatter plot between a selected protein target on the y-axis and surrogate for mitochondrial mass marker VDAC1 on the x-axis, a linear regression line and 95% predictive interval (PI) is drawn using all control myofibres. This predictive interval is used as reference upon which patient myofibre summaries are drawn, and the myofibres lying outside this predictive interval are classified as affected by mutation as described in Figure 2.6, ii) mean intensity stripchart view shows the proteins' mean intensity distributions of control and patient myofibres as described in Figure 2.7 and iii) theta stripchart view shows the proteins' theta distributions of control and patient myofibres as describe in Figure 2.8.
In addition to these three views, plotIMC also produces a Pearson's correlation plot between all protein targets for a given subject as described in Figure 2.9 and tabular summaries that present i) proportion of a given patient's myofibres lying outside 95% PI computed using control myofibres, ii) mean intensities of all proteins categorising it as 'ABOVE', 'NODIFF', 'BELOW' based on the predictive intervals, and iii) a table summing the spread of intensities of all proteins for a given patient i.e. min, max, mean, median.

[2] https://mito.ncl.ac.uk/warren_2019/

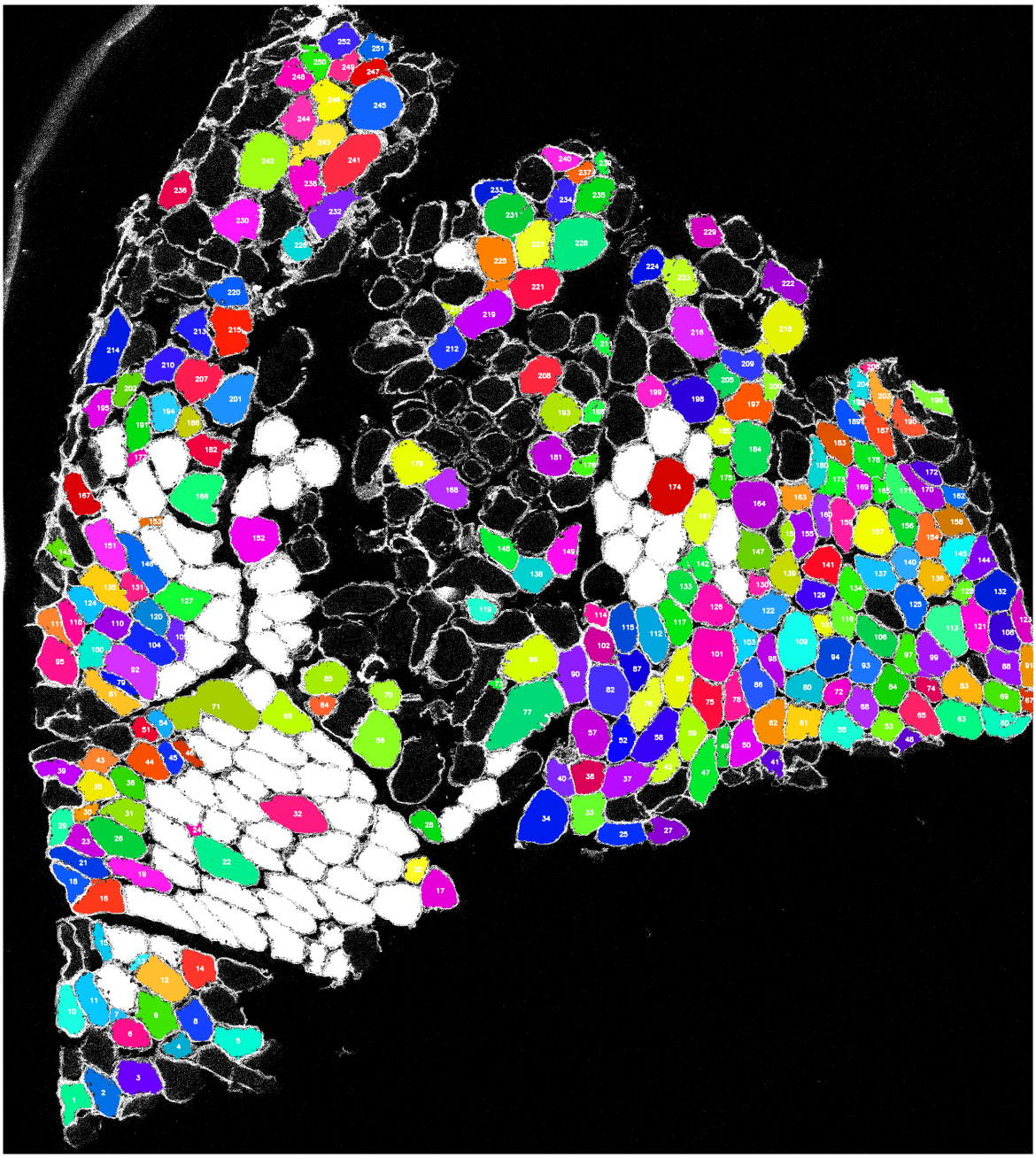

Figure 2.5 Section from P02 segmented and manually updated using mitocyto. The data exported as 16-bit TIFF files from MCD viewer has 12 tiff files per section, each corresponding to a target protein, resolution of 1 pixel per $1\mu m^2$, which is determined by the size of the laser spot. From this the myofibre membrane marker (Dystrophin) tiff file is used to construct an edge map using mitocyto. The coloured annotations (unique colour per myofibre with its number displayed in white text) are 'analysable' myofibres and white annotation are rejected myofibres.

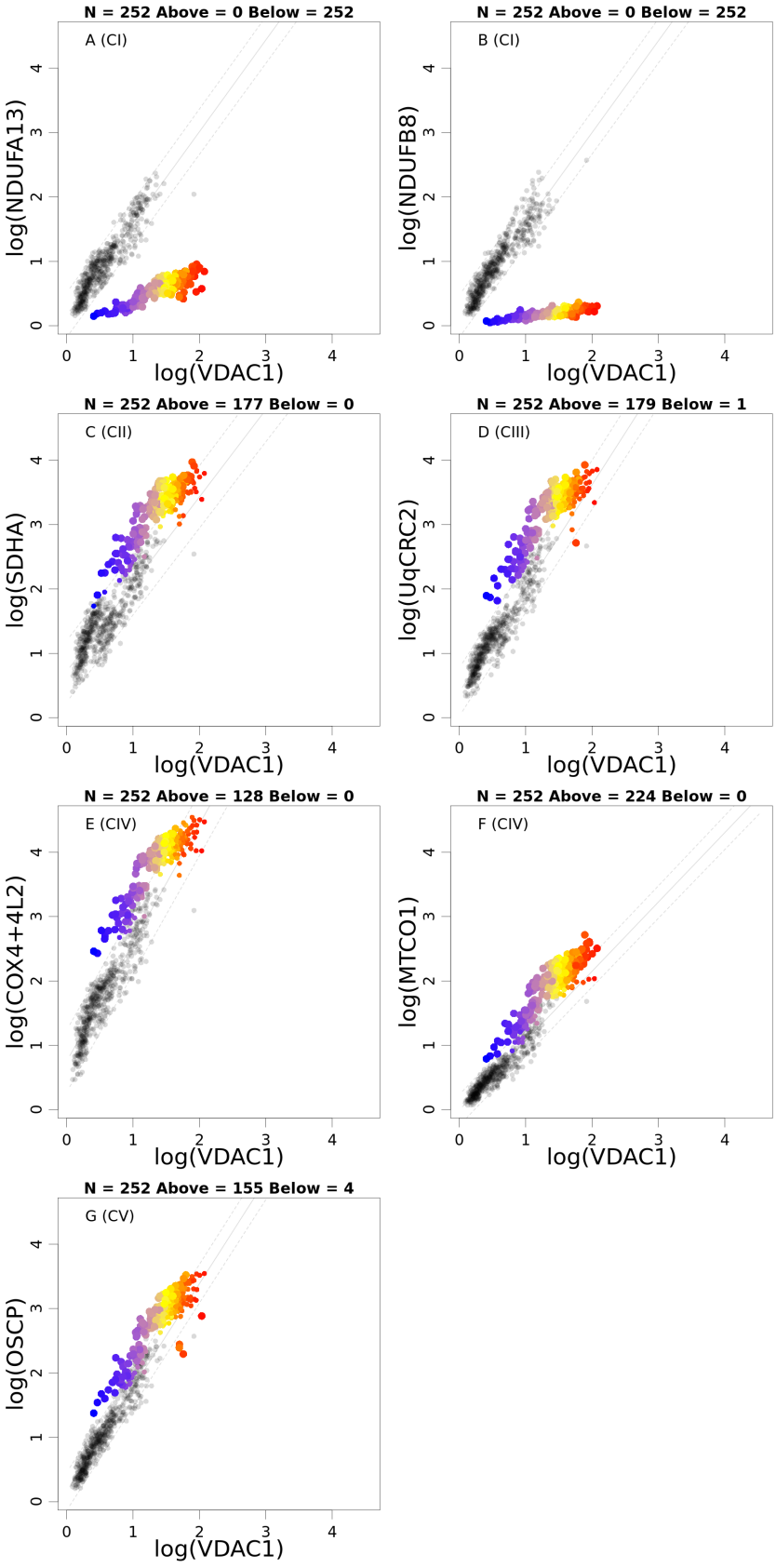

Figure 2.6 P02 myofibers 2Dmitoplot made using plotIMC. The points are coloured according to values of a selected protein (NDUFB8) , from red to blue gradient representing lower to higher values of patient's (P02) myofibres and gray points represent control myofibres. The dotted lines represent 95% PI of control myofibres.

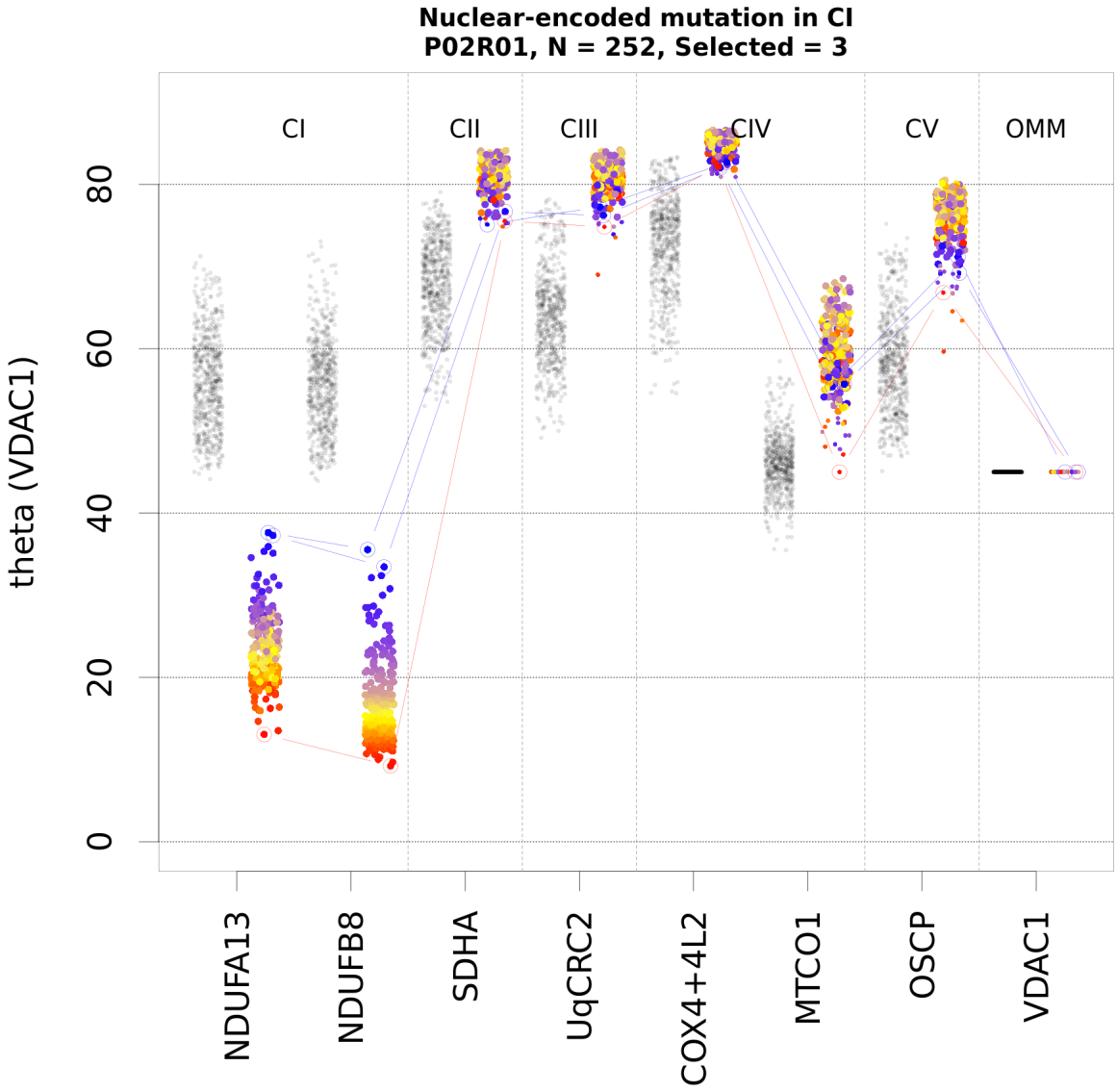

Figure 2.7 P02 myofibers stripchart mean intensities made using plotIMC. The coloured points (coloured according to NDUFB8 values) represent patient (P02) myofibres and gray points represent control myofibres. The dotted lines represent mapping of individual myofibre across various protein markers.

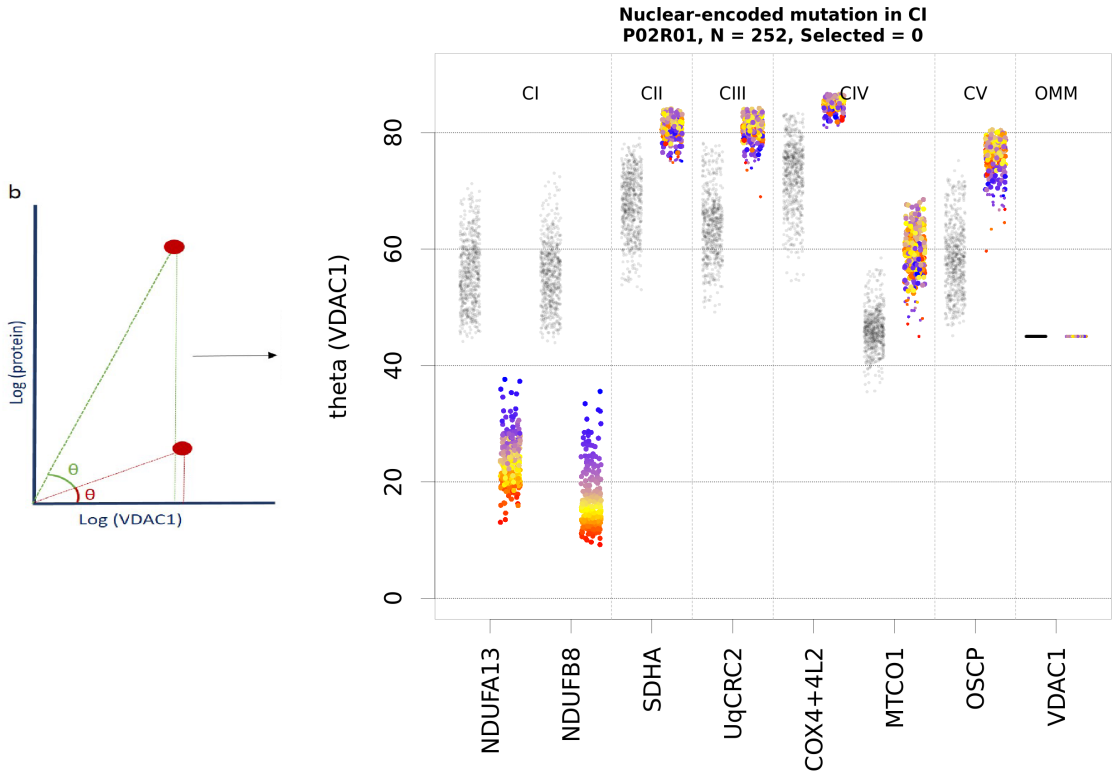

Figure 2.8 P02 myofibers strip chart theta made using plotIMC.The coloured points (coloured according to NDUFB8 values) represent patient (P02) myofibres and gray points represent control myofibres. Plot on the left defines how the theta values are computed.

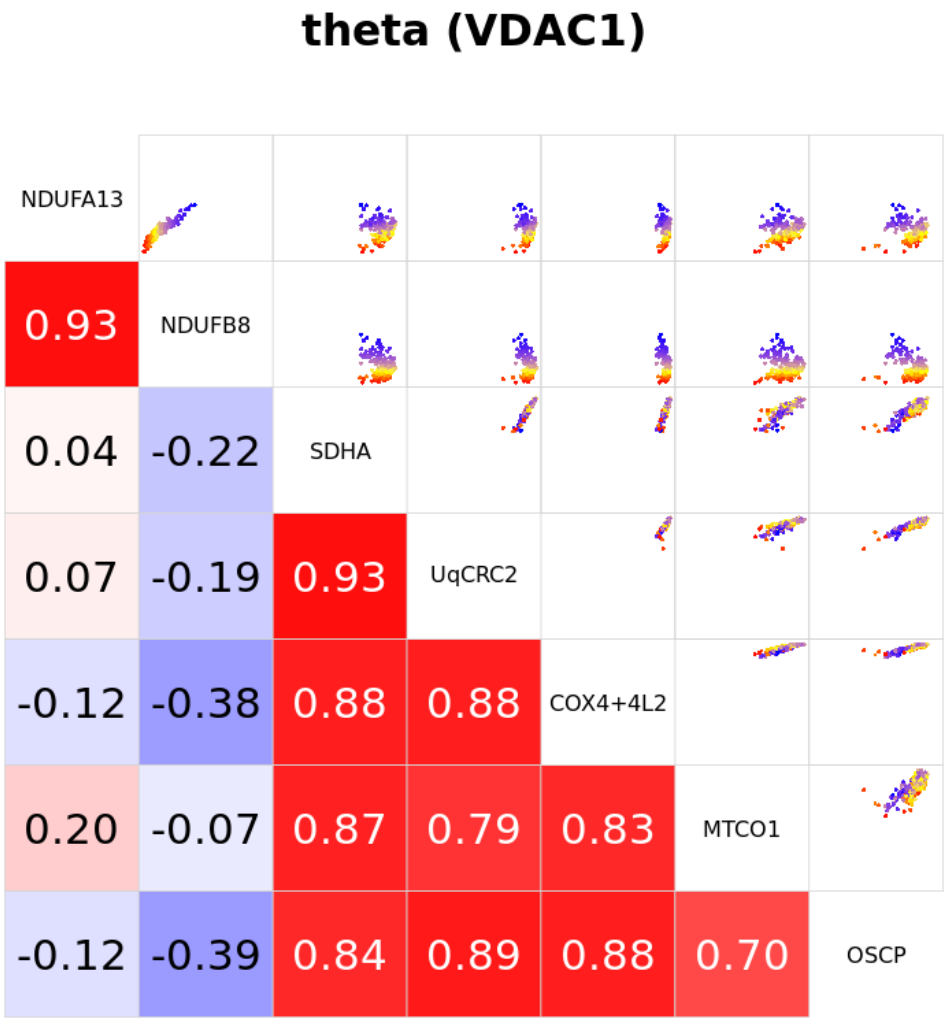

Figure 2.9 P02 myofibers correlation plot made using plotIMC. The coloured points (coloured according to NDUFB8 values) represent scatter plot between two proteins. The numbers in cells are the Pearson's correlation values.

2.7 A typical analysis of SM tissue IMC data using existing analysis pipeline

At WCMR the multiplex IMC data is analysed using mitocyto and plotIMC as performed by Warren *et al.* [1] . Here, I present a typical analysis of this data from Warren *et al.* [1] that will be used as reference/benchmark for analysis conducted using ML methods developed in this thesis.

2.7.1 Data

Skeletal muscle samples were taken from the vastus lateralis of patients with clinically and genetically-characterised mitochondrial diseases of either mtDNA or nDNA origin. For the initial cohort, patients were grouped based on the type of mutation: Nuclear-encoded mutations affecting complex I (n=2), Single, large-scale mtDNA mutations (n=2), Point mutations in mitochondrial-encoded tRNA leucine (MT-TL1) (n=3), and point mutations in other mitochondrial-encoded tRNAs (n=3). SM samples were taken from the hamstring of the healthy controls (n=3).

Ethical approval for use of mitochondrial disease patient tissue was granted by the Newcastle and North Tyneside Local Research Ethics Committee (reference 16NE/0267). Control tissue was acquired from the distal part of the hamstring muscle from individuals undergoing anterior cruciate ligament surgery following approval by the Newcastle and North Tyneside Local Research Ethics Committee (reference 12/NE/0394). Information about all patients and control cases are displayed in Table 2.1.

IMC data

The $6\mu m$ sections were transversely sectioned from frozen blocks of biopsies from subjects mentioned in Table 2.1 prepared for IMC which included air drying the frozen tissue, fixing it with paraformaldehyde, permeabilisation by dehydration and rehydration in a methanol gradient, overnight incubation with the metal-conjugated antibody panel and finally being placed in Hyperion imaging mass cytometer for ablation. After successful ablation files were exported in their native “MCD” file format and processed to single channel TIFFs (16-bit) using MCD viewer software (Fluidigm).

The antibody panel used was designed to include antibodies that specifically targeted a number of mitochondrial proteins as well as myofibre membrane and nuclear markers. Out of 12 markers, eight mitochondrial antibodies were included, seven of these targeted proteins involved in complexes I-V (CI-CV) of the mitochondrial oxidative phosphorylation

Table 2.1 Subject information: Information of patients and controls detailing gender, age at biopsy, clinical information and genetic defect. The orange rows define the genetic mutation of the subjects in the following rows.

Subject	Gender	Age	Clinical information	Genetic defect	Heteroplasmy level
Nuclear-encoded mutations affecting complex I (taken from the vastus lateralis)					
P01	M	Adult	Exercise intolerance, unable to perform sustained aerobic exercise normal resting lactate, normal CK	TMEM126B Homozygous c.635G>T, p.(Gly212Val)	NA
P02	M	Adult	Exercise intolerance, muscle cramps, elevated serum lactate	ACAD9 Compound heterozygous c.1150G>A, p.(Val384Met) and c.1168G>A, p.(Ala390Thr)	NA
Single, large-scale mtDNA mutations (taken from the vastus lateralis)					
P03	F	29yrs	CPEO and bilateral ptosis	Deletion size: 4372bp Breakpoints: 8929-13301 mtDNA deletion level: 53%	53%
P04	F	39yrs	CPEO, diplopia	Deletion size: 7498bp Breakpoints: 7130-14628 mtDNA deletion level: 28%	28%
Point mutations in mitochondrial-encoded tRNA leucine (MT-TL1) (taken from the vastus lateralis)					
P05	F	25yrs	Exercise intolerance, ptosis	m.3243A>G MT-TL1 mutation	66%
P06	F	47yrs	Modest exercise intolerance	m.3243A>G MT-TL1 mutation	34%
P07	M	53yrs	CPEO	m.3243A>G MT-TL1 mutation	74%
Point mutations in other mitochondrial-encoded tRNAs (taken from the vastus lateralis)					
P08	M	33yrs	Mitochondrial myopathy	m.10010T>C MT-TG mutation	89%
P09	F	35yrs	Mild weakness	m.14709T>C MT-TE mutation	76%
P10	M	63yrs	Exercise intolerance, prominent exertional dyspnea	m.5543T>C MT-TW mutation	NA
Healthy controls (taken from the tibialis anterior)					
C01	M	20yrs	Taken during anterior cruciate ligament surgery		
C02	M	24yrs	Taken during anterior cruciate ligament surgery		
C03	F	23yrs	Taken during anterior cruciate ligament surgery		

Table 2.2 Presented here is the list of protein targets and respective antibodies.

Protein Target	Metal label	Host and metal isotope (primary antibodies)	Secondary antibodies
Membrane marker: Dystrophin	176Yb	Mouse	
Complex I: NDUFB8	160Gd	Mouse IgG1	Anti-IgG1 biotin
Complex I: NDUFA13	164Dy	Mouse IgG2b	Anti-IgG2b Alexa Fluor 546nm
Complex II: SDHA	153Eu	Mouse IgG1	Anti-IgG1 Alexa Fluor 647nm
Complex III: UqCRC2	174Yb	Mouse IgG1	Anti-mouse IgG Alexa Fluor 488nm
Complex IV: MTCO1	172Yb	Mouse IgG2a	Anti-IgG2a Alexa Fluor 488nm
Complex IV: COX4+4L2	168Er	Mouse IgG2a	Streptavidin Alexa Fluor 647nm
Complex V: OSCP	161Dy	Mouse IgG1	
OMM marker: TOM22	158Gd	Mouse IgG	
OMM Mass marker: VDAC1	166Er	Mouse IgG2b	
DNA marker: DNA1	191lr		
DNA marker: DNA2	193lr		

machinery and one targeted a mitochondrial outer membrane protein which acted as a surrogate for mitochondrial mass, two markers targeted DNA and one myofibre membrane as detailed in Table 2.2. These targets allowed observation of all five complexes of the respiratory chain, mitochondrial outer membrane (VDAC1) can be used as a surrogate for mitochondrial mass and by extension for cytoplasm, nuclei (DNA1 and DNA2) and myofibre membrane marker (dystrophin). VDAC1 and dystrophin will be helpful in segmentation of myofibres.

2.7.2 Results using existing analysis pipeline

Myofibre segmentation

Using mitocyto with manual interventions to remove folded tissue regions and freezing damaged myofibres, an 'analysable' myofibres edge mask is made for 13 subjects resulting in 8994 myofibres across all subjects as presented in Table 2.3. However, there is no provision in mitocyto to evaluate the quality of segmentation and users have to rely on visual inspection. In addition to 'analysable' myofibres edge masks mitocyto also produces a CSV file with per myofibre statistical summaries of protein markers and myofibre morphological measurements as described in Table 2.3.

plotIMC analysis

The CSV file from mitocyto is used as input to plotIMC to analyse it. As discussed earlier plotIMC allows a user to perform exploratory data analysis using various plots as presented in Figures 2.6, 2.7, 2.8 & 2.9. For classification of control and patient myofibres, plotIMC uses 95% PI of control myofibres i.e. myofibres lying out 95% PI are classed as deficient or abnormal. It is suggested that a proportion of myofibres from a subject that lies outside 95% PI are considered deficient (dysfunctional) and patients will have a higher proportion

Table 2.3 Per myofibre statistics generated using mitocyto. MED, LOG prefix before protein marker refers to median, log of pixel intensities. Mean pixels intensities are represented by protein marker name without prefixes.

Subject_ID	Myofibre count	Input features
C01	148	SDHA, LOG_SDHA, MED_SDHA; NDUFB8, LOG_NDUFB8, MED_NDUFB8; OSCP, LOG_OSCP, MED_OSCP, NDUFA13, LOG_NDUFA13, MED_NDUFA13, VDAC1, LOG_VDAC1, MED_VDAC1, COX4+4L2, LOG_COX4+4L2, MED_COX4+4L2, MTCO1, LOG_MTCO1, MED_MTCO1, UqCRC2, LOG_UqCRC2, MED_UqCRC2, Area, AspectRatio, Perimeter, Circularity, xCoord, yCoord
C02	289	
C03	131	
P01	337	
P02	232	
P03	1361	
P04	879	
P05	1878	
P06	808	
P07	755	
P08	628	
P09	946	
P10	602	
	Total = 8994	

of deficient myofibres [1]. The results for all 13 subjects based on 95% PI for each protein marker is presented in Table 2.4. As observed in the table only for one group of patients, i.e. P01&P02 with nDNA-encoded mutations, can the 95% PI classify myofibres with >90% accuracy and for most groups using 95% PI is not a good discriminator, this can be due to more complex associations exist between protein makers that is not captured by 95% PI.

Advantage of existing analysis workflow

- Without mitocyto myofibre segmentation would be a laborious task that would take hours and days. In our experience manually annotating a myofibre can take up to two minutes per myofibre and a section can have on average 1000 myofibres and studies usually have dozens of sections.

- plotIMC allows comparison of statistical summaries across multiple subjects and proteins in both graphical and tabular format.

- Using plotIMC's 95% PI for classification certain patients' myofibres can be accurately classified as described in Table 2.4

Limitations of existing analysis workflow

Table 2.4 Classification using plotIMC. plotIMC uses 95% PI of control myofibres to classify myofibres based on their mean protein expression measured in terms of pixel intensities. The green cells are where percentage of myofibres that are either above below the 95% PI line exceed 90 %, and orange cells are where these proportions are between 70%-90% .

Subject	PI category	NDUFA13	NDUFB8	SDHA	UqCRC2	COX4	MTCO1	OSCP	VDAC1
P01	ABOVE	0.00	0.00	98.52	97.93	87.87	84.91	99.11	0.00
P01	IN	0.00	0.00	1.48	2.07	12.13	15.09	0.89	100.00
P01	BELOW	100.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
P02	ABOVE	0.00	0.00	70.24	71.03	50.79	88.89	61.51	0.00
P02	IN	0.00	0.00	29.76	28.57	49.21	11.11	36.90	100.00
P02	BELOW	100.00	100.00	0.00	0.40	0.00	0.00	1.59	0.00
P03	ABOVE	2.13	4.12	0.07	3.16	2.72	0.96	0.29	0.00
P03	IN	95.44	91.25	99.71	96.03	93.38	96.40	98.24	100.00
P03	BELOW	2.43	4.63	0.22	0.81	3.90	2.65	1.47	0.00
P04	ABOVE	1.87	0.00	36.21	59.35	76.17	45.79	55.26	0.00
P04	IN	92.64	71.73	63.79	40.30	18.22	47.43	44.74	100.00
P04	BELOW	5.49	28.27	0.00	0.35	5.61	6.78	0.00	0.00
P05	ABOVE	0.53	0.11	11.34	43.77	27.42	4.53	17.89	0.00
P05	IN	74.92	70.07	88.60	52.24	62.25	81.15	81.15	100.00
P05	BELOW	24.55	29.82	0.05	3.99	10.33	14.32	0.96	0.00
P06	ABOVE	5.62	0.85	7.81	55.43	33.21	12.82	11.48	0.00
P06	IN	81.07	76.07	92.19	41.15	60.81	77.66	86.08	100.00
P06	BELOW	13.31	23.08	0.00	3.42	5.98	9.52	2.44	0.00
P07	ABOVE	0.40	0.13	68.64	66.67	16.86	3.16	41.90	0.00
P07	IN	33.60	29.78	31.36	33.07	78.13	77.08	57.84	100.00
P07	BELOW	66.01	70.09	0.00	0.26	5.01	19.76	0.26	0.00
P08	ABOVE	0.16	0.00	89.47	5.74	3.03	0.00	90.75	0.00
P08	IN	13.56	10.85	10.53	38.44	9.41	10.69	9.25	100.00
P08	BELOW	86.28	89.15	0.00	55.82	87.56	89.31	0.00	0.00
P09	ABOVE	3.04	2.72	4.19	0.21	23.98	9.53	3.87	0.00
P09	IN	51.52	49.11	94.66	69.74	49.11	55.39	74.55	100.00
P09	BELOW	45.45	48.17	1.15	30.05	26.91	35.08	21.57	0.00
P10	ABOVE	0.17	0.00	58.08	6.01	0.86	0.34	8.93	0.00
P10	IN	14.43	11.68	41.75	34.02	15.64	13.75	74.40	100.00
P10	BELOW	85.40	88.32	0.17	59.97	83.51	85.91	16.67	0.00
C01	ABOVE	0.68	4.05	0.00	0.00	0.68	0.68	1.35	0.00
C01	IN	99.32	95.95	93.92	93.92	98.65	97.30	97.97	100.00
C01	BELOW	0.00	0.00	6.08	6.08	0.68	2.03	0.68	0.00
C02	ABOVE	1.38	0.35	1.38	0.35	0.69	0.00	2.08	0.00
C02	IN	98.62	99.65	98.62	98.96	97.23	100.00	97.92	100.00
C02	BELOW	0.00	0.00	0.00	0.69	2.08	0.00	0.00	0.00
C03	ABOVE	1.53	9.92	0.76	8.40	1.53	13.74	4.58	0.00
C03	IN	88.55	82.44	97.71	86.26	88.55	74.81	84.73	100.00
C03	BELOW	9.92	7.63	1.53	5.34	9.92	11.45	10.69	0.00

- Mitocyto segmentation needs manual updates or intervention both for segmentation and curation. More importantly the quality of annotations is not evaluated and just relies on subjective intuition. Segmentation and curation is the fundamental first step of the analysis upon which the reliability of any further analysis is dependent.
- plotIMC works on statistical summaries ignoring intra-myofibre features.
- With plotIMC it is not possible to leverage multichannel protein analysis i.e. each protein's summary is treated individually and compared parallelly.
- Using plotIMC's 95% PI for classification only myofibres from P01 & P02 can be accurately classified and to a lesser degree (70%-90%) also myofibres from P08 & P10. But for the remaining subjects the current pipeline cannot classify the myofibres.
- The association between various proteins for individual subjects or groups of genetic diagnoses cannot be explained with the current pipeline. plotIMC allows observation of the correlation between protein summaries but this does not consistently hold true across subjects which make validating any association irrelevant.
- The whole workflow requires manual interventions i.e. in segmentation, myofibre curation (selecting 'analysable' quality myofibres), moving data from mitocyto to plotIMC.

2.8 Other existing tools and methods to study multiplex IMC image data

2.8.1 Napari

Napari [86] is an open-source, multi-dimensional image viewer for Python. It is designed specifically for browsing, annotating, and analysing large multi-dimensional images, making it particularly useful for the multiplex IMC images. Napari allows development of scripts, widgets and plugins on top of it enabling the building of powerful workflows. This functionality can be leveraged to build custom image segmentation workflows such as SM tissue IMC image segmentation.

2.8.2 IMC data analysis workflow by Windhager group

Windhager *et al.* have developed a suite of tools for the analysis of multiplex data including IMC data [78]. Their end-to-end multiplex image analysis workflow consists of 35 steps and

about a dozen tools for various analysis tasks e.g. visualisation, segmentation, clustering analysis, principal component analysis (PCA). Some of these tools which are relevant for our use case are discussed below.

Napari-IMC [78] is an image viewer Napari plugin to view multi-dimensional raw IMC data from its proprietary MCD file. In addition to a multichannel view it allows the user to observe multi-regions and modes (acquisitions, panoramas), providing a holistic view of IMC raw data.

Steinbock [78] is an image segmentation tool for multiplex (including IMC) imaging data. It is a python package that incorporate various biomedical segmentation methods and models such as cellprofiler [83], Ilastik [84], DeepCell [87, 88], and Cellpose [89] into a single tool. It allows the user to select from these methods and pretrain models for image segmentation as per their requirement e.g. tissue type. In addition Steinbock allows users to generate per cell statistical and morphological summaries similar to mitocyto in CSV text files.

Cytomapper [90] is an open-source Bioconductor/R package designed for the visualisation and exploration of per cell summaries of highly multiplexed imaging data. It facilitates the analysis of spatial patterns and cell-cell interactions within the tissue. It also has an R shiny application (GUI) that allows visualization based on hierarchical gating of cells based on protein marker levels.

imcRtools [78] is an open-source Bioconductor/R package that offers helper functions for analysis of per cell summaries of highly multiplexed imaging data. Using spatial graph constructs it provides a number of analyses such as cell-cell interaction, spatial clustering.

2.8.3 Case for using these tools for SM tissue IMC data analysis to understand mitochondrial disease pathology

While these are useful methods and tools for analysis of multiplex (including IMC) data, they suffer from similar limitations to the existing pipeline in the context of our use case as discussed in Section 2.7.2.

Case for SM tissue image segmentation

Steinbock is a flexible tool that allow users to select segmentation tools/models that are appropriate to their use case. But this cannot perform the segmentation for the type of

analysis that is required for IMC in skeletal muscle i.e. precise segmentation of myofibre, removal of non-analysable regions (folded tissue) and myofibres (freezing damaged and non-transverse sliced (NTM). Furthermore, in my experience the generalised segmentation tools and models (without customised retraining) that are available within Steinbock do not produce the segmentation quality required for the use case researched in this thesis.

Case for leveraging full potential of SM IMC data

Similar to plotIMC both Cytomapper and imcRtools work with per cell summaries of IMC multiplex data and so are exposed to similar limitations i.e. ignoring intra-myofibre features.

2.9 Computer vision

Computer vision (CV) is a field of computer science that focuses on enabling computers to interpret a high-level understanding from digital images or videos. It involves the development of algorithms and models that allow computers to process, analyse, and make decisions based on visual inputs such as images [91]. CV can be classified based on tasks it accomplishes such as image segmentation, object detection, image classification, image registration. Relevant to this thesis are the following CV techniques.

2.9.1 Image segmentation

Image segmentation is a CV technique that deals with classification of pixels within an image into discrete classes with an aim to partition areas in the images that might represent objects such as cells within a tissue [92]. Depending on number of discrete classes and tagging of individual instances of the same object, image segmentation can be divided into four sub types.

Edge-based segmentation

Edge-based segmentation techniques work by identifying edges of objects within an image by employing algorithms such as watershed, threshold, Canny edge detection and finding contours [93]. These contours are processed and converted to a segmentation mask as used in mitocyto.

Semantic segmentation

Semantic segmentation techniques work by assigning pixel-level labeling of segmented areas to object classes such as segmenting all dogs and cats in an image as two discrete pixel labels [92]. A case for semantic segmentation for our use case might be for segmentation of folded tissue regions i.e. three semantic pixel labels one representing folded tissue region, normal tissue regions and background.

Instance segmentation

Instance segmentation techniques extend semantic segmentation by not only labeling each pixel but also distinguishing between different instances of the same object class [92]. A case for instance segmentation for our use case might be for myofibre segmentation i.e. each instance of myofibre within SM tissue image label with a unique pixel value/id.

Panoptic segmentation

Panoptic segmentation combines the tasks of semantic and instance segmentation. It provides a complete scene understanding by assigning a unique label to each pixel and identifying both object instances (e.g. individual cells) and semantic classes (e.g. folded tissue, normal tissue) in the image [92].

2.9.2 Image classification

Image classification is a CV technique that deals with classification of images, i.e. assigning a label or category to an entire image based on its content/features. The goal of image classification is to recognise objects, patterns, or features within an image and classify them [94]. A case for image classification for our use case might be for freezing damaged myofibre classification.

2.10 Supervised machine learning

Supervised machine learning is a paradigm of machine learning where ML models are trained with labeled data i.e. pair of input and expected output. The goal of the model is to learn associations between features/patterns in the input data and output label, so that it can correctly predict the label of unseen data [95]. This approach is “supervised” because the learning process is guided by the known ground truth labels during training. This ML

technique can be applied to a range of predictive tasks such classification, segmentation, regression.

The training of supervised ML models is performed by splitting the data into training and testing sets. The ML model is trained on the training set by minimising a loss function that measures the difference between the predicted outputs and the actual ground truth. This minimising of a loss function is achieved by adjusting the internal parameters (weights) of the models to minimize loss function which is typically done using an optimisation algorithm like stochastic gradient descent. Depending on the predictive task there are various loss functions such as binary cross-entropy for binary classification, categorical cross-entropy for multi-class classification, mean squared error, mean absolute error. The model can become too tailored to the training data, reducing its ability to generalize to new data. This is called overfitting which can be mitigated by employing various methods such as using cross validation data sets while training that allow detection of early signs of overfitting, i.e. when training loss is noticeably lower than validation loss, using data augmentation which allows introduction of artificial variation and noise in the training data, using early stopping, i.e. during training if the validation loss starts to increase while training loss continues to improve, stop the training early.

2.10.1 Machine learning for tabular data

ML for tabular data (such as multiplex summaries of myofibres) is not fundamentally different to other data modalities but a more structured format of tabular data makes many ML models applicable to it. Performance of ML models is dependent on their ability to identify relationships between input features and dependent (expected output) variable. The extent of complexity involved in this relationship needs comparable ML models that can deal with the complexity at hand.

Simple linear models perform better where the relationship between the input features and dependent variable is less complex (that can be approximated using linear functions), as this relationship get complex Tree-based and deep learning (DL) models are more useful. But it had been observed that for tabular data Tree-based models perform better than DL [96].

Generalised Linear Models (GLMs)

GLMs are a family of models that extend from traditional linear regression models that takes the general form of

$$Y = \beta_0 + X\beta \tag{2.1}$$

where Y is the dependent variable and X is the input features vector, β_0 is intercept (bias) and β is coefficient (weight) [97] .

Logistic Regression (LR)

LR is a type of GLM for binary classification i.e. when the dependent variable is binary. LR can be extend to multinomial dependent variables. The LR takes the general form of

$$Y = \frac{1}{1 + e^{-}(\beta_0 + X_i\beta)} \tag{2.2}$$

that transforms the linear regressed value to probability ranging from 0-1 using functions such as sigmoid and a threshold (usually 0.5) can be selected for binary dependent variable prediction [97] .

Tree based models

Tree or decision tree based models are a type of ML predictive model that works by recursively splitting the data into subsets based on the values of input features. This progressively increases purity, i.e. a measure of how mixed the elements are within a node (subset), a pure node is where all elements belong to a single class, i.e. leaf nodes that are terminal nodes that represent class labels of dependent variables. There are a number of parameters like maximum depth, minimum number of samples in a leaf node and optimisers that make training the tree based models possible [98].
Most of the time a single tree might be too simple to capture complex relationships/patterns in the data leading to poor performance. To address this, the concept of a tree can be extended to an ensemble of trees i.e. training number of decision trees and aggregating them to improve the performance. There are two ways of aggregating these trees – bagging: training multiple models independently and in parallel on different subsets of the training data, and combining their predictions; boosting: in contrast to bagging in boosting models are trained sequentially, where each new model tries to correct the errors made by the previous models to reduce both bias and variance by focusing on the most difficult cases in the training data.

XGBoost

XGBoost (XGB) stands for Extreme Gradient Boosting and is an extension of tree based models that builds an ensemble of decision trees by training them sequentially and selecting the trees that reduce the loss function. In addition to tree parameters, it has parameters related to boosting (adding new trees) and gradient descent to minimise the loss function [99].

2.10.2 Deep Learning

Deep learning (DL) is a sub-field of machine learning that involves training artificial neural networks (ANNs)/deep learning models that are composed of multiple processing layers to learn representation of data with multiple levels of abstraction [100]. Artificial neurons are the basic building block of ANNs that self-optimise by learning, layers of neurons are interconnected to make an ANN. DL is well suited for complex tasks such as image classification or segmentation. Use of conventional ML techniques for such tasks requires overhead of pre-processing the data to extract features, select appropriate features for training, and any sub-optimal pre-processing leads to poor ML model performance [101]. In contrast DL models have the ability to learn features automatically from inputs such as images [101].

Artificial neuron is inspired by biological neurons in the human brain, artificial neuron (neuron) is a computer programming function that receives inputs with weights and produces an output based on activation function.

$$y = f(\sum x_i w_i + \beta) \tag{2.3}$$

where f is activation function, x_i is input to neuron, w_i are weights and β is bias.

Activation function determines the output of a neuron and depending on the function can produces various outputs such as binary state outputs, i.e. 1. neuron activated state output or 2. neuron not activated state output. This introduces non-linearity into the DL model which allows it to learn complex patterns. There are various activation functions depending on DL model architecture and layers such as rectified linear unit (ReLU)(input) = max(0,input) usually employed in input and hidden layers, sigmoid(input)= $1/(1+e^{-input})$ usually employed in the output layer.

Layers in DL model are serial arrangement of neurons and interconnection between them. Typically there are three types of layer in DL models. i) Input layer is the first layer which receives raw input e.g. pixels for an image input; ii) hidden layers are intermediate layers between input and output that consist of number of neurons that transform the input into something the network can use to make a decision; iii) output layer is the final layer of the DL model that produces the prediction. The number of neurons in this layer corresponds to the number of classes in the classification problem such as predicting genetic mutation of myofibres.

Training a DL model involves adjusting weights (w) and biases (β) of neurons using backpropogation. The backpropogation process includes i) forward pass represent the flow of input through the ANN, i.e. input data passing through layers to produce output, ii) loss calculation using loss function that calculates the difference between predicted output and actual (ground truth) output; iii) backward pass is propagation of error (loss) back through the ANN from the output layer allowing the adjustment of weights (w) and biases (β) to minimise the loss.
To reduce over fitting and efficient convergence of DL models, various techniques are used, such as i) **Early stopping**: as discussed above it can be implemented by monitoring performance of the model on the validation set compared to training set and stopping the model training when its performance degrades on the validation set compared to the training set. ii) **Weight initialisation**: The selection of initial weights determine the time when the model converges (completes training). This can be selected randomly or use informed weights from pre-trained models. iii) **Drop-out**: is a mitigation to overfitting where a proportion of outputs from neurons are discarded to reduce the complexity and by extension overfitting of the model. iv) **Data augmentation**: is a process in which artificial noise and variations are introduced in the training data, e.g. rotation of image, that allow the model to generalise. v) **Batch normalisation**: is a process in which at each batch of training, the layers' input are normalised and scaled to improve training and fast convergence [102].
Optimiser are algorithms that are used to adjust weights and biases of neurons in the DL model. This plays an important role on DL model training by determining how the model parameters are updated during training. The selection of an appropriate optimiser and learning rate determine how fast the model training converges and also the predictive performance of the model. There are various options of optimisers to select from such as SGDM [103], AdaGrad [104], Adam [105], RMSProp [106] and usually during the training process selection of these with a combination of learning rates are tried.

DL model architectures

The arrangement of these layers and depth, i.e. number of hidden layers, defines DL model architecture. The DL model architecture is inspired by its application, i.e. the problem it is aiming to solve. Some of the architectures are convolutional neural networks (CNN), recurrent neural networks, transformer networks. The work accomplished in this thesis relies heavily on CNN based models.

Convolutional neural network A traditional ANN tends to struggle with the complexity involved in recognising patterns in images [107]. A CNN addresses this limitation by

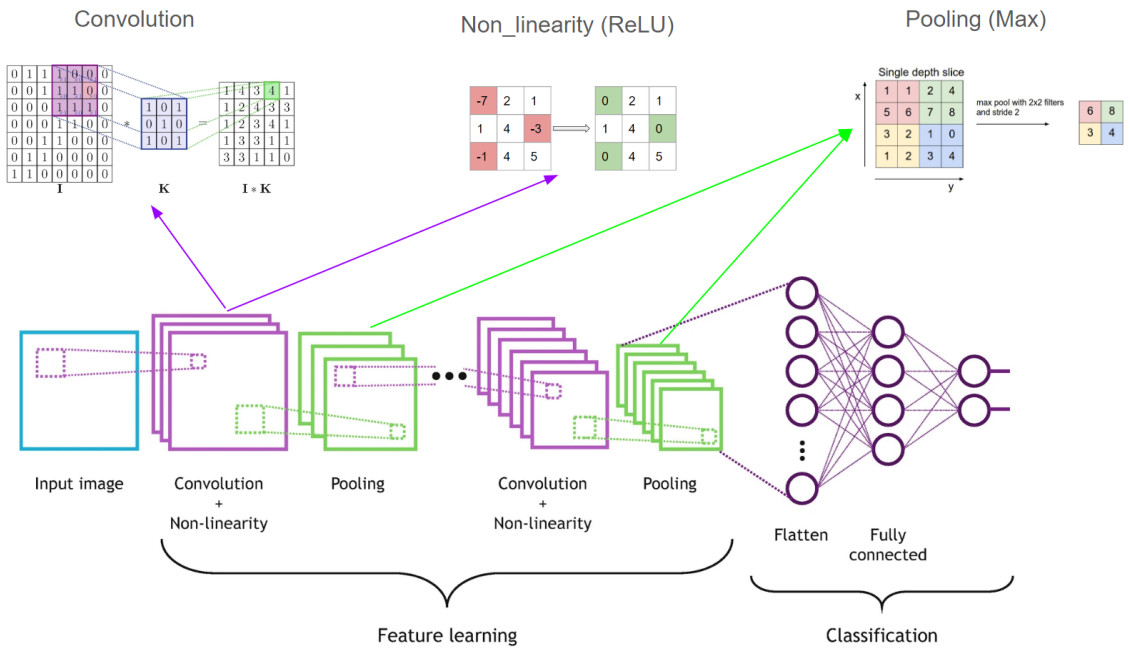

Figure 2.10 CNN architecture

Architecture of a simple CNN model. A CNN model consists of series of convolutions with non-linearity layers, pooling layers, followed by flattened and fully connected layers attached to the output layer. **Convolution layers** perform convolution operation where a kernel/filter 'K' i.e. a small matrix of (nxn) size is multiplied element-wise followed by a summation, with whole input matrix by sliding in steps, in each step a filter is moved across the image in a stride, e.g. stride=1 will move one pixel/element of matrix at a time. This is followed by padding where extra pixels/elements are added around convolution output to make it the same size as input. This is followed by application of **non-linearity** activation function such as ReLU that introduces non-linearity. The result after convolution operation is a feature map that highlights/detects features in the image. **Pooling** layers reduce the spatial dimensions of feature maps by aggregating their elements by using various techniques such as maximum pooling, average pooling where max, mean value is computed across non-overlapping regions of feature map in steps with predefined pooling size. Finally, output is flattened to 1D (for classification) and passed to a **fully connected** layer where every neuron in it is connected to every neuron of adjacent layers, including an output layer which has neurons equal to the number of classes required as output. Figure adapted from [102].

Figure 2.11 VGG16 architecture
Architecture of a VGG16 model. VGG16 models consist of a series of convolution pooling layers, followed by flattened and fully connected dense layers attached to the output layer.

improvements in the architecture of the ANN. The architecture of a simple CNN model is described in the figure 2.10.

2.10.3 Deep Learning for Computer Vision

DL models based on CNN approaches have achieved great success in image classification tasks in various domains such as bio-medicine, finance, and manufacturing [108–112]. Development of these models has been popular in the last decade leading to the invention of many new models that have considerably improved performance, setting new records for prediction accuracy on large public image datasets like ImageNet [113]. These models typically use convolutional and pooling layers with a large number of filters or other layers or use techniques such as dropout, batch normalisation, ReLU activation, inception modules and residual learning to alleviate problems such as overfitting and vanishing gradient. These initial layers are typically followed by dense fully connected layers and an output layer that uses an activation function (such as sigmoid) to convert logit into outputs [114].

One way of tracking the best models over the years has been to track the winners of ImageNet's ILSVRC challenge which highlighted models such as AlexNet, ZFNet, VGG, GoogLeNet and ResNet that are widely used in image classification applications today [115]. While the architectures of these DL models are the main cause of their improvements, these models need massive training datasets, e.g. 14 million images in the ImageNet dataset to achieve these results [114]. There are many real world cases (including ours) where availability of such a volume of data is impractical. In such cases transfer learning can often improve the performance over training only on the appropriate, but small, dataset. Transfer learning is a method where the model is first trained on a related large dataset and the trained weights are used as a starting point to further train the model with the original small dataset [114].

VGG16

VGG16 [116] is a very deep convolution neural network with 16 layers that is trained on large datasets including ImageNet [115]. It is considered an appropriate CV models for object recognition and image classification [116] and performed well on images from variety of domains [117]. VGG is an extension of CNN models that uses small (3x3) convolution filters with increased depth of 16 (also 19 for another version) weight layers, i.e. learnable parameters, as described in Figure 2.11. VGG16 can be adapted to any input image size such as in our case (height x width x 13 channels).

ResNet

ResNet [118] is another CNN based model that proposed a novel way of expanding the layers of CNN with actual performance improvements, it comes with various versions of weight layers including 18,50,101 layers. The main concept of ResNet is residual learning, i.e. instead of learning the direct mapping from input to output, ResNet learns the residual mapping. This means that each layer (or set of layers) only needs to learn the difference (or residual) between the input and the output, which can be easier to optimise. This can be formally defined as, instead of learning the F(x), i.e function mapping after the convolution and non-linear layers, they fit another mapping function H(x)=F(x)-x on which the original mapping is recast into H(x)+x. Feed forward neural networks can realise this mapping with "shortcut" residual connections by performing a simple identity mapping, and their outputs are summed to the outputs of the layers. Such "shortcut" residual connections do not add additional complexity nor parameters to the model, making this architecture very powerful and efficient [102]. ResNet50 has been used in this thesis for various classification tasks.

Hybrid models

Combining CNN based models such as VGG16, ResNet50 or a simple CNN, with tree-based models such as Random Forest, XGBoost is an interesting approach that leverages the strengths of both techniques for tasks such as image classification. The idea is to use a CNN-based model to extract features from images and then use a tree-based classifier to perform the final classification based on these features. This approach has been tried for some image classification tasks in this thesis.

UNET

UNET is a type of convolutional neural network (CNN) specifically designed for image segmentation tasks, particularly in biomedical image analysis [21]. Its architecture consists

of two main parts i) Contracting Path (Encoder): this consists of several convolution layers followed by max-pooling layers, which progressively reduce the spatial dimensions while increasing the number of feature maps. Each convolution layer typically uses ReLU (Rectified Linear Unit) activation and is followed by a pooling operation that downsamples the image. ii) Expanding Path (Decoder): this consists of upsampling operations (usually transposed convolutions) followed by convolutional layers. This path also includes skip connections that concatenate feature maps from the corresponding contracting path layers, which helps to recover fine-grained details and improve segmentation accuracy [21]. This allows UNET to be efficient in semantic segmentation tasks such as folded tissue region segmentation. UNET and its variants are used for semantic segmentation tasks in this thesis.

Mask R-CNN

This is a DL model designed for instance segmentation tasks. It consist of i) Backbone Network: A CNN used for feature extraction from the input image. Common choices for the backbone include CNN, VGG or ResNet, often combined with a Feature Pyramid Network (FPN) that allows feature extraction of objects of different sizes. ii) Region Proposal Network (RPN): This network proposes candidate object regions (RoIs). It outputs a set of proposed bounding boxes that potentially contain objects. iii) Bounding Box Regression and Classification: For each proposed RoI, the network predicts the class of the object and refines the bounding box coordinates. iv) Mask Prediction Branch: This branch is added parallel to the bounding box regression and classification branches. It predicts a binary mask for each RoI, allowing for pixel-level segmentation of the detected objects [119]. This object class, its corresponding bounding box and mask results in an instance segmentation mask such as ones required in myofibre segmentation. In this thesis mask R-CNN is experimented with alongside other instance segmentation methods for myofibre segmentation.

Stardist

StarDist [120] is a novel CNN-based instance segmentation model that is designed to predict segmentation of objects of star-convex shape such as cells or myofibres in biomedical images. Its approach represents object shapes using star-convex polygons, which simplifies the problem of predicting object boundaries, i.e by just predicting an object's center and radial distances (default set to 32) from center to boundaries, this significantly improves segmentation performance [120]. The architecture of StarDist is explained in Figure 2.12. In this thesis StarDist is experimented with alongside other instance segmentation methods for myofibre segmentation.

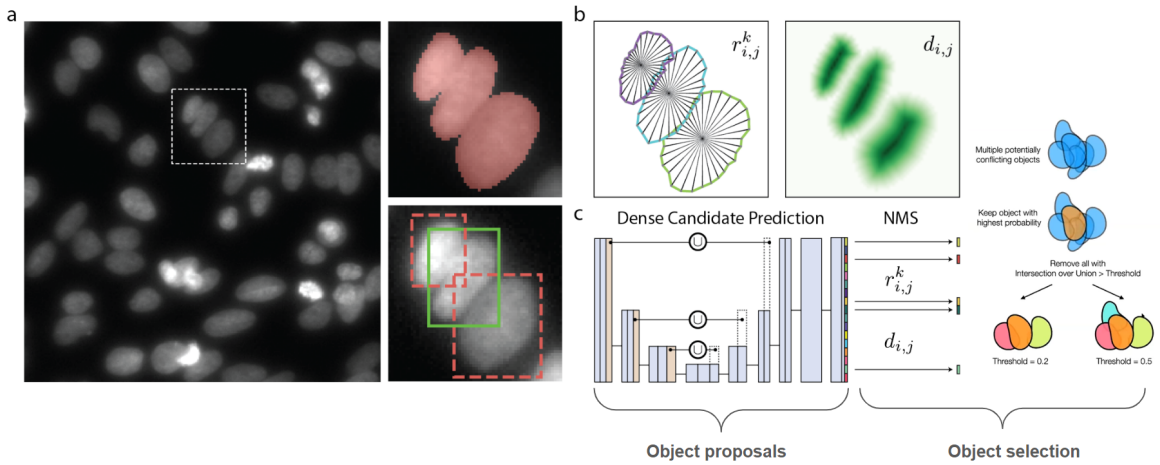

Figure 2.12 StarDist architecture. **a** is a sample input image that shows typical touching objects (cells). **b** StarDist works by learning and predicting parameters $r^k_{i,j}$ radial distances (default =32) of object k from centre (i,j) and $d_{i,j}$ is the probability of (i,j) being the centre of the object (cell). **c** Object proposals: these parameters $r^k_{i,j}$, $d_{i,j}$ are learned by a CNN-based backbone typically a UNET or mask R-CNN, for all possible candidate objects. Object selection: a non-maximum suppression algorithm selects/predicts the most likely object from proposed object candidates. Figure adapted from [120]

.

Cellpose

Cellpose [89] is another novel CNN-based instance segmentation model that is designed to predict cell (myofibre) shaped objects. It has performed well in various biomedical segmentation tasks such as cell and nuclei segmentation [83, 87, 121, 122]. Cellpose is an instance segmentation model that uses vector flow representation generated by simulated diffusion starting from the centre of the cell (in our case myofibre) toward its border in the annotation mask. It also uses a parameter to determine whether a pixel is inside or outside the cell. A neural network (typically a UNET) is then trained to predict vertical, horizontal gradients and whether a pixel belongs to any cell. These three predictions are combined to form flow mask and cell probability mask, these are then used to create final instance segmentation masks. We can optimise the final results using i) flow threshold that determines maximum allowed error of flows in flow masks and ii) cell probability threshold that determines whether pixels belong to a cell or not [89]. The architecture of cellpose is explained in Figure 2.13. In this thesis cellpose is experimented with alongside other instance segmentation methods for myofibre segmentation.

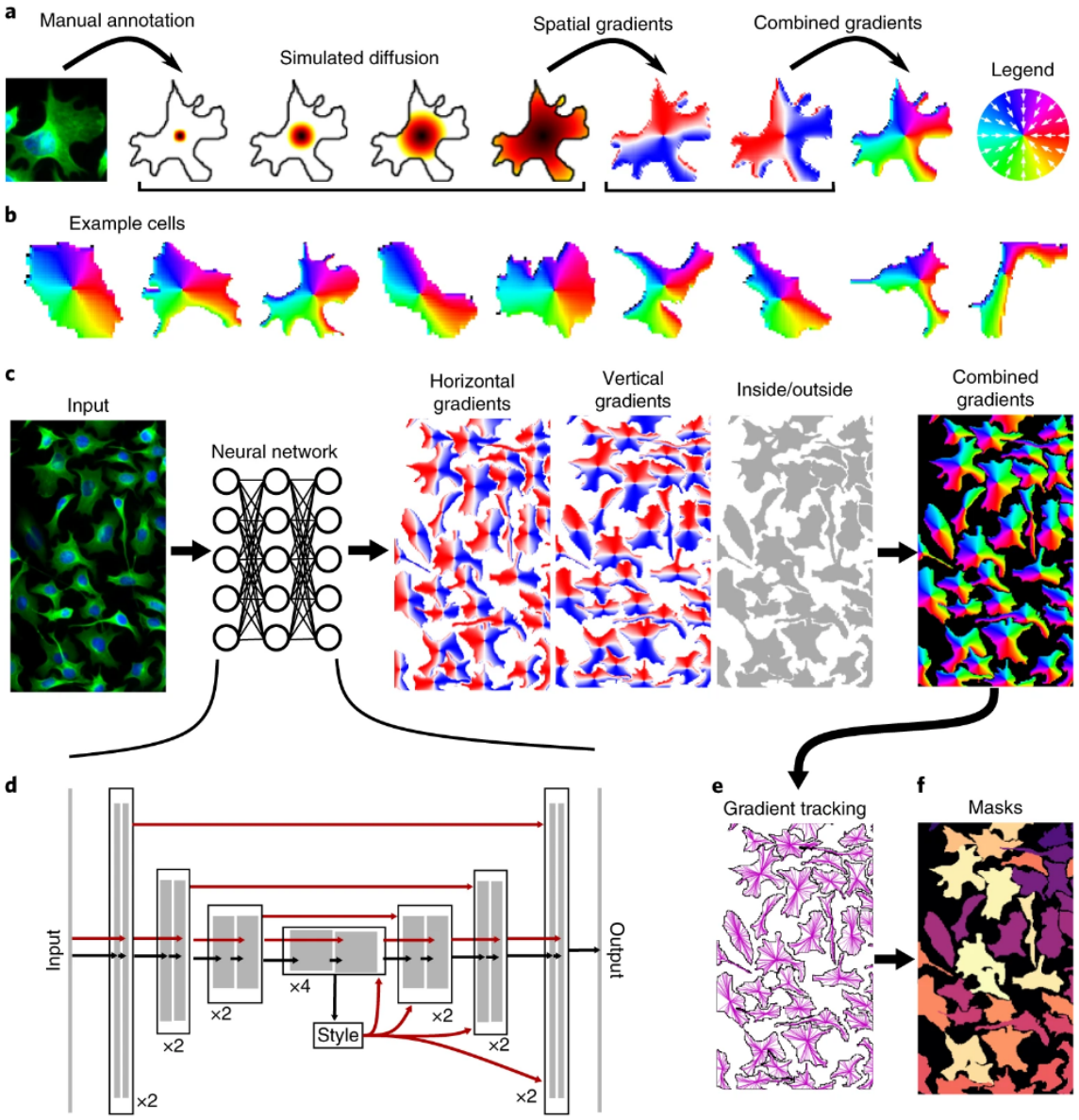

Figure 2.13 Cellpose architecture. **a** At the time of training cellpose converts manual instance segmentation masks into vector flow representations using simulated diffusion processes starting from the centre of objects (myofibres) in the mask. **b** These vector flow representations for example objects (cells) are made by combining simulated diffusion from objects' centers in both horizontal and vertical directions. **c** These vector flow representations are used to train (**c**) a UNET model that predicts vector flow representation and whether a pixel is inside or outside this vector flows. This is followed by applying (**e**) simple gradient tracing to construct (**f**) instance segmentation masks. Figure from [89]

. .

2.10.4 ML vs DL terminology in the thesis

Throughout the thesis any machine learning model that is not deep learning-based is referred to as machine learning (ML), and any model that uses deep learning is referred to as deep learning (DL).

2.11 Machine learning explainability

In machine learning and artificial intelligence (AI), the terms explainability and interpretability are frequently used interchangeably. Despite how similar they appear, it is crucial to recognise the distinctions. Rudin [123] distinguishes between interpretable and explainable AI: while explainable AI attempts to offer post-hoc explanations for currently used black-box models, which are incomprehensible to humans, interpretable AI focuses on building models that are intrinsically explainable. Lipton [124] emphasises the distinction between the questions that each family of techniques seeks to answer. Interpretability asks, "How does the model work?" while explainability seeks to respond to "What else can the model tell me?" There is no general agreement on what either interpretability or explainability means [125], however, we have used Rudin's definitions of these terms in this thesis. Throughout the thesis explainable AI methods are employed to uncover the basis of model predictions. Explainable methods (EMs) provide post-hoc explanation of the basis for a model's predictions by usually comparing input and output of the model i.e. by changing features in input to observe the effects of it on the output and attribute contribution to that feature based on amount of effect. EMs are effective methods to understand the importance of various input features such as protein markers (channels in IMC multiplex images), pixels within myofibres. But for the use case of understanding mitochondrial disease pathology using multiplex IMC data, knowing just the importance of pixels within myofibre and protein markers is not enough, it requires this importance to be defined as associations, i.e. the importance of a certain feature should be quantified in terms of its proportional importance to rest of the features and its correlation with output prediction. For an example the ideal EM should explain the importance of each protein marker quantified in proportionality terms, which allows the user to understand relative importance to other protein markers. Also it should reveal the protein marker's relationship with the output genetic mutation, e.g. if its under or over expression is linked to a genetic mutation.

SHAP

SHAP [126–129] stands for Shapley additive explanation and is a suite of EMs developed by Lundberg *et al.* that are based on Shapley values [130] from cooperative game theory. These EMs explains ML and DL model predictions by allocating optimal credit/blame to each input feature and revealing their association with the predicted class. This in theory satisfies the requirements of the use case investigated in this thesis. SHAP provides various EMs catering to explainability objectives, and models such as explaining ML models trained on tabular data, computer vision DL models trained on image data.

2.11.1 Explainable methods for tabular ML models

ML models trained on tabular data such as LR and XGB trained on the myofibres summaries in this thesis are interpretable due to their relative simple architecture. These methods also provide input feature importance scores using parameters such as logistic regression coefficients etc. that gives insights about the relative importance of each input feature. But as discussed earlier this is not enough for the use case investigated in this thesis, it requires quantified credit/blame assigned to each input feature and the relationship between input feature and the predicted class. Fortunately, SHAP provides EMs for models such as LR and XGB.

SHAP provides various EMs and corresponding plotting functions to explain tabular data trained ML models. These EMs usually explain the model's prediction output using background/reference dataset, i.e. plotting each instance in reference dataset with their SHAP values revealing both their relative proportional importance and their relationship/association/correlation direction with the predicted class. There are a number of SHAP EMs that are relevant to models used in this thesis such as i) Tree SHAP: Specifically optimized for tree-based models (e.g XGBoost), Tree SHAP efficiently computes Shapley values, providing explanation for given reference dataset, ii) Kernel SHAP: Suitable for explaining any model by approximating Shapley values using a sampling-based approach. It provides accurate explanations but may be computationally intensive for large datasets, iii) SHAP explainer: is the default EM provided by SHAP that automatically selects the most appropriate EM from the SHAP library for a given ML model[127].

SHAP also provides a range of plot functions that are very powerful in visualising the associations between input features and predicted classes. Some of the plots used in this thesis are i) SHAP waterfall plot: to look into the basis of prediction for individual instances, waterfall plots are useful to explain proportional importance and associations in terms of SHAP values, ii) SHAP bar plot: to look into the basis of prediction for a group of instances such as whole

training data, this can help to understand 'global' explanation of model predictions and can reveal proportional importance of each feature, iii) SHAP beeswarm plot also allow observation of the basis of predictions of group of instances but it reveals the associations both in terms of their proportional importance and direction of these associations, i.e. positively/negatively correlated with predicted class. iv) SHAP heatmap plot allows observation of comparative SHAP values of all features for a group of instances and a model's logit output arranged using hierarchical clustering by the similarity of explanations, i.e. SHAP values.

2.11.2 Explainable methods for CV DL models

There are EMs that provide post-hoc explanations for DL model predictions [131]. These can be categorised in many ways i) Global vs Local methods: global methods try to explain the overall decision making process of DL models by presenting explanation across all training data instances and local methods try to explain individual decisions [131]. ii) Gradient-based vs Perturbation-based methods: gradient-based methods use gradient of the output with respect to the input or extracted features to explain individual decisions; perturbation-based methods perturb input features by removing or altering their values and calculate the effect on model performance thereby finding the important features [131–133]. iii) Function, Signal and Attribution methods: This categorisation is based on the information these groups of methods present. They provide different information about model predictions that are usually complementary to each other [134].

Function methods [Grad-CAM [135] Gradients [136] or Saliency]

These are basic gradient-based methods that use gradients of output neurons with respect to input to estimate importance of input pixels in the image, i.e. greater gradient means more importance [136]. To simplify, for a linear model $y = w \times x$, these methods analyse weights (w) as gradient between output (y) and input (x), $\partial y/\partial x = w$ [134]. For a CV DL model (w) will be a tensor of gradients of the same shape as the input image.

Signal methods [DeConvNet [137], Guided Backprop [19]]

These methods aim to isolate input patterns that simulate neuron activity in the higher layers of CNN i.e. analysing the components of the input data that causes the output [134]. They do this by applying some activation function over the gradients [131, 138].

DeConvNet is a reverse CNN that maps features detected in higher layers to input pixels [137]. It does this by applying an activation function (ReLU) in the 'importance' calculation

instead of just using the gradients [138]. This allows only the positive gradients or signal to be chosen which shows the most important features in the input [137]. Guided Backprop follows a similar process as DeConvNet but also uses the backpropagated gradient in the 'importance' calculation [19].

Signals are more informative than functions in that they tell us both the regions and direction of the input image that are used by the model to predict output[114]. Signal methods like DeConvNet and Guided Backprop are used in the analysis of medical imaging data. For example, De Vos *et al.* [139] assessed coronary artery calcium for each slice of a heart or chest computed tomography (CT) image and applied deconvolution to reveal where in the slice the decision was made.

Attribution methods [Deep Taylor [140], Input Gradient [131, 138], Layer-wise Relevance propagation (LRP-Epsilon, Z, PresetAFlat, PresetBFlat)] [138, 141]

These methods attribute importance to input features/signal dimensions for the output i.e., how much the signal dimensions/features of the input contribute to the output across the neural network [134, 138]. For a linear model $y = w \times x$ attribution $r = ([w] \otimes a)$ where $[w]$ is weight vector, $\otimes$ denotes element-wise multiplication and a is the signal [134]. Attributions are built upon signals, i.e. attribution tells us the importance of each signal dimension/feature of the input image toward predicting the output. Attributions give more detailed explanation about model prediction than signal and are used to analyse many medical DL models[142].

Bohle et al. [142] employed LRP (an attribution method) to locate Alzheimer's disease-causing areas in brain MRI images. They contrasted the saliency maps produced by guided backpropagation with LRP and discovered that LRP was more accurate in detecting areas known to have Alzheimer's disease.

DeepLIFT

DeepLIFT is an EM designed for DL models. It aims to explain the contribution of each feature (input neuron) to a neural network's output prediction. DeepLIFT operates by comparing the activation of each neuron to a reference activation and assigning contributions based on how different the neuron's activation is from this reference [143].

Integrated Gradient

Integrated gradient is an EM for DL models, similar to DeConvNet and input gradient EMs it uses gradient in the network to compute attribution but instead of just taking the gradient at the input itself, Integrated gradient integrates the gradients along the path from the baseline

to the actual input. This approach accounts for the accumulated effect of each feature as it moves from a neutral state (baseline) to its actual value [144].

All these methods can provide perfect explanations for linear models but DL models are highly non-linear, which means these explanations can only be used as approximations [134]. To come back to our premise that these methods complement each other to find explanations about model prediction, as discussed earlier explanations required for the use case investigated in this thesis require knowing the association between input features (protein marker channels) and mutation class explained in terms of correlation (positive/negative) of protein markers toward prediction and relative importance/contribution of each of these protein markers quantified. This is achieved by adapting the earlier mentioned EMs into SHAP values. There are two EMs that are adapted to build attribution/explanation mask in terms of SHAP values, namely integrated gradient and DeepLIFT with their corresponding SHAP adaptions called 'GradientExplainer' and 'DeepExplainer' respectively. While many of the EMs discussed earlier have been experimented with for the work presented in this thesis, 'GradientExplainer' and 'DeepExplainer' are used for final results of profiling myofibre's genetic mutation in terms of raw multiplex IMC data.

Defining an ideal explainable ML/DL pipeline : An ideal pipeline will give a profile of myofibres associated with a genetic diagnosis. This will be in terms of relationship between all available proteins markers and leveraging the intra-myofibre morphology/ features.

2.12 Chapter summary

This chapter introduced the required background knowledge for understanding the work presented in this thesis. This includes introduction to mitochondrial biology and disease, discussion of existing analysis methods and their limitations, and introduction to ML and DL methods and the corresponding explainable methods.

Chapter 3

Explainable ML Analysis on Processed Data from mitocyto

3.1 Introduction

The current techniques for analysis of multiplex protein data (IMC) such as plotIMC and Cytomapper have limitations as explained in Section 2.7.2. These tools include i) cell (myofibre) segmentation tools such as mitocyto, Steinbock that segments and calculates statistical summaries per cell/myofibre. ii) Per myofibre summary analysis tools such as plotIMC, Cytomapper , imcRtools that allow the user to analyse and visualise multiplex IMC data. These work by plotting relative mean intensities as a colour bar per protein on section i.e. colouring each cell with mean intensity colour that allows a user to observe the difference on a section image. This can then be gated by applying a threshold to define a cell class, cell-to-cell interactions using spatial graphs with cells representing nodes and interactions representing edges.

These methods suffer from limitations as discussed in Section 2.7.2, i.e. it is not possible to leverage multichannel protein analysis i.e. each protein's summaries are treated individually and compared in parallel. Further methods were also considered to analyse multiplex data such as t-SNE, UMAP and PCA, which are all dimensionality reduction techniques. While UMAP and t-SNE did show identifiable clustering by some genetic mutations these were not helpful in profiling the myofibres in term of importance of all available protein levels.

As discussed in 2.1.2 mitochondrial disease is heterogeneous and develops independently of cell-cycle, this makes cell-to-cell interaction analysis of our data less relevant, also this analysis was previously conducted by a group in WCMR which did not yield useful results.

As discussed in Section 2.6.2 plotIMC allows biomedical scientists to analyse individual

protein channels in parallel but association between these channels to profile myofibres based on their genetic mutation is not possible, also using plotIMC's 95% PI it is not possible to classify the genetic mutation class of myofibres except myofibres linked to nuclear-encoded mutations affecting complex I. The nuclear-encoded mutations affecting complex I rely on nuclear DNA affecting all myofibres homogeneously and so the deficiencies are observed across all myofibres in these patients and allow plotIMC to classify myofibres of these patients [145]. Whereas in mtDNA mutations (the rest of the patient groups) there may exist varying proportions of wild-type and mutant mtDNA and therefore a mix of myofibres with different levels of mitochondrial dysfunction. This makes classification of cases as patient or control and profiling of myofibres of these patients difficult. Explainable ML methods can not only leverage complex patterns in the data to classify but also can reveal these patterns that can be a useful approach for the analysis of multiplex (IMC) per myofibre summaries.

3.2 Aims of this chapter

To overcome the limitation of other techniques in analysis of IMC data of mitochondrial disease patients, in this chapter explainable machine learning methods will be used to classify and profile myofibres linked to genetic mutations affecting mitochondrial dysfunction using the processed IMC data. The aims of this chapter are as follows:

- Classify the mitochondrial genetic mutations of myofibres using machine learning and per myofibre statistical summaries of multiplex mitochondrial protein markers' IMC data.
- Profile these myofibres in terms of the associations between mitochondrial protein markers and the mitochondrial genetic mutations by interrogating ML classification models using explainable ML methods.

3.3 Data and methods

3.3.1 Data

To contrast and compare the result, the same processed IMC data as in Section 2.6.2 and the same eight protein channels as presented in Table 2.4 are used. This includes three statistical summaries i) mean of pixel intensities ii) mean over log values of pixel intensities and iii) median of pixel intensities per myofibre and for each of the eight protein markers.

3.3.2 Methods:ML classification for processed IMC data

Exploratory data analysis

Exploratory data analysis (EDA) is usually the first step before any ML analysis, where all input variables/features are explored with an aim to understand the data so that informed decisions for further ML modeling can be made. EDA involves identifying preliminary patterns in the data using techniques such as statistical summaries, dimensionality reduction and correlation analysis [146].
The analysis conducted in Section 2.6.2 is EDA and does divulge many aspects about the data but nevertheless is not complete, i.e. the analysis conducted in Section 2.6.2 is controls vs a patient (tissue section). This does not give much information about various aggregates i.e. overall observations, group of patients of similar diagnosis, group of controls.
We aim to perform EDA across these aggregations and combine the knowledge gain from this and Section 2.6.2 towards selecting relevant features and ML models.

ML modelling

The aim is to perform ML classification analysis on processed IMC data. As discussed in Section 2.10.1 for tabular data usually GLMs and tree based models perform well. The complexity of relationship between input features and dependent variables decides which ML models are useful. The usual model training process was followed which include i) splitting the data into 70%: 15%: 15% for training, validation and testing, ii) optimising the parameters for the model using validation set and iii) evaluating the test results using weighted accuracy and recall which give a fuller picture in terms of both model accuracy in general but also its performance towards both classes.

Based on factors discussed earlier and the biological reasoning mentioned in Table 3.1 it was decided that the ML classification on aggregations mentioned in Table 3.1 will be performed using the sequence of EDA, GLMs (LR) and if this does not yield good classification results then to try tree based models.

3.3.3 Methods:Explainable ML methods for tabular data

As discussed in Section 2.11.1 well trained ML models are good at finding correlations which can sometimes also unravel associations [147]. To explain this let us consider a dataset of multiple input features of which if only a few (or just one) features are enough to classify the dependent variable then high-performing ML models might only use these correlations and ignore any other associations between the other input features and the dependent variable.

Table 3.1 Classes for ML classification analysis

Classes	Reasoning
nDNA encoded mutation (P01 & P02) vs Controls. This will be referred to as **class A** vs **controls**	Both patients suffer from similar mutations and are expected to present similar phenotype. i.e. protein patterns [1]
Point mutation in (MT-TL1)(P05,P06,P07) vs Controls. This will be referred to as **class B** vs **controls**	All three patients suffer from similar mutations and are expected to present similar phenotype, i.e. protein patterns [1]
Point mutation in mito encoded tRNA(P08,P09,P10) vs Controls. This will be referred to as **class C** vs **controls**	All three patients suffer from similar mutations and are expected to present similar phenotype, i.e. protein patterns
Single, Large-scale mtDNA mutation (P03 & P04) vs Controls. This will be referred to as **class D** vs **controls**	Both patients suffer from similar mutations and are expected to present similar phenotype, i.e. protein patterns [1]
P03 vs Controls	Classification of P03 myofibres is particularly challenging using plotIMC as seen in Table 2.4 i.e. > 91% of P03 myofibres were within 95% PI of control myofibres. For this reason and to include an individual patient classification case that will allow us to contrast and compare the analysis in Section 2.6, P03 vs Controls is analysed as a separate case.

This makes our objective of using ML to discover the protein patterns, i.e. all associations between these proteins with mitochondrial genetic mutation, a challenging task.

To address this issue, the following sequential strategy of applied explainable ML is adopted in this thesis.

- **Correlations between input features** from EDA identify highly correlated input features.
- **ML model training** Train models with i) all input features in cognisance of highly correlated features, expect that model will ignore some features. ii) only input features that do not have high correlation between them. iii) separately on each individual feature.
- **Apply explainable ML methods** to models trained on all relevant combinations of input features. This will distill insights that 'explain' the associations between input features and dependent variable. Where relevant and possible, to validate the associations apply multiple explainable methods, e.g. interrogate model using inbuilt importance score and SHAP.

Explainable methods for LR

Logistic regression (LR) is usually the most explainable classification ML model as the correlation and associations can be explain by linear relationships [148]. The model was interrogated using i) its coefficients for input features: it is important to understand that LR although deriving its predictions by finding a linear relationships between input features, it takes shape of log-odds as described in equation 3.1. This means the coefficient β cannot be interpreted as exact weight as in linear regression model but nevertheless its magnitude and direction are relevant in finding its correlation effect toward the predicted variable. ii) By applying SHAP explainer: SHAP values for LR as described in equation 3.2 are calculated by multiplying the coefficient β with the difference between the input feature and its expected value which in turn is calculated as means of the feature from the reference dataset (which in our case is the whole training dataset). But it should also be noted that the final SHAP values are transformed into log-odds space. This makes the SHAP values very powerful and informative not just for global explanation but also for individual local explanations [127].

$$log(P\{Y = patient\}/P\{Y = control\}) = \beta_0 + x_i\beta \tag{3.1}$$

$$\phi_i = \beta_i \cdot (x_i - E[x_i]) \tag{3.2}$$

Equation 3.1 describes LR in log-odd terms for a binary classification where x_i are input variables, Y is predicted binary variable, P is probability of Y being patient or control, β_0 is intercept/bias and β coefficients/weights; equation 3.2 describes SHAP values for a LR model where ϕ_i is the SHAP value for input feature x_i, β_i is the coefficient of x_i and $E[x_i]$ is expected values of x_i computed using the reference dataset.

Explainable methods for XGB

While simple high-bias models like LR seems easy to understand but are sensitive to model mismatch, i.e. a model's ability to capture true relationships in the data, this mismatch can sometime create artefactual relationships that do not really exist in the data [149]. Low-bias models like XGBoost (XGB) are great at capturing non-linear relationships between input features in the tabular data and so are more immune to model mismatch problems [126]. Taking the model-mismatch problem into account and using appropriate explainable methods like 'TreeExplainer' models like XGB can be more accurate representations of patterns in the data and more interpretable than even the LR model [126]. The XGB was interrogated using i) its native "importance_score": which is computed by summing the number of splittings on a feature that reduces the impurity (gain) across all the splits in the tree which is then averaged across all trees in the trained XGB model. This is helpful in deducing the correlation of the input features in magnitude but not in direction. ii) SHAP TreeExplainer: is a combination of the concept of each player's contribution from cooperative game theory, applied to tree-based models to generate local explanations, which are aggregated to provide global explanation for tree-based ML models. This retains local faithfulness to the model while still revealing global patterns, resulting in more informative, detailed explanations that are more accurate representations of the model's behaviour [126]. SHAP TreeExplainer and SHAP plots allow a user to unravel the correlation and associations between input features and predicted variable, not just in magnitude and direction but also its prevalence, most other explainable ML methods conflate magnitude of input feature importance to prevalence [126]. This is further discussed in SHAP explanation figures in Section 3.4.

Explainable methods plots

SHAP plots help explain and delineate various correlations and associations between variables that help the model predict. There are four main types of SHAP plots used in this chapter that are explained below using toy examples.

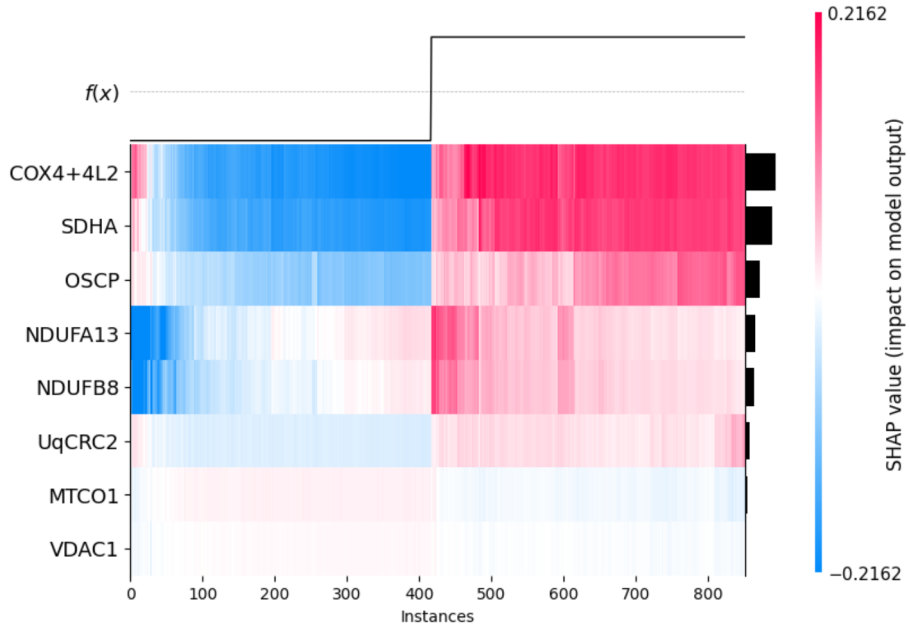

Figure 3.1 **SHAP heatmap** plot shows population of substructures in a model prediction dataset usually clustering these hierarchically such that populations of instances that have same profile are group together. This can be observed in above example i.e. instances of variables (proteins) with similar values are group together. At the top of the plot is model's predicted probability function, in this toy example taking binary values i.e. 0 or 1. The colour (blue or red) gradient of each variable highlights the contribution it makes toward predicted probability. In the heatmap this can be seen across the whole predicted dataset.

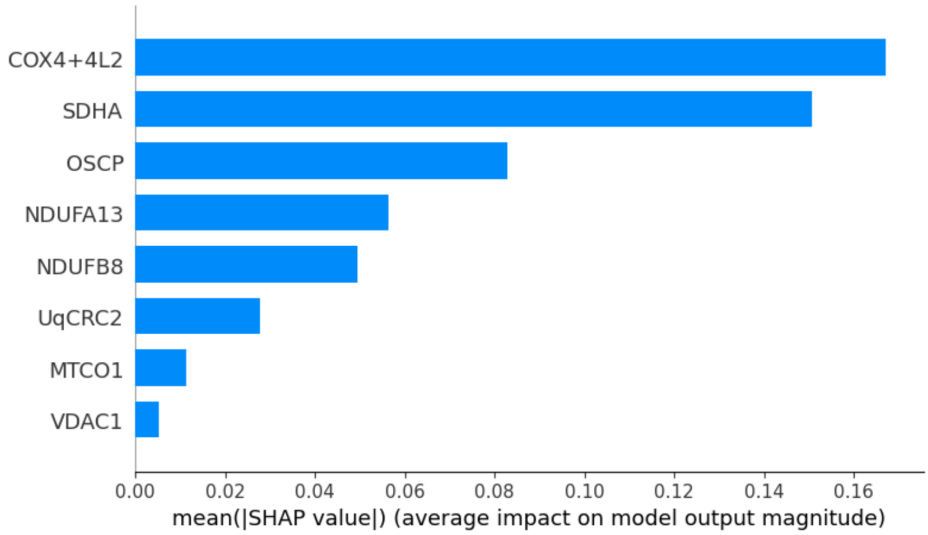

Figure 3.2 **SHAP barchart** plot shows the mean contribution of each variable towards predicted probability (disregarding sign). For the toy example above the variables with greater mean SHAP values denotes their importance toward prediction across the whole predicted dataset.

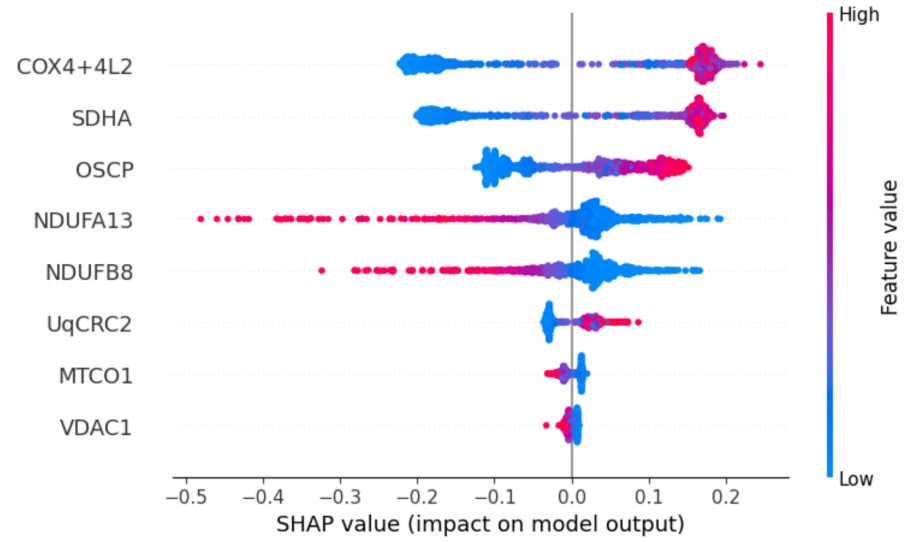

Figure 3.3 **SHAP beeswarm** plot shows prevalence of variables' contribution towards predicted probability across the whole predicted dataset. in the toy example above it can be observed that variables COX4+4L2 and SDHA's high values translates into moderately high positive (around 0.2) predicted probability contribution for many instances. On the other hand NDUFA13's high values translates into extreme high negative (around -0.5) predicted probability contribution for few instances.

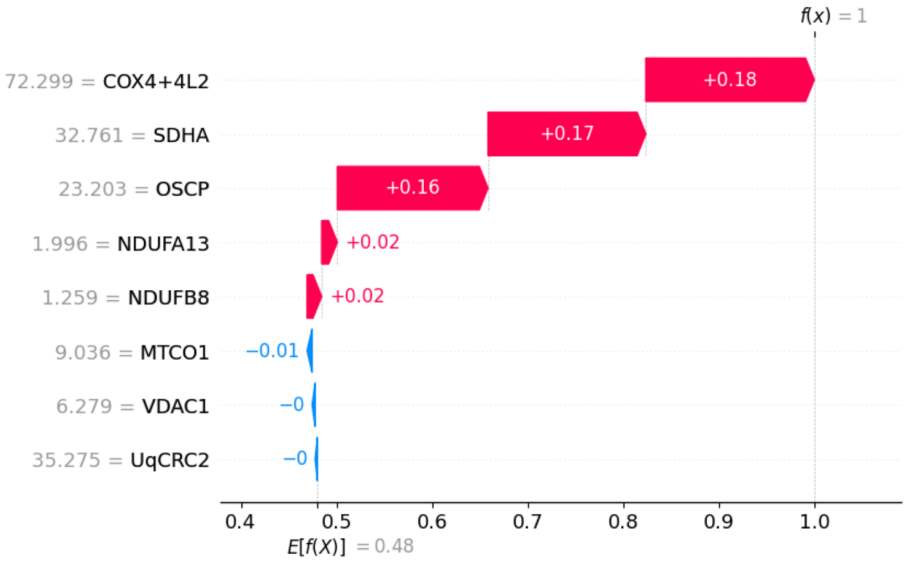

Figure 3.4 **SHAP waterfall** plot are plotted for individual predicted instances. In the toy example above it can be seen that the values of variables COX4+4L2, SDHA and OSCP heavily contribute positively toward predicted probability pushing it toward 1.

3.4 Results

3.4.1 EDA

Dataset counts

The statistical summaries presented in Table 3.2 help decide the features to be selected for ML model training. As seen in the table the relatively high variation (measure in standard deviation) exists for certain protein markers such as SDHA and COX4+4L2. Out of the three statistical summaries i.e. mean of all pixels in myofibre, mean of log of all pixels in myofibre, and median of pixels in myofibre, it was decided that mean should be used as the other two do not give any added information.

Pattern in input features

There were a number of EDAs performed including density plots that use kernel density approximation to show the probability density function of the variable; scatter plots: plot observations as dots on two coordinates represented by two variables; and correlation plots that show relationship (correlation) between variables. The most informative were the correlation plots for datasets and its subsets (i.e. various subject groups) as presented in Figure 3.5. The correlation plot gave some intuitions about highly correlated input features, which is helpful in ML model selection, feature selection and interpreting insights from explainability methods.

The following observations were made in the EDA which are presented in Figure 3.5.

- In controls a high positive correlation exists between all eight protein markers.
- In patients a high positive correlation exists between VDAC1 & SDHA, OSCP; SDHA & OSCP; NDUFB8 & NDUFA13; MTCO1 & COX4+4L2, and COX4+4L2 & UqCRC2.
- In combined (controls+patients) data a very high positive correlation exists between NDUFB8 & NDUFA13 and MTCO1 & COX4+4L2.

And taking these observations into account, the following initial intuitions about feature selections were noted:

- Having two protein markers NDUFB8 & NDUFA13 representing complex I and two protein markers MTCO1 & COX4+4L2 representing complex IV might not add any discriminating advantage for the models.

Table 3.2 Myofibre statistical summaries ranges. The first column define per myofibre features, the first four rows refer to the morphological features, followed by mean, mean(log), median intensity value for each protein. The alternating colours between group of rows is to improve readability.

Input feature	Min	Mean	Max	Std
Area	503	3105.06	17033.5	887.74
Perimeter	107.94	426.05	2089.58	202.28
AspectRatio	0.22	1.31	6.36	0.51
Circularity	0.01	0.24	0.85	0.13
NDUFB8	1.03	2.31	13.05	1.26
LOG_NDUFB8	1.02	1.95	9.60	0.96
MED_NDUFB8	1.0	1.90	10.0	1.13
NDUFA13	1.06	2.74	11.87	1.56
LOG_NDUFA13	1.04	2.26	9.25	1.20
MED_NDUFA13	1.0	2.21	10.0	1.43
SDHA	1.44	11.06	91.48	10.48
LOG_SDHA	1.31	8.97	74.57	8.48
MED_SDHA	1.0	9.62	83.0	9.22
UqCRC2	1.24	7.72	50.74	6.1
LOG_UqCRC2	1.15	6.26	43.50	4.97
MED_UqCRC2	1.0	6.66	45.0	5.41
MTCO1	1.01	2.25	15.12	1.35
LOG_MTCO1	1.01	1.91	13.05	1.05
MED_MTCO1	1.0	1.85	13.0	1.23
COX4+4L2	1.28	11.93	93.78	10.82
LOG_COX4+4L2	1.19	9.62	80.75	9.02
MED_COX4+4L2	1.0	10.40	79.5	9.73
OSCP	1.14	6.70	57.02	6.20
LOG_OSCP	1.10	5.26	41.89	4.75
MED_OSCP	1.0	5.63	46.0	5.29
VDAC1	1.05	2.42	14.8	1.33
LOG_VDAC1	1.03	2.03	11.68	1.00
MED_VDAC1	1.0	1.97	13.50	1.21

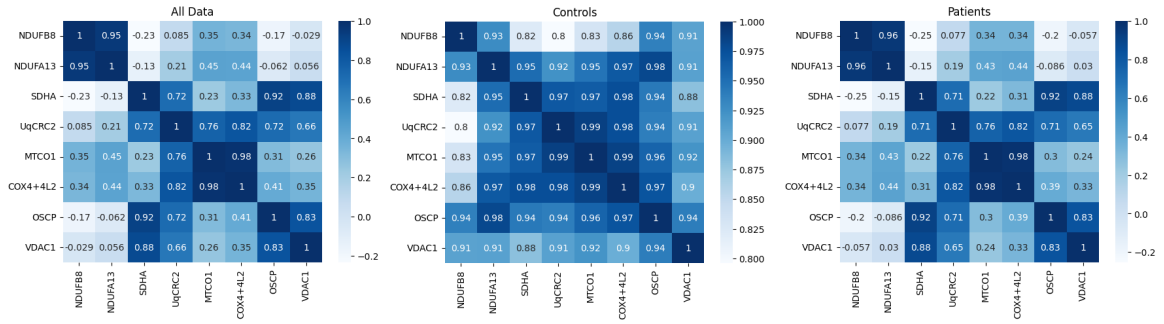

Figure 3.5 Correlation between protein markers in myofibres from all subjects and controls and patients separately. This is Pearson correlation, values range from -1 (perfect negative/inverse correlation) to 1 (perfect positive correlation), the legend to the right of each plot shows white-blue colour gradient scale with the darkest blue representing the highest value and lightest white representing the lowest value. 'All Data' refers to combined myofibres of all controls and patients.

Table 3.3 **Statistical distribution of mean intensities in controls**: These are computed using mean intensity in a myofibre for each protein marker over all control myofibres

Summary	NDUFB8	NDUFA13	SDHA	UqCRC2	MTCO1	COX4+4L2	OSCP	VDAC1
Min	1.19	1.17	1.5	1.40	1.07	1.60	1.23	1.09
Mean	2.94	3.00	5.04	4.38	1.97	7.46	3.42	1.86
Std	1.67	1.75	2.99	3.01	0.88	5.66	2.22	0.68
Max	13.05	10.69	17.73	17.51	5.72	30.53	13.05	6.81

- VDAC1 as a surrogate for myofibre mass is sensible as evident by its high positive correlation with all other markers in controls.

3.4.2 Explainable ML analysis of class A (P01 and P02) vs controls

Classification of myofibres from patients suffering from nDNA encoded mutations (class A) was the only case that the current techniques can accurately classify, as observed in Table 2.4. The classification was achieved based on 95% PI on complex I protein markers NDUFB8 and NDUFA13. In this section we apply explainable ML methods to classify these same myofibres and compare the results and insights achieved by both of these techniques.

EDA

The following observations were inferred from Figure 3.6 and Table 3.4.

- Compare to control myofibres where a high positive correlation exist between all eight protein markers. In class A myofibres a correlation similar to control myofibres is observed in six proteins i.e. excluding NDUFB8 and NDUFA13.

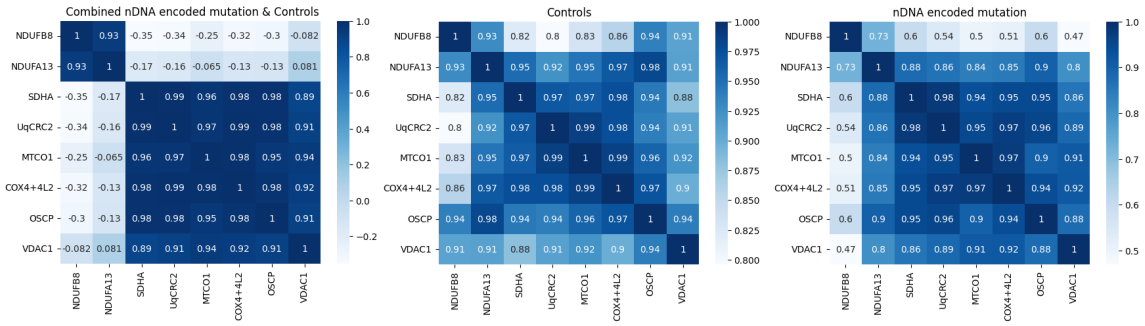

Figure 3.6 Correlation between protein markers in myofibres from class A and controls, controls, class A. In the plot nDNA encoded mutation refers to class A.

Table 3.4 **Statistical distribution of mean intensities in class A patients**: These are computed using mean intensity in a myofibre for each protein marker over all class A myofibres

Summary	NDUFB8	NDUFA13	SDHA	UqCRC2	MTCO1	COX4+4L2	OSCP	VDAC1
Min	1.05	1.16	5.66	4.59	1.95	9.87	3.96	1.45
Mean	1.22	1.7	23.54	21.96	5.68	39.63	15.35	3.45
Std	0.06	0.25	8.27	8.53	2.32	17.1	5.90	1.24
Max	1.47	2.51	53.01	50.74	15.12	93.78	34.6	7.97

- In cognisance of VDAC1 as surrogate for myofibre mass, only its correlation with SDHA, UqCRC2, MTCO1 and COX4+4L2 is comparable to controls. All other protein markers' correlation is reduced in class A myofibres relative to control myofibres.
- SDHA and UqCRC2 correlation is high in class A myofibres compared to control myofibres.

ML classification results

LR (logistic regression), tree-based models (random forest and XGB (XGboost)) were trained with various combinations of input features and the results from optimised best performing models are detailed in Table 3.6. The optimised parameters of both LR and XGB models trained with all eight markers are presented in Table 3.5. It was observed that random forest did not yield results better than XGB and so results from best performing tree-based model (XGB) are reported throughout.

As evident in Table 3.6 both LR and XGB can predict the class A myofibres with 100% accuracy. It can also be observed that on single-protein data LR outperform XGB, this is expected as in class A myofibres deficiency pattern between controls and patients is linear which the LR is leveraging. Aside from the observations presented in the the table, an accuracy of 99% was observed when trained with six out of eight protein markers i.e.

Table 3.5 Optimised models' parameters for class A vs controls myofibres. In the table 'const' refers to bias/intercept, and eight protein marker names refers to LR's coefficients for that variable (protein marker).

Method	Parameter name	Optimised parameter values
LR	const	-3.17
	NDUFB8	-1.92
	NDUFA13	-1.60
	SDHA	0.50
	OSCP	0.47
	COX4+4L2	0.37
	MTCO1	-0.34
	VDAC1	-0.34
	UqCRC2	0.18
XGB	colsample_bytree	0.7
	eta	0.1
	gamma	0.3
	max_depth	3
	min_child_weight	1

excluding NDUFB8 and NDUFA13. As discussed in the Section 3.3.3 in cases where more than one ML model produces similar accuracy, it was decided that both models would be interrogated to distill predictive inference, i.e. by applying EMs to these models.

Applying explainable ML methods to LR and XGB models for class A myofibres prediction

The results presented in Figures 3.7, 3.8, 3.9 and 3.10 are SHAP EMs applied to LR and XGB models. The global explanations generated using whole training data with SHAP EMs that are presented in Figures 3.7 and 3.9 for LR and XGB models respective shows i) both models uses slightly different patterns in the data to achieve the same 100% accuracy i.e. SHAP EM shows the top three average SHAP values for protein markers for the LR model are COX4+4L2, SDHA & OSCP, and for the XGB model are NDUFB8, SDHA & COX4+4L2. ii) The SHAP value prevalence observed in Figure 3.7**C** and 3.9**C** shows NDUFA13 & NDUFB8 proteins' high mean intensities are associated with extreme negative SHAP values in both the models, these are very consequential towards predicting instances of myofibres as negative (control) class. In addition NDUFB8's low mean intensities are associated with extreme positive SHAP values in the XGB model, these values are very consequential towards predicting instances of myofibres as a positive (class A) class. It can

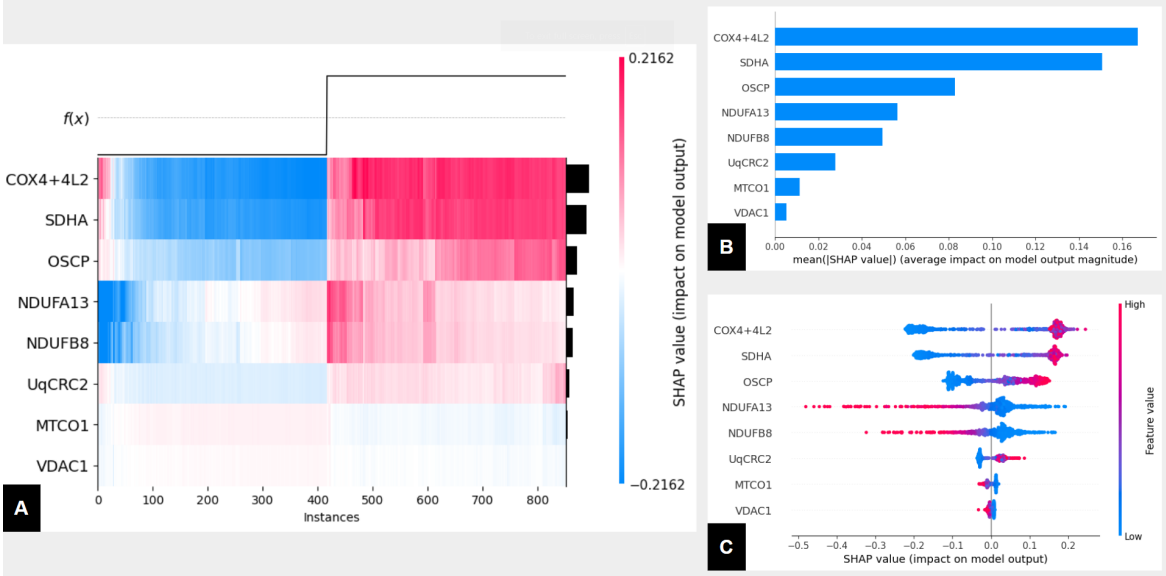

Figure 3.7 SHAP global explanations of LR model that predicts class A myofibres. The LR model interrogated here had 100% test accuracy and trained on all eight protein markers. **A**, SHAP heatmap plot, on x-axis is instances of reference dataset (which in this case is the whole training dataset of 852 instance (427 class A myofibres + 425 control myofibres)) on y-axis are SHAP values for each input feature. On the top of the plot f(x) is predicted value of the model which for LR is a value of 1 for patient myofibres and 0 for control myofibres. The SHAP values on the y-axis are encoded into colour scale. The instances are ordered using hierarchical clustering by their explanation similarity. **B**, bar chart of the average SHAP value magnitude of each input feature. **C** a set of SHAP beeswarm plots, where each dot corresponds to an individual myofibre in the analysis. The dot's position on the x axis shows the impact that feature has on the model's prediction for that instance. When multiple dots land at the same x position, they pile up to show density i.e. prevalence.

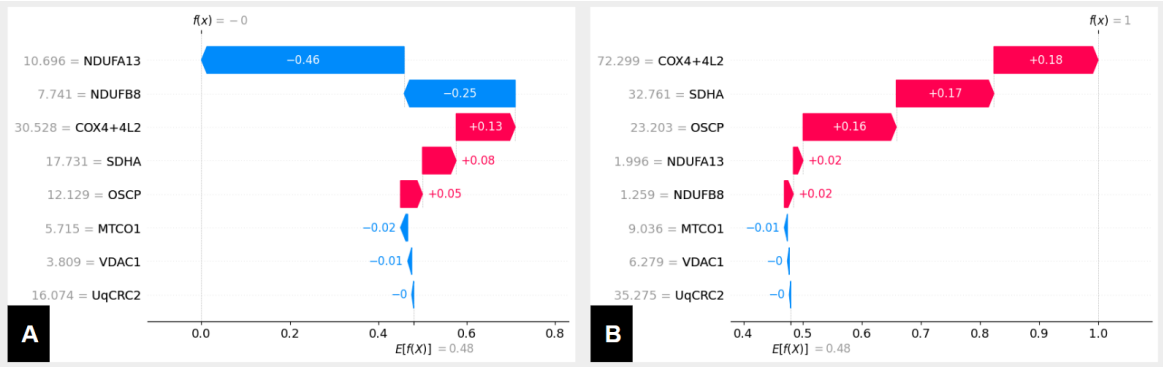

Figure 3.8 SHAP local explanations of LR model that predicts class A myofibres. **A**, SHAP waterfall plot providing local explanation for correctly predicted control myofibre instance. The bottom is E[F(x)] the expected values of the model output, then each row shows how the positive (red) or negative (blue) contribution of each feature moves the value from the expected model output over the reference (training) dataset to the model output for this prediction. **B** Another SHAP waterfall plot but for correctly predicted class A myofibre.

Table 3.6 Accuracy metrics for models trained to predict class A myofibres. Metrics in columns 2 to 5 are for models trained on all eight protein markers.

Model	Training Accuracy(%)	Test Accuracy(%)	Recall(%) (Patients)	Recall(%) (Controls)	Accuracy (single protein)(%)	Accuracy (VDAC1+Protein)(%)
LR	100	100	100	100	92.28 (NDUFB8) 92.28 (SDHA) 92.28 (UqCRC2) 90.52 (COX4+4L2) 89.82 (OSCP) 88.07 (MTCO1) 79.65 (NDUFA13) 79.65 (VDAC1)	100 (NDUFB8) 98.95 (NDUFA13) 96.14 (SDHA) 94.03 (UqCRC2) 94.03 (COX4+4L2) 93.68 (OSCP) 90.88 (MTCO1) NA (VDAC1)
XGB	100	100	100	100	78.05(OSCP) 76.60(VDAC1)) 75.15 (NDUFB8) 74.74 (COX4+4L2) 74.53 (MTCO1) 74.32 (UqCRC2) 74.12 (SDHA) 72.88 (NDUFA13)	99.65(NDUFB8) 99.65(NDUFA13) 97.19(UqCRC2) 96.84 (COX4+4L2) 96.50 (OSCP) 96.14 (SDHA) 91.93 (MTCO1) NA (VDAC1)

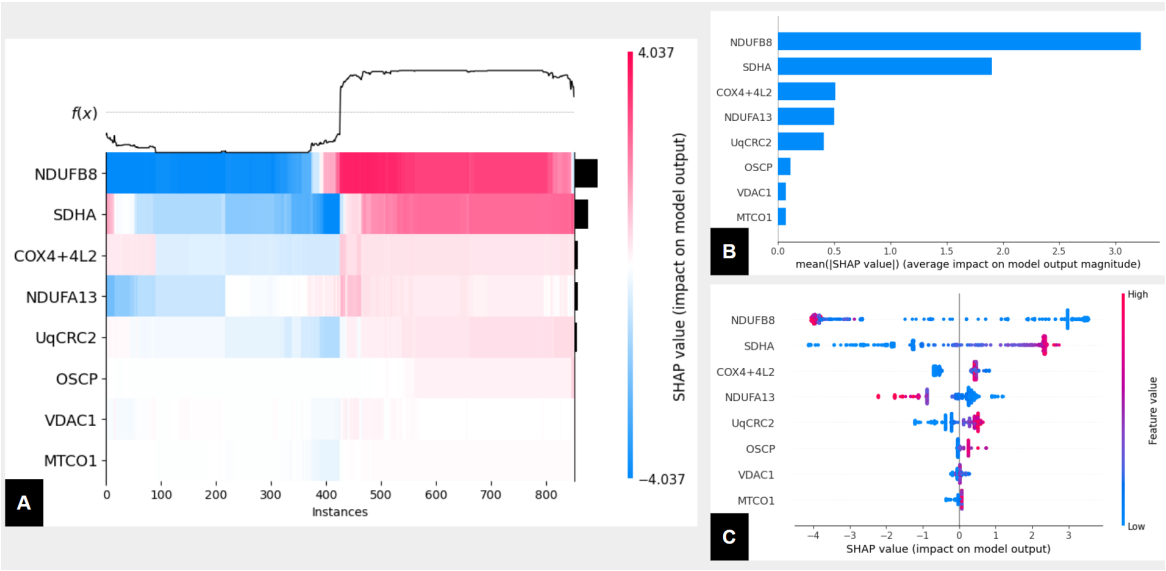

Figure 3.9 SHAP global explanations of XGB model that predicts class A myofibres. The XGB model interrogated here had 100% test accuracy and trained on all eight protein markers. **A**, SHAP heatmap plot. **B**, bar chart of the average SHAP value. **C**, a set of SHAP beeswarm plots.

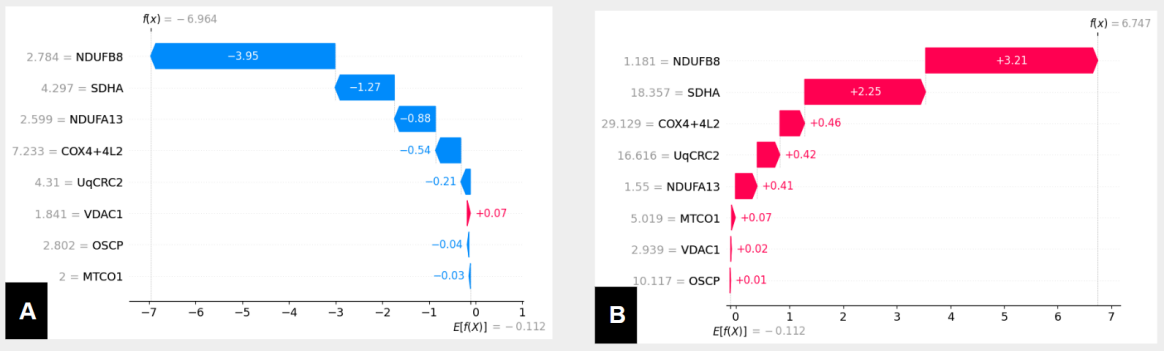

Figure 3.10 SHAP local explanations of XGB model that predicts class A myofibres. **A**, SHAP waterfall plot providing local explanation for correctly predicted control myofibre instance. **B**, Another SHAP waterfall plot but for correctly predicted class A myofibre.

also be observed that COX4+4L2 and SDHA have opposite associations, i.e. higher mean intensities of these proteins are associated with positive SHAP values that are consequential towards predicting instances of myofibres as class A. These associations are further evident in the SHAP heatmap plot in Figure 3.7**A** & 3.9**A**, where the inverse relationships between NDUFA13, NDUFB8 and COX4+4L2, SDHA is leveraged by LR model, but XGB model leverages similar inverse relationship but with NDUFA13 and SDHA.
Similar associations are also observed by studying individual instances of myofibres i.e. local explanations that are presented in Figures 3.8 & 3.10 for LR and XGB models respectively. A more detailed per protein interpretation of SHAP values combined with their predictive power and their average mean intensities across the both classes is available in Appendix .1, Tables 1 and 2.

Insights and predictive inference

The following insights are drawn from applying explainable ML methods to the dataset of class A and control myofibres.

Predictive inference from LR model insights Compared to control myofibres, in class A myofibres the levels of markers for complex IV (COX4+4L2), complex II (SDHA), complex V (OSCP) are higher and complex I markers are considerably lower, enough to predict these with 100% accuracy.

Predictive inference from XGB model insights Compared to control myofibres class A myofibres have considerably lower levels of markers for complex I (NDUFB8) and higher

levels of complex II (SDHA), these differences are enough to predict these with 100% accuracy.

Combined predictive inference

As discussed earlier ML models have a tendency to ignore associations between all input features, if some of the correlations between input features' and predictive variables is sufficient to make good predictions, it ignores other associations. This is observed in how the two models (LR and XGB) we used above behave. To overcome this shortcoming we conducted predictive power of these features with VDAC1 as a surrogate for myofibre mass. We observed when combined with VDAC1, all features have a predictive power of greater than 95%. So, taking into account the results from EDA, explainable LR model, explainable XGB model and predictive powers of all input features, the following predictive inference is generated.

In class A myofibres when compared to control myofibres, the following pattern is observed. A very considerably low level of complex I (NDUFB8 & NDUFA13) proteins, a considerably high level of complex II (SDHA), complex III (UqCRC2), complex IV (MTCO1 & COX4+4L2) and complex V (OSCP) proteins.

Biological validation

Class A myofibres, i.e. linked to nDNA encoded genetic mutations, are well understood compared to other classes of myofibres and as such present an appropriate validation case for testing the explainable ML predictive inference. In class A myofibres it is expected that the nDNA encoded complex I proteins will be down-regulated i.e. exhibit low-levels, due to this the OXPHOS electron transfer process has to rely more on the rest of the complexes resulting in upregulation of proteins in complexes II-V [1, 46, 145]. These are the exact associations presented in Section 3.4.2, the predictive inference derived from applying explainable ML methods reveal the ML models are leveraging associations that make biological sense and agree with established facts. Interestingly the methods finds many available associations as opposed to expectation that models will ignore associations that are surplus to associations sufficient to make good predictions. A slightly different associations exploited by the two models is still in line with biological basis i.e. albeit different proteins were used for differentiating the class A nevertheless these belong to same complexes.

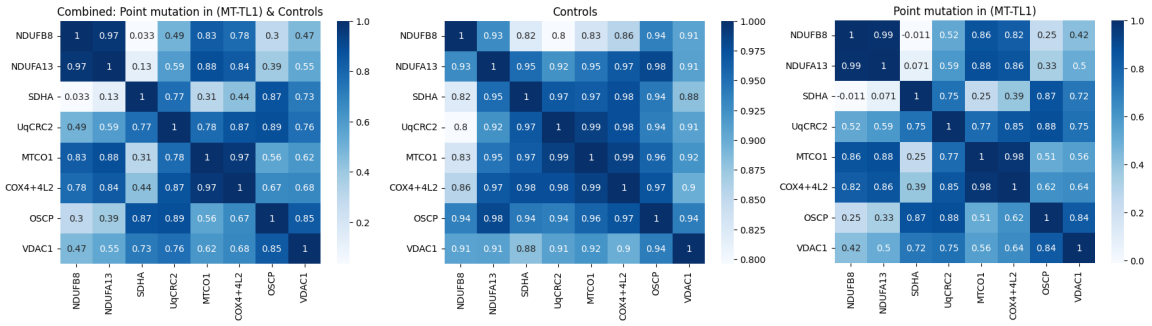

Figure 3.11 Correlation between protein markers in class B myofibres

Table 3.7 **Statistical distribution of mean intensities in class B myofibres** : These are computed using mean intensity in a myofibre for each protein marker over all class B myofibres

Summary	NDUFB8	NDUFA13	SDHA	UqCRC2	MTCO1	COX4+4L2	OSCP	VDAC1
Min	1.03	1.07	1.93	1.52	1.02	1.88	1.32	1.13
Mean	2.50	2.93	8.64	7.83	2.20	11.66	5.88	2.15
Std	1.41	1.75	6.41	4.53	0.83	7.25	3.23	0.75
Max	9.74	11.88	58.35	42.80	6.43	53.92	47.85	10.60

3.4.3 Explainable ML Analysis of class B (P05, P06, P07) vs controls

Classification of myofibres from patients carrying a point mutation in MT- TL)((class B) is not possible with current techniques as observed in Table 2.4. In this section we apply explainable ML methods to classify these same myofibres and construct predictive insights.

EDA

The following observations were inferred from Figure 3.11 and Table 3.7.

- Compared to control myofibres where a high positive correlation exists between all eight protein markers, in class B myofibres no such correlation exists except between NDUFB8 and NDUFA13; MTCO1 and COX4.
- In cognisance of VDAC1 as surrogate for myofibre mass, only its correlation with SDHA, UqCRC2, MTCO1 and COX4+4L2 is comparable to controls. All protein markers' correlation is reduced in class B myofibres compared to controls.

ML classification results (class B)

LR and XGB models for class B vs control myofibre classification were trained with various combinations of input features and the optimised parameters of both models trained with all

Table 3.8 Optimised models' parameters for class B vs controls myofibres.

Method	Parameter	Optimised parameter values
LR	const	10.42
	NDUFB8	-3.57
	NDUFA13	-3.51
	SDHA	-0.18
	OSCP	-0.17
	COX4+4L2	1.34
	MTCO1	-8.54
	VDAC1	-0.02
	UqCRC2	3.35
XGB	colsample_bytree	0.5
	eta	0.3
	gamma	0.0
	max_depth	5
	min_child_weight	1

eight markers are presented in Table 3.8. The results from optimised best performing models are detailed in Table 3.9.

As evident in the table both LR and XGB can predict the class B myofibres with 99% accuracy. As discussed in Section 3.3.3 in cases where more than one ML model produces similar accuracy, it was decided that both models will be interrogated to distill predictive inference i.e. by applying EMs to these models.

Applying explainable ML methods to LR and XGB models for class B myofibres prediction

The global explanations generated using whole training data with SHAP EMs are presented in Figures 3.12 and 3.14 for LR and XGB models respectively. In these it can be observed that i) both models use slightly different patterns in the data to achieve the same 99% accuracy i.e. SHAP EM shows the top three average SHAP values for protein markers for LR models are UqCRC2, COX4+4L2 and MTCO1, and for XGB model are UqCRC2, NDUFB8 & COX4+4L2. ii) The SHAP value prevalence observed in plot Figures 3.12**C** and 3.14**C** shows low mean intensities of UqCRC2 protein are associated with extreme negative SHAP values in both of the models, these are very consequential towards predicting instances of myofibres as negative (control) class. In addition high mean intitensities for UqCRC2 are associated with positive SHAP values in the models, these values are consequential towards predicting instances of myofibres as positive (class B) class. It can also be observed that the

Table 3.9 Accuracy metrics for models trained to predict class B myofibres

Model	Trainin Accu-racy(%	Test Accu-racy(%	Recall(%) (Patients)	Recall(%) (Controls)	Accuracy (single pro-tein)(%)	Accuracy (VDAC1+Protein)(%)
LR	99	99	100	95	61.81(NDUFB8) 57.03 (NDUFA13) 53.04 (SDHA) 63.41 (UqCRC2) 51.15 (MTCO1) 58.62 (COX4+4L2) 52.34 (OSCP) 49.85 (VDAC1)	64.30 (NDUFB8) 53.53 (NDUFA13) 55.53 (SDHA) 69.90 (UqCRC2) 50.05 (MTCO1) 59.22 (COX4+4L2) 55.83 (OSCP) NA (VDAC1)
XGB	100	99	99	93	86.34 (NDUFB8) 86.34 (NDUFA13) 85.64 (SDHA) 88.03 (UqCRC2) 86.34 (MTCO1) 86.14 (COX4+4L2) 87.44 (OSCP) 86.54 (VDAC1)	87.24 (NDUFB8) 87.14 (NDUFA13) 86.34 (SDHA) 87.64 (UqCRC2) 85.84 (MTCO1) 86.84 (COX4+4L2) 87.54 (OSCP) NA (VDAC1)

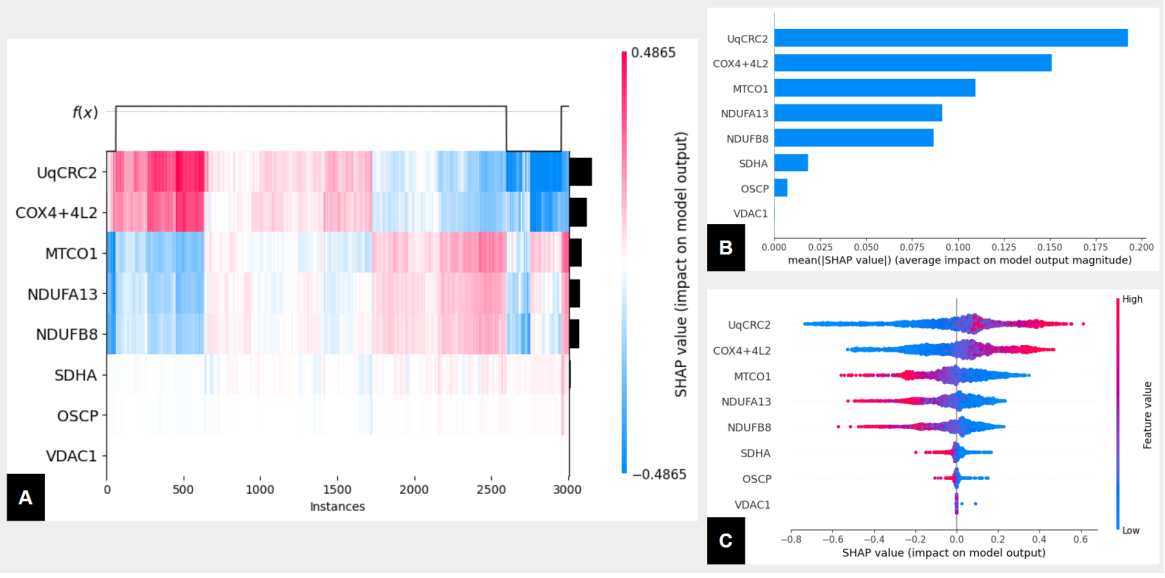

Figure 3.12 SHAP global explanations of LR model that predicts class B myofibres. The LR model interrogated here had 99% test accuracy and trained on all eight protein markers. **A**, SHAP heatmap plot. **B**, bar chart of the average SHAP value. **C**, a set of SHAP beeswarm plots

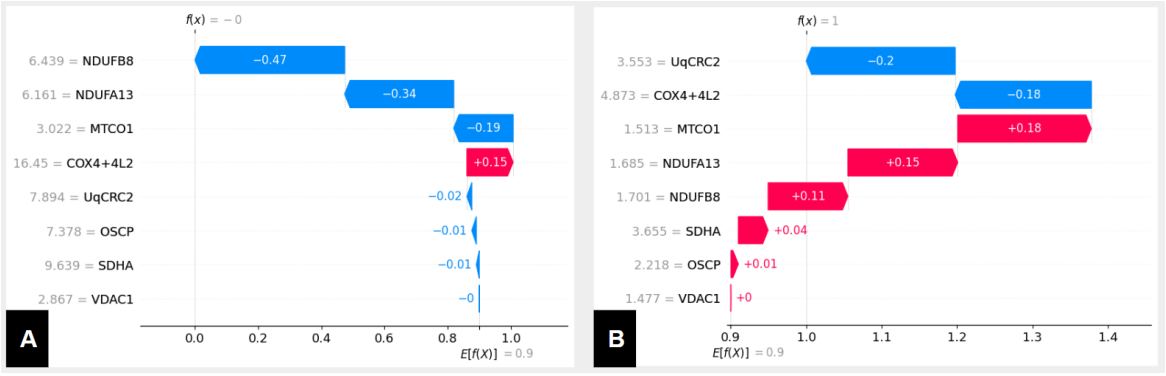

Figure 3.13 SHAP local explanations of LR model that predicts class B myofibres. **A**, SHAP waterfall plot providing local explanation for correctly predicted control myofibre instance. **B**, Another SHAP waterfall plot but for correctly predicted class B myofibre.

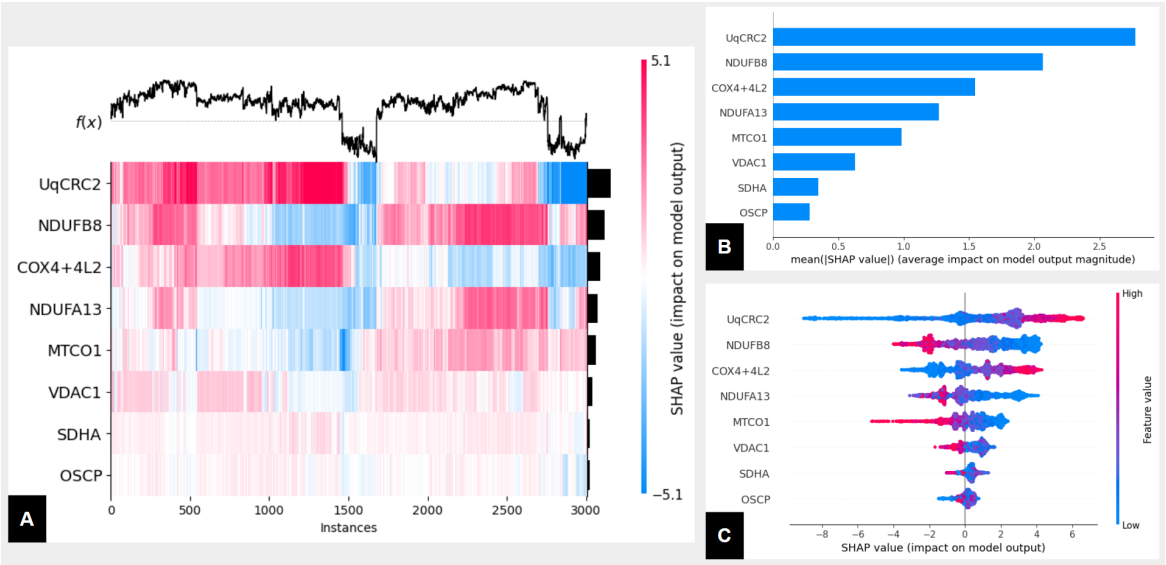

Figure 3.14 SHAP global explanations of XGB model that predicts class B myofibres. The XGB model interrogated here had 99% test accuracy and trained on all eight protein markers. **A**, SHAP heatmap plot. **B**, bar chart of the average SHAP value. **C**, a set of SHAP beeswarm plots

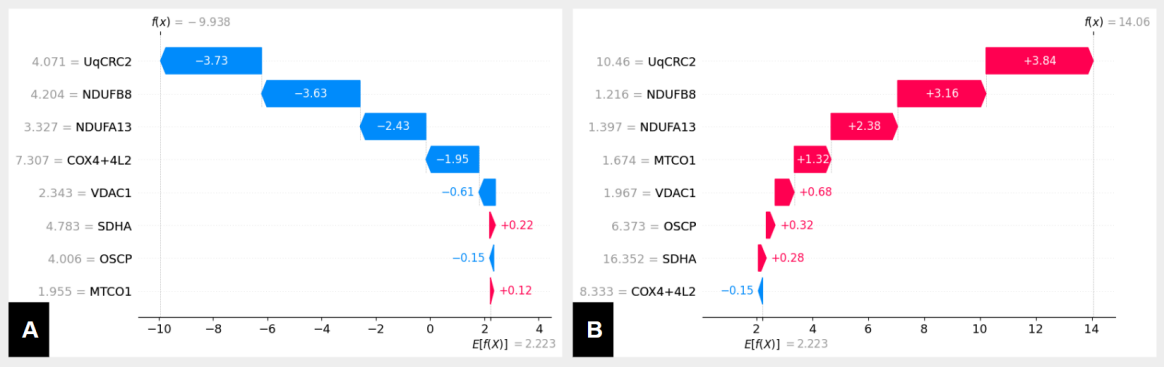

Figure 3.15 SHAP local explanations of LR model that predicts class B myofibres. **A**, SHAP waterfall plot providing local explanation for correctly predicted control myofibre instance. **B**, Another SHAP waterfall plot but for correctly predicted class B myofibre

two complex IV protein markers (COX4+4L2 and MTCO1) have inverse associations with each other, i.e. higher mean intensities of COX4+4L2 protein are associated with positive SHAP values and lower mean intensities of MTCO1 protein are associated with positive SHAP values that are consequential towards predicting instances of myofibres as positive (class B) class. These associations are further evident in the SHAP heatmap plot **A** within Figures 3.12 & 3.14, where the inverse relationship between UqCRC2 & COX4+4L2 and MTCO1, and complex I proteins (NDUFB8, NUFA13) is leveraged by the LR model, but the XGB model leverages different relationships between UqCRC2, COX4+4L2 and the rest of the proteins.

These associations are also observed by studying individual instances of myofibres that are presented in Figures 3.13 & 3.15 for LR and XGB models respectively. A more detailed per protein interpretation of SHAP values combined with their predictive power and their average mean intensities across the both classes is available in Appendix .1, Tables 3 and 4.

Insights and predictive inference

The following insights are drawn from applying explainable ML methods to dataset of class B and control myofibres.

Predictive inference from LR model insights (class B) Compared to control myofibres, in class B myofibres the levels of markers for complex III (UqCRC2) and complex IV (COX4+4L2) are higher and complex I (NDUFB8 & NDUFA13) and complex IV (MTCO1) are lower, and a strong inverse association exists between the two complex IV markers (COX4+4L2 & MTCO1), enough to predict with 99% accuracy (100% for patients & 95% for control).

Predictive inference from XGB model insights (class B) The class B myofibres differ from control myofibres in the following pattern. The levels of markers for complex III (UqCRC2) and complex IV (COX4+4L2) are higher and complex I (NDUFB8 & NDUFA13) and complex IV (MTCO1) are lower, and a strong inverse association exists between the two complex IV markers (COX4+4L2 & MTCO1), enough to predict with 99% accuracy (99% for patients & 93% for control).

Combined predictive inference (class B)

Taking into account the results from EDA, explainable LR model, explainable XGB model and predictive powers of all input features, the following predictive inference is generated.

In class B myofibres when compared to control myofibres, the following pattern is observed: higher levels of complex III (UqCRC2) and complex IV (COX4+4L2), lower levels of complex I (NDUFB8 & NDUFA13) and complex IV (MTCO1). And a strong inverse association exists between the two complex IV markers (COX4+4L2 & MTCO1). The markers of the other three complexes did not have any additive benefit for prediction accuracy i.e. leaving them out of model training did not affect accuracy.

Biological validation

The biological validation of predictive inferences discovered using explainable ML analysis for class B myofibres are not straightforward as this class of mutations is not completely understood. Nevertheless, in m.3243 point mutation which belongs to class B mutations, patients have mostly complex I deficiency (NDUFB8 and NDUFA13)[34, 36, 67, 145] which the ML models are correctly choosing . Association of high levels of UqCRC2 to class B discovered is surprising and there may well be a good biological explanation that needs further investigation by experimental validation to conclude whether it was an artefactual or novel discovery.

3.4.4 Explainable ML Analysis of class C (P08, P09, P10) vs controls

Classification of myofibres from patients carrying single point mutations in mitochondrial encoded tRNA (class C) is not possible with current techniques as observed in Table 2.4. In this section we apply explainable ML methods to classify these same myofibres and present insights achieved using explainable ML methods.

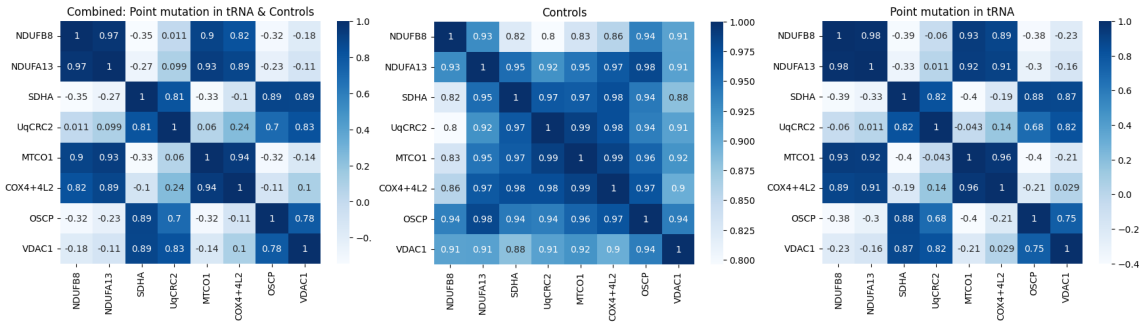

Figure 3.16 Correlation between protein markers in class C myofibres

Table 3.10 **Statistical distribution of mean intensities in class C myofibres**: These are computed using mean intensity in a myofibre for each protein marker over all class C myofibres

Summary	NDUFB8	NDUFA13	SDHA	UqCRC2	MTCO1	COX4+4L2	OSCP	VDAC1
Min	1.04	1.10	2.02	1.47	1.02	1.29	1.46	1.17
Mean	2.03	2.35	18.30	7.13	1.67	8.42	10.91	3.38
Std	1.31	1.46	14.82	5.36	0.83	6.47	8.61	2.00
Max	8.56	9.43	91.48	41.70	4.95	38.34	57.03	14.83

EDA

The following observations were inferred from Figure 3.16 and Table 3.10.

- Compared to control myofibres where a high positive correlation exists between all eight protein markers, in class C myofibres no such correlation exists except between NDUFB8 and NDUFA13; MTCO1 and COX4 and to a lesser extent UqCRC2 and SDHA.
- In cognisance of VDAC1 as surrogate for myofibre mass, in class C myofibres the correlation with SDHA, UqCRC2, MTCO1 and COX4+4L2 is comparable to controls. The correlations for all other protein markers are reduced in class C compared to controls.

ML classification results (class C)

LR and XGB models were trained with various combinations of input features and the results from optimised best performing models are detailed in Table 3.12. The optimised parameters of both LR and XGB models trained with all eight markers are presented in Table 3.11.

As evident in Table 3.12 LR and XGB can predict the class C myofibres with 98% and 99% accuracy respectively. Both models are used for predictive inference i.e. to apply explainable ML methods.

Table 3.11 Optimised models' parameters for class C vs controls myofibres.

Method	Parameter	Optimised parameter values
LR	const	5.27
	NDUFB8	-1.77
	NDUFA13	-2.62
	SDHA	1.06
	OSCP	-0.29
	COX4+4L2	2.02
	MTCO1	-7.90
	VDAC1	2.832
	UqCRC2	-0.90
XGB	colsample_bytree	0.7
	eta	0.3
	gamma	0.0
	max_depth	15
	min_child_weight	1

Table 3.12 Accuracy metrics for models trained to predict class C myofibres

Model	Training Accuracy(%)	Test Accuracy(%)	Recall(%) (Patients)	Recall(%) (Controls)	Accuracy (single protein)(%)	Accuracy (VDAC1+Protein)(%)
LR	97	98	98	95	67.93 (NDUFB8) 65.74 (NDUFA13) 71.86 (SDHA) 56.71 (UqCRC2) 63.55 (MTCO1) 44.31 (COX4+4L2) 67.05 (OSCP) 66.47 (VDAC1)	77.70 (NDUFB8) 74.19 (NDUFA13) 71.28 (SDHA) 69.82 (UqCRC2) 71.72 (MTCO1) 65.60 (COX4+4L2) 67.5 (OSCP) NA (VDAC1)
XGB	100	99	99	98	79.73 (NDUFB8) 80.03 (NDUFA13) 86.30 (SDHA) 83.09 (UqCRC2) 80.17 (MTCO1) 80.03 (COX4+4L2) 83.97 (OSCP) 84.98 (VDAC1)	85.13 (NDUFB8) 84.98 (NDUFA13) 85.13 (SDHA) 86.00 (UqCRC2) 85.57 (MTCO1) 84.69 (COX4+4L2) 86.15 (OSCP) NA (VDAC1)

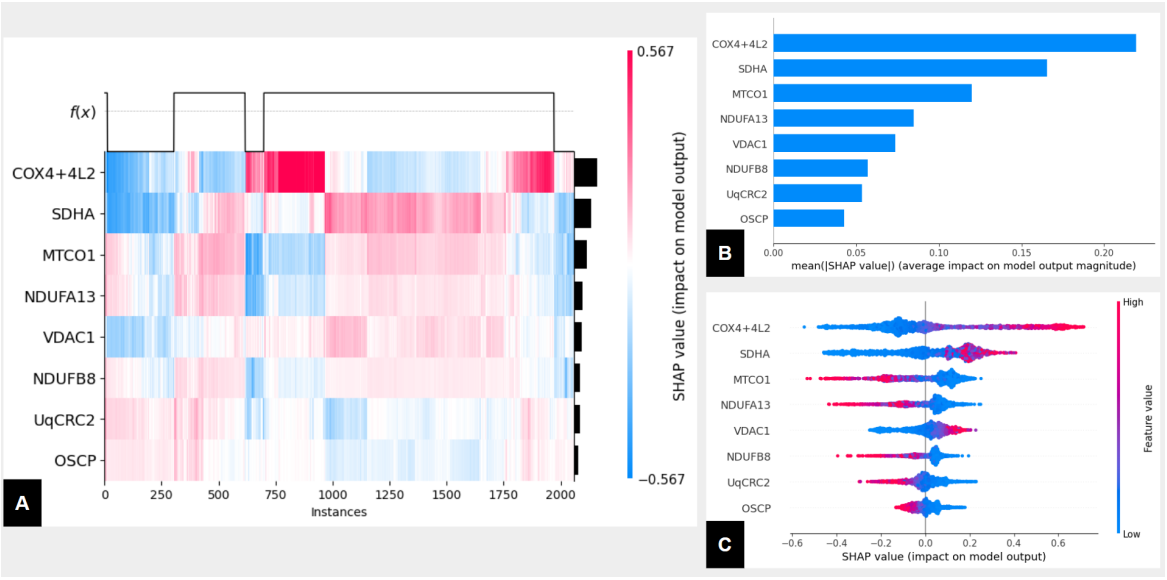

Figure 3.17 SHAP global explanations of LR model that predicts class C myofibres. The LR model interrogated here had 98% test accuracy and trained on all eight protein markers. **A**, SHAP heatmap plot. **B**, bar chart of the average SHAP value. **C**, a set of SHAP beeswarm plots

Applying explainable ML methods to LR and XGB models for class C myofibres prediction

The global explanations generated using whole training data with SHAP EMs that are presented in Figures 3.17 and 3.19 for LR and XGB models respectively show i) both models use slightly different patterns in the data to achieve 98% and 99% accuracy respectively. SHAP EM shows the top three average SHAP values for protein markers for LR model are COX4+4L2, SDHA & MTCO1, and for XGB model are NDUFB8, SDHA & COX4+4L2. ii) The SHAP value prevalence observed in plot **C** within Figures 3.17 and 3.19 shows COX4+4L2 protein high mean intensities are associated with extreme positive and vice versa SHAP values in both the models are very consequential towards predicting instances of myofibres as both classes. In addition SDHA also has a similar association to COX4+4L2 with SHAP values but results in relatively moderate SHAP values compared to COX4+4L2 . It can also be observed that the two complex IV protein markers (COX4+4L2 and MTCO1) have inverse association with each other, i.e. higher mean intensities of COX4+4L2 protein are associated with positive SHAP values and lower mean intensities of MTCO1 protein are associated with positive SHAP values that are consequential towards predicting instances of myofibres as positive (class C) class. These associations are further evident in the SHAP heatmap plot **A** within Figures 3.17 and 3.19, where the inverse relationship between COX4+4L2 and MTCO1, and complex I proteins (NDUFB8, NUFA13) is leveraged by LR

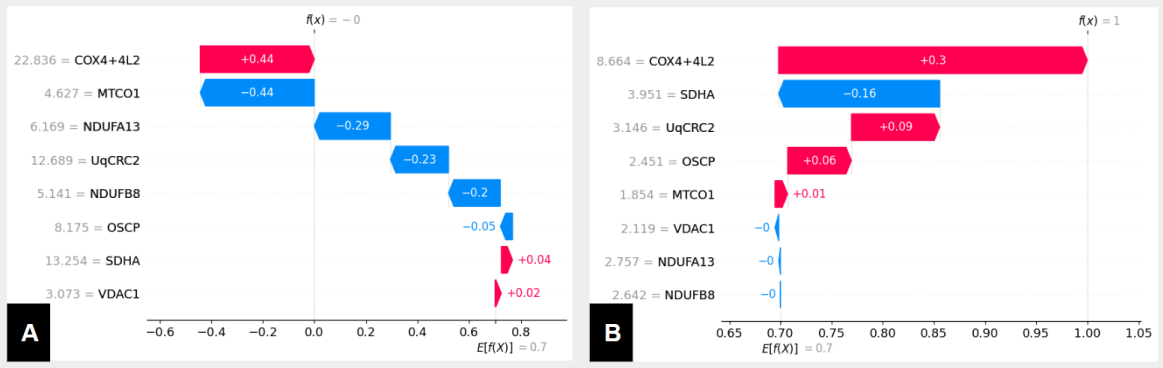

Figure 3.18 SHAP local explanations of LR model that predicts class C myofibres. **A**, SHAP waterfall plot providing local explanation for correctly predicted control myofibre instance. **B**, Another SHAP waterfall plot but for correctly predicted class B myofibre.

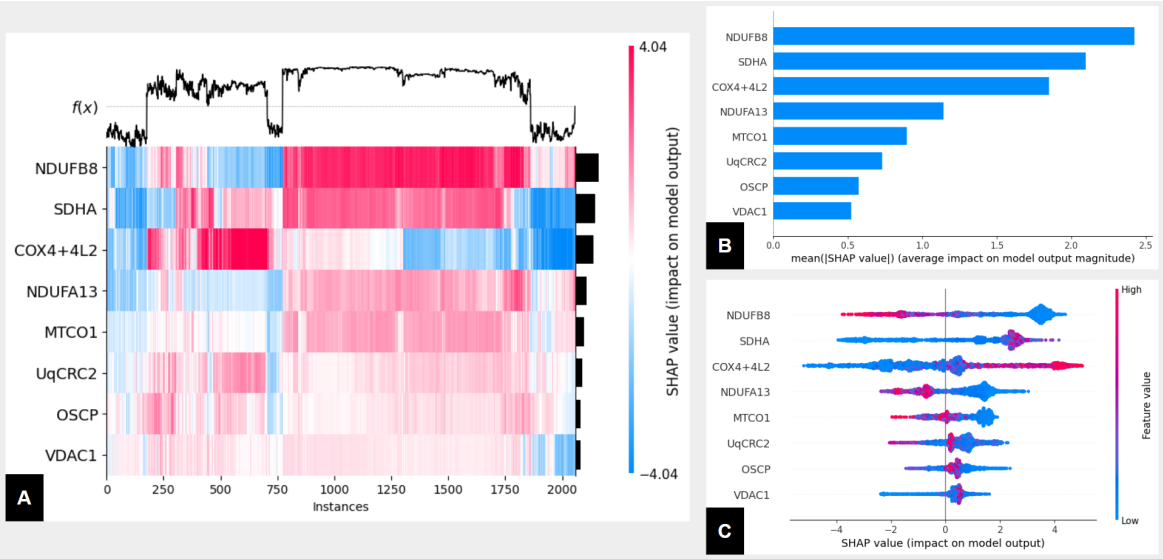

Figure 3.19 SHAP global explanations of XGB model that predicts class C myofibres. The XGB model interrogated here had 99% test accuracy and trained on all eight protein markers. **A**, SHAP heatmap plot. **B**, bar chart of the average SHAP value. **C** a set of SHAP beeswarm plots

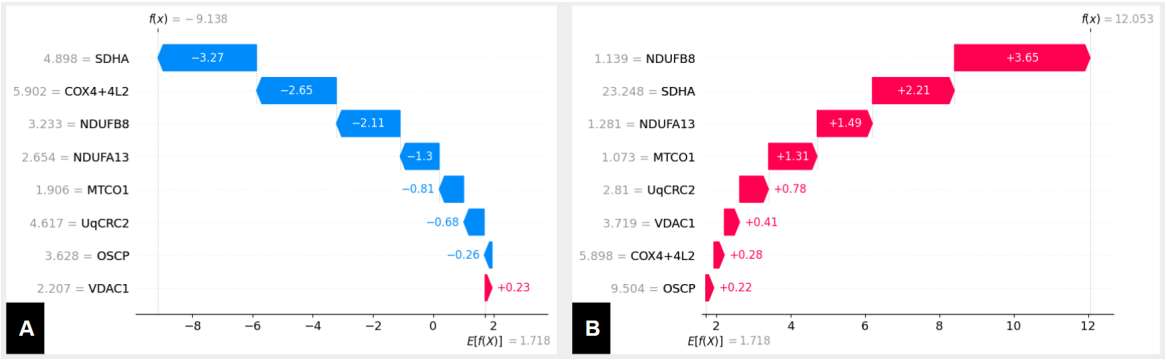

Figure 3.20 SHAP local explanations of LR model that predicts class C myofibres. **A**, SHAP waterfall plot providing local explanation for correctly predicted control myofibre instance. **B**, Another SHAP waterfall plot but for correctly predicted class C myofibre.

model, but XGB model leverages different relationships between NDUFB8 and COX4+4L2 & SDHA.

The associations observed by studying individual instances of myofibres i.e. local explanations that are presented in the Figures 3.18 & 3.20 for LR and XGB models respectively, give similar insights to 'global' explanation. A more detailed per protein interpretation of SHAP values combined with their predictive power and their average mean intensities across the both classes is available in Appendix .1, Tables 5 and 6.

Insights and predictive inference

The following insights are drawn from applying explainable ML methods to the dataset of class C and control myofibres.

Predictive inference from LR model insights (class C) The class C myofibres differ from control myofibres in the following pattern. Compared to control myofibres, in class C myofibres the levels of markers for complex IV (COX4+4L2), complex II (SDHA) are higher and complex I (NDUFB8 & NDUFA13), complex IV (MTCO1) and to a lesser extent complex III (UqCRC2), and Complex V(OSCP) are lower. An inverse association exists between the two complex IV markers (COX4+4L2 & MTCO1), enough to predict with 98% accuracy (98% for patients and 95% for control).

Predictive inference from XGB model insights (class C) The class C myofibres differ from control myofibres in the following pattern. Compared to control myofibres in class C myofibres the levels of markers for complex IV (COX4+4L2), complex II (SDHA) are higher and complex I (NDUFB8 & NDUFA13), complex IV (MTCO1) and to a lesser extent

complex III (UqCRC2) and Complex V(OSCP) are lower. An inverse association exists between the two complex IV markers (COX4+4L2 & MTCO1), enough to predict with 99% accuracy (99% for patients and 98% for control).

Combined predictive inference (class C)

Taking into account the results from EDA, explainable LR model, explainable XGB model and predictive powers of all input features, the following predictive inference is generated.

As the predictive inference from both models agrees the combine predictive inference is the same as above. It was also observed that the removal of any marker reduces the accuracy.

Biological validation

Class C mutations are not completely understood but some predictive inferences are expected such as 'deficiencies of complex I (NDUFB8 and NDUFA13), complex IV (MTCO1) and to a lesser extent complex III (UqCRC2), and Complex V(OSCP)', it is expected to see some deficiency in complex III and V but not as much as for complex I and complex IV (MTCO1). The other predictive inferences such as correlation between SDHA and UqCRC2 and predictive power of SDHA is interesting. SDHA is expected not to change or to increase and UqCRC2 to be the same if not deficient. These are interesting discoveries that warrant further investigations.

3.4.5 Explainable ML Analysis of class D (P03 and P04) vs controls

Classification of myofibres from patients suffering from single, large-scale mtDNA mutation (class D) is not possible with current techniques as observed in Table 2.4. In this section we apply explainable ML methods to classify these same myofibres and present insights achieved by explainable ML methods.

EDA

In control myofibres where a high positive correlation exists between all eight protein markers, this is also observed in class D myofibres (except with NDUFB8) but to a lesser extent. Similarly, VDAC1 have a similar correlation in class D myofibres as in control myofibres with all protein markers except NDUFB8. This is presented in Figure 3.21 and Table 3.13.

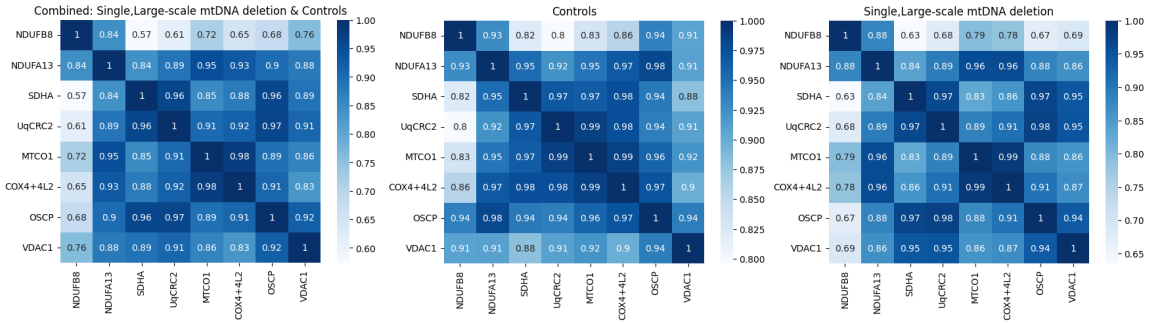

Figure 3.21 Correlation between protein markers in class D myofibres

Table 3.13 **Statistical distribution of mean intensities in class D myofibres** : These are computed using mean intensity in a myofibre for each protein marker over all class D myofibres

Summary	NDUFB8	NDUFA13	SDHA	UqCRC2	MTCO1	COX4+4L2	OSCP	VDAC1
Min	1.06	1.06	1.44	1.24	1.02	1.31	1.15	1.05
Mean	2.43	3.04	6.14	5.36	2.10	9.88	3.75	1.82
Std	0.68	1.29	3.54	2.97	0.83	7.14	2.17	0.46
Max	5.92	10.32	27.46	20.71	5.75	38.36	16.82	4.75

Table 3.14 Optimised model parameters for class D vs controls myofibres.

Method	Parameter	Optimised parameter value
LR	const	8.52
	NDUFB8	-0.56
	NDUFA13	-0.37
	SDHA	0.17
	OSCP	-1.69
	COX4+4L2	1.05
	MTCO1	-6.71
	VDAC1	-0.52
	UqCRC2	1.23
XGB	colsample_bytree	0.7
	eta	0.25
	gamma	0.0
	max_depth	5
	min_child_weight	1

Table 3.15 Accuracy metrics for models trained to predict class D myofibres

Model	Trainin Accu-racy(%	Test Accu-racy(%	Recall(%) (Patients)	Recall(%) (Controls)	Accuracy (single pro-tein)(%)	Accuracy (VDAC1+Protein)(%)
LR	86	85	97	35	57.41 (NDUFB8) 47.30 (NDUFA13) 48.72 (SDHA) 48.86 (UqCRC2) 46.01 (MTCO1) 46.44 (COX4+4L2) 45.72 (OSCP) 53.56 (VDAC1)	69.96 (NDUFB8) 60.11 (NDUFA13) 60.40 (SDHA) 72.08 (UqCRC2) 63.82 (MTCO1) 63.96 (COX4+4L2) 55.41 (OSCP) NA (VDAC1)
XGB	100	93	97	78	82.05 (NDUFB8) 80.12 (NDUFA13) 79.06 (SDHA) 79.34 (UqCRC2) 79.06 (MTCO1) 79.06 (COX4+4L2) 78.35 (OSCP) 81.48 (VDAC1)	82.33 (NDUFB8) 80.12 (NDUFA13) 82.47 (SDHA) 82.47 (UqCRC2) 80.34 (MTCO1) 80.91 (COX4+4L2) 81.05 (OSCP) NA (VDAC1)

ML classification results (class D)

LR and XGB models were trained with various combinations of input features and the results from optimised best performing models are detailed in Table 3.15. The optimised parameters of both LR and XGB models trained with all eight markers are presented in Table 3.14.

As evident in Table 3.15 only XGB can predict the class D myofibres with accuracy > 90% i.e. 93% more important to note is the recall for controls for both models i.e. 35% by LR and 78% by XGB. This inaccuracy of the LR model made it unfit for predictive inference. Whilst XGB model's control recall is better but not > %90, but because it achieved 97% on class D myofibres, it was decided that it would be used for predictive inference, i.e. to apply explainable ML methods.

Applying explainable ML methods to XGB model for class D myofibres prediction

The global explanations generated using whole training data with SHAP EMs that are presented in Figure 3.22 for the XGB model shows i) the patterns in the data it uses to achieve 93% accuracy i.e. SHAP EM shows the top three average SHAP values for protein markers are UqCRC2, COX4+4L2 & OSCP. ii) The SHAP value prevalence observed in plot **C** within Figure 3.22 shows low mean intensities for UqCRC2 protein are associated with extreme negative SHAP values and to lesser extent vice versa, these are very consequential towards predicting instances of myofibres as control class and to a lesser extent as class D. In

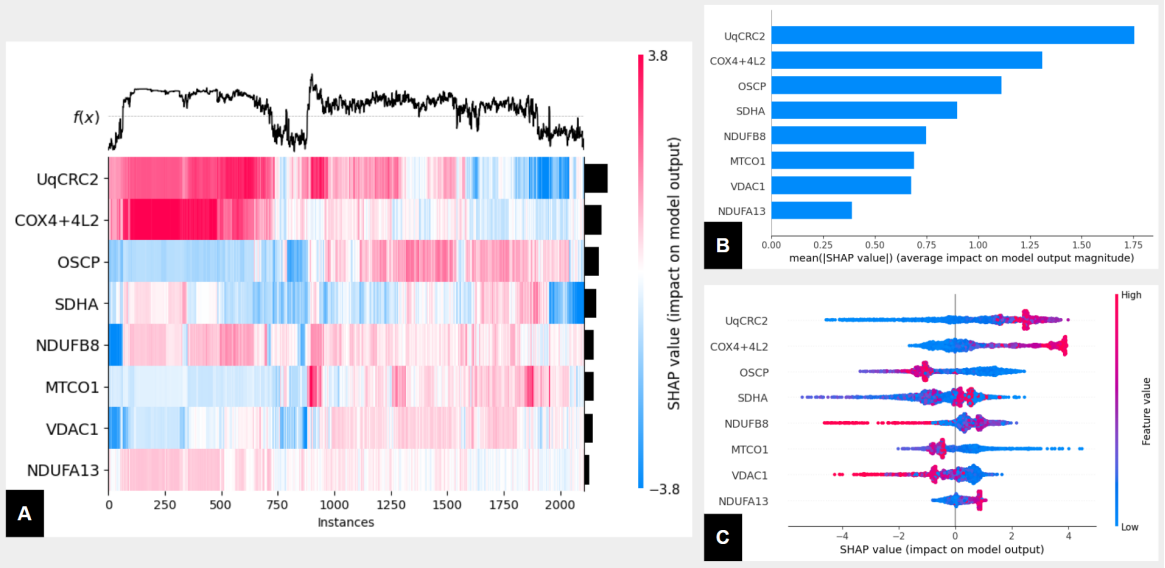

Figure 3.22 SHAP global explanations of XGB model that predicts class D myofibres. The XGB model interrogated here had 93% test accuracy and trained on all eight protein markers. **A**, SHAP heatmap plot. **B**, bar chart of the average SHAP value. **C**, a set of SHAP beeswarm plots

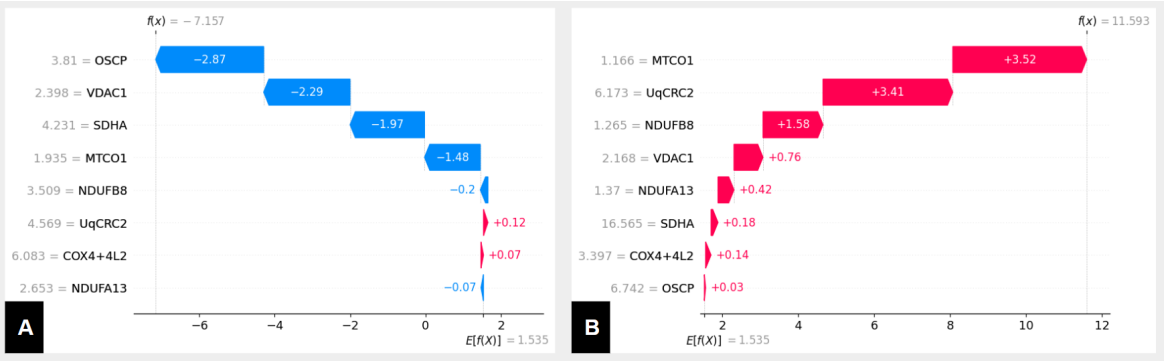

Figure 3.23 SHAP local explanations of XGB model that predicts class D myofibres. **A**, SHAP waterfall plot providing local explanation for correctly predicted control myofibre instance. **B**, Another SHAP waterfall plot but for correctly predicted class D myofibre.

addition COX4+4L2 also has a similar association to UqCRC2 with SHAP values. It can also be observed that the two complex IV protein markers (COX4+4L2 and MTCO1) have inverse association with each other, i.e. higher mean intensities of COX4+4L2 protein are associated with positive SHAP values and lower mean intensities of MTCO1 protein are associated with positive SHAP values that are consequential towards predicting instances of myofibres as positive (class D) class. These associations are further evident in the SHAP heatmap plot **A** within Figure 3.19, where the inverse relationship between UqCRC2,COX4+4L2 and OSCP and MTCO1 is leveraged by the XGB model.
Similar associations are also observed by studying individual instances of myofibres i.e. local explanations that are presented in Figure 3.23. A more detailed per protein interpretation of SHAP values combined with their predictive power and their average mean intensities across the both classes is available in Appendix .1, Table 7.

Insights and predictive inference

The following insights are drawn from applying the explainable XGB ML model to a dataset of class D and control myofibres.

Predictive inference from XGB model insights (class D) The predictive inference for this model should be seen with acknowledgement of its low accuracy on control myofibres. The class D myofibres differ from control myofibres in the following pattern. Compared to control myofibres, the levels of markers for complex III (UqCRC2) and complex IV (COX4+4L2) are higher and complex I (NDUFB8) is lower in class D myofibres. Further, an inverse association exists between the two complex IV markers (COX4+4L2 & MTCO1), enough to predict with 93% accuracy (97% for patients & 78% for control).

Biological validation

It should be noted that biologically validating predictive inferences for myofibres linked with Class D mutations is not possible as these mutations' relationship with OXPHOS proteins is not fully understood. Predictive inference finds higher levels of complex III (UqCRC2) and complex IV (COX4+4L2), lower levels of complex I (NDUFB8), which make biological sense but need to be investigated further experimentally. The inverse association between complex IV proteins is surprising, a finding that warrants further investigation.

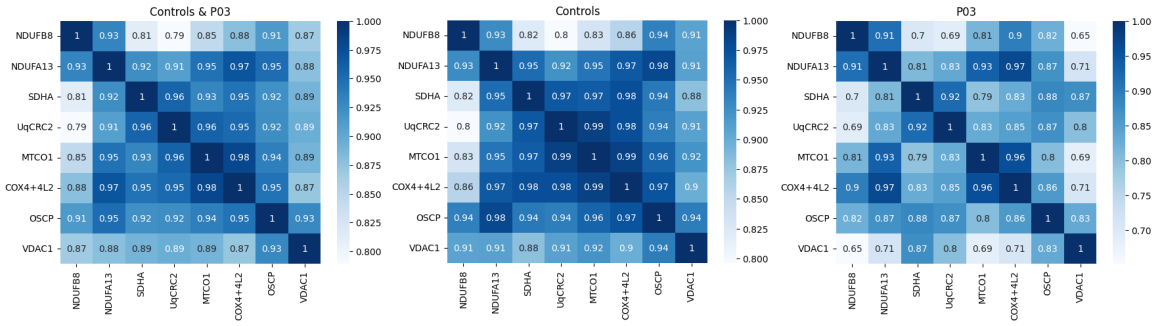

Figure 3.24 Correlation between protein markers in P03 myofibres

Table 3.16 **Statistical distribution of mean intensities in P03 myofibres** : These are computed using mean intensity in a myofibre for each protein marker over all P03 myofibres

Summary	NDUFB8	NDUFA13	SDHA	UqCRC2	MTCO1	COX4+4L2	OSCP	VDAC1
Min	1.06	1.06	1.44	1.24	1.02	1.31	1.15	1.05
Mean	2.17	2.25	3.75	3.30	1.56	5.07	2.22	1.51
Std	0.62	0.64	1.13	0.92	0.28	1.94	0.52	0.22
Max	5.64	6.08	16.92	8.34	3.54	19.37	5.61	4.75

3.4.6 Explainable ML Analysis of P03 vs Controls

Classification of myofibres from P03 was one of the cases that was not possible with the current techniques as observed in Table 2.4. For this reason it was decided that to use it as a challenging benchmark to provide contrast. The following observations were made in the EDA which is presented in Figure 3.24. In controls a high positive correlation exists between all eight protein markers. To a lesser extent it is also observed in control and P03 data and to an even lesser extent in just P03 data. VDAC1's correlation with SDHA is comparable to controls. The correlation for all other protein markers is reduced in P03 myofibres compared to controls.

ML classification results (P03)

LR and XGB models were trained with various combinations of input features and the results from optimised best performing models are detailed in Table 3.18. The optimised parameters of both LR and XGB models trained with all eight markers are presented in Table 3.14.

As evident in Table 3.18 only XGB can predict the P03 myofibres with an accuracy of 95%; more important to note is the recall for controls for both models i.e. 55% by LR and 91% by XGB. The inaccuracy of the LR model makes it unfit for predictive inference. Based on accuracy and recalls it was decided that XGB should be used for predictive inference, i.e. to apply explainable ML methods.

Table 3.17 Optimised model parameters for P03 vs controls myofibres.

Method	Parameter	Optimised parameter value
LR	const	6.62
	NDUFB8	1.28
	NDUFA13	1.19
	SDHA	-0.71
	OSCP	-4.38
	COX4+4L2	0.33
	MTCO1	-4.84
	VDAC1	-0.93
	UqCRC2	2.84
XGB	colsample_bytree	0.7
	eta	0.15
	gamma	0.1
	max_depth	8
	min_child_weight	1

Table 3.18 Accuracy metrics for models trained to predict P03 myofibres

Model	Training Accuracy(%)	Test Accuracy(%)	Recall(%) (Patients)	Recall(%) (Controls)	Accuracy (single protein)(%)	Accuracy (VDAC1+Protein)(%)
LR	82	84	95	55	70.60 (NDUFB8) 67.50 (NDUFA13) 68.33 (SDHA) 62.73 (UqCRC2) 69.15 (MTCO1) 66.46 (COX4+4L2) 72.67 (OSCP) 71.43 (VDAC1)	73.50 (NDUFB8) 73.91 (NDUFA13) 73.29 (SDHA) 74.53 (UqCRC2) 73.71 (MTCO1) 74.12 (COX4+4L2) 74.32 (OSCP) NA (VDAC1)
XGB	100	95	97	91	75.15 (NDUFB8) 72.88 (NDUFA13) 74.12 (SDHA) 74.32 (UqCRC2) 74.53 (MTCO1) 74.74 (COX4+4L2) 78.05 (OSCP) 76.60 (VDAC1)	76.60 (NDUFB8) 76.40 (NDUFA13) 76.81 (SDHA) 76.60 (UqCRC2) 76.40 (MTCO1) 77.02 (COX4+4L2) 75.57 (OSCP) NA (VDAC1)

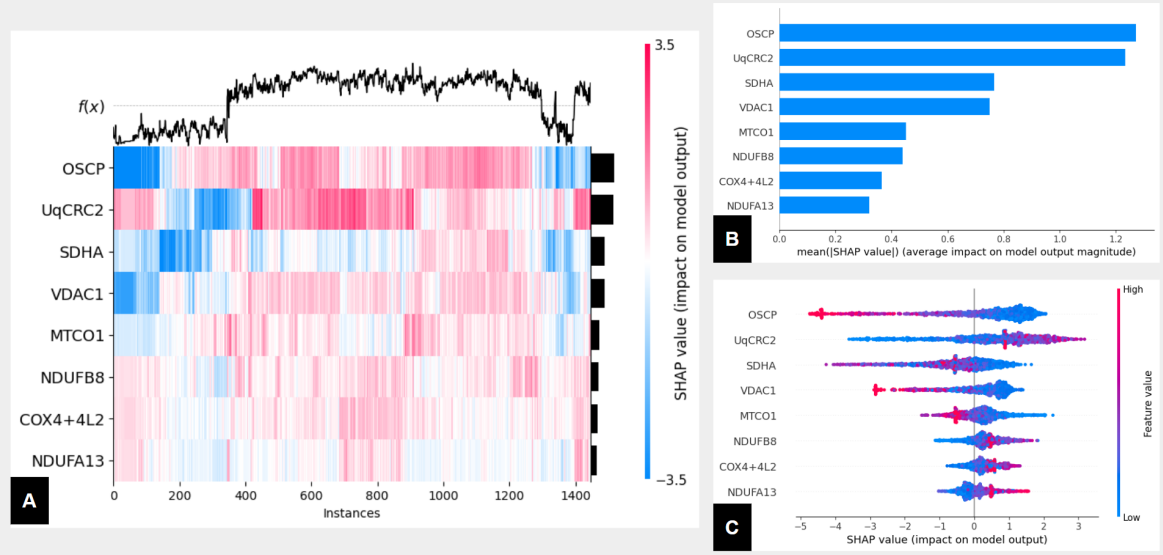

Figure 3.25 SHAP global explanations of XGB model that predicts P03 myofibres. The XGB model interrogated here had 95% test accuracy and trained on all eight protein markers. **A**, SHAP heatmap plot. **B**, bar chart of the average SHAP value. **C**, a set of SHAP beeswarm plots

Applying explainable ML methods to XGB model for P03 myofibres prediction

The results are presented in Figures 3.25 and 3.26. The global explanations generated using whole training data with SHAP EMs that are presented in Figure 3.25 for XGB model show i) the patterns in the data it uses to achieve 95% accuracy i.e. SHAP EM shows the top three average SHAP values for protein markers are OSCP, UqCRC2 and SDHA. ii) The SHAP value prevalence observed in plot **C** within Figure 3.25 shows OSCP protein high mean intensities are associated with extreme negative SHAP values and to a lesser extent vice versa, these are very consequential towards predicting instances of myofibres as control

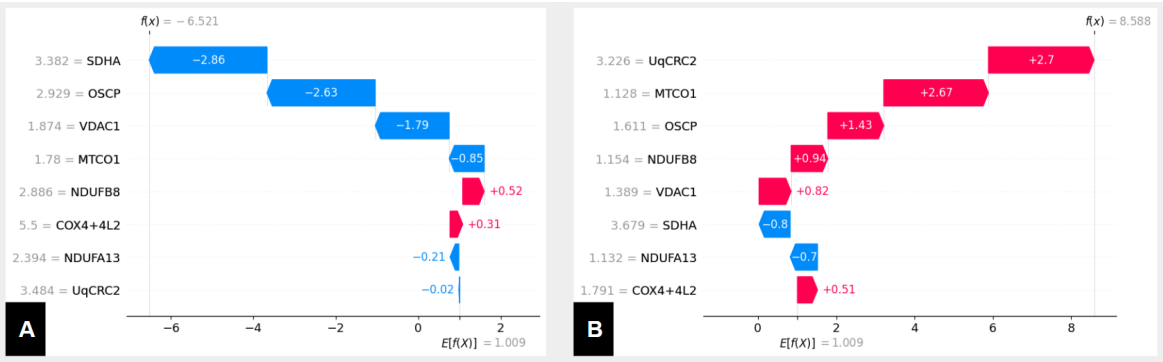

Figure 3.26 SHAP local explanations of LR model that predicts P03 myofibres. **A**, SHAP waterfall plot providing local explanation for correctly predicted control myofibre instance. **B**, Another SHAP waterfall plot but for correctly predicted P03 myofibre.

class and to a lesser extent as P03 class. In addition SDHA also has a similar association to OSCP with SHAP values, but inverse association is observed with UqCRC2. It can also be observed that the two complex IV protein markers (COX4+4L2 and MTCO1) have inverse association with each other, i.e. higher mean intensities of COX4+4L2 protein are associated with positive SHAP values and lower mean intensities of MTCO1 protein are associated with positive SHAP values that are consequential towards predicting instances of myofibres as positive (P03) class. These associations are further evident in the SHAP heatmap plot **A** within Figure 3.25, where the inverse relationship between (OSCP& SDHA) and UqCRC2 is leveraged by the model.
Similar associations are also observed by studying individual instances of myofibres i.e. local explanations that are presented in Figure 3.26. A more detailed per protein interpretation of SHAP values combined with their predictive power and their average mean intensities across both classes is available in Appendix .1, Table 8.

Insights and predictive inference

The following insights are drawn from applying the explainable XGB ML model to a dataset of P03 and control myofibres.

Predictive inference from XGB model for P03 myofibres should be seen with acknowledgement of its lower accuracy on control myofibres. The associations used by the model are complex and combinatory with only complex V(OSCP) marker lower level's association with P03 being clear. But other markers are good discriminators too, evident by their respective predictive powers and the model with all eight markers was able to predict with 95% accuracy (97% for patients and 91% for control).

Biological validation

The predictive inference for P03 myofibres is limited which makes biological validation difficult.

3.5 Discussion

3.5.1 Classification accuracy of explainable ML methods

As observed in Section 3.4, the predictive accuracy of the methods developed in this chapter ranges from 100% for class A myofibres to 93% for class D myofibres. These are better than

the current technique as presented in Table 2.4. These classification accuracies from a perfect 100% to relatively good 93% implies there exist a pattern in the dataset which ML models can exploit to achieve these accuracies. It can also be concluded that the models selected for this task, i.e. LR and XGB, were appropriate, this is evident by classification results and their ability to capture linear and non linear associations between all the input features and the target variable, this was evident in their explanations plots.

It was also observed low-biased XGB models provided complementary and additive results to high-biased LR as discussed in Section 3.3.3.

3.5.2 Insights from combination of predictive power of individual features and explainable methods

The predictive inferences presented in tables in Appendix .1 e.g. Table 1, help to understand how each feature was used by the model to make predictions in terms of their overall impact, correlation of their value towards target outcome, contribution prevalence (i.e. proportion of predictions on a scale of contribution), predictive power of the feature on its own, and its statistical shape (mean and standard deviation) compared to control myofibres. All these collective and combined with other features explanations allow biomedical scientists to infer associations used by the model. In three out of five cases analysed in this chapter the predictive inference of two models was used, sometimes this provided additive insights.

For the first three cases (class A, B and C) both LR and XGB models produced >95% accuracy; the explanations were much easier to interpret, as can be observed in the explanation plots. This was due to associations used by models being linear which resulted in explanation plots showing linear correlations between features and target variables. This ultimately allowed me to generate hypotheses that are simple, i.e. description of target class myofibre patterns in terms of up or down regulation compared to control class.

For the last two cases (class D and P03) the models did not used linear associations as evidenced by poor performance of LR and better performance of XGB models with increased 'max_depth' which implies the bases of prediction are more complex. This resulted in SHAP explanation plots which were inconclusive about a feature's correlation with target variable, i.e. the strength of a feature's pixel intensity cannot be easily decomposed as contributing positively or negatively towards target prediction on the global level. In this scenario studying local SHAP explanations provided insights into associations. In addition to results presented earlier, SHAP interaction analysis [127] was also performed but the results from it did not add any additional insights.

Biological validation and novel insights

Class A myofibres are instrumental in validating the ML models as these have well defined biologically expected patterns that can be used for validation of predictive insights. These expected patterns such as downregulation of complex I proteins and compensatory upregulation of complex II-V in class A myofibres has been used to validate the derived class A predictive inference insights. All the insights discovered by the explainable ML pipeline for class A myofibres are in accord with the biological expectations as discussed in Section 3.4.2. There are a number of predictive insights that are potential novel pathology explaining discoveries such as association of high levels of UqCRC2 to class B mutations, correlation between SDHA and UqCRC2 in class C mutations and inverse association between complex IV proteins in class D mutations.

3.5.3 Limitations

While the classification accuracy and explanations where models were able to exploit linear associations were useful but there are a number of limitations of the ML analysis conducted in this chapter.

Reliability of processed data

The methods used in this chapter assume that the processed data derived from mitocyto segmentation to be reliable. This assumption is not backed by any evidence or evaluation of the quality of segmentation and curation of myofibres. This is a substantial weakness of the study conducted in this chapter, i.e. if segmentation quality affected the per myofibre mean protein intensities it will have impacted the model training.

Ignoring intra-myofibre features

The ML analysis conducted in this chapter used statistical summaries that ignored all intra-myofibre features. There are many hypotheses [2, 3] that theorise the existence of differential features within myofibre in various mitochondrial dysfunctions, this is further discussed in Section 6.1. The ML analysis conducted in this chapter would not be useful in this regard.

Interpretations of complex associations

As observed in classification of Class D and P03 myofibres where the model exploited complex associations, the interpretations of explanation plots were complex and deriving a

reliable hypothesis backed by hard evidence was difficult. The local explanation plots are helpful but studying hundreds of them to decipher associations is impractical.

Small patient cohort

The analysis conducted in this study has a small number of patients per genetic mutation class. This is because patients suffer from a disease that is rare; nevertheless, the conclusion derived from this analysis should be viewed considering this limitation.

3.5.4 Conclusion

In this chapter predictive inference of four classes of mitochondrial mutations using processed single-cell (myofibre) data was performed. The classification results exceeded the current techniques and explanations provided by SHAP plots were helpful in recognising the patterns/associations between input features and target features, especially when models were able to exploit multi-linear relationships/associations. The reliability of processed data and impracticality of deciphering complex association need further research.

Chapter 4

NCL-SM: A Fully Annotated Dataset of Images from Human Skeletal Muscle Biopsies

4.1 Introduction

As discussed in Section 3.5.3 reliability of myofibre segmentation and myofibre curation i.e. selecting the myofibres that are not affected by any artefactual defects, is an unresolved issue with the current analysis techniques. Subsequently the quality of myofibre segmentation and curation achieved using mitocyto and others were evaluated as presented in Chapter 5 and were found to be imprecise. It was then decided to use deep learning for these two tasks and it was found that vanilla/generalised DL models such as UNET, Mask R-CNN, Cellpose did not produce the segmentation quality required and explained in Chapter 5 .

To train a bespoke model for precise myofibre segmentation and curation, requires a precisely segmented and curated dataset of SM tissue images for training. To the best of my knowledge there is no such publicly available dataset for myofibre segmentation or classification. Making this dataset available and clearly defining the challenge involved in myofibre segmentation and curating the usable myofibres is a crucial first step towards development of an automatic tool for this problem.

High quality annotated datasets are critical for development of relevant ML/DL models or pipelines. This has been evident since the early days of modern ML with datasets such as MNIST [150], COCO [151], ImageNet [115] and more recently SA-1B [152] enabling the construction of some seminal ML models like ResNet [118], VGG [153], vision transformer [154] and SAM [152]. It was decided to undertake a project to build Newcastle Skeletal

Muscle (NCL-SM), a dataset of precisely segmented and curated SM tissue images, and make it publicly available to nurture open science and allow transparency of any solutions we build using this dataset.

4.1.1 SM tissue segmentation for analysis

Microscopy imaging and cytometry-based pseudo imaging techniques allow us to observe protein expressions within individual cells within tissue images, when a single cell segmentation approach is used. For many biological and disease processes, this is the most appropriate spatial scale for understanding mechanisms. Further, the spatial arrangement of cells of different classes within tissues *in vitro* is often informative about biology and disease pathology. There are many diseases affecting SM tissue, including amyotrophic lateral sclerosis [155], multiple sclerosis [156], muscular dystrophy [157] and a wide range of mitochondrial diseases [158]. The dataset presented in this chapter is collected from healthy human control subjects and from patients suffering from genetically diagnosed muscle pathology, including mitochondrial diseases.
Any analysis of SM images at the single myofibre level requires 1) precise segmentation of individual myofibres, as regions close to the myofibre membrane are known to exhibit differential features [2]. 2) removal of myofibres damaged by freezing that might occur in the process of storage and thawing, as the subcellular patterns in protein expression, or indeed per-cell mean expression, will be impacted by the technical artefact, masking the target biology. 3) removal of SM myofibres that are not sliced in transverse orientation or are partially sectioned, as the presence of such myofibres does not allow for a standard or uniform comparison across all the myofibres in a tissue and 4) removal of folded tissue. Tissue can fold in on itself during tissue handling and slide preparation. Such folding artificially amplifies apparent protein expression in affected regions and is again a purely technical artefact, not related to target biology.

Currently most single-myofibre SM analysis is carried out using custom built semi-automatic pipelines like mitocyto [1], using general image analysis tools like Ilastik [84] or cellprofiler [83], or using vanilla ML models like StarDist [120] or Cellpose [89]. None of these approaches produce the segmentation quality required for analysis of SM out of the box, without suitable training in my experience. Custom built pipelines like mitocyto are used more often than general ML models (which are not trained on SM data) as these require relatively fewer corrections than general ML models. I evaluate and discuss mitocyto segmentation quality in Chapter 5 . To improve the segmentation quality and remove compromised myofibres and SM regions, biomedical scientists spend hours manually correcting the issues in tissue section images and segmentation masks such as manually correcting individual

myofibre annotations and classifications, before doing downstream quantitative analysis. This can be an inefficient use of scientists' time but also the corrections are subjective and not reproducible.

The main barrier to the development of a suitable ML tool/model for fully automatic segmentation and curation of SM myofibres is the lack of any high quality, manually segmented and curated SM myofibre image data on which to train appropriate ML models. And to account for subjectivity in annotations there need to be be duplicate annotations to reveal the level of subjectivity.

4.2 Aims

The aims of this chapter are

- To develop evaluation metrics that detail the quality of segmentation and curation required for SM tissue image analysis.
- To build a dataset of precisely segmented and curated SM tissue sections that capture diverse subject groups, i.e. controls and patients, and imagining techniques, i.e. IF (microscopy) and IMC and evaluate the quality of this dataset using the developed evaluation metrics.

4.3 Methods

4.3.1 SM tissue segmentation and curation protocol

The SM tissue segmentation and curation protocol developed under the guidance of biomedical experts from WCMR can be divided into four tasks.

Myofibre segmentation

The protocol for myofibre segmentation is i) include all areas within a myofibre that had mitochondrial mass signal, ii) exclude any areas within a myofibre that had myofibre membrane signal and iii) prioritise signal from within myofibre when membrane signal is weak. Point iii) is necessary because noise is common in these types of data resulting in some overlapping mass and membrane pixels. In such scenarios, we consider mitochondrial mass signals within a myofibre to be the most reliable indicator of myofibre morphology as presented in Figure 4.1.

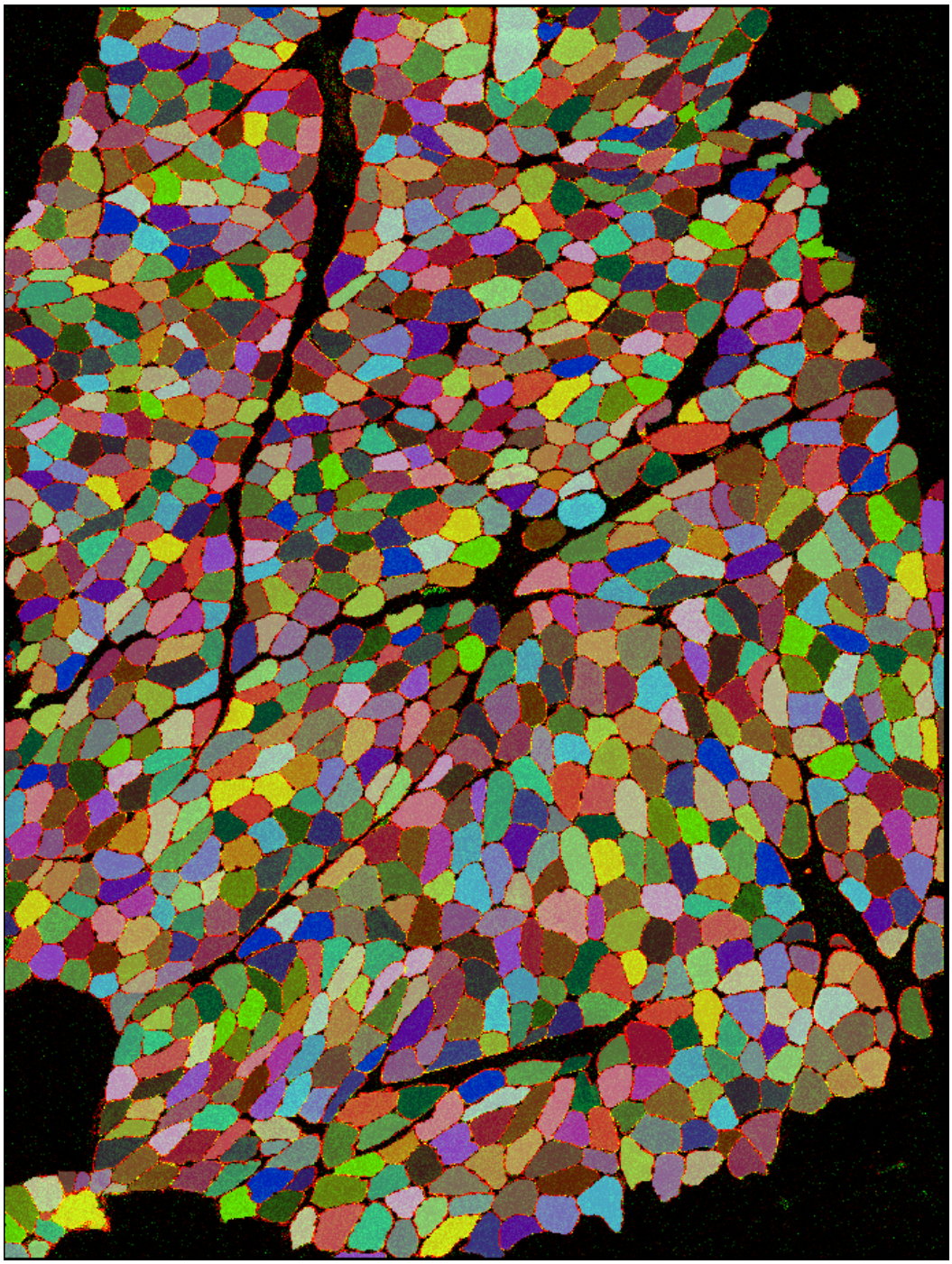

Figure 4.1 **Typical Manual Segmentation of a SM Tissue Section**: An IMC SM tissue section image from subject P17. Consists of 1,068 myofibres manually segmented following the protocol, i.e. include all myofibre mass and exclude membrane. Each colour is an unique pixel value per myofibre in the annotated mask.

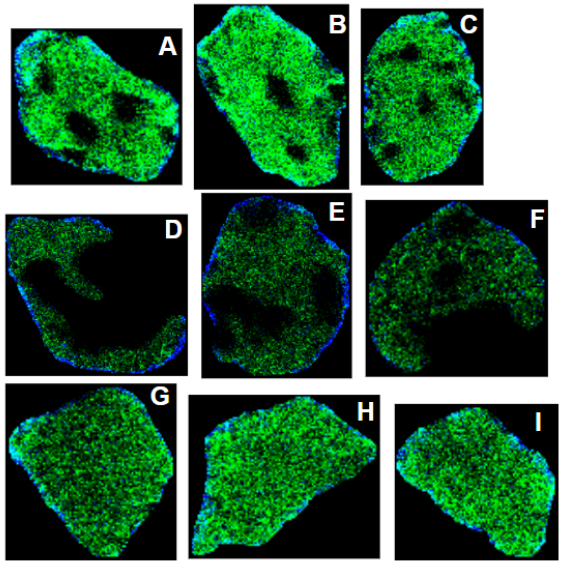

Figure 4.2 **Identifying Myofibre Freezing Artefacts** All myofibres from section P02. A (id:190), B (id:138) and C (id:257) are typical freezing damaged myofibres resulting in a leopard spot pattern, D (id:448), E (id:117) and F (id:398) are partial myofibres that are more severely damaged by freezing, G (id:303), H (id:415) and I (id:288) are examples of myofibres without any freezing defects. Figure adapted from Khan *et al.* [41].

Freezing Artefact Myofibre (FAM) classification

The protocol for identifying myofibres with freezing artefacts was i) look for a leopard spot pattern within myofibres that are typical of freezing damage and ii) look for partial myofibres i.e. large part of myofibres missing as a result of freezing. These patterns are demonstrated in Figure 4.2.

Non-Transverse Sliced Myofibre (NTM) classification

The protocol for identifying non-transverse sliced myofibres was to look for i) myofibres with skewed aspect ratio, e.g. elongated; ii) all myofibres at the border of image: these are partial observations; iii) segmented objects which are too small or too big; and iv) myofibres with unusual convexity. NTM shapes are demonstrated in Figure 4.3.

Folded tissue Regions (FR) segmentation

Folded tissue regions were segmented by an expert biomedical scientist, these were identified by looking for overlapping membrane signals that result in mesh patterns as demonstrated in Figure 4.4.

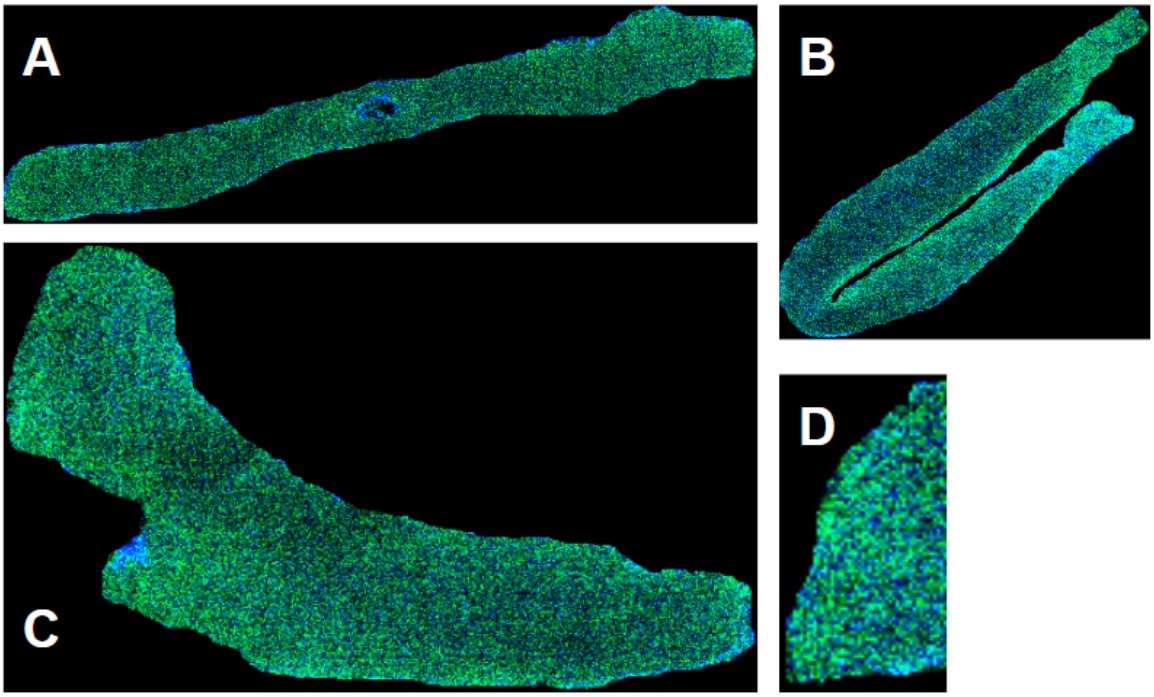

Figure 4.3 **Typical Non-Transverse Sliced Myofibres**: Myofibres from tissue section C04. A (ID: 39) is a typical elongated myofibre cross-section which has not been cut in the transverse orientation B is myofibre (ID:6) with unusual convexity, C is myofibre (ID:24) of large area, and D is a partial myofibre (ID:12) that has been truncated by the border of the image

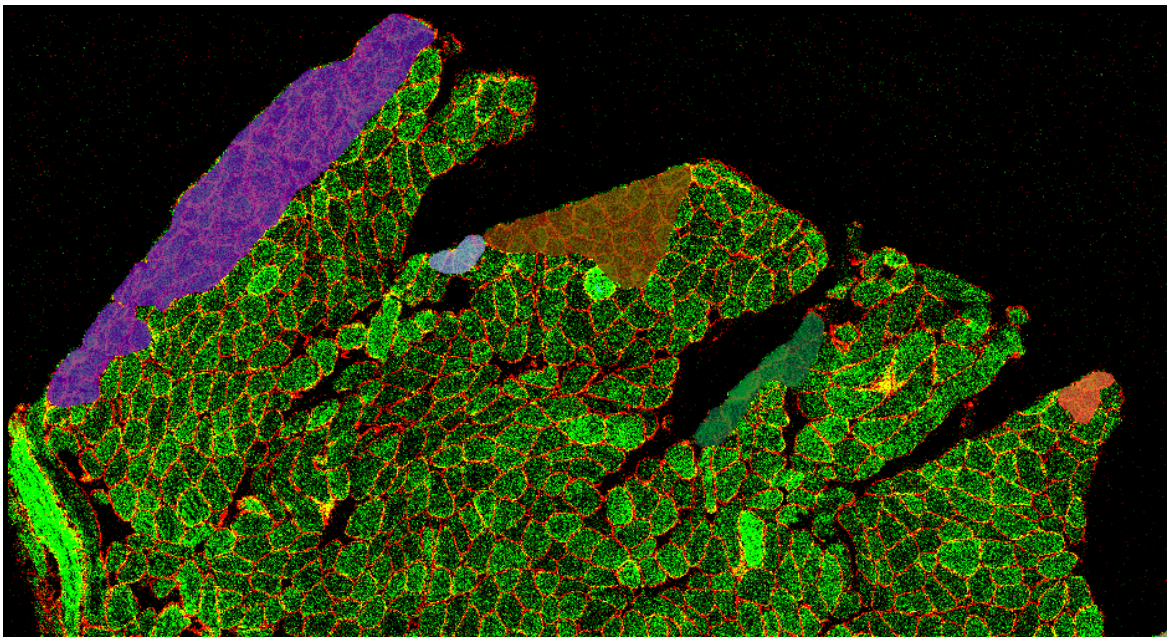

Figure 4.4 **Typical SM Tissue Folding** Segmentation of folded regions (each instance of folding is coloured differently) in a tissue section P06

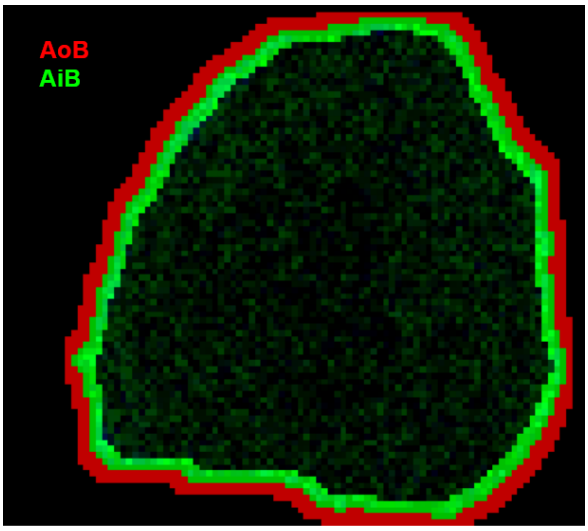

Figure 4.5 **Area Near the Membrane** IMC image of a fibre (ID:723) from tissue section P17 illustrating Area outside the Border (AoB, red) and Area inside the Border (AiB, green) on either side of the border of an annotated myofibre. These areas are identified by eroding and dilating the border using 5x5 and 9x9 pixel kernels for IMC and IF images respectively.

4.3.2 Evaluation metrics for SM tissue image segmentation and curation

In order to evaluate the quality of myofibre segmentation annotations we need quantitative metrics that capture the nuances required during SM analysis. In this section these nuances are defined and the quality evaluation metrics are introduced.

Myofibre segmentation evaluation metrics

Intersection over Union (IoU) is a widely used evaluation metric to measure the quality of annotation/segmentation in computer vision tasks. But for myofibre segmentation examining IoU of each myofibre alone will not reveal important aspects about segmentation quality, i.e. the areas missed or included matter, as emphasised in Section 4.1, in other words we want any automatic pipeline to have high accuracy segmenting areas on either side of the border of each myofibre as illustrated in Figure 4.5. To measure this, we developed two quantitative metrics: myofibre mass missed correlation (r_{AoB}) and myofibre membrane included correlation (r_{AiB}) along with the aforementioned IoU. All three metrics are defined below:

- **Myofibre mass missed correlation** (r_{AoB}): This is the Pearson correlation between the proportion of myofibre mass pixels missed in the Area outside the myofibre Border (AoB) by annotator x, compared to annotator y, across all available myofibres $i \in 1...n$.

$$r_{AoB} = \frac{\sum_{i=1}^{n}(x_{AoBi} - \overline{x_{AoB}})(y_{AoBi} - \overline{y_{AoB}})}{\sqrt{\sum_{i=1}^{n}(x_{AoBi} - \overline{x_{AoB}})^2 \sum_{i=1}^{n}(y_{AoBi} - \overline{y_{AoB}})^2}} \tag{4.1}$$

where x_{AoBi} is the proportion of myofibre mass pixels in AoB in myofiber i as annotated by annotator x; y_{AoBi} is the proportion of myofibre mass pixels in AoB in myofiber i as annotated by annotator y; overbar represents the mean across all n myofibres; and n is the number of myofibres assessed.

- **Myofibre membrane included correlation (r_{AiB})**: Defined as above, but for AiB:

$$r_{AiB} = \frac{\sum_{i=1}^{n}(x_{AiBi} - \overline{x_{AiB}})(y_{AiBi} - \overline{y_{AiB}})}{\sqrt{\sum_{i=1}^{n}(x_{AiBi} - \overline{x_{AiB}})^2 \sum_{i=1}^{n}(y_{AiBi} - \overline{y_{AiB}})^2}} \tag{4.2}$$

where x_{AiBi} is the proportion of myofibre membrane pixels in AiB in myofiber i as annotated by annotator x; y_{AiBi} is the proportion of myofibre membrane pixels in AiB in myofiber i as annotated by annotator y.

- **IoU (IoU_i)**: This is defined as the intersection of overlapping pixels divided by union of all pixels between two annotations of myofibre i. This is measured per myofibre and $\overline{IoU}$ is the mean across all n myofibres assessed.

$$IoU_i = \frac{|X_i \cap Y_i|}{|X_i \cup Y_i|} \tag{4.3}$$

where X_{i} and Y_{i} are annotations of myofibre i by annotators X and Y.

Myofibre curation evaluation metrics for FAM and NTM

As discussed in Section 4.1.1, inclusion of FAMs and NTMs in the analysis can impact the subcellular patterns in protein expression, or indeed per-cell mean expression, masking the target biology. At the same time it is vital to emphasise that the tissue collected for the analysis come from muscle biopsies of patients suffering from rare mitochondrial diseases, making these biopsies a rare and precious resource that needs to be utilised without any wastage i.e. salvaging the myofibres wherever possible. For this reason striking a balance between full utilisation of SM tissue and preserving the analysis from any artefactual factors is important. It was decided that *sensitivity*, *specificity* and F1 score as metrics will allow us to measure the myofibre curation quality and help find a balance as discussed above.

$$sensitivity = \frac{true\ positives}{true\ positives + false\ negatives} \tag{4.4}$$

$$specificity = \frac{true\,negatives}{true\,negatives + false\,positives} \tag{4.5}$$

$$F1\ score = \frac{2 * true\,positives}{2 * true\,positives + false\,positives + false\,negative} \tag{4.6}$$

FR segmentation evaluation metrics

Folded tissue section affects regions of tissue which means its affect needs to be calculated on the whole tissue section. It was decided that the best metrics to evaluate FR in a tissue section will be IoU, i.e. intersection of overlapping pixels divided by union of all pixels between two annotations of all FRs in a tissue section.

4.4 Data

As part of this thesis a high quality fully manually segmented and mostly manually curated SM imaging dataset, i.e. NCL-SM,[1] and its quality evaluation code[2] is made publicly available. The images are collected using two different imaging technologies: Imaging Mass Cytometry (IMC) and ImmunoFluorescence (IF) amounting to 50,434 myofibres in 46 tissue sections, 30,794 of which are classed as 'analysable', 18,102 classed as 'not-analysable-due-to-shape', 1,538 myofibres classed as 'not-analysable-due-to-freezing-damage' and 405 annotations of folded tissue regions.
The NCL-SM dataset includes images of 46 tissue sections that capture spatial variation in protein expression within tissue (including within myofibres). We use microscopy-based techniques, i.e. IF and advanced protein expression measurement techniques like IMC that allow us to observe the spatial variation in the expression of up to 40 proteins in tissue simultaneously.

4.4.1 Capturing images

The following is the sequential process of collecting this data.

Biopsies

Following ethical approval from the Newcastle and North Tyneside Local Research Ethics Committee and informed consent from control and patient subjects, biopsies of SM were

[1] https://doi.org/10.25405/data.ncl.24125391
[2] www.github.com/atifkhanncl/NCL-SM

collected from patients and healthy control subjects as described in Table 4.1. However, ethical approval for this project was not required as it consisted of a re-analysis of existing image data collected as part of studies conducted in [1, 2, 47, 159]. This original data was obtained from approved studies on samples from Newcastle Brain Tissue Resource (Approval Ref:2021031) and Newcastle Mitocondrial Research Biobank (approved by the Newcastle and North Tyneside Research Ethics committee reference 16NE/0267) and Newcastle Biobank (approved by the Newcastle and North Tyneside Research Ethics committee 12/NE/0394).

Table 4.1 Subject information: Information of patients and controls detailing gender, age at biopsy, clinical information, genetic defect, heteroplasmy level, tissue section (TS) name and Imaging type. The subject IDs that start with P and C are patients and controls whose tissues were imaged using IMC, and the subject IDs that start with S are subjects whose tissues were imaged using IF .The orange rows define the genetic mutation of the subjects in the following rows.

Subject	Gender	Age	Clinical information	Genetic defect	Heteroplasmy level	TS names	Imaging type
Nuclear-encoded mutations affecting complex I (taken from the vastus lateralis)							
P01	M	Adult	Exercise intolerance, unable to perform sustained aerobic exercise normal resting lactate, normal CK	TMEM126B Homozygous c.635G>T, p.(Gly212Val)	NA	P01	IMC
P02	M	Adult	Exercise intolerance, muscle cramps, elevated serum lactate	ACAD9 Compound heterozygous c.1150G>A, p.(Val384Met) and c.1168G>A, p.(Ala390Thr)	NA	P02	IMC
Single, large-scale mtDNA mutations (taken from the vastus lateralis)							

P03	F	29yrs	CPEO and bilateral ptosis	Deletion size: 4372bp Breakpoints: 8929-13301 mtDNA deletion level: 53%	53%	P03	IMC
Point mutations in mitochondrial-encoded tRNA leucine (MT-TL1) (taken from the vastus lateralis)							
P05	F	25yrs	Exercise intolerance, ptosis	m.3243A>G MT-TL1 mutation	66%	P05	IMC
P06	F	47yrs	Modest exercise intolerance	m.3243A>G MT-TL1 mutation	34%	P06	IMC
P07	M	53yrs	CPEO	m.3243A>G MT-TL1 mutation	74%	P07	IMC
Point mutations in other mitochondrial-encoded tRNAs (taken from the vastus lateralis)							
P08	M	33yrs	Mitochondrial myopathy	m.10010T>C MT-TG mutation	89%	P08	IMC
P09	F	35yrs	Mild weakness	m.14709T>C MT-TE mutation	76%	P09	IMC
P10	M	63yrs	Exercise intolerance, prominent exertional dyspnea	m.5543T>C MT-TW mutation	NA	P10	IMC
DYSF mutation							
S09	NA	Adult	NA	DYSF	NA	S09	IF
S11	NA	Adult	NA	DYSF	NA	S11	IF

S13	NA	Adult	NA	DYSF	NA	S13	IF
S15	NA	Adult	NA	DYSF	NA	S15	IF
S18	NA	Adult	NA	DYSF	NA	S18	IF
POLG mutation							
S43	NA	Adult	NA	POLG	NA	S43	IF
S60	NA	Adult	NA	POLG	NA	S60	IF
RRM2B mutation							
S23	NA	Adult	NA	RRM2B	NA	S23	IF
Healthy controls (taken from the tibialis anterior)							
C01	M	20yrs	Taken during anterior cruciate ligament surgery			C01	IMC
C03	F	23yrs	Taken during anterior cruciate ligament surgery			C03	IMC
C03	F	23yrs	Taken during anterior cruciate ligament surgery			C03-2	IMC
C04	F	Adult	Taken during anterior cruciate ligament surgery			C04	IMC

C04	F	Adult	Taken during anterior cruciate ligament surgery			C04-2	IMC
C11	NA	Adult	Taken during anterior cruciate ligament surgery			C11	IMC
S02	NA	Adult	NA			S02	IF
S26	NA	Adult	Taken during anterior cruciate ligament surgery			S26	IF
S37	NA	Adult	Taken during anterior cruciate ligament surgery			S37	IF
Mitochondrial disease patients							
P11	NA	Adult	NA	NA	NA	P11	IMC
P12	NA	Adult	NA	NA	NA	P12	IMC
P13	NA	Adult	NA	NA	NA	P13	IMC
P14	NA	Adult	NA	NA	NA	P14	IMC
P15	NA	Adult	NA	NA	NA	P15	IMC
P16	NA	Adult	NA	NA	NA	P16	IMC
P17	NA	Adult	NA	NA	NA	P17	IMC
P18	NA	Adult	NA	NA	NA	P18	IMC
P19	NA	Adult	NA	NA	NA	P19	IMC
P20	NA	Adult	NA	NA	NA	P20	IMC
P21	NA	Adult	NA	NA	NA	P21	IMC
P22	NA	Adult	NA	NA	NA	P22	IMC

S38	NA	Adult	NA	NA	NA	S38	IF
S41	NA	Adult	NA	NA	NA	S41	IF
S44	NA	Adult	NA	NA	NA	S44	IF
S45	NA	Adult	NA	NA	NA	S45	IF
S48	NA	Adult	NA	NA	NA	S48	IF
S53	NA	Adult	NA	NA	NA	S53	IF
S54	NA	Adult	NA	NA	NA	S54	IF
S67	NA	Adult	NA	NA	NA	S67	IF

Tissue imaging

IMC described in Section 2.4 is a relatively new pseudoimaging technique that allows observation of up to 40 protein markers simultaneously but IF described in Section 2.3 is more widely used due to cost implications. For this reason it was decided that the segmentation tool would be more widely useful, if it can cater for images captured using both of these techniques. This meant the training data needed to include images captured using both IMC and IF.
The NCL-SM consist of 22,979 myofibres in 27 tissue sections imaged used IMC and 27,455 myofibres in 19 tissue sections imaged using IF. Together these total 50,434 myofibres in 46 tissue sections from 44 subjects.

Annotation

The 46 tissue section images in NCL-SM are made by arranging grayscale images of a myofibre membrane protein marker, i.e. Dystrophin and of a mitochondrial mass protein marker VDAC1 into an RGB image where R = membrane protein marker, G = mass protein marker and B = 0. Each channel of the images is contrast stretched (5th to 95th percentile) to improve image contrast for the segmentation task but raw images without contrast stretching are also included in NCL-SM. The annotation protocols discussed in Section 4.3.1 were followed to create NCL-SM.

Myofibre segmentation: Following the annotation protocol mentioned in Section 4.3.1 we trained annotation specialists from Gamaed[3] working under the close oversight of expert biomedical scientists from the WCMR to segment each myofibre in all 46 tissue sections, amounting to 50,434 myofibres, using the online Apeer platform[4] for

[3] www.gamaed.com/
[4] www.apeer.com/

all manual annotations. All segmentation went through rigorous visual inspection and a number of random myofibres in the data were segmented separately by expert biomedical scientists for quality assurance (QA). This is further discussed in Section 4.5.

Freezing artefact myofibre (FAM) classification: Following the annotation protocol mentioned in Section 4.3.1 FAM classification annotation was duplicated by two experts from WCMR and any disagreement was resolved by discussion among a panel of experts. This resulted in 1,538 myofibres with freezing artefacts where both annotators were in agreement. All annotations which differed between annotators were reviewed and resolved by a third expert after discussion.

Non-transverse sliced myofibre (NTM) classification: Following the annotation protocol mentioned in Section 4.3.1, for NTM classification annotations the following approach was used: 1) two experts working together identified up to 1,500 such myofibres in the data; 2) using these 1,500 myofibres, thresholds for area, convexity and aspect ratio were calculated; 3) these thresholds and a function to detect any myofibre on the edge were then applied on whole data, resulting in two classes of myofibres (NTM and non-NTM); 4) finally both classes of myofibres were rigorously visually inspected to detect and correct any mis-classification. These resulted in 18,102 NTMs.

Folded tissue region (FR classification): Following the annotation protocol mentioned in Section 4.3.1 FR were segmented by a biomedical scientist resulting in annotations for 405 different tissue regions affecting 37 out of 46 sections. A duplicate FR segmentation of tissue sections was performed by another expert biomedical scientist for measuring the reference human-to-human segmentation quality i.e. IoU score.

4.5 Results

4.5.1 NCL-SM counts

All the detailed myofibre segmentation and classification annotation counts in the NCL-SM are presented in Table 4.2.

Table 4.2 Annotation counts in the Newcastle skeletal muscle (NCL-SM) dataset. The table reports the counts of tissue section (TS), analysable myofibre (AM), non-transverse myofibre (NTM), freezing artefact myofibre (FAM) and folded region (FR).

Imaging Technique	TS Count	Myofibre Count	AM Count	NTM Count	FAM Count	FR Count
IMC	27	22,979	14,841	7,358	780	84
IF	19	27,455	15,953	10,744	758	321
Total	46	50,434	30,794	18,102	1,538	405

Table 4.3 Reported are the annotation quality metrics' values for QA human-to-human annotation comparison as mention in Section 4.4.1. In the table MF-A, r_{AoB}, r_{AiB}, $\overline{IoU}$, (A%(IoU >0.80), A%(IoU >0.90), A%(IoU >0.95)), QA-IMC, QA-IF stands for 'Myofibres Assessed', 'Myofibre Mass Missed Correlation', 'Myofibre Membrane Included Correlation', 'Mean IoU', ('Accuracy in terms of % of myofibres meeting IoU threshold of 0.8, 0.9 and 0.95'), 'QA for IMC images' and 'QA for IF images' respectively.

Annotations	MF-A	r_{AoB}	r_{AiB}	$\overline{IoU}$	A%(IoU >0.80)	A%(IoU >0.90)	A%(IoU >0.95)
QA-IMC	53	0.99	0.77	0.96	100	100	77.4
QA-IF	23	0.92	0.94	0.96	100	100	74

4.5.2 NCL-SM myofibre segmentation evaluation

The myofibre segmentation quality of NCL-SM evaluated by duplicate manual annotation performed to check for quality assurance is presented in Table 4.3. The lower r_{AiB} observed for IMC images are likely due to low resolution of IMC compared to IF.

4.5.3 NCL-SM FAM classification evaluation

The FAM classification quality of NCL-SM evaluated by duplicate manual annotations performed to check inter-annotator variability are presented in Table 4.4. The ground truth (GT) annotations are verified and corrected by a panel of experts as described in Section 4.4.1. As seen in the table there is some inter annotator variability (IAV) reflected by sensitivity scores of 0.79 and 0.87 between ANT1 and ANT2 annotations.

Table 4.4 NCL-SM FAM classification inter annotator variability (IAV). ANT1, ANT2, GT, SENS, SPEC and F1 stands for annotator 1, annotator 2, ground truth annotation, sensitivity, specificity and F1 score.

Annotations	TS	GT vs ANT1			GT vs ANT2			ANT1 vs ANT2		
		SENS	SPEC	F1	SENS	SPEC	F1	SENS	SPEC	F1
IAV(IMC)	P02, P06	0.79	0.99	0.88	0.87	0.99	0.92	0.89	0.98	0.84
IAV(IF)	S11, S60	0.88	0.99	0.90	0.85	0.98	0.86	0.81	0.98	0.82

Table 4.5 NCL-SM FR segmentation quality evaluation. TS, IoU and Mean(IoU) stands for tissue section, intersection over union and mean intersection over union of 19 TS respectively.

TS	C03	C04	P02	P03	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P18	P19	P20	P21	Mean(IoU)
IoU	0.51	0.12	0.16	0.28	0.47	0.61	0.12	0.34	0.38	0.46	0.53	0.60	0.04	0.16	0.45	0.52	0.52	0.22	0.36

4.5.4 NCL-SM NTM classification evaluation

As discussed in Section 4.4.1 myofibre morphological features were used to classify NTMs. The morphological features identified for annotating NTMs following the protocol developed by expert biomedical scientists as mentioned in Section 4.4.1 are converted into logical functions to classify NTMs/non-NTMS that is described in Algorithm 1. The NTM results, i.e. quality assessment is straightforward, i.e. any myofibre that has these morphological thresholds is classified as a NTM.

Algorithm 1 NTM classification based on myofibre morphological features. 'm' prefix before every morphological feature stands for myofibre and 'ImageResolution' of IMC is 1 μm and IF is 0.33 μm.

NTMClassification $(MyofibreMorphologicalFeatures, ImageResolution)$:

Input : Myofibre morphological features: $mArea$, $mOnBorder$, $mConvexity$, $mAspectRatio$, $mLengthSquaredByArea$ and $ImageResolution$

Output : $NTMclassification()$

if $((mArea < 210/ImageResolution) or (mArea > 27000/ImageResolution) or (mOnBorder = True) or ((mConvexity > 1.19) and (mLengthSquaredByArea > 3.10)) or ((mAspectRatio > 2.2) and (mLengthSquaredByArea > 1.8)))$ **then**
    return $True$;
else
    return $False$;
end

4.5.5 NCL-SM FR segmentation evaluation

The FR segmentation quality of NCL-SM evaluated by duplicate manual annotation performed to check for quality assurance are presented in table 4.5. As can be seen there is a high degree of IAV reflected by poor IoU score across all TS i.e. IoU < 0.61.

4.6 Discussion

The NCL-SM annotation quality results presented in Section 4.5 reflects its usefulness as a resource that addresses the lack of SM tissue datasets that are i) precisely segmented as

evident in Table 4.3, ii) curated as evident in Tables 4.4, 4.5, and iii) large enough as detailed in Table 4.2 for training DL models.

4.6.1 NCL-SM utility

The contribution of NCL-SM as a resource can be described as follows:

Defines the process of SM tissue segmentation and curation – to best of my knowledge there is no literature that defines the detailed process of preparing multiplex SM tissue data for single-cell analysis, as it has been done here with NCL-SM.

Evaluation metrics – As part of building NCL-SM evaluation metrics were introduced that allow users to uncover and inspect the nuance quality requirements for SM tissue image analysis as described in Section 4.3.2.

Large precisely segmented dataset – NCL-SM is fully manually segmented and mostly manually curated, allowing it to be used as a benchmark dataset of SM tissue. It is large, i.e. with >50k myofibres, making it suitable to train many DL models.

Publicly available – NCL-SM in its entirety with evaluation code is publicly available for open science. This will help development of a wide variety of SM tissue related applications and research.

4.6.2 Limitations

Whilst, as describe earlier, the NCL-SM is a useful resource it does have some limitations.

FAM IAV as observed in Table 4.4 there is noticeable inter annotator variation in classifying FAMs. The specificity which reveals more about misclassification of non-FAM, is around 99%. But the sensitivity which reveals more about misclassification of FAM, is between 79% and 87%. The relatively poor sensitivity is contributing toward reduction of respective F1 scores. This highlights that there is noticeable disagreement between domain experts to classifying FAMs, which in turn is due to ambiguity or subjectivity in detecting FAMs in edge cases were freezing damage is not obvious. This should be acknowledged in any tool/pipeline build for SM tissue image segmentation and curation.

FR poor IoU as observed in Table 4.5 the mean IoU for FR calculated across various SM tissue sections is observed to be 0.36, which is quite poor to be used as a benchmark. This highlights the difficulty in precisely segmenting the FRs in a TS, while it was

observed that presence of folding is much easier to observe,precisely segmenting the region affected by folding is very difficult and involves subjective intuition and so is reflected in poor IoU in duplicate annotation. This should be addressed appropriately by any tool developed using NCL-SM.

Low FR count folding is an undesirable artifact that biomedical scientists strive to avoid in tissue preparation. In our dataset there are 405 FR which is a lower count for training DL models.

4.6.3 Conclusion

In this chapter NCL-SM dataset is introduced and described in detail, i.e. defining the protocols for SM tissue segmentation and curation for single cell analysis, defining evaluation metrics that reveal the quality of NCL-SM. These evaluation metrics revealed that for three out four annotation tasks, i.e. excluding FR segmentation, the annotation quality is good as reflected by minimal IAV. The chapter also described the limitations of this useful resource.

Chapter 5

myocytoML: An Automatic Segmentation Pipeline for Muscle Fibres

5.1 Introduction

As discussed in Chapter 4, precise segmentation and curation of myofibres in images of SM tissue cross-sections is a non-trivial and important part of many downstream analyses to understand diseases affecting muscles. Presently segmentation and curation involve either using imprecise annotations or significant manual interventions that introduces subjectivity, both of these in turn affect reliability of the analysis.

The pace of development of ML/DL models to solve various computer vision (CV) tasks is rapid [87, 92]. The same applies to the development of DL models for CV tasks in a biomedical domain [89, 120]. These DL models address tasks of segmentation, object detection and classification, individually or simultaneously i.e. panoptic segmentation. As described in Section 4.3.1 the process of preparing SM imaging data for single cell analysis can be divided into four CV tasks, i) myofibre segmentation, ii) Freezing artifact myofibre (FAM) classification, iii) non-transverse sliced myofibre (NTM) classification and iv) folded tissue regions (FR) segmentation. Quality evaluation for each of these tasks, considering the nuance requirements of single cell analysis of SM tissue, required appropriate metrics as described in Section 4.3.2. There exist a number of applications, both domain specific such as quadruple immunofluorescence analyser [160], mitocyto [1], MiCAT [77] and Steinbock [78], and generalised such as cellprofiler [83] , Fiji ImageJ [161], or using vanilla ML models like Cellpose [89] and StarDist [120]. But none of these applications/pipelines are built to address all four CV tasks mentioned earlier and are not optimised to produce the precision required for SM tissue image single-cell analysis. Furthermore, there are no tools/models

to segment FR in SM tissue section images, or to classify FAM and NTM. Based on these shortcomings of previous methods and to address these, it was decided to build a ML pipeline that can address these challenges by leveraging NCL-SM [41] to train appropriate ML/DL models.

5.2 Aims

The aim of this chapter is to build an application or pipeline that addresses the following requirements.

- Precise segmentation and classification of SM tissue images guided by precision achieved by duplicate human to human annotations in NCL-SM.
- The application should address the issue of observed ambiguity in FAM classification and FR segmentation tasks as reported in Section 4.6.2.
- The application/pipeline should provide a graphical user interface (GUI), which should allow the user flexibility to amend masks acknowledging the ambiguity mentioned above.

5.3 Background: myocytoML design decisions

The NCL-SM manual annotation protocols discussed in Section 4.3.1 give insights into different aspects of performing SM tissue image segmentation. Taking inspiration from the manual annotation process, the following design decisions were made as discussed below.

5.3.1 Panoptic model vs separate models

The decision over selection of the appropriate type of ML model was guided by separation of four CV tasks in manual annotation of NCL-SM. It is observed that there exist additive and sometime competitive complexity with addition of each CV task, i.e. segmenting myofibre and folded regions simultaneously or classifying FAMs and NTMs simultaneously is virtually impossible, as a myofibre can be both FAM and NTM, i.e. a non-transverse sliced myofibre that is also damaged by freezing, also FRs consist of myofibres. For these reasons it was decided that each of the four identified CV tasks will be addressed separately by myocytoML.

5.3.2 Order of execution

In the manual annotation process the following order is applied: 1. Segmenting FRs and removing them from the image; 2. Using segmentation tools such as mitocyto the remaining SM tissue image is segmented; 3. FAMs and NTMs are identified and removed; and 4. The remaining segmented myofibres are manually corrected where required, resulting in the final 'analysable' myofibre mask.

Adopting this same order of execution will not be ideal, as this will not afford the flexibility required especially considering ambiguity/subjectivity in FAM classification or FR segmentation. To accommodate the required flexibility the following order of execution of four CV tasks was decided: 1. The myofibre segmentation and FR segmentation are separately executed resulting in myofibres_mask and FR_mask; 2. Using image and/or myofibres_mask NTM and FAM classification is separately executed resulting in NTM_mask & FAM_mask.3. Finally non-zero pixels that exist in FR_mask, NTM_mask and FAM_mask are removed from myofibres_mask, resulting in an 'analysable' myofibres mask.

5.3.3 Quality information

One of the limitations of all existing applications for SM tissue image segmentation is that these do not give any description or evaluation about the quality of annotations they produce. In the absence of such annotation quality information, users have to rely on visual inspection which is subjective. To address this it was decided that myocytoML will accompany annotation quality information in terms of metrics defined in Section 4.3.2. These metrics will be generated by comparing selected/random handfuls of myofibres annotated by the user to the myocytoML output.

5.3.4 Flexibility

As discussed in Section 4.6.2 there is noticeable IAV in FAM classification and little agreement between manual annotators about FR segmentation. This is reflected in the ground truth duplicate annotations, which means the models trained using this ground truth data most likely will have similar issues. To address this it was decided that myocytoML should have a GUI that should allow users to amend masks generated by all four CV tasks with minimal effort, moreover it should automatically update the final 'analysable' myofibres mask reflecting any amendments made to any of the four CV masks.

5.4 Methods

The following models and methods as appropriate for each of the four CV tasks were selected based on the latest literature.

5.4.1 Methods of myofibre segmentation task

As discussed in Section 5.3, the expected myofibres segmentation mask is an instance segmentation mask. There are a number of ML models and traditional CV methods that can accomplish instance segmentation [21, 82–84, 89, 119, 120, 152, 161–166]. But it has been observed that DL models [89, 120] demonstrate an unparalleled performance compared to any traditional CV methods. The approaches to solve instance segmentation for biomedical CV tasks by ML/DL models can be categorised into three broad categories: 1. pixel-classification based approaches, where supervised ML models are trained to predict the class of each pixel in an image such as belonging to a cell vs background, then post processing techniques such as contour detection are applied to create an instance segmentation mask. There is a spectrum of ML models such as tree-based XGB [99], DL based ResNet [118], and VGG [116], that can be employed for pixel-based classification. But these models have limitations especially segmenting densely packed images i.e. where objects to segment share borders this approach performs poorly [167]; 2. Feature-based segmentation, where DL models are designed such that they can leverage an object (e.g. cell) features to perform more precise segmentation. This can be achieved by encoder-decoder based models such as UNET [21] as discussed in Section 2.10.3 and RPN based models such as Mask R-CNN as discussed in Section 2.10.3; 3. Distance-map based segmentation approaches are more popular for single cell segmentation where models predict distance [89, 120] for each pixel to/from centre and/or border of the object. This results in a distance or flow map which is processed to produce precise single-cell instance segmentation masks [167]. Distance-map based models such as Cellpose [89] and StarDist [120] as discussed in Section 2.10.3 are built upon feature-based models i.e. the backbone for such models is either UNETs or mask R-CNNs. Cellpose and StarDist performed comparative analysis of segmentation quality with feature based models (UNETs & mask R-CNN) and concluded that the distance-map based models outperform feature-based models [89, 120].

Windhager *et al.* [78] has built the most relevant pipeline to our problem, it is a comprehensive pipeline/workflow for multiplex tissue image single cell end-to-end analysis, a part of this pipeline deals with single-cell segmentation in which they provide options to use either pixel-based classification models, feature-based or distance-map based models.

But as discussed in Section 5.1 this pipeline and tools do not acknowledge or address the four CV tasks that are specific to SM tissue image analysis as discussed earlier. Moreover, these vanilla models; i.e. without customised retraining models with NCL-SM, were found to perform poorly as observed in Table 5.3, i.e. masks produced using these models do not make it to the top five (or indeed the top 20) best performing models.

Based on the factors discussed above it was decided that distance-map based models, Cellpose and StarDist should be used. These models with various combinations of transfer learning (pre-training weights), training data (IMC and/or IF) and optimisation of hyperparameters were experimented with.This is discussed in more detail in Section 5.5.1

5.4.2 Methods of FAM classification task

In the manual annotation process FAM are classified based on damage caused by freezing as described in Section 4.3.1. An ideal method would detect this feature, i.e. freezing damage and classify the myofibre as FAM. A range of traditional CV based approaches were experimented with such as pixel intensity distribution i.e. a freezing damage (whole or partial myofibre damage) should reflect as abnormal peaks at zero in intensity distributions for FAM. But the classification results were poor, it was observed that for various patient groups' myofibres there exist zero intensity peaks in their distributions that are not linked with freezing damage but may be a phenotype of the disease. It was then decided to use ML for FAM classification. There are a number of classification ML models [109, 118, 168–170], as discussed in Section 2.10.3; some of these are DL models such as vanilla CNN, ResNet and VGG and others are hybrid models constructed by fusing a DL model for feature extraction and a tree-based model such as XGB trained on extracted features to predict class. It was decided to experiment with CNN, VGG16, ResNet50, hybrid CNN-XGB, VGG16-XGB and ResNet50-XGB and select the model with the best FAM classification performance.

5.4.3 Methods of NTM classification task

As discussed in Section 4.3.1 there already exists a morphological based function as defined in algorithm 1 that classifies NTMs with perfect precision. Based on this it was decided that the same NTM function should be used to classify NTMs in myocytoML.

5.4.4 Methods of FR segmentation task

The characteristics of folded regions segmentation are distinct to myofibre segmentation in that FR are groups of myofibres folded on themselves. This means that these are larger in

area, have a higher intensity for the mass marker and usually have a mesh membrane pattern as describe in Figure 4.2. This means both pixel-based classification models and feature based models are good candidates for the FR segmentation task.

But it should be acknowledged that annotator agreement in terms of IoU metric in the ground truth data is found to be poor i.e. 0.36 across 19 TS images. This means it is unlikely that any model can precisely segment FR.

Given these factors it was decided feature based models, i.e. UNETs and Mask R-CNN, shall be used for experiments.

5.5 Experiments and results

5.5.1 Myofibre segmentation task

Experiment data for myofibre segmentation

The training and test data is taken from NCL-SM which consists of two sub datasets for IMC and IF images. All ML models were trained using images built by arranging grayscale images of a cell membrane protein (Dystrophin) marker and mitochondrial mass protein (VDAC1) marker into an RGB image where R = membrane protein marker, G = mass protein marker and B = 0. As discussed earlier, these two markers, i.e. Dystrophin and VDAC1, are good markers/identifiers of myofibre membrane and its mass respectively. However, single channel, i.e. only membrane marker and only mass marker, were also experimented with, but these gave poor results. As described in Table 5.1 one or two TS images were withheld from IMC and IF datasets for testing the trained models. The DL models used here required the training images to be of uniform size, for this all training images were split into patches. It was decided that three different patch sizes as described in Table 5.1 would be experimented with, making sure that the patch sizes is at least twice the average diameter of myofibre.

Experiment ML model parameters for myofibre segmentation

As discussed in Section 5.4.1 distance-map based models StarDist and Cellpose were selected to experiment for myofibre segmentation. These models were trained with various combinations of datasets, patch sizes, initialised weights and model parameters as described in Table 5.2. The data was split into 70%: 15%: 15% for training, validation and testing. Experiments were conducted on a machine with the following specification: GPU: 4 x NVIDIA Tesla V100 (16GB), CPU: 24 cores, RAM:448GB.

Table 5.1 myocytoML training data details.

Imaging Technique	Channels	Tissue sections		Average myofibre diameter (pixels)	Training patch sizes
		Test	Training		
IMC	R= membrane marker; G= mass marker ; B= zero	2 (P02 & P06)	25 (all except P02 & P06)	65	256x256; 512x512; 1024x1024
IF	R= membrane marker; G= mass marker ; B= zero	1 (S48)	18 (all except S48)	200	512x512; 1024x1024; 2048x2048

Table 5.2 Details of myofibre segmentation models experiments. 'Datasets & patch sizes' column describes the image type, i.e. IMC or IF and various patch sizes that were experimented with. The column 'Initialised weights' describes the names of pre-trained weights that were used for transfer learning, 'From scratch' means random initialised weights were used. The 'Parameters' column details all parameters relevant to the model being trained, that were being experimented with, for each parameter the values after the colon denotes the options that were experimented with. In the table diam_mean, mae, mse stands for mean diameter of the myofibre, mean absolute error, mean squared error respectively.

Model	Datasets & patch sizes	Initialised weights	Parameters
StarDist	1) IMC & (256x256; 512x512; 1024x1204); 2) IF & (512x512; 1024x1204; 2048x2048); 3) IMC+ resized_IF to 1/3 size & (256x256; 512x512; 1024x1204)	1) From scratch; 2) '2D_versatile_he'	1) Backbone: UNET; 2) n_ray: 32, 64, 128; 3) grid: (2,2) , (4,4), (8,8); 4) train_learning_rate: 1e-5, 1e-4; 5) train_batch_size: 16, 32; 6) training_metric: dist_iou_metric, mae, mse; 7) epochs >2000 & patience 200 epochs
Cellpose	1) IMC & (256x256; 512x512; 1024x1204); 2) IF & (512x512; 1024x1204; 2048x2048); 3) IMC+ resized_IF to 1/3 size & (256x256; 512x512; 1024x1204)	1) From scratch; 2) 'Tissue_Net'; 3) 'cyto'; 4) 'cyto2'	1) Backbone: UNET; 2) diam_mean: 65 (IMC & IMC+resized_IF), 200 (IF); 3) learning_rate: 0.1, 0.2; 4) weight_decay =1e-5, 1e-4, 1e-3; 5) train_batch_size: 8,16, 32; 6) training_metric: Cellpose_loss; 7) epochs >2000 & patience 200 epochs

Results for myofibre segmentation

The results for top performing models are presented in Table 5.3. The list does include top performing StarDist models that are not in the top 10 but are presented for comparison. As can be observed in the table, the best performing model for IMC images was a Cellpose model that was trained with random initialised weights i.e. from scratch trained on patch size of 512x512 pixels. This model has performed the best across five out of six evaluation metrics i.e. excluding the metrics 'A%(IoU>0.95)' which defines percentage of myofibres with >0.95 IoU. It is also seen that this model has considerably outperformed the existing workflow of using mitocyto with manual correction (myocytoML vs mitocyto+ r_{AoB} : 0.95 vs 0.90; r_{AiB} : 0.78 vs -.015; $\overline{IoU}$: 0.94 vs 0.91; A%(IoU >0.80): 98.3 vs 95.24; A%(IoU >0.90):92.2 vs 74; A%(IoU >0.95): 47.35 vs 59.4). This same model performance when compared to "gold standard" manual annotations is comparable but slightly lower in evaluation metrics as seen in the table.
The best performing model for IF images was again a Cellpose model that was trained with random initialised weights on training data that included IMC images and resized IF images with patch size of 512x512 pixels. This model has performed the best across four out of six evaluation metrics, i.e. excluding the metrics r_{AiB} & 'A%(IoU>0.80)' . It is also seen that this model has comparable albeit slightly lower performance than compared to the "gold standard" manual annotations i.e. myocytoML vs benchmark r_{AoB} : 0.92 vs 0.92; r_{AiB} : 0.87 vs 0.94; $\overline{IoU}$: 0.92 vs 0.96; A%(IoU >0.80): 95.41 vs 100; A%(IoU >0.90):89.66 vs 100; A%(IoU >0.95): 49.54 vs 74.
Cellpose models outperformed all other models and tools across both the image types, which suggests the flow map based approach is particularly appropriate for myofibre segmentation. In addition to results reported in Table 5.3 other tools and models including vanilla StarDist, Cellpose models, Ilastik where experimented with but all of these achieved $\overline{IoU} < 0.85$.

5.5.2 FAM classification task

Experiment data

The training and test data for FAM classification models is as describe in Table 5.1 except for splitting the images into patches. Instead, for training FAM models the RGB images were split into individual myofibre images using NCL-SM segmentation masks. These were then separated into two classes 'FAM' and 'non_FAM' using the 'FAM' and 'all_myofibres' masks from NCL-SM. This resulted in >20k (IMC) & >25k (IF) myofibres for training. But there exists a substantial class imbalance as there are no more than 700 'FAM' class myofibres in each IMC and IF sub-datasets. For this reason it was was decided that a class

Table 5.3 Results of myofibre segmentation models experiments. Please note the benchmark here differ slightly as these are calculated for only test TS images as opposed to Table 4.3 which was accessed across more TS images. The green cells are top metrics score across column per I_type.

I_type	Top 3	Metrics						Model parameters
		r_{AoB}	r_{AiB}	$\overline{IoU}$	A%(IoU >0.80)	A%(IoU >0.90)	A%(IoU >0.95)	
	Benchmark	0.95	0.77	0.95	100	100	59.4	NA
	mitocyto+	0.90	-0.15	0.91	95.24	74	11.3	NA
IMC	CP_IMC_Scratch	0.95	0.78	0.94	98.3	92.2	47.35	Model: Cellpose; 1) training dataset: IMC; 2) initialised weight: random; 3) patch size: 512x512; 4) learning rate & weight decay: 0.1 1e-4; 5) diam_mean: 65; 6) batch size: 8
	CP_IMC_cyto2	0.95	0.77	0.94	97.9	89.77	48.44	Model: Cellpose; 1) training dataset: IMC; 2) pre-trained weights: 'cyto2'; 3) patch size: 512x512; 4) learning rate & weight decay: 0.1 1e-4; 5) diam_mean: 65; 6) batch size: 8
	CP_IMC_TN	0.95	0.77	0.93	97.9	89.86	49.25	Model: Cellpose; 1) training dataset: IMC; 2) pre-trained weights: 'Tissue_Net'; 3) patch size: 512x512; 4) learning rate & weight decay: 0.1 1e-4; 5) diam_mean: 65; 6) batch size: 8
	SD_IMC_Scratch	0.86	0.65	0.85	80.47	57.63	11.58	Model: StarDist 1) training dataset: IMC; 2) initialised weights: random; 3) patch size: 512x512; 2) n_ray: 64 ; 3) grid: (2,2) ; 4) train_learning_rate: 1e-4; 5) train_batch_size: 16; 6) training_metric: dist_iou_metric
	Benchmark	0.92	0.94	0.96	100	100	74	NA
IF	CP_Mix_Scratch	0.92	0.87	0.92	95.41	89.66	49.54	Model: Cellpose; 1) training dataset: IMC+ resized_IF; 2) initialised weights: random; 3) patch size: 512x512; 4) learning rate & weight decay: 0.1 1e-4; 5) diam_mean: 65; 6) batch size: 8
	CP_IF_Scratch	0.90	0.92	0.92	98.01	88.65	18.25	Model: Cellpose; 1) training dataset: IF; 2) initialised weights: random; 3) patch size: 1024x1024; 4) learning rate & weight decay: 0.1 1e-4; 5) diam_mean: 200; 6) batch size: 8
	CP_IMC_TN	0.90	0.83	0.90	94.02	71.16	10.75	Model: Cellpose; 1) training dataset: IMC; 2) pre-trained weights: 'Tissue_Net'; 3) patch size: 512x512; 4) learning rate & weight decay: 0.1 1e-4; 5) diam_mean: 65; 6) batch size: 8
	SD_Mix_Scratch	0.76	0.61	0.73	66.32	32.1	0	Model: StarDist 1) training dataset: IIMC+ resized_IF; 2) initialised weights: random; 3) patch size: 512x512; 2) n_ray: 64 ; 3) grid: (2,2) ; 4) train_learning_rate: 1e-4; 5) train_batch_size: 16; 6) training_metric: dist_iou_metric

Table 5.4 Details of FAM models experiments.'Datasets' column describes the image type i.e. IMC or IF . The column 'Initialised weights' describes the names of pre-trained weights that were used for transfer learning, 'None' means random initialised weights were used. The 'Parameters' column details all parameters relevant to the model being trained. For each parameter the values after the colon denote the options that were experimented with. In the table CNN-XGB, VGG16-XGB, ResNet50-XGB, SGD, RMSprop stands for hybrid CNN-XGB model, hybrid VGG16-XGB model, hybrid ResNet50-XGB model, stochastic gradient descent, root mean square propagation respectively. The architecture of the CNN model used is described in the Table 6.2

Model	Datasets	Initialised weights	Parameters
CNN	IMC; IF	None	1) Image resized: 128x128, 224x224; 2)optimizer: 'Adam', 'SGD', 'RMSprop'3) learning rate: 1e-3,1e-4 4) Training metrics: sensitivity, specificity, F1 score, Accuracy; 5) epochs >1000 with patience = 100
VGG16	IMC; IF	'ImageNet'	1) Image resized: 224x224; 2) optimizer: 'Adam', 'SGD', 'RMSprop'3) learning rate: 1e-3,1e-4 4) Training metrics: sensitivity, specificity, F1 score , Accuracy; 5) epochs >1000 with patience = 100
ResNet50	IMC; IF	'ImageNet'	same as for VGG16
CNN-XGB	IMC; IF	None	1) Image resized: 128x128, 224x224; 2) optimizer: 'Adam', 'SGD', 'RMSprop'3) learning rate: 1e-3,1e-4 4) Training metrics: sensitivity, specificity, F1 score, Accuracy; 5) epochs >1000 with patience = 100; 6) XGB ("eta": [0.05, 0.10, 0.15, 0.20, 0.25, 0.30], "max_depth": [3, 4, 5, 6, 8, 10, 12, 15], "min_child_weight": [1, 3, 5, 7], "gamma": [0.0, 0.1, 0.2 , 0.3, 0.4],"colsample_bytree" : [0.3, 0.4, 0.5 , 0.7])
VGG16-XGB	IMC; IF	'ImageNet'	1) Image resized: 224x224; 2) optimizer: 'Adam', 'SGD', 'RMSprop'3) learning rate: 1e-3,1e-4 4) Training metrics: sensitivity, specificity, F1 score, Accuracy; 5) epochs >1000 with patience = 100; 6) XGB :same as above
ResNet50-XGB	IMC; IF	'ImageNet'	same as for VGG16-XGB

balanced training data of around 1300 myofibres for IMC and 1200 for IF should be created by selecting all available 'FAM' class myofibres and randomly selecting a similar count of 'non_FAM' class myofibres from available 'non_FAM' class myofibres.

Experiment ML model parameters for FAM classification

As discussed in Section 5.4.2 the experiments were conducted by training the six pixel-based classification models. The data was split into 70%: 15%: 15% for training, validation and testing. The details of training are presented in Table 5.4. Experiments were conducted on a machine with the following specification: GPU: 1 x NVIDIA Tesla V100 (16GB), CPU: 6 cores, RAM:112GB.

Results for FAM classification

The results for top performing FAM classification models are presented in Table 5.5. As seen in the table the results are compared with two sets of benchmark metrics, this is because the benchmark consists of three sets of FR annotation, i.e. one "ground truth" and two duplicate annotations, this resulted in two benchmarks. These two benchmarks were arranged in descending order, i.e. higher followed by lower. The top three models that produced the best performance across the four evaluation metrics i.e. FAM_sensitivity, FAM_specificity, F1 score and accuracy are listed in the table for both the image types.
The best performing model in both datasets, i.e. IMC and IF, was a ResNet50 model that was trained on 'ImageNet' pre-trained weights, with patch size of 224x224 pixels. For IMC images the ResNet50 model is closer to the lower benchmark metrics but this is not the case for IF images where all models underperform compared to the benchmark.

5.5.3 FR segmentation task

Experiment data

The training and test data for FR segmentation models is as describe in Table 5.1 except for sizes of split image patches, i.e. FR are usually larger than myofibre as folding usually affects multiple myofibres, for this reason patch sizes of 1024x1024, 2048x2048 and 4096x4096 pixels were selected instead.

Table 5.5 Results FAM classification models experiments. In the table I_type, RN_IMC, RN_XGB_IMC, VGG_IMC, SENS, SPEC stands for image type, ResNet50 model trained on IMC, hybrid ResNet50-XGB model trained on IMC, VGG model trained on IMC, sensitivity, specificity respectively .*Note the readings for FAM classification evaluation in Chapter 4 Table 4.4 will differ as these were calculated across all TS images, whereas the results presented here are only for test TS images as the remaining images were used for training. The green cells are top metrics score across column per I_type.

I_type	Top 3	Metrics				Model parameters
		SENS	SPEC	F1	Accuracy	
	Benchmark (high)	0.87	0.99	0.92	0.97	NA
	Benchmark (low)	0.80	0.98	0.84	0.95	NA
IMC	RN_IMC	0.83	0.91	0.74	0.89	Model: ResNet50; 1) training dataset: IMC; 2) pre-trained weight: 'ImageNet'; 3) image resize: 224x224; 4) learning rate: 1e-3; 5) training metric: accuracy
	RN_XGB_IMC	0.78	0.90	0.70	0.87	Model: ResNet50+XGB; 1) training dataset: IMC; 2) pre-trained weight: 'ImageNet'; 3) image resize: 224x224; 4) learning rate: 1e-3; 5) training metric: accuracy; 6) XGB: (eta=0.3,gamma=0,max_depth=6)
	VGG_IMC	0.5	0.94	0.57	0.86	Model: VGG16; 1) training dataset: IMC; 2) pre-trained weight: 'ImageNet'; 3) image resize: 224x224; 4) learning rate: 1e-3; 5) training metric: accuracy
	Benchmark (high)	0.87	0.99	0.82	0.99	NA
	Benchmark (low)	0.64	0.98	0.63	0.97	NA
IF	RN_IF	0.83	0.62	0.3	0.64	Model: ResNet50; 1) training dataset: IMC; 2) pre-trained weight: 'ImageNet'; 3) image resize: 224x224; 4) learning rate: 1e-3; 5) training metric: accuracy
	VGG_IF	0.8	0.6	0.28	0.62	Model: VGG16; 1) training dataset: IMC; 2) pre-trained weight: 'ImageNet'; 3) image resize: 224x224; 4) learning rate: 1e-3; 5) training metric: accuracy
	RN_XGB_IF	0.88	0.55	0.27	0.57	Model: ResNet50+XGB; 1) training dataset: IMC; 2) pre-trained weight: 'ImageNet'; 3) image resize: 224x224; 4) learning rate: 1e-3; 5) training metric: accuracy; 6) XGB: (eta=0.3,gamma=0,max_depth=8)

Table 5.6 Details of FR segmentation models experiments.'Datasets % patch sizes' column describes the image type, i.e. IMC or IF and split image patch sizes. The column 'Initialised weights' describes the names of pre-trained weights that were used for transfer learning, 'From scratch' means random initialised weights were used.'Parameters' column details all parameters relevant to the model being trained. For each parameter the values after the colon denote the options that were experimented with. In the table IMC+ resized_IF to 1/3 size refers to a training dataset where IMC images were split and combined with IF images that were first resized to 1/3 and then split.

Models	Datasets & patch sizes	Initialised weights	Parameters
UNET [21], UNET++[165], R2_UNET_2D [171], Attention_UNET [172], UNET_3plus_2d [164], Mask_RCNN [119]	1) IMC & (512x512; 1024x1204; 2048x2048); 2) IF & (1024x1204; 2048x2048;4096x4096); 3) IMC+ resized_IF to 1/3 size & (512x512; 1024x1204; 2048x2048)	1) From scratch; 2) 'ImageNet'	1) Backbone: 'VGG16'; 2) learning rate: 1e-3,1e-4; 3)Training metrics: IoU, 'losses.dice_coef'; 4) epochs >2000 with patience = 200

Experiment ML model parameters for FR segmentation

As discussed in Section 5.4.4 for FR segmentation it was decided that feature-based models are selected for experiments. The selected models are presented in Table 5.6. The data was split into 70%: 15%: 15% for training, validation and testing.

Results for FR segmentation

The results for 'top' performing FR segmentation models are presented in Table 5.7. As seen in the table, UNET_2D model performance evaluated using image wide IoU of folded regions exceeded the benchmark metric. But this should be seen in the context of the poor IoU in the duplicate manually annotated "ground truth" data. Although the UNET_2D model outperformed the benchmark metric but IoU of 0.21, 0.64 for P02, P06 respectively is poor by normal standards.

Table 5.7 Results of FR segmentation models experiments. *Note the readings for FR segmentation evaluation in Chapter 4 Table 4.5 will differ as these were calculated across many TS images where folding exists, whereas the results presented here are only for test TS images as the remaining images were used for training. The golden row is benchmark annotation score.

Model	IoU		Parameters
	P02	P06	
Benchmark	0.16	0.61	NA
UNET_2d	0.21	0.64	1) Backbone: 'VGG16'; 2) patch size: 1024x1024; 3) learning rate: 1e-3; 4) optimiser: 'Adam'; 5) training metric: 'losses.dice_coef'
UNET+	0.11	0.54	1) Backbone: 'VGG16'; 2) patch size: 1024x1024; 3) learning rate: 1e-3; 4) optimiser: 'Adam'; 5) training metric: 'losses.dice_coef'
UNET_3plus_2d	0.04	0.66	1) Backbone: 'VGG16'; 2) patch size: 1024x1024; 3) learning rate: 1e-3; 4) optimiser: 'Adam'; 5) training metric: 'losses.dice_coef'

5.6 myocytoML

The design considerations for myocytoML are based on the discussion in Section 5.3. For each of the four CV tasks discussed in Section 5.4 the respective top performing model presented in Tables 5.3, 5.5, 5.7 in the Section 5.5 is selected.

5.6.1 myocytoML architecture

myocytoML[1] is a Python application and the architecture of it is described in Figure 5.1.

5.6.2 myocytoML graphical user interface

As can be observed in Section 5.5, for most of the CV tasks required for SM tissue segmentation and curation, myocytoML precision is close to the benchmark metrics measured between duplicate annotations by humans. But it can also be observed as discussed in Section 4.6 the 'ground truth' annotations are not perfect which is reflected in the benchmark metrics. This means the segmentation and classification masks produced by myocytoML will not be perfect. And the evaluation metrics produce by myocytoML by comparing the myocytoML masks with random manual annotations by the user will give a quantitative measure of quality of segmentation, which can be helpful in identifying the mistakes in masks that need correction.

[1] https://github.com/atifkhanncl/mitoML_segmentation_pipeline

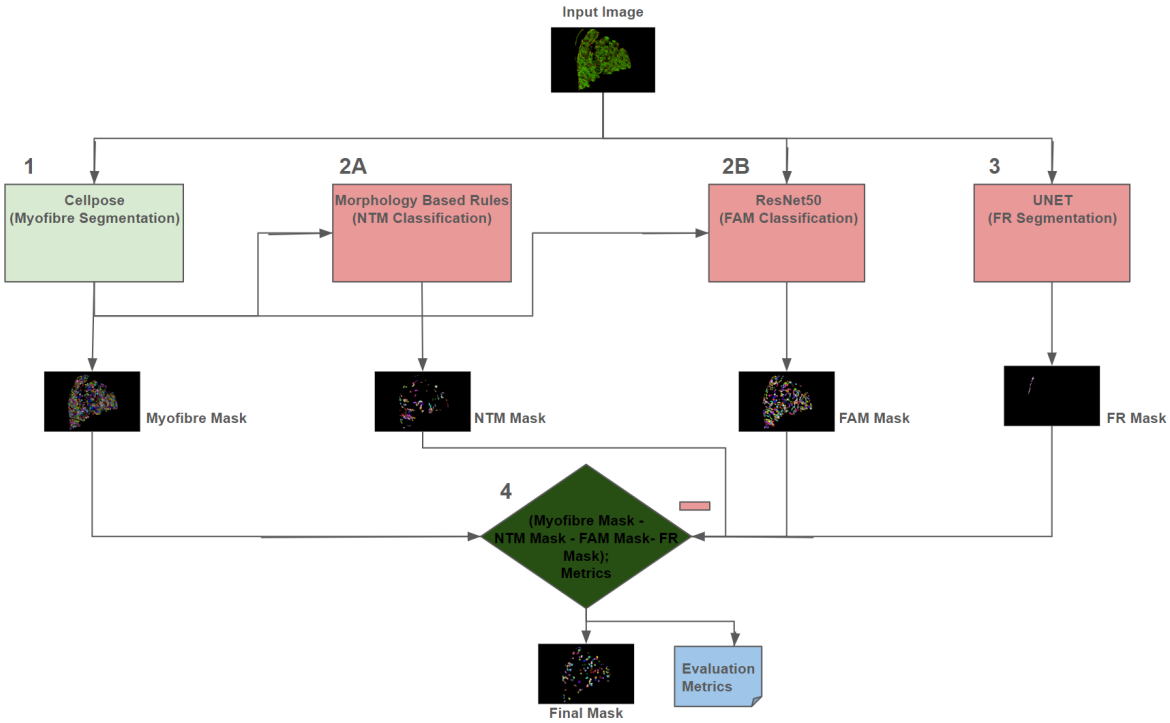

Figure 5.1 myocytoML Design. **Input Image**: SM image made by arranging greyscale images of a cell membrane protein marker and myofibre mass protein/any cytoplasm marker into an RGB image where R = membrane protein marker, G = myofibre mass marker and B = 0; **1**: Custom trained Cellpose model predicts 'Myofibre Mask' for input image; **2A**: Each myofibre in 'Myofibre Mask' is classified as NTM/non-NTM based on morphological features, producing 'NTM Mask'; **2B**: Using 'Myofibre Mask' each fibre in the input image is segmented into individual myofibre images, each of these is fed to a custom trained ResNet50 model that classify myofibres as either FAM/non-FAM, producing a 'FAM Mask' ; **3**: Input image is fed to a trained UNET model that predicts 'FR Mask'; **4**: Final instance segmentation mask of 'Analysable' myofibres is made by removing 'NTM Mask', 'FAM Mask' and 'FR Mask' from 'Myofibre Mask'. In this step quality evaluation metrics (as discussed in Section 4.3.2) are also computed if a 'ground truth' manual annotation mask is provided.

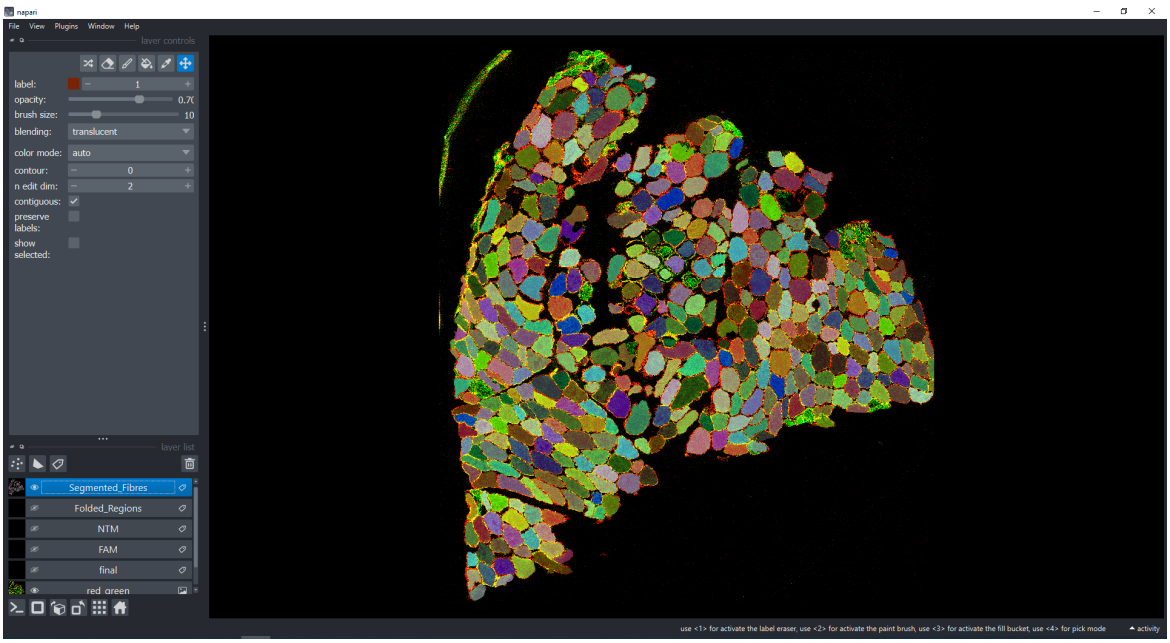

Figure 5.2 myocytoML allows users to review and amend each mask with minimal effort. Segmentation masks can be amended using erase, brush, pencil and color fill tools. Classification masks can be amended by double clicking on myofibre to change its binary class.

To make myocytoML a useful tool, the mask correction process needs to be simple and quick. Taking these points into consideration a GUI for myocytoML review was built on top of a image viewer package Napari [86] using its widgets and scripts functionality, this allows users to view and amend the myocytoML output masks with ease, as describe in Figure 5.2.

5.6.3 myocytoML standard operating procedure (SOP)

The following SOP of myocytoML is envisioned to be efficient.

Inputs The user provides (i) images of cell membrane marker and myofibre mass marker for each SM TS; ii) Manual segmentation masks where a selected few myofibres are segmented for QA, this can be done using Napari.

Output myocytoML produces (i) masks: raw 'Myofibre Mask', 'NTM Mask', 'FAM Mask', 'FR Mask'; ii) csv files: 1) include all morphological detail of each myofibre and 2) evaluation metrics measured by comparing myocytoML masks with manual annotated masks.

Review Using myocytoML review on Napari users can review and amend each mask by i) for segmentation mask: amend using erase, brush, pencil and color fill tool; ii) for

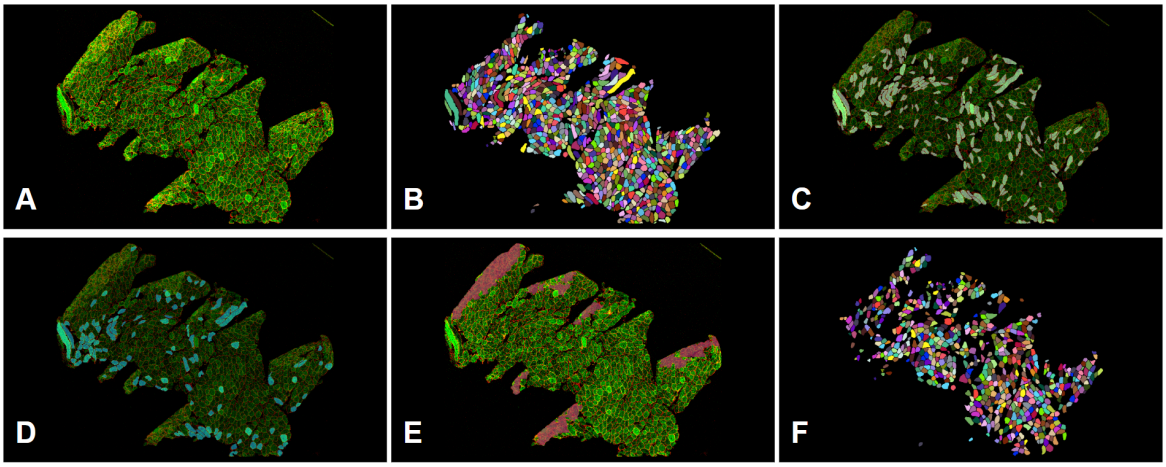

Figure 5.3 myocytoML SOP. End-to-end SM tissue section (TS) image annotation process for SM TS 'P06'. **A**: User provides two greyscale SM tissue images, one a cell membrane protein (Dystrophin) marker and other a myofibre mass protein (VDAC1) marker. myocytoML arranges grayscale images into an RGB image where R = membrane protein marker, G = mass protein marker and B = 0; **B**: myocytoML myofibre instance segmentation mask for A; **C**: myocytoML FAM mask for A, overlaid on the image; **D**: myocytoML NTM mask for A, overlaid on the image; **E**: myocytoML FR mask for A, overlaid on the image; **F**: Final instance segmentation mask of 'Analysable' myofibres made by removing objects identified in C, D and E from B.

classification masks: amend by double clicking on myofibre to change its binary class i.e. from NTM to non-NTM or from FAM to non-FAM and vice versa.

5.7 Discussion

Myofibre segmentation : if done manually then myofibre segmentation is the most time consuming of the four CV task involved in SM image segmentation and curation. As observed in Table 5.3 myocytoML exceeds in precision when compared to the current process, i.e. using mitocyto plus manually correcting masks. It is also observed that its precision measured in terms of evaluation metrics are close to the benchmark duplicate human annotations. It was observed that Cellpose outperformed StarDist for my use case i.e. precise segmentation of SM images.

NTM classification : NTM was the simplest of the four CV tasks to solve with the development of algorithm 1.

FAM segmentation : As observed FAM classification performance is closer to the lower bound of the benchmark for IMC images as observed in Table 5.5. But for IF images the

results were poorer, mostly due to poor specificity by the models. This can be due to number of factors but mostly due to subjectivity in FAM annotation. It was decided to acknowledge that there will be subjectivity in this CV task since biomedical scientists need to balance between full utilisation of damaged tissue and avoiding any adverse affect on analysis. For this reason it was decided that a good enough FAM model with an easy process to correct misclassification would be the most useful solution.

FR segmentation has the poorest IAV as discussed in Section 4.6.2 which limits the segmentation accuracy that can be achieved using this training data. As observed in Table 5.7 the IoU score achieved using myocytoML exceeds the benchmark metrics. But it is clear the IoU scores of 0.21 and 0.61 are poor by normal standards. By observing prediction of the FR models it was noted that the model predictions were usually correct about large folding areas. For this reason in the myocytoML, it was decided that only large FR (>1500 pixels diameters) should be retained from the model segmentation mask. This processed mask with an easy process to correct the segmentation mask was deemed to be a suitable solution.

5.7.1 myocytoML utility

Studying the pathology of mitochondrial diseases or other disorders affecting SM usually requires analysis of SM tissue images. For a reliable single-cell analysis of the SM tissue images, the segmentation and curation of these SM cells (myofibres) need to be precise. myocytoML is an end-to-end SM tissue image segmentation and curation pipeline that produces segmentation and curated masks that are close to the ones achieved by domain expert human annotators. myocytoML is also flexible to accommodate the subjectivity that exists in SM image segmentation/curation with a simple mask amendment process.

Biomedical scientists at WCMR have recently started using myocytoML replacing other tools and methods.

5.7.2 Limitations

myocytoML is a useful tool for SM tissue image segmentation as discussed earlier but it has a few limitations.

- As it is a prototype it is not optimised, i.e. the time to process a large IMC image can be up to 45 minutes. This is because the pipeline processes different CV tasks sequentially. By optimising the code and employing multi threading this can be improved.

- Its installation can be simplified by packaging it as a Napari plugin. This is a feature that can enhance the usability of myocytoML.
- While flexibility in amending masks is useful, the amendments can be used for continuous learning i.e. retraining models to learn from these corrections. But this feature is not implemented in the current version of myocytoML. This might improve the annotation quality by minimising the subjectivity over time as models improve by learning from corrections.

5.7.3 Conclusion

In this chapter I describe the process of building myocytoML: a SM tissue image segmentation and curation pipeline that is flexible to address subjectivity involved in this process. The chapter described and discussed the design decision of myocytoML, including the choice of models for the four CV tasks and order of execution of these CV tasks. The results presented in this chapter show that myocytoML meets almost all the benchmark metrics for IMC images and to a lesser extent IF images. The chapter also describes myocytoML GUI built on top of Napari and the SOP of using myocytoML for multiplex image segmentation and curation of myofibres in SM TS images.The utility and limitations of myocytoML are also discussed.

Chapter 6

Explainable DL Analysis to Classify Mitochondrial Disease in Myofibre and SM Tissue

6.1 Introduction

As discussed in Chapters 2 and 3 there exist a number of techniques such as plotIMC, Cytomapper, imcRTools [76–80, 82, 173–175] for analysis of multiplex protein data (IMC). These techniques usually employ single cell segmentation, followed by applying processing techniques to the multiplex data with an aim to i) extract statistical summaries, or features or ii) reduce dimensions, this is followed by comparative, neighbourhood, clustering, cell-to-cell interaction analysis, and visualisation.

These existing approaches are essentially attempting to resolve the curse of dimensionality of multiplex data by reducing the dimensions (usually spatially in the form of statistical summaries or features, sometimes channel-wise, i.e. reducing by combining multiple channels). This effectively leads to ignoring features in the dimensions where reduction is applied. This is more of an issue in the analysis of SM tissue of mitochondrial disease patients, where some of the associations between various OXPHOS proteins (captured in channels) and theories proposing existence of differential features within cells (captured in pixels), might be lost through the reduced dimensions. To explain this further let us take two cases: 1) applying dimensionality reduction channel-wise will result in merging of channels' signals which will limit observation of actual delineated importance/association of each channel involved in the analysis; 2) Applying dimensionality reduction spatially, e.g. by using per myofibre mean pixel intensities that ignore intra-myofibre features, and introduce limitations on performing

analyses to test certain hypotheses proposing the existence of differential features within myofibres, such as the perinuclear niche hypothesis by Vincent *et al.* [2] that propose the existence of differential features near nuclei of myofibres with primary or secondary mtDNA mutations.
Explainable DL, i.e. a combination of high performing DL classification models [108–112] that are setting new records for prediction accuracy [113] and state-of-the-art explainable DL methods [19, 127, 128, 134, 136, 137, 143] that help explain the basis of model predictions, can be an approach to classify myofibres and SM tissue sections and profile these in terms of their spatial and channel-wise features. To the best of my knowledge there are no such studies for multiplex SM TS image analysis and more broadly for any multiplex biomedical data.

6.2 Aims

To overcome the limitations of existing approaches where intra-myofibre spacial features are ignored, in this chapter I will analyse (classify) and profile segmented multiplex (IMC) myofibres. The aims of this chapter are as follows:

- Predict the mitochondrial genetic mutations of myofibres using DL and raw multichannel images of segmented myofibres.
- Profile these myofibres based on their mitochondrial genetic mutation using explainable DL methods.
- Extend this analysis to unsegmented raw multichannel images of SM tissue sections.

6.3 Data and methods

6.3.1 Data

To perform both comparative and complementary analysis between the studies performed in this chapter and Chapters 2 and 3, the data, i.e. 13 TS from 13 subjects as described in Table 2.1, imaged using IMC that consist of 12 channels/protein markers as described in Table 2.2, and grouped into five groups as described in Table 2.6 remains same.
As described in Section 2.7.1 a TIFF (16-bit) file corresponding to each of 12 protein markers for every subject was used as raw data.

Table 6.1 Myofibre count for explainable DL analysis

Genetic mutation	TS count	Myofibre count
Controls – healthy control subjects	3	645
ClassA – nDNA encoded mutation (P01 & P02)	2	400
ClassB – Point mutation in (MT-TL1)(P05,P06,P07)	3	2927
ClassC – Point mutation in mito encoded tRNA(P08,P09,P10)	3	1632
ClassD – Single,Large-scale mtDNA mutation (P03 & P04)	2	1753
P03 – Single, large-scale mtDNA mutations	1	1262

Myofibre segmentation and curation

Following myocytoML SOP as described in Section 5.6.3 myofibres in all 13 TS IMC images were segmented and curated. The exception was the TS images that were used for training myocytoML models, in this case these manually annotated and curated myofibres are used instead. This resulted in the myofibre count as detailed in Table 6.1.

Single-myofibre multiplex images

The 12 protein marker images and a binary (0 or 255) 'Analysable' myofibre mask are combined to make a 13 channel array (height x width x 13 channels) for each TS image. Using 'Analysable' myofibre mask individual analysable myofibres images were created. Extra padding of pixels (value=zero) were added to these images to make them all of uniform (200x200x13) shape with myofibre in the center, these dimensions were decided based on the average cross-section diameter of myofibre in IMC images found to be 65 pixels. The inclusion of the myofibre mask as the 13th channel was done to observe the predictive power of myofibre morphology which is represented by this channel. An additional dataset of single-myofibre multiplex data with a different padding resulting in shape (224x224x13) is also made to allow the training of models with 'ImageNet' weights.

6.3.2 Methods

As discussed in Chapter 3, ML models trained on statistical summaries of myofibre had predictive accuracy ranging from 93% for class D myofibres to 100% for class A myofibres. This was followed by application of explainable ML methods to these ML models to extract the ML model's basis of prediction leading to decomposing this into the correlations and association between the various protein markers involved. The associations were then used to create reports of insights and predictive inference. In this chapter the aim is to achieve high

Table 6.2 Simple CNN model architecture

Layer	Output shape
Input	(height,width,channels)
Convolution (CONV2D)	(None,198,198,32)
Max-pooling (MAXPOOL2D)	(None,99,99,32)
CONV2D	(None,97,97,64)
MAXPOOL2D	(None,48,48,64)
Flatten	(None,147456)
Dropout (0.5)	(None,147456)
Dense	(None, 1)

predictive accuracy using the raw segmented multiplex myofibre images and apply relevant explainable ML methods to extract predictive inference and insights.

DL classification models

As discussed in Section 2.10.3 there are a number of high performing models for CV tasks including image classification [108–112]. The task of classifying the genetic mutation class of myofibres using multiplex segmented IMC data is an image classification task albeit consisting of more channels than the usual RGB or grayscale images. CNN-based models are good candidates for multiplex myofibre image classification and leveraging transfer learning i.e. using pre-trained weights from models trained on large datasets such as ImageNet, can improve model training. It was decided to experiment with DL models: i) Simple CNN: A stack of convolution, max-pooling, dropout and dense layers as describe in Table 6.2; ii) VGG16 with and without 'ImageNet' weights; iii) ResNet50 with and without 'ImageNet' weights.

Explainable methods for DL models

As discussed in Section 2.11.2 there are number of explainable methods [19, 134, 136–142, 144, 176, 177] for DL models that use different approaches to produce explainable masks. Explainable masks shows the attribution/importance of pixel/s and features in an image towards making a prediction. The faithfulness of these EMs is evaluated by the properties and axioms they satisfy. Such as

Sensitivity : An EM satisfies this property if for two given inputs x_i and x which only differ in one feature/pixel, the EM assigns a non-zero attribution to that feature/pixel [144].

Implementation invariance : If two different DL models are functionally equivalent i.e. produce the same predictions, then an EM satisfies an implementation invariance property if the attributions it produces are the same for the two DL models [144].

Completeness : An EM satisfies this property if the EM attributions add up to the difference between the output of the EM at the input x_i and baseline (reference) input x [144].

Linearity : An EM satisfies this property if the linear combination of two DL models represented as f_1, f_2 and their linear combination model represented as f_3= a . f_1+b .f_2 then the EM attributions for f_3 should be the weighted sum of the attributions for f_1 and f_2 with weights a and b respectively[144].

Local accuracy : An EM satisfies this property if for a DL model which is functionally represented as f(x) the EM approximation matches the output $f(x\prime)$ where $x\prime$ is simplified/reference input, e.g. a blank image [127].

Missingness : An EM satisfies this property if a feature/pixel can be toggled off/removed from input and this does not affect the prediction of a DL model and a corresponding explanation/attribution map has zero contribution assigned to that feature/pixel[127].

Consistency : An EM satisfies this property if a DL model changes so that if some simplified input's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease [127].

There are many EMs such as layer-wise relevance propagation[138, 141], DeConvNet[137], Deep Taylor [140], input Gradient [131, 138], integrated gradient [144] and DeepLIFT [177] that satisfy many of the properties discussed above and can be appropriate EMs for DL models. But the requirement for our use-case has an addition factor to consider, i.e. associations: knowing attribution or importance alone of pixels/features is not useful to understand mitochondrial disease pathology in terms of OXPHOS proteins. It requires the association between OXPHOS proteins and mutation class explained in terms of correlation (positive/negative) of protein markers toward prediction, and the relative importance/contribution of each of these protein markers quantified.

This is achieved by adapting the earlier mentioned EMs into Shapley additive explanations (SHAP values). There are two EMs that satisfy most of the EM evaluation properties and are adapted to build attribution/explanation masks in terms of SHAP values, namely integrated gradient and DeepLIFT with their corresponding SHAP adaptations called 'GradientExplainer' and 'DeepExplainer' respectively.

Integrated Gradient (GradientExplainer) For a DL model represented as function F(x) for some current input x, and a baseline input $x\prime$ e.g. a blank image with all zeros. Consider the straight line path from the baseline $x\prime$ to the input x, and compute the gradients at all points along the path. Instead of just taking the gradient at the input itself, Integrated Gradients integrate the gradients along the path from the baseline to the actual input. This approach accounts for the accumulated effect of each feature as it moves from a neutral state (baseline) to its actual value [144].
GradientExplainer is a SHAP adaptation of integrated gradient: Integrated gradients require a single baseline/reference value (e.g. a blank image) to integrate from, as an adaptation to generate approximate SHAP values using integrated gradient, the expected gradients reformulate the integral as an expectation and combines that expectation with sampling reference values from the background dataset (provided by users). This leads to a single combined expectation of gradients that converges to attributions that sum to the difference between the expected model output and the current output [178].

DeepLIFT (DeepExplainer) Intuitively it works by comparing the activation of each neuron for the current input to a reference activation, i.e. when a reference (average) input was passed, and assigning contributions based on how different the neuron's activation is from this reference. This difference from the reference contribution is computed by propagating these differences backwards through the network, from the output layer to the input layer, assigning a contribution score to each input feature. The idea is to distribute the difference in the output back to the input features in a way that considers both the weights and activations of the intermediate neurons. It uses a multiplicative approach to propagate these contribution scores, which considers both the gradient and the activation difference. This approach helps in handling cases where gradients might be small or zero. The contributions from all input features sum to the total difference between the actual output and the baseline output [143].
DeepExplainer is a SHAP adaptation of DeepLIFT. DeepExplainer approximates the conditional expectations of SHAP values using a selection of background samples (provided by the user). Contribution score attribution rules of DeepLIFT can be chosen to approximate Shapley values; by integrating over many background samples DeepExplainer estimates approximate SHAP values such that they sum to the difference between the expected model output on the passed background samples and the current model output [178].

It was decided that both 'GradientExplainer' and 'DeepExplainer' should be used as EMs for classification models. This will allow me to validate the faithfulness of these methods i.e. if the explanation masks generated by these two methods are similar in i) pixel importance, observed as a strength of SHAP value for a given pixel toward a prediction, and ii) channel

importance, observed as relative importance of all channels, i.e. the order of importance score attributed to each channel.

6.4 Experiments and results (myofibre)

6.4.1 Experiment design

Experiment data

As discussed in Section 6.3.1 multichannel (IMC) segmented myofibre images are used in the experiments. Training was conducted with i) all 13 channels (including mask), ii) eight OXPHOS channels as in Table 2.4 iii) individual channel, iv) selected channels identified in the predictive inference insights from Section 3.4. All models are trained for binary classification, the classes are defined in Table 3.1.

DL classification models experiments

Based on the strategy discussed in Section 6.3.2 the experiments were conducted by training i) CNN, ii) VGG16 and iii) ResNet50 models. These models were trained as binary classifiers with various combinations of pre-trained weights and model parameters as described in Table 6.3. The ImageNet pre-trained weights were used for transfer learning and a range of parameter ranges were experimented with based on DL model classification literature. The data was split into 70%: 15%: 15% for training, validation and testing. It was decided to observe accuracy holistically, i.e. across both classes, so the following metrics were recorded i) accuracy ii) recall (patients y=1) iii) recall (controls y=0). All experiments were conducted on a machine with the following specification: GPU: 1 x NVIDIA Tesla V100 (16GB), CPU: 6 cores, RAM:112GB.

Explainable methods experiments

As discussed in Section 6.3.2 'GradientExplainer' and 'DeepExplainer' were used to make explanation masks. For 'GradientExplainer' whole training data was used as 'background'/reference but for 'DeepExplainer' 100 training samples (50 from each class) were used as background, this was due to limitation of GPU memory. The 'DeepExplainer' would not run when exceeding 100 samples on available GPUs and it was infeasible to source a machine with high end GPU memory. Where comparable results were obtained by VGG16 and ResNet50, VGG16 was selected to apply EMs due to its relative simple architecture. It was consistently

Table 6.3 Details of myofibre classification experiments. The 'Datasets & patch dimensions' column describes the various patch dimensions that were experimented with. The column 'Initialised weights' describes the names of pre-trained weights that were used for transfer learning, 'None' means random initialised weights were used. The 'Parameters' column details all parameters relevant to the model being trained, that were being experimented with, for each parameter the values after the colon denote the options that were experimented with.

Model	Datasets & patch dimensions	Initialised weights	Parameters
CNN	multiplex myofibre images i) (200,200,13)- All channels ii)(200,200,8)- OXPHOS channels, iii) (200,200,1)- Individual channels	None	1) optimizer: 'adam', 'SGD', 'RMSprop'; 2) learning rate: 1e-3,1e-4; 3) Training metrics: Accuracy, Recall (patients), Recall (controls); 4) epochs >1000 with patience = 100
VGG16	multiplex myofibre images i) (200,200,13), ii) (200,200,8), iii)(200,200,1), iv) (224,224,13) v)(224,224,8), vi) (224,224,1)	None, 'ImageNet'	1) optimizer: 'adam', 'SGD', 'RMSprop'; 2) learning rate: 1e-3,1e-4; 3) Training metrics: Accuracy, Recall(patients), Recall(controls); 4) epochs >1000 with patience = 100
ResNet50	multiplex myofibre images i)(200,200,13), ii) (200,200,8), iii)(200,200,1), iv)(224,224,13), v) (224,224,8), vi) (224,224,1)	None, 'ImageNet'	1) optimizer: 'adam', 'SGD', 'RMSprop'; 2) learning rate: 1e-3,1e-4; 3) Training metrics: Accuracy, Recall(patients), Recall(controls); 4) epochs >1000 with patience = 100

observed that for all experiments conducted the explanation masks by both GradientExplainer and DeepExplainer were equivalent in both the pixels identified as important and the order of channel's absolute SHAP values (ASV): calculated as the sum of all SHAP values (irrespective of sign) in a channel's explanation mask.

6.4.2 Results

Explainable DL analysis of class A vs controls

Patients suffering from nDNA encoded mutations (class A) was the only case that the current techniques can accurately classify as observed in Table 2.4. ML models presented in Section 3.4.2 also predicted this myofibre with 100% accuracy. In this section I apply explainable DL methods to classify these same myofibres but using raw segmented data and report the results and insights achieved by these techniques.

DL classification results CNN, VGG16 and ResNet50 models were trained and results are reported in Table 6.4. In addition to the models mentioned in Table 6.4 other models also produced 100% accuracy, including one model that was trained on seven selected channels (NDUFB8, NDUFA13, SDHA, UqCRC2, MTCO1, COX4+4L2 and OSCP), i.e. selected because of their predictive power as discussed in Section 3.4.2.

Explainable methods for class A vs controls models As there were many models with 100% predictive accuracy, it was decided that the model with simpler architecture, i.e. CNN should be used to apply EM due to lower computation cost. Three CNN models trained on i) all 13 channels, ii) 8 OXPHOS channels and iii) seven selected channels were used to apply EMs. Both GradientExplainer and DeepExplainer were applied and SHAP values explanation masks produced by both methods are very similar. However, with DeepExplainer only 100 multiplex myofibres images as reference/background were possible due to limited GPU memory. Hence all explanation masks reported here were generated using GradientExplainer.

ASV values were used to identify the channel importance/attribution and pixel colours in the explanation masks represent SHAP values, i.e. red represents positive SHAP values that contribute positively towards positive class (class A) and blue represents negative SHAP values that contribute negatively towards negative class (control). As seen in Figures 6.1 and 6.2, explanation masks for both models report the top four ASV values for OSCP, NDUFB8, UqCRC2 and SDHA. Figure 6.3 explanation mask shows the models trained on 13 channels have the highest ASV values for myofibre mask and OSCP. OSCP and NDUFB8 channels have the highest ASV in both Figures 6.1 and 6.2, observing the positive and negative SHAP

Table 6.4 Classification metrics for DL models trained to predict class A myofibres. In the table Chns,Acc,R_P,R_C and NA stands for channels, accuracy, recall (patients' myofibres), recall (controls' myofibres) and not available respectively. Note: individual channel training was performed for the top 2 performing models.

Top (3) models	Metrics (13 Chns)			Metrics (8 Chns)			Acc (single protein)(%)
	Acc(%)	R_P(%)	R_C(%)	Acc(%)	R_P(%)	R_C(%)	
CNN	100	100	100	100	100	100	92.28 (NDUFB8) 99.36 (NDUFA13) 83.44 (SDHA) 94.26 (UqCRC2) 89.81 (MTCO1) 94.26 (COX4+4L2) 94.26 (OSCP) 83.44 (VDAC1) 87.26 (TOM22) 85.99 (Dystrophin) 67.51 (DNA1) 64.97 (DNA2) 54.14 (Mask)
VGG16 initialised weights (random)	100	100	100	100	100	100	98.73 (NDUFB8) 92.36 (NDUFA13) 97.45 (SDHA) 98.73 (UqCRC2) 89.81 (MTCO1) 94.9 (COX4+4L2) 97.45 (OSCP) 88.53 (VDAC1) 61.78 (TOM22) 92.36 (Dystrophin) 85.35 (DNA1) 88.53 (DNA2) 55.41 (Mask)
ResNet50 initialised weights (random)	100	100	100	100	100	100	NA

values within these masks reveal that high intensity pixels in NDUFB8 are associated with control class predictions and similarly it can be seen the high intensity pixels, especially near the membrane, are assosiated with control class predictions.
Figure 6.4 is made by merging the four channels OSCP, NDUFB8, UqCRC2 and SDHA with the highest ASV into a RGB image. The four channels were scaled, i.e. divided by the number of channels (4) and sequentially added to red, green, blue channels. As observed in the figure 6.4 this approach profiles the two classes into visually distinct colour groups, i.e. class A myofibres are bluish-purple and control myofibres are greenish-white.

Biological validation The predictive power of complex I protein markers such as NDUFA13 (99.36 % CNN model), NDUFB8 (98.36 % VGG16 model) is expected as discussed earlier in Section 3.4.2 in class A myofibres nDNA encoded complex I proteins will be down-regulated, i.e. exhibit low- evels, and models are leveraging this feature of class A myofibres. Similarly, the predictive power of other OXPHOS proteins such as UqCRC2, SDHA, OSCP and COX4+4L2 is also expected due to upregulation as discussed in Section 3.4.2. The explanation masks of the two myofibres reveals OSCP, NDUFB8, SDHA and UqCRC2 have the four highest ASV. This means for these two myofibres the model's predictions were influenced more by these four markers/channels. As discussed earlier these four markers should provide the differential features and the models seems to be leveraging this. Furthermore, the profiling of these two myofibres based on the highest four ASV markers in Figure 6.4 is an interesting visualisation of classifying these myofibres.

Explainable DL analysis of class B vs controls

ML models presented in Section 3.4.3 predicted these myofibres with 99% accuracy. In this section I apply explainable DL methods to classify these same myofibres but using raw segmented data and report the results and insights achieved by these techniques.

DL classification results CNN, VGG16 and ResNet50 models were trained and results are reported in Table 6.5. In addition to the models mentioned in Table 6.5 a model trained on five channels (NDUFB8, NDUFA13, UqCRC2, MTCO1 and COX4+4L2) that were selected because of predictive inference insights in Section 3.4.3 recorded accuracy of 99%; recall (class B myofibres) of 99%; recall(control myofibres) of 98% .

Explainable methods for class B vs controls models Based on predictive accuracy it was decided that the VGG16 model should be used to apply EM. Two VGG16 models trained

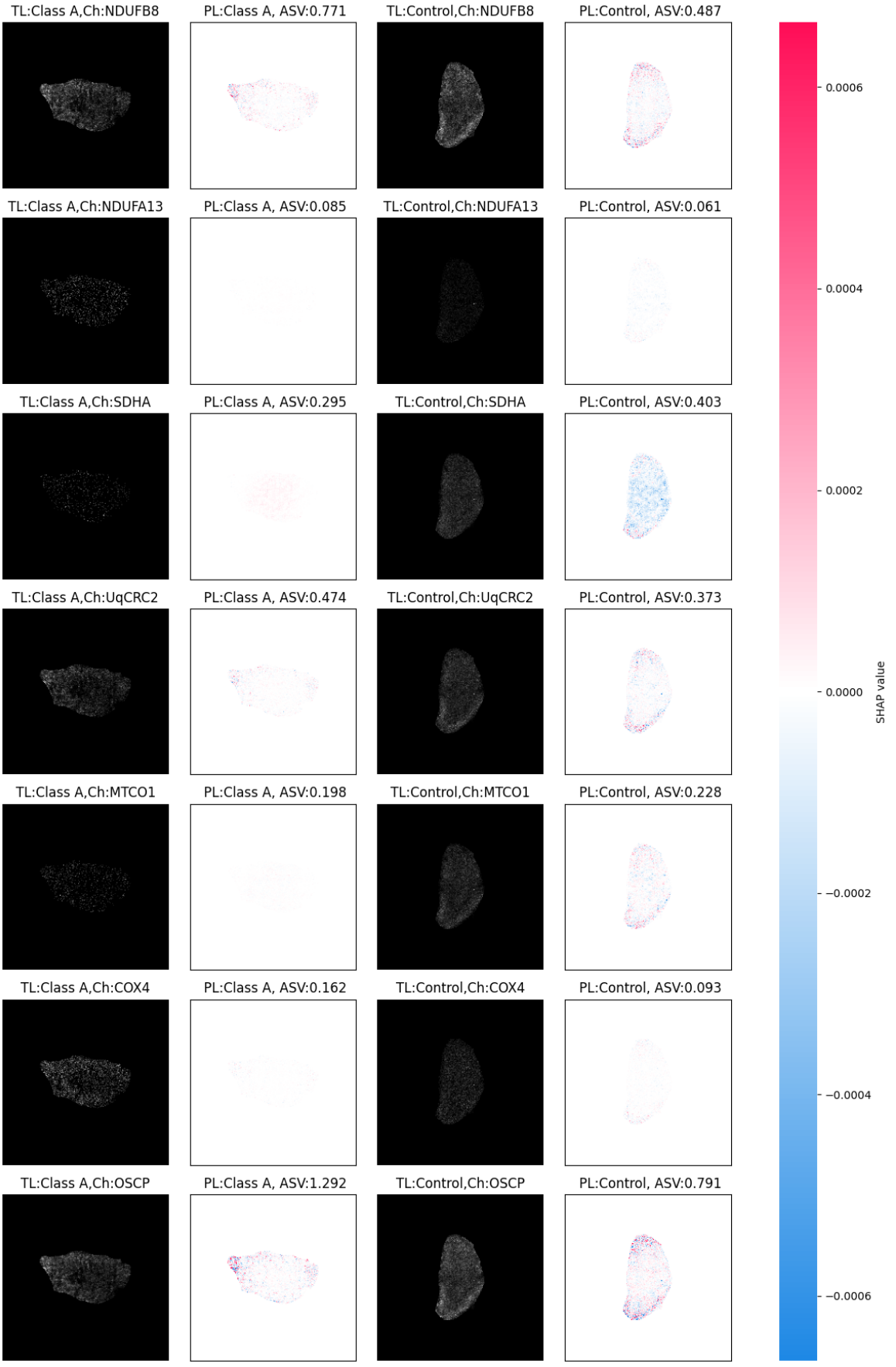

Figure 6.1 GradientExplainer applied to CNN model (class A) trained on 7 OXPHOS channels selected because of the predictive inference insights in Section 3.4.2. The first & third columns are raw channel images, second and third columns are their respective explanation masks. In the figure TL,PL,Ch,ASV stands for true label, predicted label, channel, absolute shap value respectively.

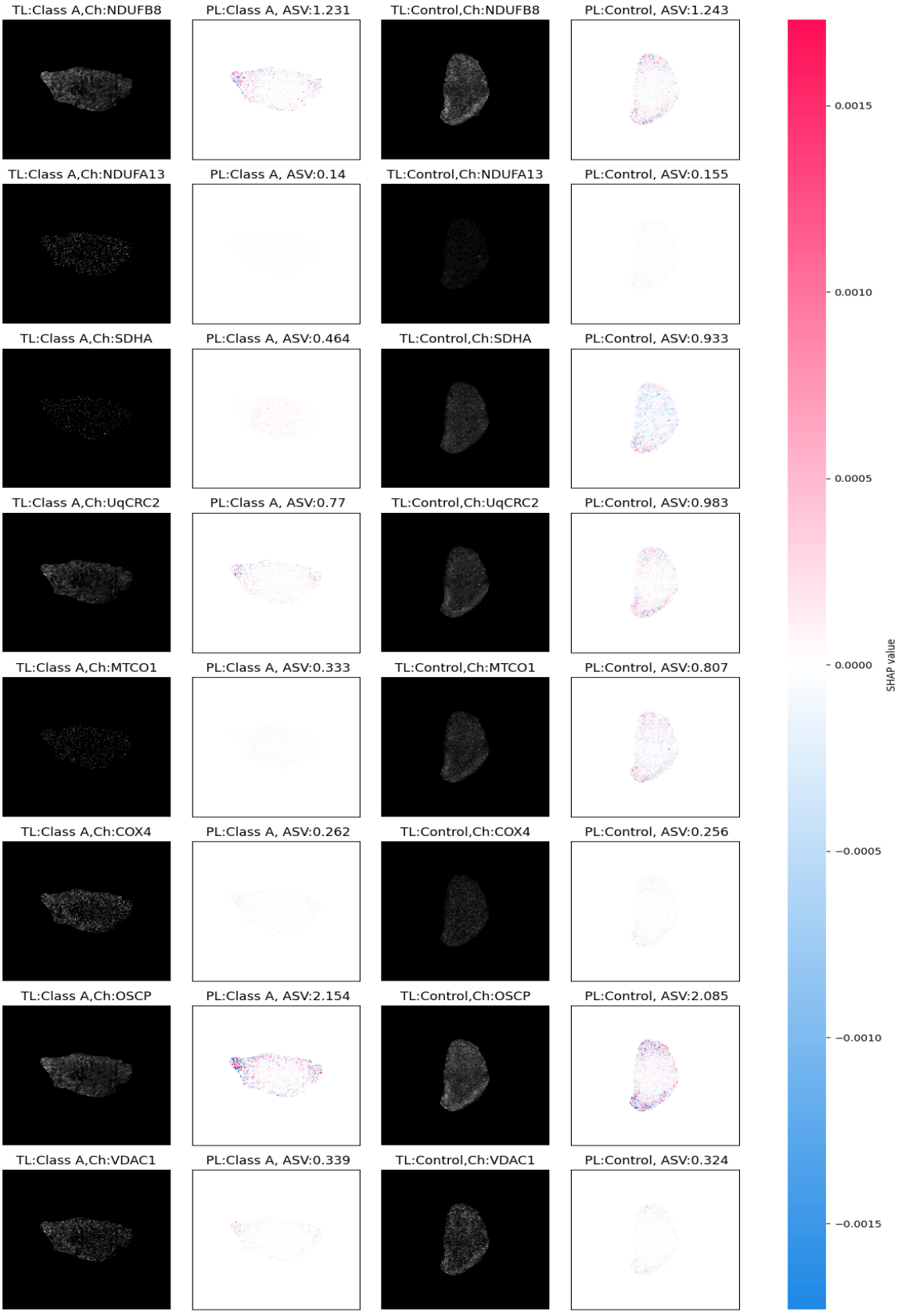

Figure 6.2 GradientExplainer applied to CNN model (class A) trained on 8 OXPHOS channels.

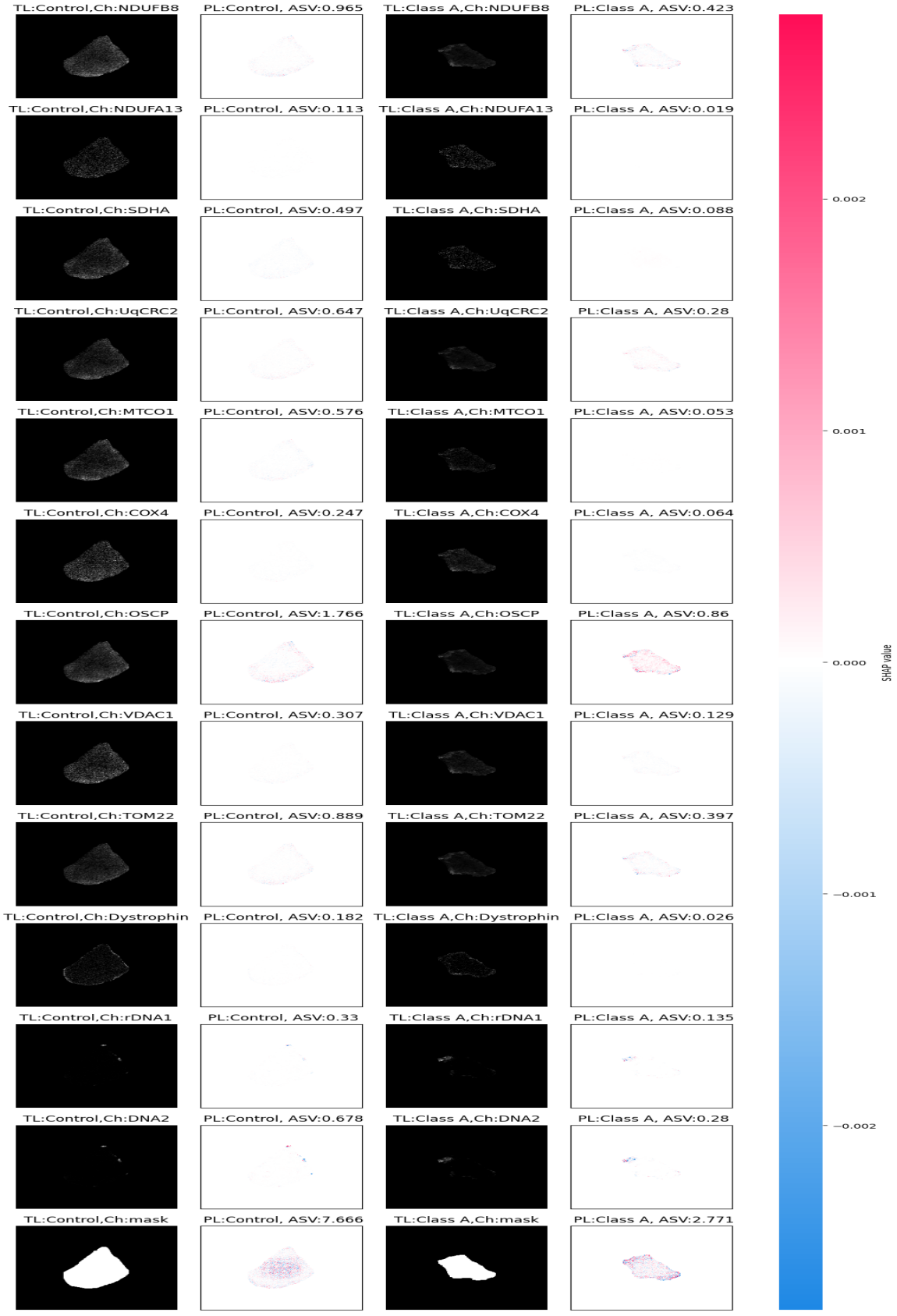

Figure 6.3 GradientExplainer applied to CNN model (class A) of trained on all 13 channels.

Table 6.5 Classification metrics for DL models trained to predict class B myofibres. Note: individual channel training was performed for the top 2 performing models.

Top (3) models	Metrics (13 Chns)			Metrics (8 Chns)			Acc (single protein)(%)
	Acc (%)	R_P (%)	R_C(%)	Acc (%)	R_P (%)	R_C (%)	
VGG16 initialised weights (random)	99	99	96	98	99	95	89.36 (NDUFB8) 98.62 (NDUFA13) 87.13 (SDHA) 86.57 (UqCRC2) 83.58 (MTCO1) 86.19 (COX4+4L2) 87.68 (OSCP) 86.01 (VDAC1) 86.57 (TOM22) 97.95 (Dystrophin) 87.5 (DNA1) 89.36 (DNA2) 87.69 (Mask)
ResNet50 initialised weights (random)	98	100	90	98	99	94	86.75 (NDUFB8) 84.32 (NDUFA13) 86 (SDHA) 86.58 (UqCRC2) 82.46 (MTCO1) 84.7 (COX4+4L2) 85.63 (OSCP) 84.88 (VDAC1) 85.07 (TOM22) 89.36 (Dystrophin) 85.44 (DNA1) 87.87 (DNA2) 77.62 (Mask)
CNN	91	97	63	96	99	85	NA

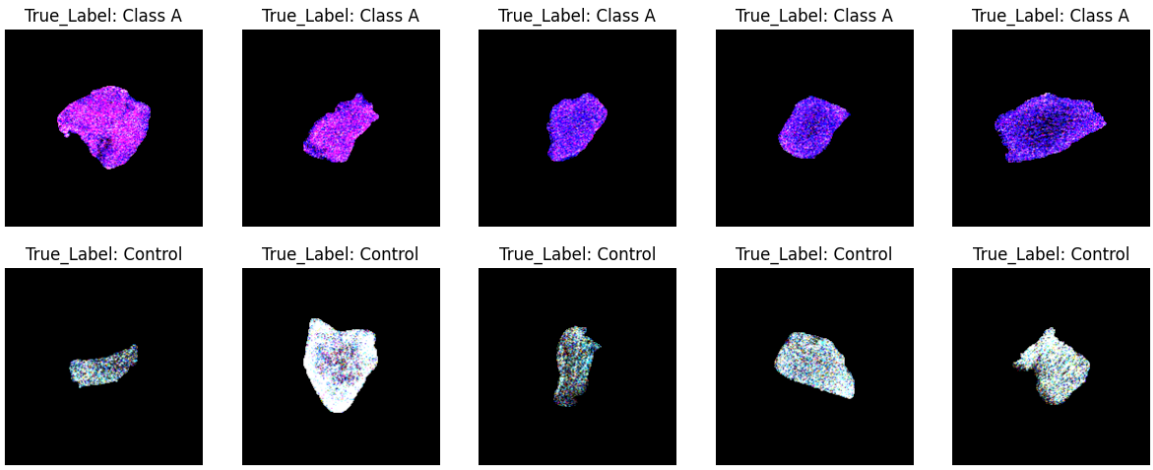

Figure 6.4 RGB image made by weighted stacking of input channels of class A and control myofibres that had highest 4 ASV values in Figure 6.1, i.e. OSCP, NDUFB8, SDHA and UqCRC2. On the top row are class A myofibres and on the bottom row are control myofibres.

on i) eight OXPHOS channels and ii) five select channels were used to apply EMs. All explanation masks reported here were generated using GradientExplainer.

ASV values were used to identify the channel importance/attribution and pixel colours in the explanation masks represent SHAP values. As seen in Figures 6.5 and 6.6 explanation masks for both models report the top four ASV values for NDUFB8, OSCP, UqCRC2 and SDHA. Observing the positive and negative SHAP values within these high ASV explanation masks reveals that high intensity pixels in the centre of the myofibre in NDUFB8 are associated with control class predictions; it can be seen the that low intensity pixels in SDHA are associated with class B myofibres predictions.
Figure 6.7 is made by merging the four channels NDUFB8, OSCP, UqCRC2 and SDHA with highest ASV into a RGB image. As observed in Figure 6.7 this approach profiles the two classes into visually distinct colour groups, i.e. class B myofibres are reddish-purple and control myofibres are whitish.

Biological validation Above 92% accuracy observed for complex I marker trained VGG16 single-protein models is expected as discussed in Section 3.4.3, but the highest predicted accuracy observed being just using Dystrophin (membrane marker) is a surprising finding. This could be artifactual and warrants an experimental validation. The highest ASV for the NDUFB8 and the high positive SHAP values for pixels with low intensities being within the NDUFB8 channel as presented in Figures 6.5 and 6.6, implies that models are leveraging the expected downregulation of complex I markers. The SDHA downregulation's association with class B is an intriguing finding that needs experimental validation.

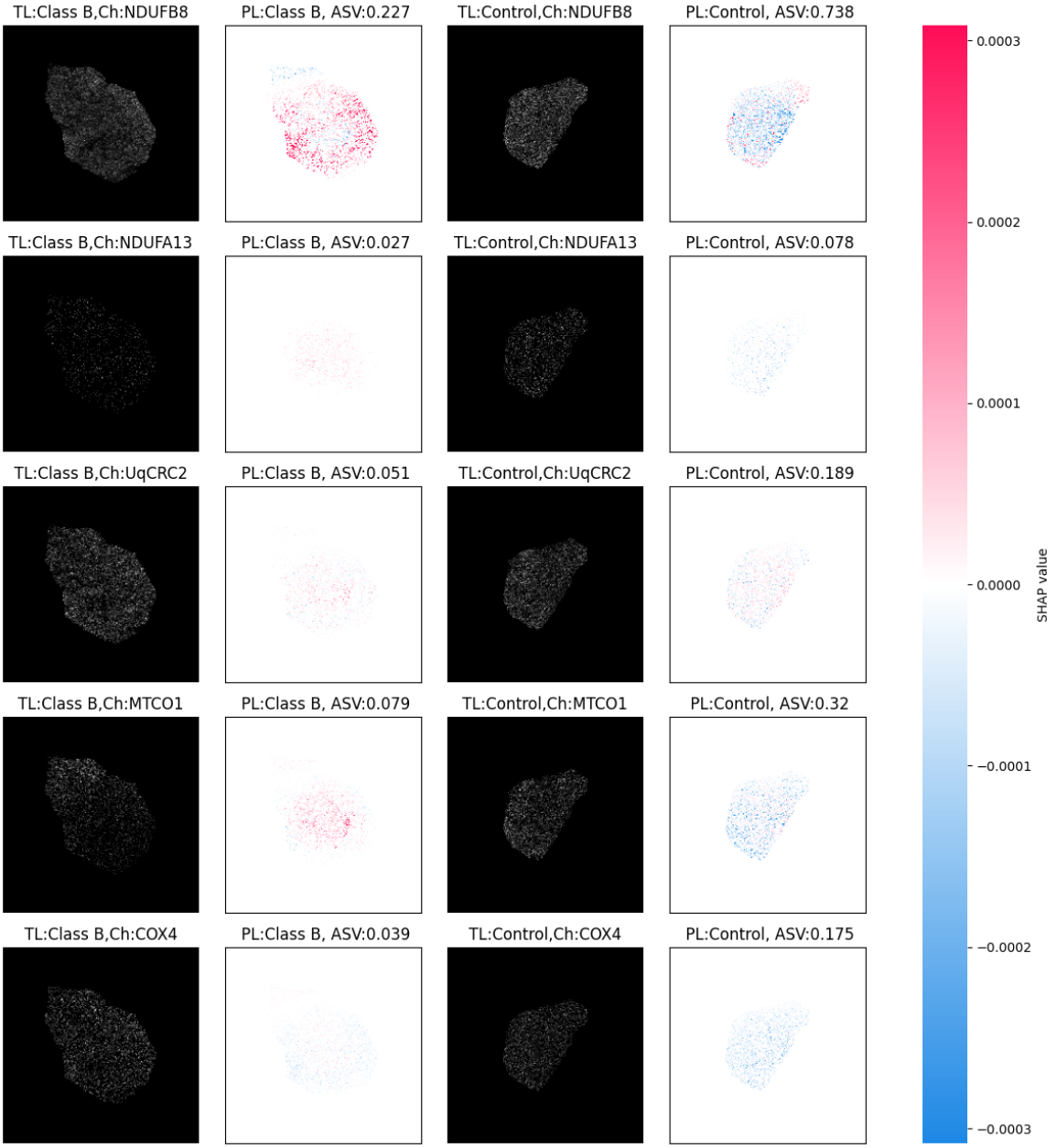

Figure 6.5 GradientExplainer applied to VGG16 model trained on 5 OXPHOS channels of myofibres from class B patients and controls, selected because of the predictive inference insights in Section 3.4.3.

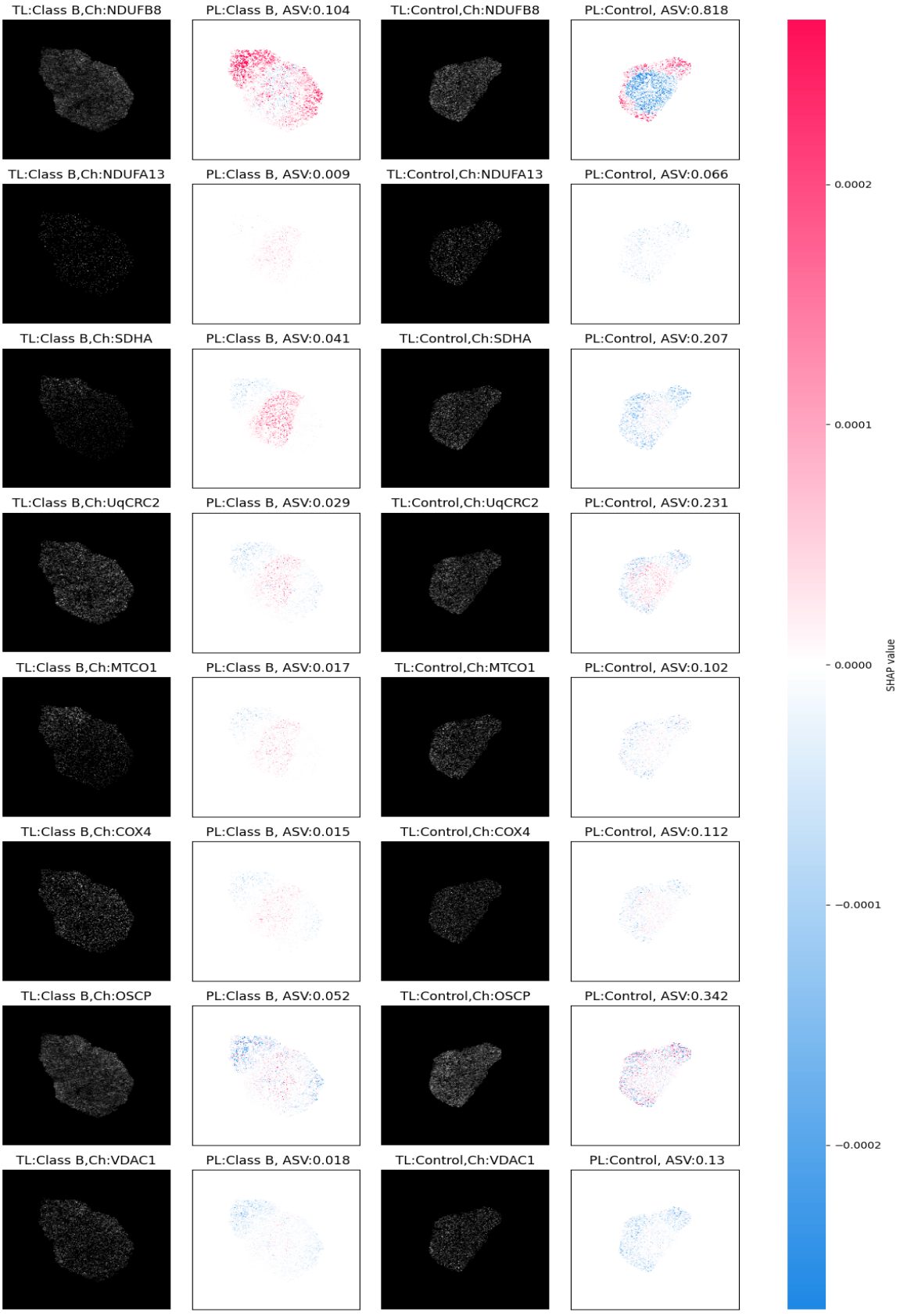

Figure 6.6 GradientExplainer applied to VGG16 model trained on 8 OXPHOS channels of myofibres from class B patients and controls.

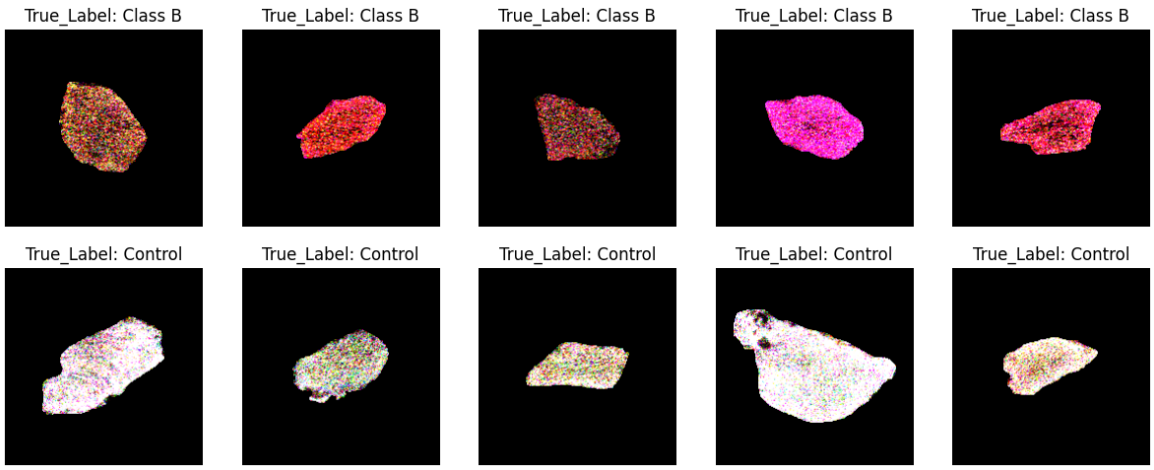

Figure 6.7 RGB image made by weighted stacking of input channels of class B and control myofibres that had highest 4 ASV values in Figure 6.5 i.e. NDUFB8, OSCP, UqCRC2 and SDHA. On the top row are class B myofibres and on the bottom row are control myofibres.

Explainable DL analysis of class C vs controls

ML models presented in Section 3.4.4 predicted these myofibres with 99% accuracy. In this section I apply explainable DL methods to classify these same myofibres but using raw segmented data and report the results and insights achieved by these technique.

DL classification results CNN, VGG16 and ResNet50 models were trained and results are reported in Table 6.6. In addition to the models mentioned in Table 6.6 a model trained on five channels (NDUFB8, NDUFA13, SDHA, MTCO1 and COX4+4L2) that were selected because of predictive inference insights in Section 3.4.4 recorded accuracy of 99%; recall(class B myofibres) of 99%; recall(control myofibres) of 100% .

Explainable methods for class C vs controls models Based on predictive accuracy it was decided that the VGG16 model should be used to apply EMs. Two VGG16 models trained on i) eight OXPHOS channels and ii) five selected channels were used to apply EMs. All explanation masks reported here were generated using GradientExplainer.

ASV values were used to identify the channel importance/attribution and pixel colours in the explanation masks represent SHAP values. As seen in Figures 6.8 and 6.9 explanation masks for both models report a difference in the highest four ASV values: for the five channel model it is NDUFB8, NDUFB13, SDHA, and MTCO1, whereas for the eight channel model it is NDUFB8, OSCP, UqCRC2 and SDHA. Observing the positive and negative SHAP values within these high ASV explanation masks reveals that high intensity pixels

Table 6.6 Classification metrics for DL models trained to predict class C myofibres. Note: individual channel training was performed for top 2 performing models.

Top (3) models	Metrics(13 Chns)			Metrics (8 Chns)			Acc (single protein) (%)
	Acc (%)	R_P (%)	R_C (%)	Acc (%)	R_P (%)	R_C (%)	
VGG16 initialised weights (random)	96	96	95	99	99	98	93.57 (NDUFB8) 92.69 (NDUFA13) 91.22 (SDHA) 87.43 (UqCRC2) 76.9 (MTCO1) 90.64 (COX4+4L2) 92.4 (OSCP) 89.77 (VDAC1) 88.89 (TOM22) 98.24 (Dystrophin) 89.47 (DNA1) 87.72 (DNA2) 71.05 (Mask)
CNN	88	89	85	99	99	99	85.38 (NDUFB8) 79.53 (NDUFA13) 84.5 (SDHA) 74.85 (UqCRC2) 77.78 (MTCO1) 78.65 (COX4+4L2) 85.67 (OSCP) 83.92 (VDAC1) 90.35 (TOM22) 96.2 (Dystrophin) 71.2 (DNA1) 74.27 (DNA2) 70.76 (Mask)
ResNet50 initialised weights (random)	91	97	76	99	100	97	NA

near the membrane of the myofibre in NDUFB8 are associated with class C myofibre class predictions; it can be seen that the high intensity pixels in SDHA are associated with control class myofibres predictions.
Figure 6.10 is made by merging the four channels NDUFB8, NDUFB13, SDHA and OSCP with the two highest ASV from each explanation mask into a RGB image. As observed in Figure 6.10 this approach does not profile the two classes clearly.

Biological validation The highest predicted accuracy being observed in both models trained with just Dystrophin (membrane marker) is a surprising finding. This could be artefactual and warrants an experimental validation. The same applies for DNA1 and DNA2 markers unless it is that there are more nuclei in either classes of myofibres that is leading the models to exploit this as a differential feature or even the location of nuclei within myofibres, this warrants further experimental validation. The above 90% accuracy with single-protein models trained on OSCP,SDHA and COX4+4L2 is an interesting finding that also warrants experimental validation. The highest ASV for the complex I markers and the high positive SHAP values for pixels with low intensities within NDUFA13 channel as presented in Figures 6.8 and 6.9, implies that models are leveraging the expected downregulation of complex I markers.

Explainable DL analysis of class D vs controls

ML models presented in Section 3.4.5 predicted these myofibres with 93% accuracy. In this section I apply explainable DL methods to classify these same myofibres but using raw segmented data and report the results and insights achieved by these technique.

DL classification results CNN, VGG16 and ResNet50 models were trained and results are reported in Table 6.7. In addition to the models mentioned in Table 6.7 a model trained on three channels (NDUFB8, UqCRC2 and COX4+4L2) that were selected because of predictive inference insights in Section 3.4.5 recorded accuracy of 89% ;recall(class B myofibres) of 98%; recall(control myofibres) of 65%.

Explainable methods for class D vs controls models The VGG16 model is used to apply EMs due to its high predictive accuracy. A VGG16 model trained on eight OXPHOS channels was used to apply EMs. All explanation masks reported here were generated using GradientExplainer.

ASV values were used to identify the channel importance/attribution and pixel colours in the explanation masks represent SHAP values. As seen in Figure 6.11 the explanation mask

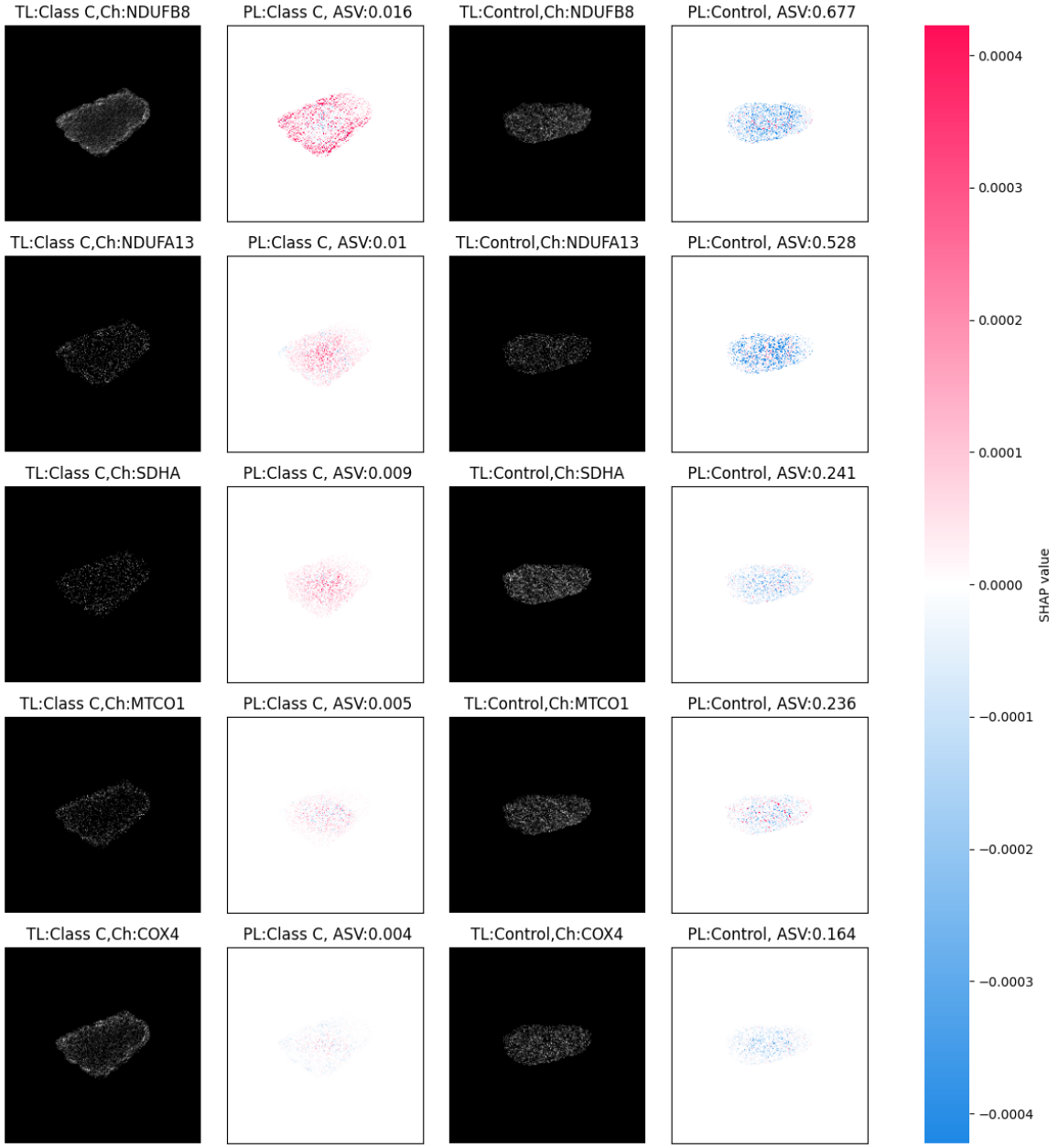

Figure 6.8 GradientExplainer applied to VGG16 model trained on 5 OXPHOS channels of myofibres from class C patients and controls, selected because of the predictive inference insights in Section 3.4.4.

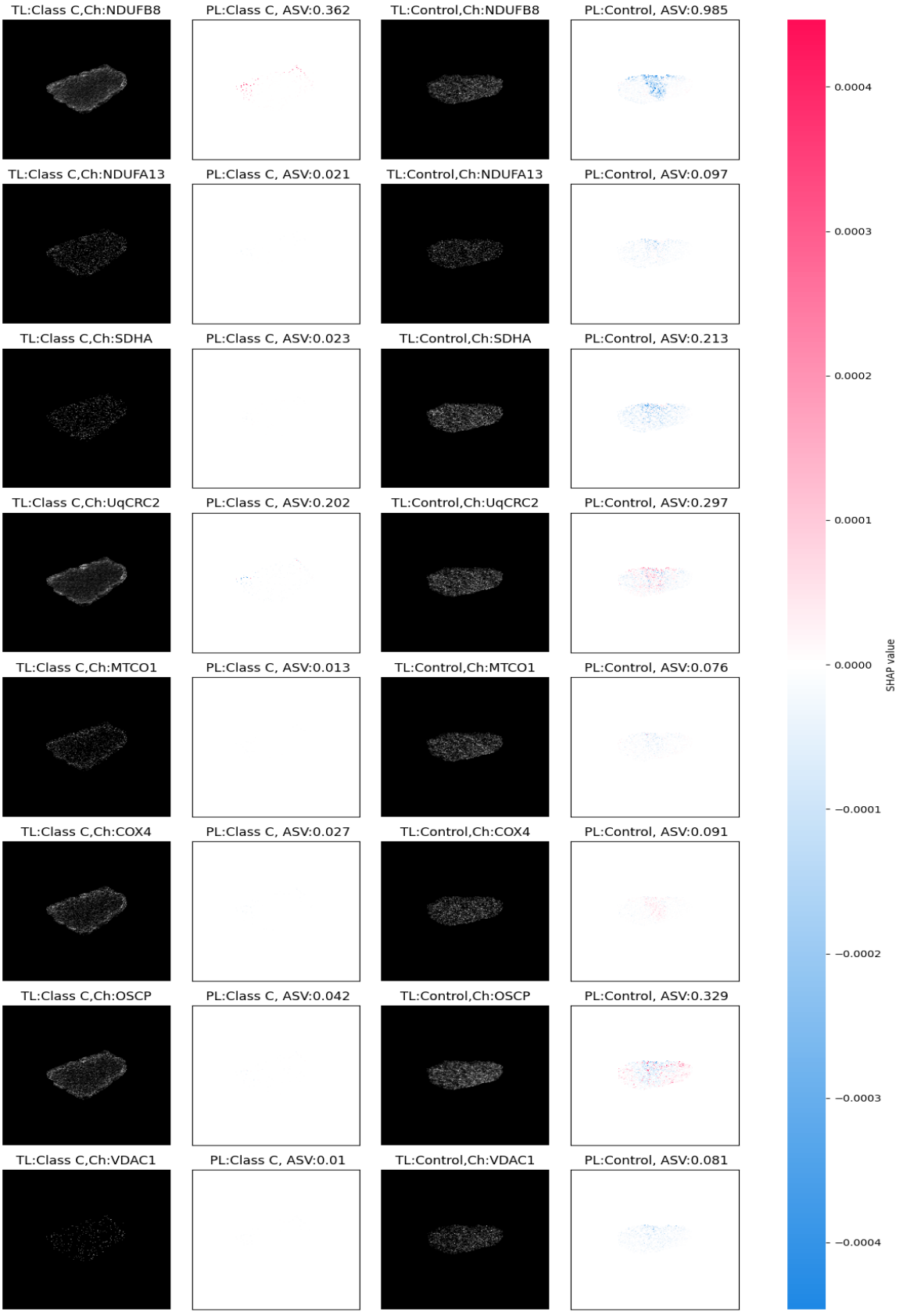

Figure 6.9 GradientExplainer applied to VGG16 model trained on 8 OXPHOS channels of myofibres from class C patients and controls.

Table 6.7 Classification metrics for DL models trained to predict class D myofibres. Note: individual channel training was performed for top 2 performing models.

Top(3) models	Metrics(13 Chns)			Metrics(8 Chns)			Acc (single protein)(%)
	Acc(%)	R_P(%)	R_C(%)	Acc(%)	R_P(%)	R_C(%)	
VGG16 initialised weights (random)	93.3	95	89	99.8	100	99	81.11 (NDUFB8) 80.83 (NDUFA13) 80.23 (SDHA) 83.33 (UqCRC2) 79.72 (MTCO1) 84.72 (COX4+4L2) 85.23 (OSCP) 85.83 (VDAC1) 90 (TOM22) 93.61 (Dystrophin) 84.72 (DNA1) 83.05 (DNA2) 79.17 (Mask)
ResNet50 initialised weights (random)	94.4	98	84	99	99	97	81.39 (NDUFB8) 76.94 (NDUFA13) 77.77 (SDHA) 81.39 (UqCRC2) 80.27 (MTCO1) 75.28 (COX4+4L2) 80 (OSCP) 74.17 (VDAC1) 84.44 (TOM22) 87.78 (Dystrophin) 85 (DNA1) 83.33 (DNA2) 78.05 (Mask)
CNN	92	93	89	93	94	89	NA

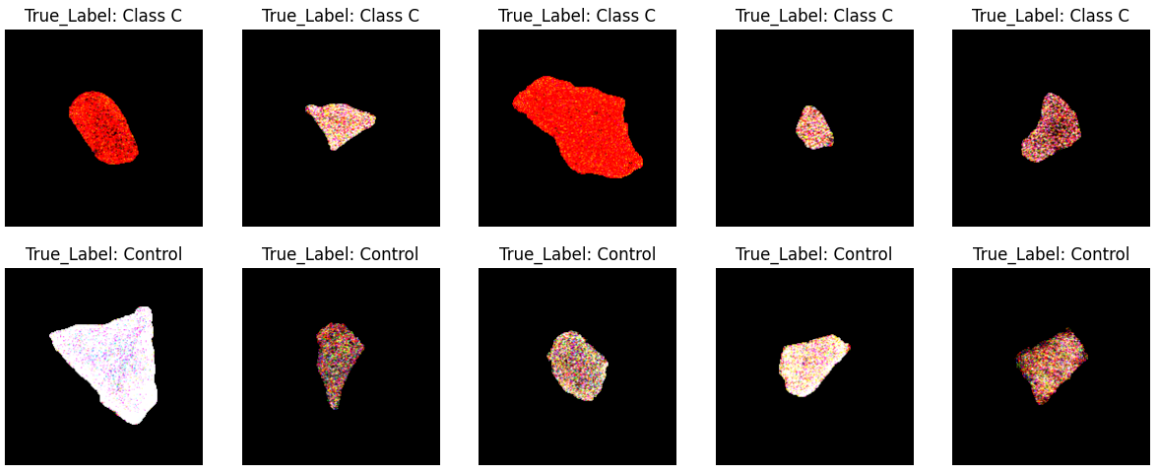

Figure 6.10 RGB image made by weighted stacking of input channels of class C and control myofibres that had the highest 4 ASV values in Figure 6.8 i.e. NDUFB8, OSCP, UqCRC2 and SDHA. On the top row are class C myofibres and on the bottom row are control myofibres.

for the model reported the highest four ASV values for channels NDUFB8, UqCRC2, OSCP and NDUFB13. Observing the positive and negative SHAP values within these highest ASV explanation masks revealed that high intensity pixels in UqCRC2 are associated with control myofibre class predictions; it can be seen that low intensity pixels in NDUFA13 are associated with class D myofibre predictions.
Figure 6.12 is made by merging the four channels NDUFB8, UqCRC2, OSCP and NDUFB13 that have the highest ASV in the explanation mask into a RGB image. As observed in Figure 6.12, this approach profiles the two classes into visually distinct colour groups, i.e. class D myofibres are bluish and control myofibres are whitish.

Biological validation The highest predicted accuracy being observed in VGG16 models trained with just Dystrophin;TOM22 is a surprising finding. This could be artifactual and warrants an experimental validation. The ASV and individual SHAP values per pixel and channel are not straightforward to validate as class D myofibres are not completely understood yet.

Explainable DL analysis of P03 vs controls

ML models presented in Section 3.4.6 predicted these myofibres with 95% accuracy. In this section I apply explainable DL methods to classify these same myofibres but using raw segmented data and report the results and insights achieved by these technique.

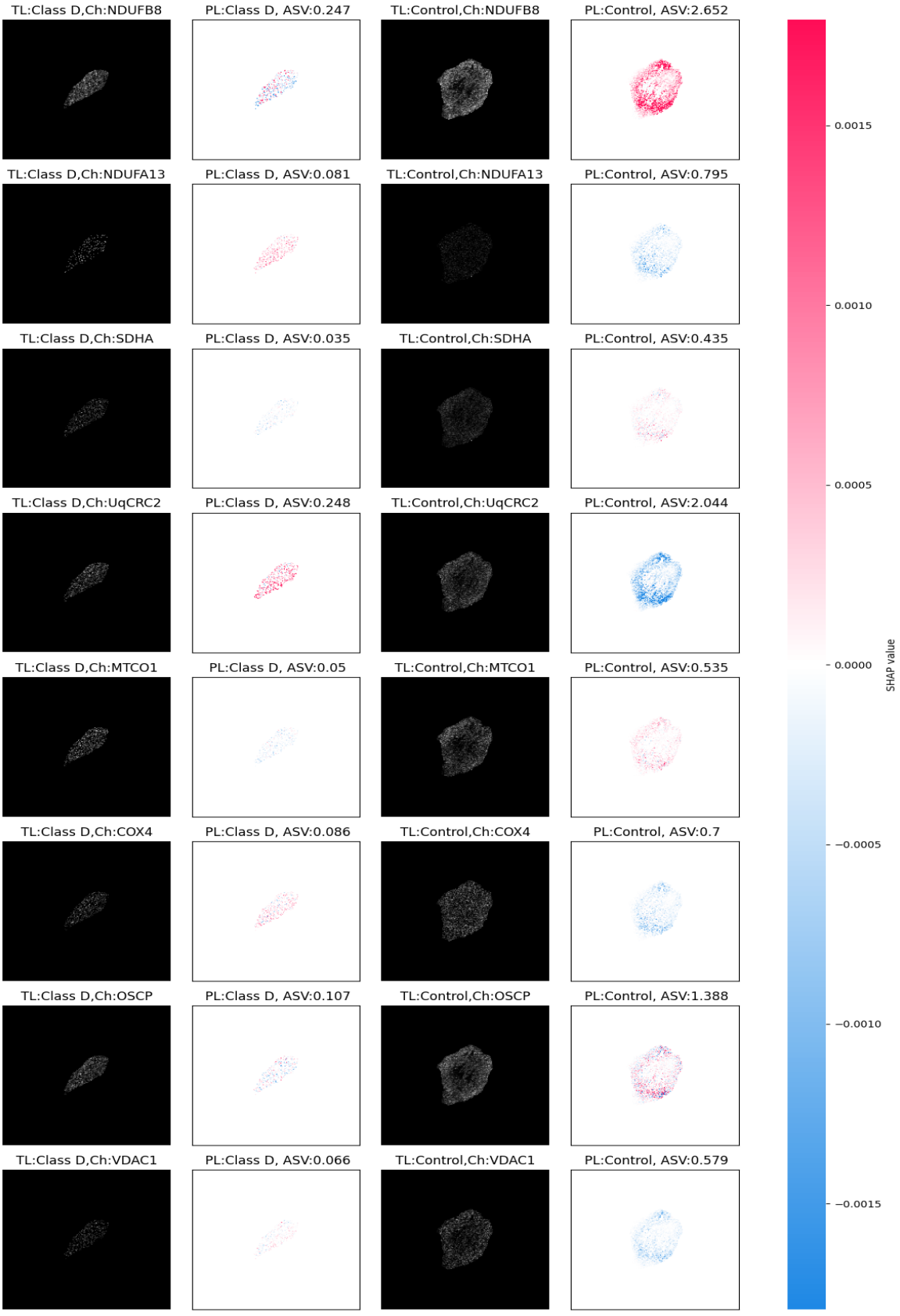

Figure 6.11 GradientExplainer applied to VGG16 model trained on 8 OXPHOS channels of myofibres from class D patients and controls.

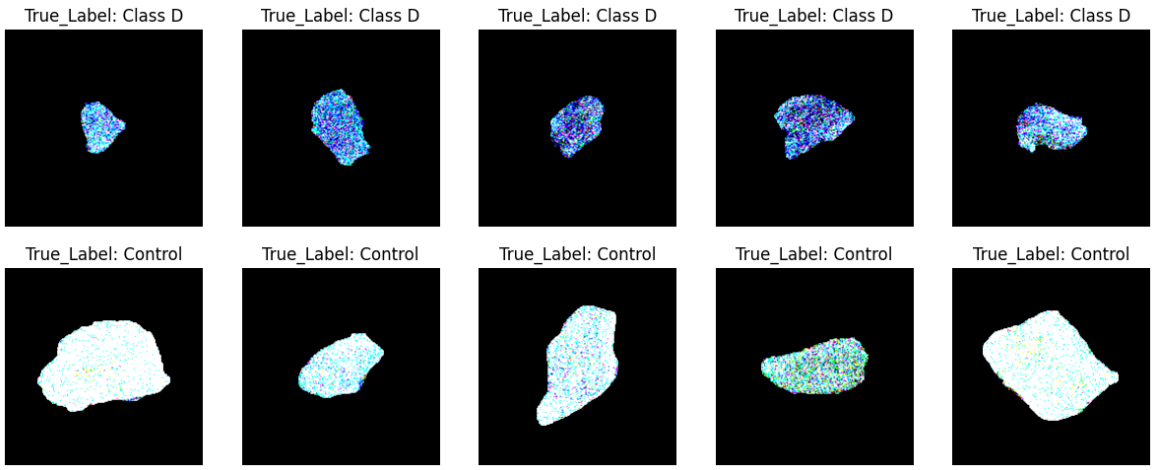

Figure 6.12 RGB image made by weighted stacking of input channels of class D and control myofibres that had the highest 4 ASV values in Figure 6.11 i.e. NDUFB8, UqCRC2, OSCP and NDUFB13. On the top row are class D myofibres and on the bottom row are control myofibres.

DL classification results CNN, VGG16 and ResNet50 models were trained and results are reported in Table 6.8.

Explainable methods for P03 vs controls models The VGG16 model is selected to apply EMs as it achieved the highest predictive accuracy. A VGG16 model trained on eight OXPHOS channels was used to apply EMs. All explanation masks reported here were generated using GradientExplainer.

ASV values were used to identify the channel importance/attribution and pixel colours in the explanation masks represent SHAP values. As seen in Figure 6.13, the explanation mask for the model reported the highest four ASV values for channels UqCRC2, NDUFB8, NDUFA13 and COX4+4L2. Observing the positive and negative SHAP values within these highest ASV explanation masks reveal that high intensity pixels in UqCRC2 and NDUFA13 are associated with control myofibre class predictions; it can be seen that low intensity pixels in COX4+4L2 are associated with P03 class myofibre predictions.
Figure 6.14 is made merging the four channels UqCRC2, NDUFB8, NDUFA13 and COX4+4L2 that have the highest ASV in the explanation mask into a RGB image. As observed in Figure 6.14, this approach profiles the two classes into visually distinct colour groups, i.e. P03 myofibres are greenish and control myofibres are whitish.

Table 6.8 Classification metrics for DL models trained to predict P03 myofibres. Note: individual channel training was performed for top 2 performing models.

Top (3) models	Metrics (13 Chns)			Metrics (8 Chns)			Acc (single protein) (%)
	Acc (%)	R_P (%)	R_C (%)	Acc (%)	R_P (%)	R_C (%)	
VGG16 initialised weights (random)	89	93	84	98	100	99	88.15 (NDUFB8) 84.67 (NDUFA13) 86.06 (SDHA) 87.11 (UqCRC2) 83.27 (MTCO1) 89.2 (COX4+4L2) 90.24 (OSCP) 89.55 (VDAC1) 90.59 (TOM22) 93.73 (Dystrophin) 88.85 (DNA1) 83.62 (DNA2) 91.99 (Mask)
ResNet50 initialised weights (random)	90	93	84	95	99	87	80.84 (NDUFB8) 85.02 (NDUFA13) 78.74 (SDHA) 81.53 (UqCRC2) 82.58 (MTCO1) 82.93 (COX4+4L2) 85.36 (OSCP) 83.97 (VDAC1) 90.59 (TOM22) 90.59 (Dystrophin) 85.02 (DNA1) 81.18 (DNA2) 83.97 (Mask)
CNN	79	81	75	86	90	79	NA

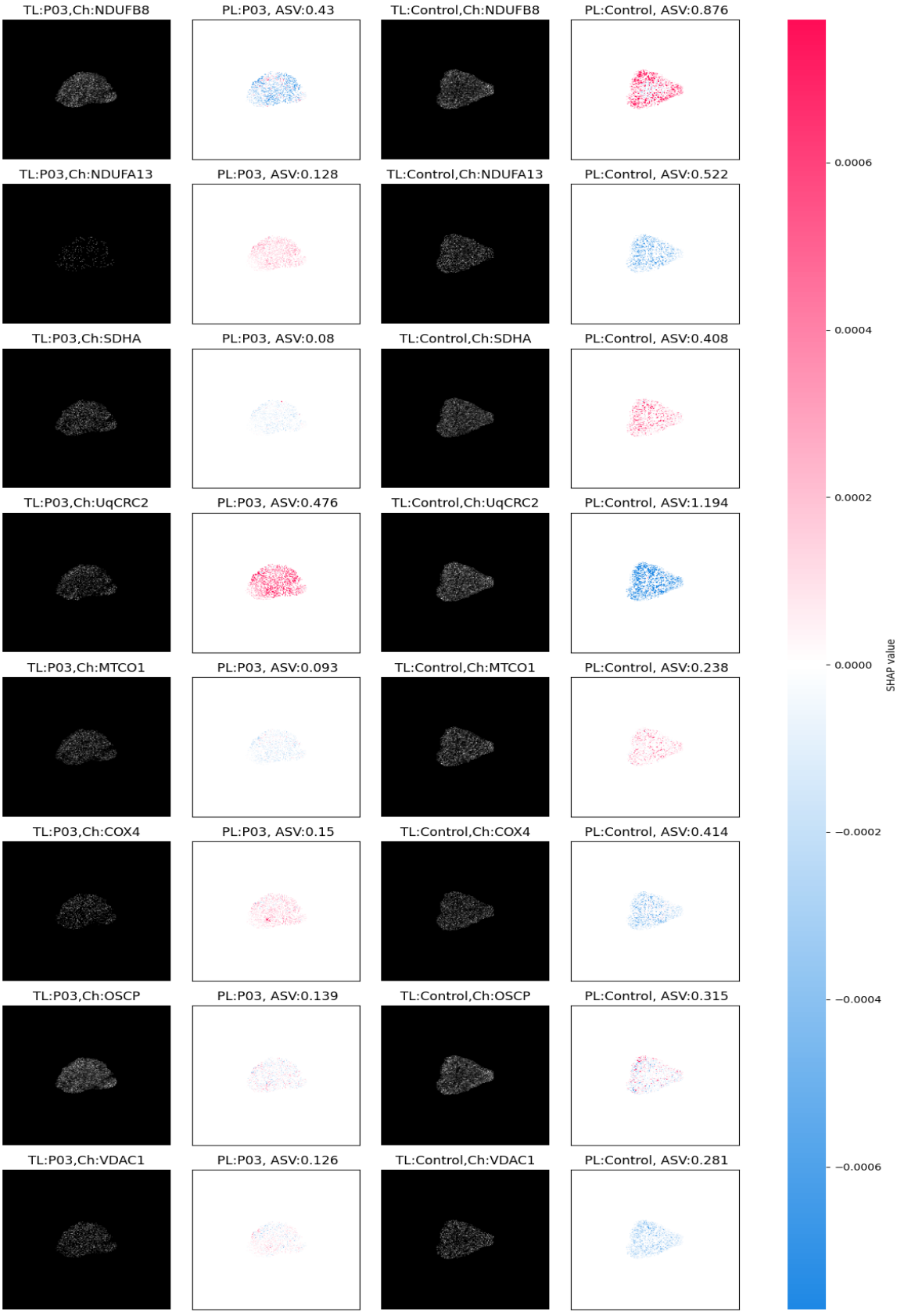

Figure 6.13 GradientExplainer applied to VGG16 model trained on 8 OXPHOS channels of myofibres from P03 patient and controls.

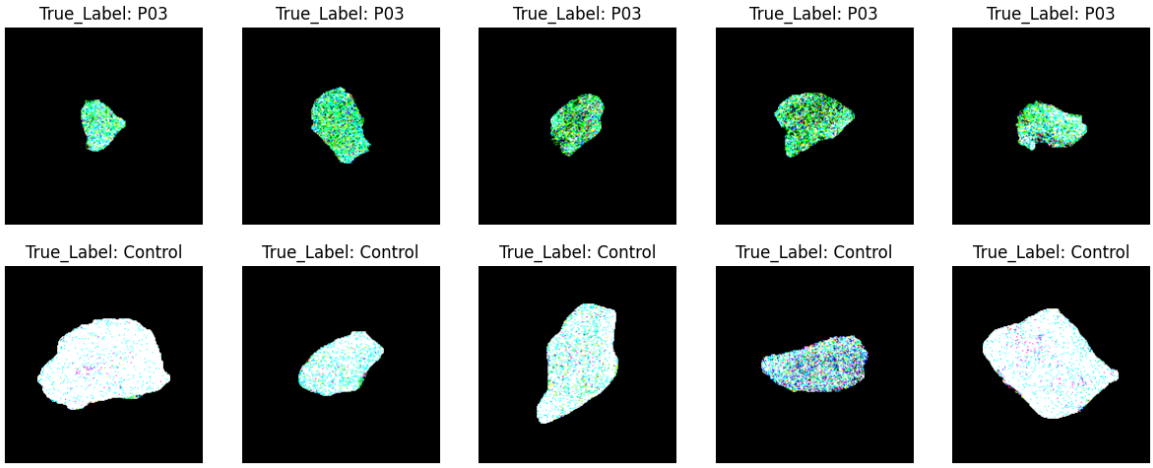

Figure 6.14 RGB image made by weighted stacking of input channels of P03 and control myofibres that had the highest 4 ASV values in figure 6.13, i.e. UqCRC2, NDUFB8, NDUFA13 and COX4+4L2. On the top row are P03 myofibres and on the bottom row are control myofibres.

EM applied to Dystrophin channel model

It was observed in all class models that Dystrophin has high (>90%) individual predictive accuracy. GradientExplainer was applied to a class C Dystrophin model that recorded 98.2% accuracy as reported in Table 6.6.

Observing the SHAP values within the Dystrophin explanation mask reveal that high intensity pixels within myofibres are associated with control myofibre class predictions.

6.5 Experiments and results (SM TS)

Another study to investigate if it is possible to classify multiplex (IMC) SM TS images without segmentation was also conducted. The work was led by me in collaboration with two Masters students: S. Pilla and S. Ramesh.

6.5.1 Data

The data used in this study is the same as that described in Section 6.3.1 with additional TS multiplex IMC images from control subject 'C04', bringing total TS to 14 (10 patients and 4 controls). With this limited data it was realised that class-wise (A, B, C and D) is not possible to train DL models. It was decided a binary classification analysis would be performed between control vs patient classes.

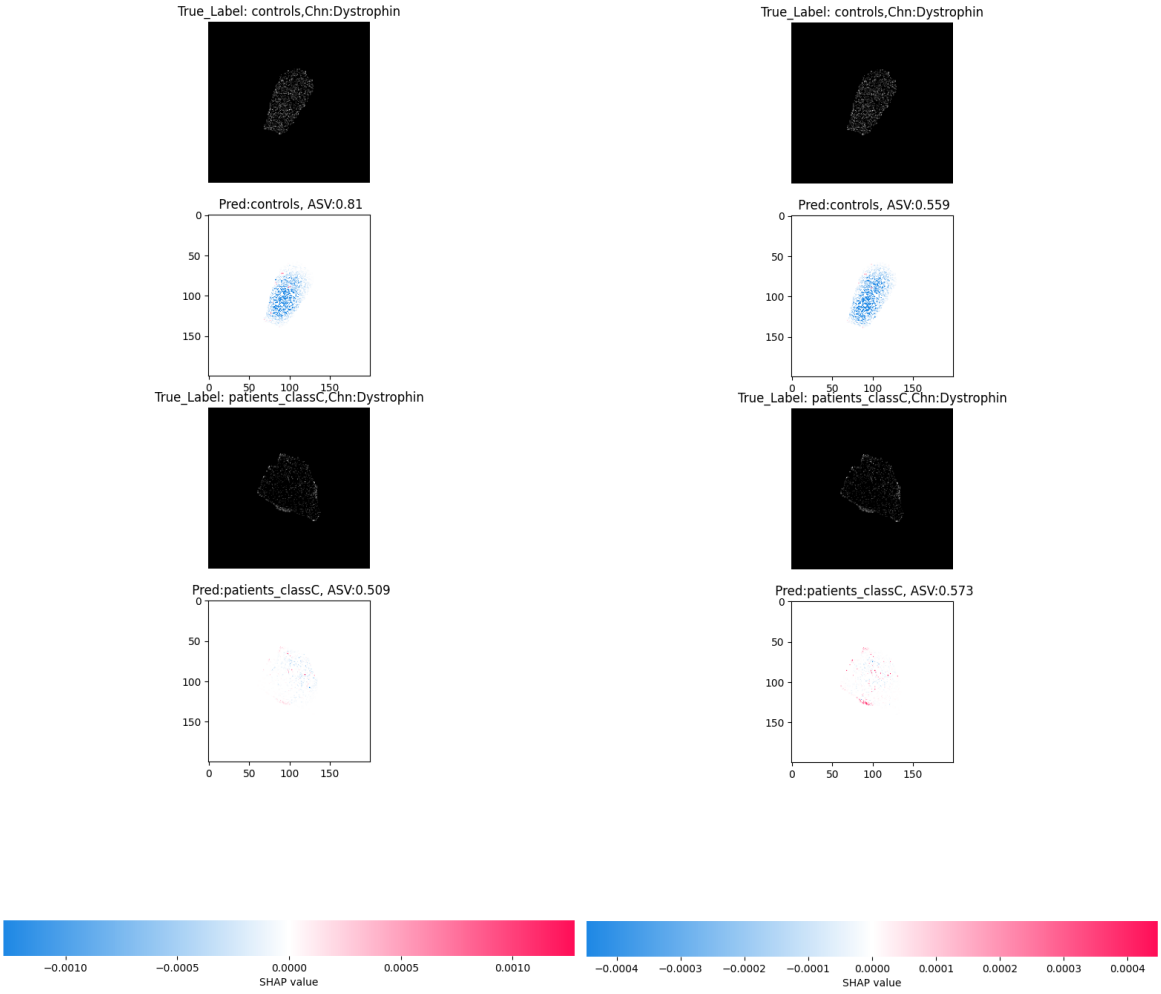

Figure 6.15 GradientExplainer (left) & DeepExplainer (right) explanation masks for the VGG16 model trained on Dystrophin channel of myofibres from class C patient and controls.

Data processing

Ten channels i.e. NDUFB8, NDUFA13, SDHA, UqCRC2, MTC01, COX4+4L2, OSCP, VDAC1, TOM22, Dystrophin were select and all TS muliplex images were split into 512x512 pixel patches. This resulted in 228 control class images and 721 patient class images that were used for training and testing the models.

Experiment design

VGG16 and ResNet50 DL models were used for training using i) all 10 channels with adapted input layers of the models and ii) each channel individually. The data was split into 70%: 15%: 15% for training, validation and testing. A range of EMs such as Gradients [136], DeConvNet [137], Guided Backprop [19], Deep Taylor [140], Input Gradient [131, 138], Layer-wise Relevance propagation [138, 141] were used but due to requirements of the use-case investigated in this thesis, as discussed in Section 6.3.2 GradientExplainer and DeepExplainer are presented.

6.5.2 Results

DL classification of unsegmented IMC data

CNN, VGG16 and ResNet50 models were trained and top 10 results are reported in Table 6.9. The results from CNN model were poorer than other models and so did not make it to the table. In addition to the models mentioned in Table 6.9 a model trained on eight OXPHOS channels that were selected to match the analysis mentioned in Section 2.6.2, recorded a test accuracy of 98% .
The high predictive accuracy of these models prompted us to experiment/train the models four times with different random seeds to split training, validation and test data. The mean accuracy over these four runs still remains high as seen in Table 6.9.

Explainable methods applied to model trained on unsegmented IMC data

A number of EMs were applied as mentioned earlier, to six selected models with high predictive accuracy, most of these were not easy to interpret as it is difficult to generate metrics using them which are equivalent to ASV from SHAP that inform channel importance. So, DeepExplainer and GradientExplainer were used but as these EMs produced similar explanation masks, in this section GradientExplainer explanation masks are presented with ASV, applied to the VGG16 model trained on the patches of unsegmented IMC data consists of eight OXPHOS protein channels.

Table 6.9 Unsegmented data model ranking: models trained on various channels ordered by mean test accuracy over 4 different training runs which were distinguished by random seeds. * Chn, TA, RS, SD and VAR stands for Channel, Test Accuracy, Random Seed, Standard Deviation and Variance respectively. The green cells highlights models where mean TA exceeded 90%.

Model	Chn	TA(%)RS-A	TA(%)RS-B	TA(%)RS-C	TA(%)RS-D	Mean TA	SD TA	Var TA
VGG16	10 chns	100	98.95	98.95	98.95	99.21	0.52	0.27
ResNet50	UqCRC2	100	92.86	100	96.43	97.32	3.41	11.68
ResNet50	NFUFA13	100	92.86	92.86	96.43	95.53	3.41	11.68
ResNet50	Dystrophin	96.43	96.43	92.86	92.86	94.64	2.06	4.24
ResNet50	OSCP	92.86	96.43	92.86	92.86	93.75	1.78	3.18
ResNet50	COX4	100	96.43	85.71	89.29	92.85	6.52	42.53
ResNet50	SDHA	89.29	82.14	92.86	96.43	90.18	6.10	37.22
ResNet50	NDUFB8	85.71	85.71	92.86	85.71	87.49	3.57	12.78
ResNet50	VDAC1	85.71	78.57	89.29	85.71	84.82	4.49	20.20
ResNet50	TOM22	71.43	85.71	89.29	85.71	83.03	7.91	62.70
ResNet50	MTCO1	67.86	82.14	75	92.86	79.46	10.66	113.73

ASV values were used to identify the channel importance/attribution and pixel colours in the explanation masks represent SHAP values. As seen in Figure 6.16 explanation masks for the model report the high ASV values for channels NDUFB8, OSCP, UqCRC2 and SDHA. Unlike a clean trend of negative correlation of SHAP values between the two binary classes that exist in the analysis of segmented myofibre, in unsegmented TS images it is difficult to find such trends. As seen in Figure 6.16 the SHAP values looks similarly spread in both the explanation masks representing patient and control TS patch. This makes extraction of insights from these explanation masks difficult.

6.6 Discussions

6.6.1 DL models classification

The predictive accuracy of DL models for all five cases i.e. Class A, B, C & D, and P03 ranged from 98% for P03 to 100% for class A myofibres. This surpasses both the ML analysis performed in Chapter 3 which ranged from 93% to 100%, and analysis conducted using existing tools discussed in Section 2.6.2 which had classification accuracy exceeding 90% for one of the five cases mentioned above. It is clear that DL models are leveraging some differential features that are available in the raw multiplex data which is accessible to these models because of their convolution architecture that allows these models to detect beyond linear relationships. The high predictive accuracy of these models make these eligible

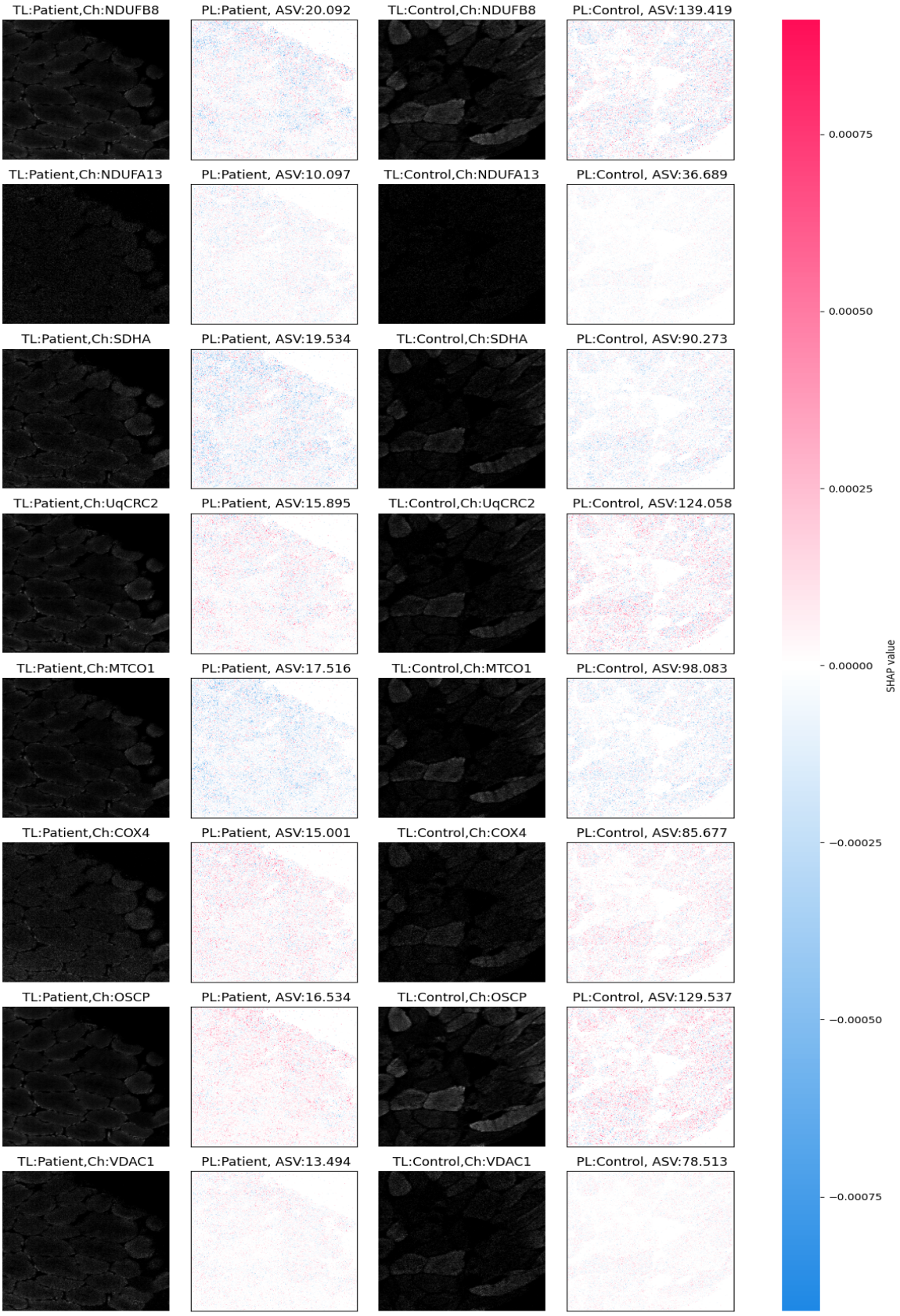

Figure 6.16 GradientExplainer applied to VGG16 model trained on 10 channels of myofibres from patient and control classes.

candidates to apply EMs that can reveal the basis of predictions of these models, that can be used to derive predictive inferences and validate that they are giving the right answer for the right reason.

Individual channel models Across all five cases it is observed that models trained on individual channels related to myofibre morphology i.e. Dystrophin (myofibre membrane marker), TOM22 and VDAC1 (both surrogates for myofibre mass) recorded high predictive accuracy relative to other channels, the highest model trained to predict class C myofibres using only Dystrophin achieved predictive accuracy of 98.2%. The use of myocytoML for segmentation results in 'analysable' myofibres that should be of uniform quality that limits morphological artifacts creeping in due to user bias. Nevertheless this is an interesting finding.

6.6.2 EM to profile myofibres

It is clear the EMs selected i.e. GradientExplainer and DeepExplainer are useful and faithful, as evidenced by the similarity of explanation masks produce by these two EMs. These explanation masks revealed the 'proportional' importance of each channel in terms of ASV towards a class prediction, and SHAP value of pixels revealed the trend linking pixel intensities to class predictions. This allowed in some cases, e.g. class C, identification of spatial features within myofibre that are important for model prediction. The channel-wise importance identified using ASV was useful in profiling four out five cases of myofibres analysed in this thesis. These profiles constructed by combining the four channels with highest ASV into a weighted RGB image is a useful finding. This essentially allows us to define a class of myofibre in terms of a threshold range (of colour) that is composed of a combination of thresholds of associated channels. It should be noted that this is not a unique solution i.e. there may be many other associations between channels that might produce similar differential colour thresholds. When compared to explanations derived in chapter 3 the explanations derived using DL models and myofibre images divulge inter myofibre features' associations with mutations.

EM on Dystrophin model The application of GradientExplainer to the class C VGG16 model trained only with Dystrophin reveals that the differential feature leveraged by the models seems to be the relatively elevated presence of Dystrophin within myofibre in control class myofibres than class C myofibres. Dystrophin is usually present in the membrane, the elevated presence of Dystrophin inside the myofibre is an interesting finding that might be linked to differences in processing tissue from controls and class C patients. It is noted

that due to scarcity of control tissue, it undergoes relatively more rounds of freeze/thaw than patients as it is used as control in many studies. This might introduce some artefactual spreading of Dystrophin, however this is a hypothesis that needs a separate experimental validation.

Validation of predictive inference insights The main purpose of the explainable DL pipeline developed in this chapter is to discover novel spatial and channel-wise associations that can help understand mitochondrial disease pathology. To this end, the pipeline discovers various associations that are presented as predictive inference insights. These insights are essentially the differential patterns in the data that the DL models are leveraging to classify myofibres with good accuracy i.e. >98% across all classes. But these differential patterns can be novel or artefactual associations, and differentiating these requires some validation methods. Biological validation is used to substantiate the predictive inference insights i.e. by comparing these insights to the biologically expected associations/patterns. For this, a class A mutations myofibre case is useful as this class is relatively better understood and provides well defined expected biological patterns. These expected patterns such as downregulation of complex I proteins and compensatory upregulation of complex II-V in class A myofibres have been used to validate the derived class A predictive inference insights. All the insights discovered by explainable DL pipeline for class A myofibres are in line with the biological expectations as discussed in Section 6.4.2. However, for other classes where the pathology is poorly understood, validating the predictive insights requires designing and conducting biomedical experiments which is out of the scope of this thesis.

6.6.3 Unsegmented SM multiplex image classification

The high predictive accuracy attained by models trained on unsegmented patches of multiplex TS is interesting, especially with single-protein models such as UqCRC2, as there are no biologically known associations for classifying these myofibres. This implies there exist unknown differential features that models are exploiting to predict the class of the patch. However, applying EMs to these models does not reveal many useful predictive insights due to complex explanation masks that point to various combinations of SHAP values for pixels within both classes of SM TS image patches.

6.6.4 Limitations of explainable DL analysis

Data The studies conducted in this chapter have used small cohort of subjects with some classes' models trained with less than 400 myofibres from two subjects. This makes the

analysis vulnerable to subject specific biases. The dataset must be representative of the population. All tissue used in this analysis was processed and imaged at WCMR which exposes it to acquisition of process/machine related artifacts.

Explainable methods The EMs used here essentially show difference from reference. If the reference is not representative of reality then the explanation produced will be flawed. For reference data, a class balanced training dataset was used but with a small dataset it is difficult to guarantee that the reference is representative of reality. The predictive inference was built here by studying small number of myofibres from test data. It should be noted the explanations are locally accurate of the myofibre studied and should not be conflated to represent a typical class of myofibre. It should also be noted that explanation masks are a way to understand model predictions, the predictive insights from these is based on the associations the model used to produce this prediction. This again should not be conflated as evidence of biological associations.

Flawed binary class design of unsegmented IMC data Due to scarcity of images the analysis conducted in the study used unsegmented IMC data for controls vs patients which is not based on biological curiosity.

Limited insights from models trained with unsegmented IMC data The best scenario expectation for this study was explanation masks that reveal dysfunction myofibres within TS. But the explanation masks are hard to interpret and do not lead to derivation of predictive insights.

6.6.5 Scope for improvements and future work

In this thesis a small number of DL models and EM techniques were leveraged to build myofibre profiles. This by no means reveals all possible associations. The current pipeline is script based which may deter biomedical domain users. The actual usefulness of these methods can be fully exploited by the domain users. To enable domain users to leverage DL for finding association this needed to be built as a tool installed on a server with appropriate GPUs.

In this thesis there was no research done on development of methods for validation of the association found using explainable DL/ML pipelines. Similar to significance testing in statistics, there is a need for validation testing specific to predictive insights built using explainable DL.

6.7 Conclusion

This chapter describes the process of building an explainable DL pipeline that has predictive accuracy exceeding 98% across all five myofibre mutation cases. This DL classification pipeline is also applied to unsegmented SM TS images that produce 99% accuracy but for control vs all patients classification. To interrogate these highly accurate predictive DL models various EMs were tried but GradientExplainer and DeepExplainer were selected because these provide both the association between OXPHOS proteins and mutation class explained in terms of correlation (positive/negative) of protein markers toward prediction, and the relative importance/contribution of each of these protein markers quantified. These EMs generated explanation masks that are used to profile myofibres in four out of the five cases studied in this chapter. The utility and limitation of this explainable DL pipeline in the context of discovering mitochondrial disease pathology are also discussed.

Chapter 7

Final Discussion

Mitochondrial diseases are metabolic genetic disorders that can cause severe disabilities and adversely affect the life expectancy of patients [35], their affects are pronounced in high energy demanding cells such as myofibres in the SM tissue. Mitochondrial diseases are currently untreatable due to limited understanding of their pathology which is complex and highly heterogeneous in presentation. One way of studying these diseases is by profiling the affected cells (myofibres) in terms of levels of OXPHOS proteins within the myofibre. IMC allows observation of selected OXPHOS protein markers by imaging the SM tissue as multiplex image data. The analysis of this multiplex data can potentially allow classification and profiling of myofibres affected by mitochondrial disease causing genetic mutations. With previous techniques classification of only one group of genetic mutations was possible, out of the five that were studied. This is due to limitations of the previous techniques that includes: i) inability to analyse (classify) high dimensional multiplex data without employing dimensionality reduction, which essentially ignores features in the dimensions were the reduction is applied; ii) imprecise segmentation of myofibres within the multiplex image data, myofibre segmentation is the fundamental first step upon which the reliability of any analysis conducted depends and this also limits validation of theories such as the existence of differential features in perinuclear regions within myofibre, that require precise myofibre segmentation. The work undertaken in this thesis addresses these limitations with the contributions discussed in the following section.

7.1 Main contributions of this thesis

7.1.1 NCL-SM

NCL-SM, a fully manually segmented dataset of more than 50k myofibres, is developed as part of this thesis. It is a useful resource for developing and evaluating any tool or pipeline that deals with segmentation of SM tissue images, as evident in subsequent development of myocytoML. To the best of my knowledge there is no public dataset of precisely segmented SM TS images, the related dataset such as TissueNet [179] which consists of various tissue segmentation data does not contain annotations required of the use case studied in this thesis. Furthermore, models such as Cellpose trained with just TissueNet produced the segmentation quality that was subpar to Cellpose trained with NCL-SM as presented in Table 5.3.

NCL-SM is not just a dataset, it also describes the issues that must be addressed for reliable myofibre segmentation. It prescribes protocols to filter 'analysable' myofibres, by identifying and removing FAMs, NTMs and FRs from the SM tissue image. It defines the evaluations metrics such as (r_{AoB}),(r_{AiB}) that are specific to the task of single-cell SM tissue image segmentation and curation of 'analysable' myofibres.

It also provides a benchmark duplicate manual annotations to evaluate against. This highlights the level of subjectivity involved in various parts of the annotation process with some tasks such as FR segmentation experiencing a high level of IAV. This helps any developers of new SM segmentation tools to set the expectation and remedies for subjectivity in the annotation process.

NCL-SM is a useful resource but its main limitation is its limited diversity, i.e. it consists of only frozen tissue, from dozens of patients suffering mainly from mitochondrial disease and a wider variety of other neuromuscular disorders, all of whose biopsies were processed in Newcastle. To be more representative NCL-SM should be expanded to include precisely annotated SM tissue images of fixed tissue, from other institutes or centres around the world.

7.1.2 myocytoML

myocytoML is single-cell SM tissue image segmentation pipeline that produces the segmentation and curation of myofibres that virtually meets the quality achieved by the 'gold standard' duplicate manual annotations, as described in Section 5.5. It was developed by recognising the subjectivity that exists in various tasks of SM tissue image segmentation, it addresses this by providing a flexible GUI that allows users to amend all produced masks with ease and speed. The annotation masks produced by myocytoML are accompanied by evaluation metrics that informs the user about the quality of annotations produced. Other

related methods for single-cell segmentation such as MiCAT [77], Steinbock [78], generalised such as cellprofiler [83], DeepCell [88], mitocyto [1] do not address the four CV tasks i.e. myofibre segmentation, NTM classification, FAM classification and FR classification, required for SM tissue image analysis. Furthermore these methods also lack the precising of myofibre segmentation compared to one achieved using myocytoML as discussed in Section 5.5.

myocytoML is a useful tool which is already being used in WCMR, nonetheless, it is a prototype that has limitations. The code of myocytoML is not optimised, this results in large IMC images taking up to 45 minutes to be processed. The installation process of myocytoML can be improved by packaging it as a Napari plugin.

7.1.3 Explainable machine learning analysis of multiplex image data: raw segmented myofibre images; statistical summaries per myofibre

Explainable methods applied to ML and DL models are potent tools to classify and profile myofibres, evident in the research conducted in Chapters 3 and 6. The classification accuracy exceeds 98% across all cases studied using DL on raw segmented multiplex images and 93% for the same using ML on per myofibre statistical summaries. These are substantial improvements in accuracy compared to the previous method, i.e. plotIMC for which classification accuracy did not exceed 90% for most cases studied. But more importantly these models with high predictive accuracy make them eligible to apply EMs to know their basis of predictions, i.e. there is no point in interrogating an inaccurate model. EMs applied to both ML models trained on per myofibre statistical summaries and DL models trained on raw segmented myofibre images provide predictive inferences which were insightful in profiling myofibres linked to the five cases of genetic mutation studied. EMs based on Shapley values were used as these quantify the contribution of each input feature towards model prediction, this helps ascertain relative importance in both directions i.e. spatially in terms of pixels; channel-wise in terms of OXPHOS proteins. It needs to be emphasised that only a broad case explainable DL was carried out in this thesis, where a high predictive accuracy model trained with selected protein markers of interest is interrogated mainly to understand associations between these proteins and the predicted class, which was further extended to profile myofibres. But there exist many niche and biologically curious use cases of this approach such as constraining the input to small areas within myofibre e.g. perinuclear region and equal area non-perinuclear region, looking for associations between channels, and relative importance of these small areas towards class predictions.

In four out of five cases studied it was possible to profile myofibres in terms of weighted RGB images made by combining the channels with the highest four ASV scores. In these four cases a clear differential colour threshold that can classify classes of myofibre was observed. This threshold can be decomposed in to threshold combinations of the protein markers. When EMs were applied to ML models trained on per myofibre statistical summaries, this revealed the associations between protein markers in SHAP plots. These plots were generated using whole training data and in a sense give a picture of associations across the samples used in training. The combination of explanation masks from DL models that allows observation of individual myofibres and SHAP plots from ML models that allow observation of associations across the population of myofibres, is an informative approach in understanding the association spatially and channel-wise, that can be helpful in understanding the mitochondrial disease pathology in the SM tissue.

While this approach of using EMs has been useful in profiling myofibres in the use cases investigated in this thesis, there are factors that need to be considered. That is, i) the explanations provided by this methods are the association that the ML/DL model is using for accurate predictions, this should not be conflated as biological discovery of this association, ii) the explanations distilled using the approach developed in this thesis are not exhaustive i.e. they pick the most prevalent associations but not all associations the model is using, so the association discovered should not be misunderstood for this to be the unique solution of associations used by the model.

The validation of predictive inference insights were carried out by comparing the insights from explainable ML/DL pipelines to biologically expected associations/patterns. For this, class A myofibres were instrumental, as this class is relatively better understood and provides well defined expected biological patterns for class A myofibres. Comparing the class A predictive insights to biologically expected patterns showed that all insights were in accord with the expected biological associations. But for other classes of mutations for which the pathology is not fully understood, performing such validation is not possible, in these cases the validation requires some sort of biomedical experiments, which require resources and time that might not always be feasible.

The main limitation of this approach is differentiating the novel insights from artefactual insights. ML/DL models are powerful classifiers that try to exploit any available differential pattern in the training data. Especially when the data is high dimensional multiplex images, these patterns can be complex such as a differential feature that is a complex combination of spatial and channel-wise associations. Distilling and decomposing such complex patterns using the explainable ML/DL approach introduced in this thesis require analysis of many individual myofibre cases and building an evidence base hypothesis using biological experi-

mentation. This is time consuming and there may be cases where experimental validation is not possible.

7.2 Future work

7.2.1 myocytoML

As discussed earlier myocytoML is a useful tool that is already being used but it is a prototype. To allow it to be used across the SM image analysis domain it needs to packaged as a Napari plugin. Its design is uniquely suited to be built as a continuous learning platform, i.e. its flexibility where it allows users to amend masks can be leveraged to improve model performance and over time improve the IAV across all four CV tasks. To build myocytoML as a continuous learning platform it needs be resourced with appropriate compute (GPU) and web GUI.

7.2.2 A unified pipeline for multiplex biomedical data

To democratise the use of AI for biomedical discoveries there is a need for a unified pipeline for multiplex biomedical data with appropriate user inference that appeals to the domain users, i.e. biomedical scientists. Specific to the use case studied in this thesis, this will be a combination of myocytoML for segmentation, curation and extraction per myofibre summaries; explainable ML/DL to find spatial and channel-wise associations and to profile myofibres; and traditional statistical analysis tools such as plotIMC, imcRTools. This can be built upon existing applications such as Napari and designed to leverage GPUs. From the experience conducted in this research it was observed that with typical GPUs used in this research such as NVIDIA Tesla V100 (16GB) it takes hours and days for some models to be trained. This needs to be addressed or mitigated by optimising code, employing high end GPUs or design adjustments such that different tasks are executed in parallel.

7.2.3 Validation methods for associations discovered using explainable ML

There are no validation methods specific to the associations discovered using EMs. The standard significance testing or randomised control trial-A/B testing are dependent on strength of observations used, i.e. subjects. For the studies in this thesis that used only 13 subjects across four classes with some classes represented by two subjects, this is not an appropriate validation test.

The purpose of validation tests for the studies discussed in this thesis is to i) differentiate between novel insights and artefactual insights. Artefactual insights are usually the result of artifacts in the data that discriminately affect classes, but it can also be due to nuances in the ML/DL models and explainable methods. The required validation method needs to address the detection of artefactual insights by assigning a confidence score for each predictive insight that takes into account all sources from where artifacts can creep in. ii) Inform the strength of the predictive insight in terms of number of observations that back it up across the whole of the training data. This is an open research question that requires more research and the results of this can benefit ML/DL applications across all domains.

7.3 Conclusion

The research conducted in this thesis introduced a novel approach for understanding mitochondrial disease pathology. The explainable ML/DL analysis of SM tissue images to discover mitochondrial disease pathology has introduced a paradigm shift in the approach to understanding mitochondrial disease pathology using artificial intelligence models. The tremendous potential of ML/DL models in resolving various tasks that help with analysis of SM tissues for understanding their pathology is evident throughout this thesis. This includes addressing the limitations with previous methods, i.e. imprecise segmentation of myofibres within the multiplex image data that affects the reliability of downstream analysis; inability to analyse (classify) high dimensional multiplex data without employing dimensionality reduction. The development of myocytoML allows biomedical scientists to precisely segment and curate myofibres in multiplex SM tissue images. The segmentation quality achieved using myocytoML is close to human-level. Moreover, it provides a convenient GUI that allows users to amend masks. As part of this, NCL-SM, a large dataset of precisely segmented SM tissue images, is released for public use. This benchmark dataset lays foundations for evaluation and training of future myofibre segmentation pipelines. The explainable DL/ML pipelines developed in this thesis were able to classify mutation classes of myofibres, discover associations and profile myofibres using the high dimensional multiplex image data. This is a exciting new approach that opens a multitude of opportunities not just for mitochondrial disease pathology but for biomedical discoveries at large.

Appendix

.1 Predictive inference tables from Chapter 3

Table 1 Insights from explainable LR model trained for class A vs control prediction

Input feature	**SHAP magnitude and direction**	**SHAP values prevalence**	**Predictive power**	**Control ǀ Patients (mean & STD values)**

COX4+4L2	On global model level COX4+4L2 has the greatest magnitude of all, as seen in figure 3.7 B. And the direction of correlation with SHAP values is positive as seen in figure 3.7 C, this implies higher COX4+4L2 values mostly result in positive SHAP values which inturn push the prediction towards positive (class A) class.	As observed in figure 3.7 C it has moderately long tails in both directions with higher density on both extremes. This implies for a majority of predictions its SHAP values (both positive & negative) was a major contributor towards model making a prediction. This is further evident in figure 3.7 A as darker (higher) positive & negative SHAP values relative to other features over most of 852 training instances.	90.52(COX4+4L2); 94.03(COX4+4L2 with VDAC1)	7.46 & 5.66 | 39.63 & 17.1
SDHA	The contribution of SDHA towards model prediction is very similar to COX4+4L2 in all respects albeit slightly reduced in magnitude as observer in figure 3.7 B, C, A.	As observed in the 3.7 C & A again it has similar SHAP values prevalence pattern to COX4+4L2	92.28(SDHA); 96.14(SDHA with VDAC1)	5.04 & 2.99 | 23.54 & 8.27

OSCP	The contribution of OSCP towards model predictions is similar to COX4+4L2 & SDHA in all respects albeit reduced in magnitude as observer in figure 3.7 B, C & A.	As observed in figure 3.7 C & A again it has similar SHAP values prevalence pattern to COX4+4L2 & SDHA	89.82(OSCP); 93.68(OSCP with VDAC1)	3.42 & 2.22 | 15.35 & 5.90
NDUFA13	The global level contribution of NDUFA13 towards the model prediction is the 4th highest out of the 8 features as seen in figure 3.7 B. And as seen in figure 3.7 C the direction of correlation with SHAP values is negative, this implies higher NDUFA13 values mostly result in negative SHAP values which in turn push the prediction towards negative (control) class.	As observed in figure 3.7 C it has a high density of SHAP values that are marginally greater than zero but a very long left tail. Which implies for most predictions its contribution to model prediction is not the most consequential but for some predictions higher values of NDUFA13 lead to a very high negative SHAP values, that are very consequential towards predicting them as negative (control) class, as evident in the figure 3.8 A.	79.65 (NDUFA13); 98.95(NDUFA13 with VDAC1	3.00 & 1.75 |1.7 & 0.25

NDUFB8	The contribution of NDUFB8 towards model prediction is very similar to NDUFA13 in all respects albeit slightly reduced in magnitude as observer in figure 3.7 B, C & A.	As observed in figure 3.7C & A it has similar SHAP values prevalence pattern to NDUFA13.	92.28 (NDUFB8); 100 (NDUFB8 with VDAC1)	2.94 &1.67|1.22 & 0.06
UqCRC2	As observed in figure 3.7 B, C & A, its contributions towards the model predictions are modest and slight positively correlated to SHAP values	As observed in the figure 3.7 C the highest density of SHAP values of UqCRC2 is marginally less than zero and no extreme values, which implies for all predictions its contribution towards model prediction is not consequential.	92.28 (UqCRC2); 94.03 (UqCRC2 with VDAC1)	4.38 & 3.01 |21.96 & 8.53
MTCO1	As observed in figure 3.7 B, C & A, its contributions towards the model predictions are very modest.	As observed in figure 3.7 C, the highest density of SHAP values for MTCO1 is marginally greater than zero and no extreme values which implies for all predictions its contribution towards model prediction is not consequential.	88.07 (MTCO1); 90.88 (MTCO1 with VDAC1)	1.97 & 0.88 | 5.68 & 2.32

VDAC1	As observed in figures 3.7 B, C & A, its contributions towards the model predictions are very modest.	As observed in the figure 3.7 C, the highest density of SHAP values for VDAC1 is near zero with no extreme values which implies for all predictions its contribution towards model prediction is not consequential.	79.65 (VDAC1)	3.45 & 1.24 | 1.86 & 0.68

Table 2 Insights from explainable XGB model trained for class A vs control prediction

Input feature	**SHAP magnitude and direction**	**SHAP values prevalence**	**Predictive power**	**Control | Patients (mean & STD values)**

NDUFB8	The contribution of NDUFB8 towards model prediction has the greatest magnitude of all, as seen in the figure 3.9 B. And the direction of correlation with SHAP values is negative as seen in figure 3.9 C, this implies higher NDUFB8 values mostly result in negative SHAP values which in turn push the prediction towards negative (control) class and vice versa.	As observed in figure 3.9 C it has very long tails in both directions with higher density on both extremes. This implies for a majority of predictions its SHAP values (both positive & negative) was major contributor towards model predictions. This is further evident in figure 3.9 A: as darker (higher) positive & negative SHAP values relative to other features over most of the training instances. Also evident in figures figure 3.10 A & B.	75.15 (NDUFB8) ;99.65(NDUFB8 with VDAC1)	2.94 &1.67|1.22 & 0.06

SDHA	The contribution of SDHA towards model prediction has the second greatest magnitude of all but considerably smaller than NDUFB8, as seen in the figure 3.9 B. The direction of correlation with SHAP values is positive as seen in the figure 3.9 C.	As observed in the figure 3.9 B. The direction of correlation with SHAP values is positive as seen in the figure 3.9 C it has very long tails in both directions with higher density on right extreme. This implies for a majority of predictions its SHAP values (both positive & negative but more so positive) was a major contributor towards model predictions. This is further evident in the figure 3.9 A as darker (higher) positive & negative SHAP values relative to all other features except NDUFB8 over most of the instances.	74.12 (SDHA); 96.14 (SDHA with VDAC1)	5.04 & 2.99 | 23.54 & 8.27

COX4+4L2	As observed in the figure 3.9 B, C & A, COX4+4L2 contributions towards the model predictions are modest and slight positively correlated to SHAP values	As observed in the figure 3.9 C, the highest density of SHAP values of COX4+4L2 is around zero and there are not many extreme values, which implies for most predictions its contribution was not consequential.	74.74 (COX4+4L2); 96.84 (COX4+4L2 with VDAC1)	7.46 & 5.66 | 39.63 & 17.1
NDUFA13	The contribution of NDUFA13 towards model prediction is similar to NDUFB8 in all respects albeit reduced in magnitude as observed in the figure 3.9 B, C & A.	As observed in the figure 3.9 C & A NDUFA13 has slightly similar SHAP value prevalence to NDUFB8 but its densities are much closer toward zero this implies it has modest contribution compare to the top 2 features towards the model predictions.	72.88 (NDUFA13); 99.65(NDUFA13)	3.00 & 1.75 |1.7 & 0.25

UqCRC2	As observed in the figure 3.9 B, C & A, its contributions towards the model predictions are modest and slight positively correlated to SHAP values.	As observed in the figure 3.9 C, the highest density of SHAP values of UqCRC2 is around zero and not many extreme values, which implies for most predictions its contribution was not consequential.	74.32 (UqCRC2); 99.65 (UqCRC2 with VDAC1)	4.38 & 3.01 |21.96 & 8.53
OSCP	As observed in the figure 3.9 B, C & A, its contributions towards the model predictions are very modest.	As observed in the figure 3.9 C, the highest density of SHAP values for OSCP is around zero with no extreme values, which implies its contribution towards model predictions is not consequential.	78.05(OSCP); 96.50(with VDAC1)	3.42 & 2.22 | 15.35 & 5.90
VDAC1	As observed in the figure 3.9 B, C & A, its contributions towards the model predictions are very modest.	As observed in the figure 3.9 C, the highest density of SHAP values for VDAC1 is around zero with no extreme values which implies its contribution towards model predictions is not consequential.	76.60 (VDAC1)	3.45 & 1.24 | 1.86 & 0.68

MTCO1	As observed in the figure 3.9 B, C & A, its contributions towards the model predictions are very modest.	As observed in the figure 3.9 C, the highest density of SHAP values for MTCO1 is around zero and no extreme values which implies its contribution towards model predictions is not consequential.	74.53 (MTCO1); 91.93 (MTCO1 with VDAC1)	1.97 & 0.88 ǀ 5.68 & 2.32

Table 3 Insights from explainable LR model trained for class B vs control prediction

Input feature	**SHAP magnitude and direction**	**SHAP values prevalence**	**Predictive power**	**Control ǀ Patients (mean & STD values)**

UqCRC2	UqCRC2 has the greatest mean SHAP values over all predictions as seen in the figure 3.12 (B). As seen in the figure 3.12 (C) its direction of correlation with SHAP values is positive.	As observed in the figure 3.12 (C) it has very long tails in both directions with higher density around zero. This implies for a good proportion of predictions its SHAP values made modest contribution towards model predictions but very long and thick tails implies for another good number of predictions its contribution were consequential for both classes . This is further evident in the figure 3.12 (A) as there exist mix of darker (higher) and lighter(lower) SHAP values.	63.41 (UqCRC2) ; 69.90(UqCRC2 with VDAC1)	4.38 & 3.01 | 7.83 & 4.53

COX4+4L2	The contribution of COX4+4L2 towards model prediction has the second greatest magnitude of all as seen in the figure 3.12 (B). And the direction of correlation with SHAP values is positive as seen in the figure 3.12 (C).	As observed in figures 3.12 (C) & (A) it has similar SHAP values prevalence pattern to UqCRC2 but with reduced magnitude.	58.62 (COX4+4L2); 59.22 (COX4+4L2 with VDAC1)	7.46 & 5.66 | 11.66 & 7.25
MTCO1	The contribution of MTCO1 is roughly inverse to COX4+4L2 in direction but with reduced magnitude as seen in figures 3.12 (B) & (C).	As observed in figures 3.12 (C) & (A) it has roughly similar prevalence SHAP values pattern to COX4+4L2 but in reverse. The inverse relationship can also be observed in the figure 3.13 (B).	51.15 (MTCO1); 50.05 (MTCO1 with VDAC1)	1.97 & 0.88 | 2.20 & 0.83

NDUFA13	The contribution of NDUFA13 is similar to MTCO1 in direction but with reduced magnitude.	As observed in figures 3.12 (C) & (A) it has long left tail and higher density around zero, this implies for majority of predictions its contributions were not great but for some negative (control) it has consequential contribution towards model predictions.	57.03 (NDUFA13); 53.54(NDUFA13)	3.00 & 1.75 | 2.93 & 1.75
NDUFB8	The contribution of NDUFB8 is similar to NDUFA13 in both magnitude and direction.	As observed in figures 3.12 (C) & (A) it has similar prevalence SHAP values pattern to NDUFA13	61.81(NDUFB8); 64.31 (NDUFB8 with VDAC1)	2.94 &1.67| 2.50 & 1.41
SDHA	As observed in figures 3.12 (B), (C) & (A), its contributions towards the model predictions are very modest.	As observed in the figure 3.12 (C), the highest density of SHAP values for SDHA is around zero and no extreme values which implies its contribution towards model prediction is not consequential.	53.04(SDHA); 55.53(SDHA with VDAC1)	5.04 & 2.99 | 8.64 &6.41

OSCP	As observed in figures 3.12 (B), (C) & (A), its contributions towards the model predictions are very modest.	As observed in the figure 3.12 (C), the highest density of SHAP values for OSCP is around zero and no extreme values which implies its contribution towards model prediction is not consequential.	52.34 (OSCP); 55.83 (OSCP with VDAC1)	3.42 & 2.22 | 5.88 & 3.23
VDAC1	As observed in figures 3.12 (B), (C) & (A), its contributions towards the model predictions are very modest.	As observed in the figure 3.12 (C), the highest density of SHAP values for VDAC1 is around zero and no extreme values which implies its contribution towards model prediction is not consequential.	49.85 (VDAC1)	3.45 & 1.24 | 2.15 & 0.75

Table 4 Insights from explainable XGB model trained for class B vs control predictions

Input feature	**SHAP magnitude and direction**	**SHAP values prevalence**	**Predictive power**	**Control | Patients (mean & STD values)**

UqCRC2	UqCRC2 has the greatest mean SHAP value over all predictions as seen in the figure 3.14 (B). As seen in the figure 3.14 (C) its direction of correlation with SHAP values is positive.	As observed in the figure 3.14 (C) it has long tails but longer still on left with higher density around [0-4]. This implies for a majority of predictions its SHAP values made decent contribution towards model predictions but very long left tail implies for a good number of predictions its contribution were consequential for negative (control) class . This is further evident in the figure 3.14 (A) as there exist darker (higher) SHAP values compare to other features.	63.41 (UqCRC2) ; 69.90(UqCRC2 with VDAC1)	4.38 & 3.01 | 7.83 & 4.53

NDUFB8	NDUFB8 has second biggest mean SHAP value over all predictions as seen in the figure 3.14 (B). As seen in the figure 3.14 (C) its direction of correlation with SHAP values is negative.	As observed in the figure 3.14 (C) it has it a moderate tails with highest density around [0-5] & -2. This implies its contributions for predicting positive (class B) class were decent but had moderate contributions toward predicting negative (control) class.	63.41 (NDUFB8) ; 69.90(NDUFB8 with VDAC1)	2.94 &1.67| 2.50 & 1.41
COX4+4L2	COX4+4L2 has third biggest mean SHAP value over all predictions as seen in the figure 3.14 (B). As seen in the figure 3.14 (C) its direction with SHAP values is positive.	As observed in the figure 3.14 (C) it has similar SHAP value prevalence patterns to NDUB8 but in reverse.	63.41 (COX4+4L2) ; 69.90(COX4+4L2 with VDAC1)	7.46 & 5.66 | 11.66 & 7.25
NDUFA13	NDUFA13 has same correlation with SHAP values as NDUFB8.	As observed in the figure 3.14 (C) it has similar SHAP value prevalence patterns to NDUB8 with reduced magnitude.	63.41 (NDUFA13) ; 69.90(NDUFA13 with VDAC1)	3.00 & 1.75 | 2.93 & 1.75

MTCO1	The contribution of MTCO1 is roughly inverse to COX4+4L2 in direction but with reduced magnitude as seen in figures 3.14(B) & (C).	As observed in figures 3.14 (C) & (A) it has roughly similar prevalence SHAP values pattern to COX4+4L2 but in reverse, except it has a longer left tail. This implies it affected the model decision more consequentially for some negative (control) class.	63.41 (MTCO1) ; 69.90(MTCO1 with VDAC1)	1.97 & 0.88 | 2.20 & 0.83
VDAC1	As observed in figures 3.14 (B), (C) & (A), its contributions towards the model predictions are very modest.	As observed in the figure 3.14 (C), the highest density of SHAP values for VDAC1 is around zero and no extreme values which implies its contribution towards model prediction is not consequential.	63.41 (VDAC1)	3.45 & 1.24 | 2.15 & 0.75

SDHA	As observed in figures 3.14 (B), (C) & (A), its contributions towards the model predictions are very modest.	As observed in the figure 3.14 (C), the highest density of SHAP values for SDHA is around zero and no extreme values which implies its contribution towards model prediction is not consequential.	63.41 (SDHA) ; 69.90(SDHA with VDAC1)	5.04 & 2.99 | 8.64 &6.41
OSCP	As observed in figures 3.14 (B), (C) & (A), its contributions towards the model predictions are very modest.	As observed in the figure 3.14 (C), the highest density of SHAP values for OSCP is around zero and no extreme values which implies its contribution towards model prediction is not consequential.	63.41 (OSCP) ; 69.90(OSCP with VDAC1)	3.42 & 2.22 | 5.88 & 3.23

Table 5 Insights from explainable LR model trained for class C vs control prediction

Input feature	**SHAP magnitude and direction**	**SHAP values prevalence**	**Predictive power**	**Control | Patients (mean & STD values)**

COX4+4L2	COX4+4L2 has the greatest mean SHAP values over all predictions as seen in the figure 3.17 (B). As seen in the figure 3.17 (C) its direction of correlation with SHAP values is positive.	As observed in the figure 3.17 (C) it has very long tails in both directions with higher density around -0.2. This implies for a good proportion of predictions its SHAP values made average contribution towards model predictions but very long tails especially right one, implies for some predictions its contribution were consequential for both classes but more so for positive (class C) class . This is further evident in the figure 3.17 (A)as there exist mix of darker (higher) and lighter(lower) SHAP values.	44.31(COX4+4L2) ;65.60(COX4+4L2 with VDAC1)	7.46 & 5.66 | 8.42 & 6.47

SDHA	The contribution of SDHA towards model prediction has the second greatest magnitude of all but smaller than COX4+4L2, as seen in the figure 3.17(B). The direction of correlation with SHAP values is positive as seen in the figure 3.17 (C).	As observed in the figure 3.17 (C)it has long left tail with higher density around 0.2. This implies for most prediction its contributions were average especially for positive (class C) class. But long left tail implies its contribution for some negative (control) predictions were consequential.	71.9(SDHA) ; 71.23(SDHA with VDAC1)	5.04 & 2.99 |18.30 & 14.82
MTCO1	The contribution of MTCO1 is roughly inverse to COX4+4L2 in direction but with reduced magnitude as seen in figure 3.17 (B) & (C).	As observed in figures 3.17 (C) & (A) it has roughly similar prevalence SHAP values pattern to COX4+4L2 but in reverse. The inverse relationship can also be observed in the figure 3.18 (A).	63.55(MTCO1) ; 71.72(MTCO1 with VDAC1)	1.97 & 0.88 | 1.67 & 0.83

NDUFA13	The contribution of NDUFA13 is similar to MTCO1 in direction but with reduced magnitude.	As observed in figures 3.17 (C) & (A) it has long left tail and higher density around 0.1, this implies for majority of predictions its contributions were modest but for some negative (control) class it has consequential contribution towards model predictions.	65.74(NDUFA13) ; 74.20(NDUFA13 with VDAC1)	3.00 & 1.75 | 2.35 & 1.46
VDAC1	As observed in figures 3.17 (B),(C) & (A), its contributions towards the model predictions are very modest.	As observed in the figure 3.17 (C), the highest density of SHAP values for VDAC1 is around 0.1 and no extreme values which implies its contribution towards model predictions were modest.	66.47(VDAC1)	3.45 & 1.24 | 3.38 & 2.00
NDUFB8	The contribution of NDUFB8 is similar to NFUFA13 in direction but with reduced magnitude.	As observed in the figure 3.17 (C), it has similar SHAP value prevalence pattern to NDUFB8	67.93(NDUFB8) ; 77.7(NDUFB8 with VDAC1)	2.94 &1.67| 2.03 & 1.31

UqCRC2	As observed in figures 3.17 (B),(C) & (A), its contributions towards the model predictions are very modest.	As observed in the figure 3.17 (C), it has roughly similar SHAP value prevalence pattern to NDUFB8	56.70(UqCRC2) ; 69.83(UqCRC2 with VDAC1)	4.38 & 3.01 | 7.13 & 5.36
OSCP	As observed in figures 3.17 (B),(C) & (A), its contributions towards the model predictions are very modest.	As observed in the figure 3.17 (C), the highest density of SHAP values for OSCP is around zero and no extreme values which implies its contribution towards model predictions were modest.	67.05(OSCP) ; 67.5(OSCP with VDAC1)	3.42 & 2.22 | 10.91 & 8.61

Table 6 Insights from explainable XGB model trained for class C vs control prediction

Input feature	**SHAP magnitude and direction**	**SHAP values prevalence**	**Predictive power**	**Control | Patients (mean & STD values)**

NDUFB8	NDUFB8 has the greatest mean SHAP values over all predictions as seen in the figure 3.19 (B). As seen in the figure 3.19 (C) its direction of correlation with SHAP values is negative.	As observed in the figure 3.19 (C)it has long tails with higher density around [3-4]. This implies for a majority of predictions its SHAP values made decent contribution towards model predictions were consequential for both (control) classes but more so for positive (class C) class . This is further evident in the figure 3.19 (A) as there exist darker (higher) SHAP values compare to other features.	79.73 (NDUFB8) ;85.13(NDUFB8 with VDAC1)	2.94 &1.67| 2.03 & 1.31

SDHA	The contribution of SDHA towards model prediction has the second greatest magnitude of all but smaller than NDUFB8, as seen in the figure 3.19 (B). The direction of correlation with SHAP values is roughly positive as seen in the figure 3.19 (C).	As observed in the figure 3.19 (C)it has long left tail with higher density around 2. This implies for most prediction its contributions were decent for positive (class C) class. But long left tail implies its contribution for some negative (control) predictions were consequential. This is further evident in the figure 3.20	86.3 (SDHA) ;85.13(SDHA with VDAC1)	5.04 & 2.99 |18.30 & 14.82

COX4+4L2	COX4+4L2 has the third greatest mean SHAP values over all predictions as seen in fig B. As seen in the figure 3.19 (C) its direction of correlation with SHAP values is positive.	As observed in the figure 3.19 (C) it has very long tails in both directions with higher density around [-2 to 1]. This implies for a good proportion of predictions its SHAP values made average contribution towards model predictions but very long tails implies for some predictions its contribution were consequential for both classes . This is further evident in the figure 3.19 (A) as there exist mix of darker (higher) and lighter(lower) SHAP values.	80.03 (COX4+4L2) ; 84.69(COX4+4L2 with VDAC1)	7.46 & 5.66 | 8.42 & 6.47
NDUFA13	The direction of correlation of NDUFA13 with Shap values seems to be negative.	As observed in the figure 3.19 (C) , it has tails in both directions with higher density around 1.2. This implies for most model predictions its contributions were average.	80.03 (NDUFA13) ;84.99(NDUFA13 with VDAC1)	3.00 & 1.75 | 2.35 & 1.46

MTCO1	The contribution of MTCO1 is roughly negative in direction as seen in figures 3.19 (B) &(C).	As observed in figures 3.19 (C) & (A) it has roughly similar SHAP values prevalence pattern to NDUFA13.	80.17 (MTCO1) ; 85.57(MTCO1 with VDAC1)	1.97 & 0.88 | 1.67 & 0.83
UqCRC2	As observed in figures 3.19 (B), (C) & (A), its contributions towards the model predictions are modest.	As observed in the figure 3.19 (C), it has roughly similar SHAP value prevalence pattern to MTCO1 but compressed towards zero.	83.09 (UqCRC2) ; 86.00(UqCRC2 with VDAC1)	4.38 & 3.01 | 7.13 & 5.36
OSCP	As observed in figures 3.19 (B), (C) & (A), its contributions towards the model predictions are modest.	As observed in the figure 3.19 (C), it has roughly similar SHAP value prevalence pattern to UqCRC2	83.97(OSCP) ; 86.15(OSCP with VDAC1)	3.42 & 2.22 | 10.91 & 8.61
VDAC1	As observed in figures 3.19 (B), (C) & (A), its contributions towards the model predictions are modest.	As observed in the figure 3.19 (C), the highest density of SHAP values for VDAC1 is around 1 and a left tail which implies its contribution towards model predictions were modest except in some prediction of negative (control) class.	84.99 (VDAC1)	3.45 & 1.24 | 3.38 & 2.00

Table 7 Insights from explainable XGB model trained for class D vs control prediction

Input feature	**SHAP magnitude and direction**	**SHAP values prevalence**	**Predictive power**	**Control | Patients (mean & STD values)**
UqCRC2	UqCRC2 has the greatest mean SHAP values over all predictions as seen in the figure 3.22 (B). As seen in the figure 3.22 (C) its direction of correlation with SHAP values is positive.	As observed in the figure 3.22 (C) it has long tails with higher density around [2-4]. This implies for a majority of predictions its SHAP values made decent contribution towards model predictions for both the classes .	79.34 (UqCRC2) ; 82.48(UqCRC2 with VDAC1)	4.38 & 3.01 | 5.36 & 2.97
COX4+4L2	The contribution of COX4+4L2 towards model prediction has the second greatest magnitude of all as seen in the figure 3.22 (B). The direction of correlation with SHAP values is positive as seen in the figure 3.22 (C).	As observed in the figure 3.22 (C) it has long right tail with higher density at right extreme and around zero. This implies for some prediction its contributions were modest for negative (control) class. But long right tail implies its contribution for some positive (class D) predictions were consequential.	79.05 (COX4+4L2) ; 80.91(COX4+4L2 with VDAC1)	7.46 & 5.66 | 9.88 & 7.14

OSCP	OSCP has the third greatest mean SHAP values over all predictions as seen in the figure 3.22 (B). As seen in the figure 3.22 (C) its direction of correlation with SHAP values is negative.	As observed in the figure 3.22 (C) it has tails in both directions with higher density around [-2 to -1]. This implies for a good proportion of predictions its SHAP values made average contribution towards model predictions but tails implies for some predictions its contribution were consequential for both classes . This is further evident in the figure 3.22 (A) as there exist mix of darker (higher) and lighter(lower) SHAP values.	78.34 (OSCP) ; 81.05(OSCP with VDAC1)	3.42 & 2.22 | 3.75 & 2.17
SDHA	As seen in the figure 3.22 (C) the direction of correlation of SDHA with Shap values is inconclusive.	As observed in the figure 3.22 (C) , it has long left tail with higher density around zero. This implies for most model predictions its contributions were modest but for some negative (control) class predictions it was consequential.	79.06 (SDHA) ; 82.47(SDHA with VDAC1)	5.04 & 2.99 | 6.14 & 3.54

NDUFB8	The direction of correlation of NDUFB8 with SHAP values is inconclusive .	As observed in figures 3.22 (C) & (A) it has long left tail with high density around 1. this implies modest contribution for most predictions but for some negative prediction its contributions was consequential.	82.05(NDUFB8) ; 82.34(NDUFB8 with VDAC1)	2.94 &1.67| 2.43 & 0.68
MTCO1	As observed in figures 3.22 (C) & (A), its contributions towards the model predictions are similar to NDUFB8 but in reverse.	As observed in the figure 3.22 (C) , it has roughly similar SHAP value prevalence pattern to NDUFB8 but in reverse.	79.06 (MTCO1) ; 80.34(MTCO1 with VDAC1)	1.97 & 0.88 | 2.10 & 0.83
VDAC1	As observed in figures 3.22 (B),(C) & (A), its contributions towards the model predictions are similar to NDUFB8	As observed in the figure 3.22 (C), it has roughly similar SHAP value prevalence pattern to NDUFB8	81.48(VDAC1)	3.45 & 1.24 | 1.82 & 0.46
NDUFA13	As observed in figures 3.22 (B),(C) & (A), its contributions towards the model predictions are very modest.	As observed in the figure 3.22 (C), the highest density of SHAP values for NDUFB13 is around 1 which implies its contribution towards model predictions were modest for both classes.	80.20 (NDUFA13) ; 80.20(NDUFA13 with VDAC1)	3.00 & 1.75 | 3.04 & 1.29

Table 8 Insights from explainable XGB model trained for P03 vs control prediction

Input feature	**SHAP magnitude and direction**	**SHAP values prevalence**	**Predictive power**	**Control | Patients (mean & STD values)**
OSCP	OSCP has the greatest mean SHAP values over all predictions as seen in the figure 3.25 (B). As seen in the figure 3.25 (C) its direction of correlation with SHAP values is negative.	As observed in the figure 3.25 (C) it has long left tail with higher density at around 2. This implies for most its contribution towards the prediction is average but for some (mostly) negative (control) class predictions, its contribution is consequential.	78.05 (OSCP) ; 75.57(OSCP with VDAC1)	3.42 & 2.22 | 2.22 & 0.52
UqCRC2	The contribution of UqCRC2 towards model prediction has the second greatest magnitude of all as seen in the figure 3.25 (B). The direction of correlation with SHAP values is roughly positive as seen in the figure 3.25 (C).	As observed in the figure 3.25 (C) it has long tails with higher density around [2-4]. This implies for a majority of predictions its SHAP values made decent contribution towards model predictions for both classes .	74.33 (UqCRC2) ; 76.60(UqCRC2 with VDAC1)	4.38 & 3.01 |3.30 & 0.92

SDHA	As seen in the figure 3.25 (C) the direction of correlation of SDHA with Shap values is roughly negative.	As observed in the figure 3.25 (C) it has long left tail with higher density around 0. This implies for a majority of predictions its SHAP values made modest contribution towards model predictions . But for some negative (control) class predictions its contribution was consequential.	74.12 (SDHA) ;76.81(SDHA with VDAC1)	5.04 & 2.99 | 3.75 & 1.13
VDAC1	As seen in figure 3.25 (B) & (C) the direction of correlation of VDAC1 with Shap values is roughly negative with similar magnitude to SDHA.	As observed in the figure 3.25 (C) it has left tail with higher density around 1. This implies for a majority of predictions its SHAP values made modest contribution towards model predictions . But for some negative (control) class predictions its contribution was consequential.	76.60(VDAC1)	3.45 & 1.24 | 1.51 & 0.22

MTCO1	As seen in the figure 3.25 (C) the direction of correlation of MTCO1 with Shap values is negative.	As observed in the figure 3.25 (C) it has right tail with higher density around 0-1. This implies for a majority of predictions its SHAP values made modest contribution towards model predictions . But for some positive (P03) class predictions its contribution was above average.	74.53 (MTCO1) ; 76.40(MTCO1 with VDAC1)	1.97 & 0.88 | 1.56 & 0.28
NDUFB8	As seen in the figure 3.25 (C) the direction of correlation of NDUFB8 with SHAP values is roughly positive.	As observed in the figure 3.25 (C) it has small tails with higher density around 0-1. This implies for a majority of predictions its SHAP values made modest contribution towards model predictions .	75.15(NDUFB8) ;76.60(NDUFB8 with VDAC1)	2.94 &1.67| 2.17 & 0.62

COX4+4L2	As seen in the figure 3.25 (C) the direction COX4+4L2 has similar magnitude and direction to NDUFB8.	As observed in the figure 3.25 (C) it has small tails with higher density around 0-1. This implies for a majority of predictions its SHAP values made modest contribution towards model predictions	74.74(COX4+4L2) ; 77.02(COX4+4L2 with VDAC1)	7.46 & 5.66 | 5.07 & 1.94
NDUFA13	As seen in the figure 3.25 (C) the direction NDUFA13 has similar magnitude and direction to NDUFB8.	As observed in the figure 3.25 (C) it has small tails with higher density around -0.5 to 0.5. This implies for a majority of predictions its SHAP values made modest contribution towards model predictions.	72.88 (NDUFA13) ; 76.40(NDUFA13 with VDAC1)	3.00 & 1.75 | 2.25 & 0.64

Bibliography

[1] C. Warren, D. McDonald, R. Capaldi, D. Deehan, R. W. Taylor, A. Filby, D. M. Turnbull, C. Lawless, and A. E. Vincent, "Decoding mitochondrial heterogeneity in single muscle fibres by imaging mass cytometry," *Scientific Reports*, 2020.

[2] A. E. Vincent, H. S. Rosa, K. Pabis, C. Lawless, C. Chen, A. Grünewald, K. A. Rygiel, M. C. Rocha, A. K. Reeve, G. Falkous, V. Perissi, K. White, T. Davey, B. J. Petrof, A. A. Sayer, C. Cooper, D. Deehan, R. W. Taylor, D. M. Turnbull, and M. Picard, "Subcellular origin of mitochondrial DNA deletions in human skeletal muscle," *Annals of Neurology*, vol. 84, pp. 289–301, 8 2018.

[3] A. E. Vincent, H. S. Rosa, C. L. Alston, J. P. Grady, K. A. Rygiel, M. C. Rocha, R. Barresi, R. W. Taylor, and D. M. Turnbull, "Dysferlin mutations and mitochondrial dysfunction.," *Neuromuscular disorders : NMD*, vol. 26, pp. 782–788, 11 2016.

[4] Q. Su, Q. Liu, R. I. Lau, J. Zhang, Z. Xu, Y. K. Yeoh, T. W. Leung, W. Tang, L. Zhang, J. Q. Liang, Y. K. Yau, J. Zheng, C. Liu, M. Zhang, C. P. Cheung, J. Y. Ching, H. M. Tun, J. Yu, F. K. Chan, and S. C. Ng, "Faecal microbiome-based machine learning for multi-class disease diagnosis," *Nature Communications 2022 13:1*, vol. 13, pp. 1–8, 11 2022.

[5] Y. K. Lee, D. Ryu, S. Kim, J. Park, S. Y. Park, D. Ryu, H. Lee, S. Lim, H. S. Min, Y. K. Park, and E. K. Lee, "Machine-learning-based diagnosis of thyroid fine-needle aspiration biopsy synergistically by Papanicolaou staining and refractive index distribution," *Scientific Reports 2023 13:1*, vol. 13, pp. 1–9, 6 2023.

[6] X. Zhong, Y. Lin, W. Zhang, and Q. Bi, "Predicting diagnosis and survival of bone metastasis in breast cancer using machine learning," *Scientific Reports 2023 13:1*, vol. 13, pp. 1–20, 10 2023.

[7] D. Doudesis, K. K. Lee, J. Boeddinghaus, A. Bularga, A. V. Ferry, C. Tuck, M. T. Lowry, P. Lopez-Ayala, T. Nestelberger, L. Koechlin, M. O. Bernabeu, L. Neubeck, A. Anand, K. Schulz, F. S. Apple, W. Parsonage, J. H. Greenslade, L. Cullen, J. W. Pickering, M. P. Than, A. Gray, C. Mueller, N. L. Mills, A. M. Richards, C. Pemberton, R. W. Troughton, S. J. Aldous, A. F. Brown, E. Dalton, C. Hammett, T. Hawkins, S. O'Kane, K. Parke, K. Ryan, J. Schluter, K. Wild, D. Wussler, Miró, F. J. Martin-Sanchez, D. I. Keller, M. Christ, A. Buser, M. R. Giménez, S. Barker, J. Blades, A. R. Chapman, T. Fujisawa, D. M. Kimenai, J. Leung, Z. Li, M. McDermott, D. E. Newby, S. D. Schulberg, A. S. Shah, A. Sorbie, G. Soutar, F. E. Strachan, C. Taggart, D. P. Vicencio, Y. Wang, R. Wereski, K. Williams, C. J. Weir, C. Berry, A. Reid, D. Maguire, P. O. Collinson, Y. Sandoval, and S. W. Smith, "Machine learning for diagnosis of

myocardial infarction using cardiac troponin concentrations," *Nature Medicine 2023 29:5*, vol. 29, pp. 1201–1210, 5 2023.

[8] S. S. Al-Zaiti, C. Martin-Gill, J. K. Zègre-Hemsey, Z. Bouzid, Z. Faramand, M. O. Alrawashdeh, R. E. Gregg, S. Helman, N. T. Riek, K. Kraevsky-Phillips, G. Clermont, M. Akcakaya, S. M. Sereika, P. Van Dam, S. W. Smith, Y. Birnbaum, S. Saba, E. Sejdic, and C. W. Callaway, "Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction," *Nature Medicine 2023 29:7*, vol. 29, pp. 1804–1813, 6 2023.

[9] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C. C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature 2024 630:8016*, vol. 630, pp. 493–500, 5 2024.

[10] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature 2021 596:7873*, vol. 596, pp. 583–589, 7 2021.

[11] S. N. Chandrasekaran, H. Ceulemans, J. D. Boyd, and A. E. Carpenter, "Image-based profiling for drug discovery: due for a machine-learning upgrade?," *Nature Reviews Drug Discovery 2020 20:2*, vol. 20, pp. 145–159, 12 2020.

[12] S. Ekins, A. C. Puhl, K. M. Zorn, T. R. Lane, D. P. Russo, J. J. Klein, A. J. Hickey, and A. M. Clark, "Exploiting machine learning for end-to-end drug discovery and development," *Nature Materials 2019 18:5*, vol. 18, pp. 435–441, 4 2019.

[13] E. Meijering, "A bird's-eye view of deep learning in bioimage analysis," *Computational and Structural Biotechnology Journal*, vol. 18, p. 2312, 1 2020.

[14] M. Allam, S. Cai, and A. F. Coskun, "Multiplex bioimaging of single-cell spatial profiles for precision cancer diagnostics and therapeutics," *npj Precision Oncology 2020 4:1*, vol. 4, pp. 1–14, 5 2020.

[15] J. Engelmann, A. D. McTrusty, I. J. MacCormick, E. Pead, A. Storkey, and M. O. Bernabeu, "Detecting multiple retinal diseases in ultra-widefield fundus imaging and data-driven identification of informative regions with deep learning," *Nature Machine Intelligence 2022 4:12*, vol. 4, pp. 1143–1154, 12 2022.

[16] M. Doan and A. E. Carpenter, "Leveraging machine vision in cell-based diagnostics to do more with less," *Nature Materials 2019 18:5*, vol. 18, pp. 414–418, 4 2019.

[17] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives," *Neurocomputing*, vol. 444, pp. 92–110, 7 2021.

[18] Y. Liu, Z. Ma, X. Liu, S. Ma, and K. Ren, "Privacy-Preserving Object Detection for Medical Images with Faster R-CNN," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 69–84, 2022.

[19] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[20] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. De Almeida, H. Sirelkhatim, G. Richard, M. Skwark, K. Beguir, M. Lopez, and T. Pierrot, "Nature Methods nature methods Nucleotide Transformer: building and evaluating robust foundation models for human genomics,"

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9351, pp. 234–241, 2015.

[22] "An artificial intelligence tool that can help detect melanoma | MIT News | Massachusetts Institute of Technology."

[23] A. Petkidis, V. Andriasyan, L. Murer, R. Volle, and U. F. Greber, "A versatile automated pipeline for quantifying virus infectivity by label-free light microscopy and artificial intelligence," *Nature Communications 2024 15:1*, vol. 15, pp. 1–11, 6 2024.

[24] T. Royal Society, "THE AI REVOLUTION IN SCIENTIFIC RESEARCH,"

[25] I. Georgescu, "How machines could teach physicists new scientific concepts," *Nature Reviews Physics 2022 4:12*, vol. 4, pp. 736–738, 7 2022.

[26] A. H. Nielsen, A. Iosifidis, and H. Karstoft, "Forecasting large-scale circulation regimes using deformable convolutional neural networks and global spatiotemporal climate data," *Scientific Reports 2022 12:1*, vol. 12, pp. 1–12, 5 2022.

[27] J. Zhou, B. Huang, Z. Yan, and J. C. G. Bünzli, "Emerging role of machine learning in light-matter interaction," *Light: Science & Applications 2019 8:1*, vol. 8, pp. 1–7, 9 2019.

[28] S. J. Newman and R. T. Furbank, "Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data," *Nature Plants 2021 7:10*, vol. 7, pp. 1354–1363, 10 2021.

[29] N. Zobeiry and K. D. Humfeld, "AN ITERATIVE SCIENTIFIC MACHINE LEARNING APPROACH FOR DISCOVERY OF THEORIES UNDERLYING PHYSICAL PHENOMENA A PREPRINT," 2019.

[30] J. Giorgio, W. J. Jagust, S. Baker, S. M. Landau, P. Tino, and Z. Kourtzi, "A robust and interpretable machine learning approach using multimodal biological data to predict future pathological tau accumulation," *Nature Communications 2022 13:1*, vol. 13, pp. 1–14, 4 2022.

[31] F. Wong, E. J. Zheng, J. A. Valeri, N. M. Donghia, M. N. Anahtar, S. Omori, A. Li, A. Cubillos-Ruiz, A. Krishnan, W. Jin, A. L. Manson, J. Friedrichs, R. Helbig, B. Hajian, D. K. Fiejtek, F. F. Wagner, H. H. Soutter, A. M. Earl, J. M. Stokes, L. D. Renner, and J. J. Collins, "Discovery of a structural class of antibiotics with explainable deep learning," *Nature 2023 626:7997*, vol. 626, pp. 177–185, 12 2023.

[32] T. Naito, K. Inoue, S. Namba, K. Sonehara, K. Suzuki, B. Japan, K. Matsuda, N. Kondo, T. Toda, T. Yamauchi, T. Kadowaki, and Y. Okada, "Machine learning reveals heterogeneous associations between environmental factors and cardiometabolic diseases across polygenic risk scores," *Communications Medicine 2024 4:1*, vol. 4, pp. 1–12, 9 2024.

[33] J. D. Janizek, A. B. Dincer, S. Celik, H. Chen, W. Chen, K. Naxerova, and S. I. Lee, "Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models," *Nature Biomedical Engineering 2023 7:6*, vol. 7, pp. 811–829, 5 2023.

[34] G. S. Gorman, A. M. Schaefer, Y. Ng, N. Gomez, E. L. Blakely, C. L. Alston, C. Feeney, R. Horvath, P. Yu-Wai-Man, P. F. Chinnery, R. W. Taylor, D. M. Turnbull, and R. McFarland, "Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease," *Annals of Neurology*, vol. 77, pp. 753–759, 5 2015.

[35] M. Barends, L. Verschuren, E. Morava, V. Nesbitt, D. Turnbull, and R. McFarland, "Causes of Death in Adults with Mitochondrial Disease," *JIMD Reports*, vol. 26, p. 103, 2016.

[36] R. McFarland, R. W. Taylor, and D. M. Turnbull, "A neurological perspective on mitochondrial disease," 2010.

[37] "Wellcome Trust Centre for Mitochondrial Research Newcastle UK."

[38] V. Di Leo, C. Lawless, M.-P. Roussel, T. B. Gomes, G. S. Gorman, O. M. Russell, H. A. L. Tuppen, E. Duchesne, and A. E. Vincent, "Journal of Neuromuscular Diseases xx (2023) x-xx," 2023.

[39] A. Khan, "atifkhanncl/mitoML_segmentation_pipeline: Machine learning pipeline to segment muscle fibers from IMC and IF images."

[40] A. Khan, C. Lawless, A. E. Vincent, S. Pilla, S. Ramesh, and A. S. McGough, "Explainable Deep Learning to Profile Mitochondrial Disease Using High Dimensional Protein Expression Data," *Proceedings - 2022 IEEE International Conference on Big Data, Big Data 2022*, pp. 4375–4384, 2022.

[41] A. Khan, C. Lawless, A. E. Vincent, C. Warren, V. D. Leo, T. Gomes, and A. S. McGough, "NCL-SM: A Fully Annotated Dataset of Images from Human Skeletal Muscle Biopsies," in *2023 IEEE International Conference on Big Data (BigData)*, pp. 3704–3710, 2023.

[42] E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider, "An estimation of the number of cells in the human body," *Annals of Human Biology*, 2013.

[43] C. L. Alston, M. C. Rocha, N. Z. Lax, D. M. Turnbull, and R. W. Taylor, "The genetics and pathology of mitochondrial disease," 2017.

[44] C. Lawless, L. Greaves, A. K. Reeve, D. M. Turnbull, and A. E. Vincent, "The rise and rise of mitochondrial DNA mutations," *Open Biology*, vol. 10, no. 5, 2020.

[45] P. MITCHELL, "Coupling of Phosphorylation to Electron and Hydrogen Transfer by a Chemi-Osmotic type of Mechanism," *Nature*, vol. 191, no. 4784, pp. 144–148, 1961.

[46] C. F. Warren, "Quantitative analysis of oxidative phosphorylation dysfunction in mitochondrial myopathy and ageing," 2019.

[47] A. E. Vincent, "Investigating the pathogenesis of mitochondrial dysfunction in mitochondrial and other myopathies," 2017.

[48] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. L. De Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden, and I. G. Young, "Sequence and organization of the human mitochondrial genome," *Greenberg. R. As'r. 1*, vol. 290, no. 6, pp. 338–346, 1981.

[49] M. J. Young and W. C. Copeland, "Human mitochondrial DNA replication machinery and disease," *Current opinion in genetics & development*, vol. 38, pp. 52–62, 6 2016.

[50] H. A. Tuppen, E. L. Blakely, D. M. Turnbull, and R. W. Taylor, "Mitochondrial DNA mutations and human disease," *Biochimica et biophysica acta*, vol. 1797, pp. 113–128, 2 2010.

[51] R. Rossignol, B. Faustin, C. Rocher, M. Malgat, J. P. Mazat, and T. Letellier, "Mitochondrial threshold effects," *The Biochemical journal*, vol. 370, pp. 751–762, 3 2003.

[52] D. C. Wallace, G. Singh, M. T. Lott, J. A. Hodge, T. G. Schurr, A. M. Lezza, L. J. Elsas, and E. K. Nikoskelainen, "Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy," *Science (New York, N.Y.)*, vol. 242, no. 4884, pp. 1427–1430, 1988.

[53] I. J. Holt, A. E. Harding, and J. A. Morgan-Hughes, "Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies," *Nature 1988 331:6158*, vol. 331, no. 6158, pp. 717–719, 1988.

[54] M. Mancuso, D. Orsucci, C. Angelini, E. Bertini, V. Carelli, G. P. Comi, C. Minetti, M. Moggio, T. Mongini, S. Servidei, P. Toninc, A. Toscano, G. Uziel, C. Bruno, E. C. Ienco, M. Filosto, C. Lamperti, D. Martinelli, I. Moroni, O. Musumeci, E. Pegoraro, D. Ronchi, F. M. Santorelli, D. Sauchelli, M. Scarpelli, M. Sciacco, M. Spinazzi, M. L. Valentino, L. Vercelli, M. Zeviani, and G. Siciliano, "Phenotypic heterogeneity of the 8344A>G mtDNA "MERRF" mutation," *Neurology*, vol. 80, pp. 2049–2054, 5 2013.

[55] E. A. Schon, R. Rizzuto, C. T. Moraes, H. Nakase, M. Zeviani, and S. DiMauro, "A direct repeat is a hotspot for large-scale deletion of human mitochondrial DNA," *Science (New York, N.Y.)*, vol. 244, no. 4902, pp. 346–349, 1989.

[56] M. Magner, H. Kolářová, T. Honzik, I. Švandová, and J. Zeman, "Clinical manifestation of mitochondrial diseases.," *Developmental period medicine*, vol. 19, no. 4, pp. 441–449, 2015.

[57] S. E. Calvo, K. R. Clauser, and V. K. Mootha, "MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins," *Nucleic Acids Research*, vol. 44, p. D1251, 1 2016.

[58] E. Ylikallio and A. Suomalainen, "Mechanisms of mitochondrial diseases," *Annals of medicine*, vol. 44, pp. 41–59, 2 2012.

[59] A. H. Hakonen, P. Isohanni, A. Paetau, R. Herva, A. Suomalainen, and T. Lönnqvist, "Recessive Twinkle mutations in early onset encephalopathy with mtDNA depletion," *Brain*, vol. 130, pp. 3032–3040, 11 2007.

[60] C. Rouzier, S. Bannwarth, A. Chaussenot, A. Chevrollier, A. Verschueren, N. Bonello-Palot, K. Fragaki, A. Cano, J. Pouget, J. F. Pellissier, V. Procaccio, B. Chabrol, and V. Paquis-Flucklinger, "The MFN2 gene is responsible for mitochondrial DNA instability and optic atrophy 'plus' phenotype," *Brain : a journal of neurology*, vol. 135, no. Pt 1, pp. 23–34, 2012.

[61] A. R. Stiles, M. T. Simon, A. Stover, S. Eftekharian, N. Khanlou, H. L. Wang, S. Magaki, H. Lee, K. Partynski, N. Dorrani, R. Chang, J. A. Martinez-Agosto, and J. E. Abdenur, "Mutations in TFAM, encoding mitochondrial transcription factor A, cause neonatal liver failure associated with mtDNA depletion," *Molecular genetics and metabolism*, vol. 119, pp. 91–99, 9 2016.

[62] H. Tyynismaa, R. Sun, S. Ahola-Erkkilä, H. Almusa, R. Pöyhönen, M. Korpela, J. Honkaniemi, P. Isohanni, A. Paetau, L. Wang, and A. Suomalainen, "Thymidine kinase 2 mutations in autosomal recessive progressive external ophthalmoplegia with multiple mitochondrial DNA deletions," *Human molecular genetics*, vol. 21, pp. 66–75, 1 2012.

[63] P. Amati-Bonneau, M. L. Valentino, P. Reynier, M. E. Gallardo, B. Bornstein, A. Boissière, Y. Campos, H. Rivera, J. G. De La Aleja, R. Carroccia, L. Iommarini, P. Labauge, D. Figarella-Branger, P. Marcorelles, A. Furby, K. Beauvais, F. Letournel, R. Liguori, C. La Morgia, P. Montagna, M. Liguori, C. Zanna, M. Rugolo, A. Cossarizza, B. Wissinger, C. Verny, R. Schwarzenbacher, M. Martín, J. Arenas, C. Ayuso, R. Garesse, G. Lenaers, D. Bonneau, and V. Carelli, "OPA1 mutations induce mitochondrial DNA instability and optic atrophy 'plus' phenotypes," *Brain*, vol. 131, pp. 338–351, 2 2008.

[64] R. K. Naviaux and K. V. Nguyen, "POLG mutations associated with Alpers' syndrome and mitochondrial DNA depletion," *Annals of neurology*, vol. 55, pp. 706–712, 5 2004.

[65] C. Fratter, P. Raman, C. L. Alston, E. L. Blakely, K. Craig, C. Smith, J. Evans, A. Seller, B. Czermin, M. G. Hanna, J. Poulton, C. Brierley, T. G. Staunton, P. D. Turnpenny, A. M. Schaefer, P. F. Chinnery, R. Horvath, D. M. Turnbull, G. S. Gorman, and R. W. Taylor, "RRM2B mutations are frequent in familial PEO with multiple mtDNA deletions," *Neurology*, vol. 76, p. 2032, 6 2011.

[66] N. G. Larsson and A. Oldfors, "Mitochondrial myopathies," *Acta Physiologica Scandinavica*, vol. 171, pp. 385–393, 3 2001.

[67] G. S. Gorman, P. F. Chinnery, S. DiMauro, M. Hirano, Y. Koga, R. McFarland, A. Suomalainen, D. R. Thorburn, M. Zeviani, and D. M. Turnbull, "Mitochondrial diseases," *Nature Reviews Disease Primers 2016 2:1*, vol. 2, pp. 1–22, 10 2016.

[68] A. Munnich, A. Rötig, D. Chretien, J. M. Saudubray, V. Cormier, and P. Rustin, "Clinical presentations and laboratory investigations in respiratory chain deficiency," *European Journal of Pediatrics*, vol. 155, no. 4, pp. 262–274, 1996.

[69] J. Rahman and S. Rahman, "Mitochondrial medicine in the omics era," *The Lancet*, vol. 391, no. 10139, pp. 2560–2574, 2018.

[70] W. K. Engel and G. G. Cunningham, "Rapid examination of muscle tissue: An improved trichrome method for fresh-frozen biopsy sections," *Neurology*, vol. 13, no. 11, pp. 919–923, 1963.

[71] S. L. Old and M. A. Johnson, "Methods of microphotometric assay of succinate dehydrogenase and cytochrome c oxidase activities for use on human skeletal muscle," *The Histochemical Journal*, vol. 21, pp. 545–555, 9 1989.

[72] M. C. Rocha, J. P. Grady, A. Grünewald, A. Vincent, P. F. Dobson, R. W. Taylor, D. M. Turnbull, and K. A. Rygiel, "A novel immunofluorescent assay to investigate oxidative phosphorylation deficiency in mitochondrial myopathy: understanding mechanisms and improving diagnosis," *Scientific Reports 2015 5:1*, vol. 5, pp. 1–17, 10 2015.

[73] F. Corporation, "MCD Viewer User Guide (FLDM-400317)," 2019.

[74] C. Giesen, H. A. O. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P. J. Schüffler, D. Grolimund, J. M. Buhmann, S. Brandt, Z. Varga, P. J. Wild, D. Günther, and B. Bodenmiller, "highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry," *Articles nAture methods* |, vol. 11, no. 4, p. 417, 2014.

[75] Warren C, "plotIMC - interactive visualisation of imaging mass cytometry data-https://mito.ncl.ac.uk/warren_2019/," 2019.

[76] D. Schapiro, H. W. Jackson, S. Raghuraman, J. R. Fischer, V. R. Zanotelli, D. Schulz, C. Giesen, R. Catena, Z. Varga, and B. Bodenmiller, "histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data," *Nature Methods 2017 14:9*, vol. 14, pp. 873–876, 8 2017.

[77] D. Schapiro, H. W. Jackson, S. Raghuraman, J. R. Fischer, V. R. Zanotelli, D. Schulz, C. Giesen, R. Catena, Z. Varga, and B. Bodenmiller, "miCAT: A toolbox for analysis of cell phenotypes and interactions in multiplex image cytometry data," *Nature Methods*, vol. 14, pp. 873–876, 8 2017.

[78] J. Windhager, V. Riccardo, T. Zanotelli, D. Schulz, L. Meyer, M. Daniel, B. Bodenmiller, and N. Eling, "An end-to-end workflow for multiplexed image processing and analysis," *Nature Protocols* |, vol. 18, pp. 3565–3613, 2023.

[79] C. R. Stoltzfus, J. Filipek, B. H. Gern, B. E. Olin, J. M. Leal, Y. Wu, M. R. Lyons-Cohen, J. Y. Huang, C. L. Paz-Stoltzfus, C. R. Plumlee, T. Pöschinger, K. B. Urdahl, M. Perro, and M. Y. Gerner, "CytoMAP: A Spatial Analysis Toolbox Reveals Features of Myeloid Cell Organization in Lymphoid Tissues," *Cell Reports*, vol. 31, p. 107523, 4 2020.

[80] E. T. McKinley, J. Shao, S. T. Ellis, C. N. Heiser, J. T. Roland, M. C. Macedonia, P. N. Vega, S. Shin, R. J. Coffey, and K. S. Lau, "MIRIAM: A machine and deep learning single-cell segmentation and quantification pipeline for multi-dimensional tissue images," *Cytometry Part A*, vol. 101, pp. 521–528, 6 2022.

[81] B. Englert and S. Lam, "The Caltech-UCSD Birds-200-2011 Dataset," *IFAC Proceedings Volumes (IFAC-PapersOnline)*, vol. 42, no. 15, 2009.

[82] "CnrLwlss/mitocyto: Automatic and manual image analysis of cells in serial sections."

[83] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini, "CellProfiler: image analysis software for identifying and quantifying cell phenotypes," *Genome Biology*, vol. 7, no. 10, p. R100, 2006.

[84] S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, and A. Kreshuk, "ilastik: interactive machine learning for (bio)image analysis," *Nature Methods*, 2019.

[85] I. Culjak, D. Abram, T. Pribanic, H. Dzapo, and M. Cifrek, "A brief introduction to OpenCV," 2012.

[86] C.-L. Chiu, N. Clack, and t. n. community, "napari: a Python Multi-Dimensional Image Viewer Platform for the Research Community," *Microscopy and Microanalysis*, vol. 28, pp. 1576–1577, 8 2022.

[87] N. F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, T. Dougherty, C. Camacho Fullaway, B. J. McIntosh, K. Xuan Leow, M. Sarah Schwartz, C. Pavelchek, S. Cui, I. Camplisson, O. Bar-Tal, J. Singh, M. Fong, G. Chaudhry, Z. Abraham, J. Moseley, S. Warshawsky, E. Soon, S. Greenbaum, T. Risom, T. Hollmann, S. C. Bendall, L. Keren, W. Graf, M. Angelo, and D. Valen, "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning," *Nature Biotechnology*, 2022.

[88] D. Bannon, E. Moen, M. Schwartz, E. Borba, T. Kudo, N. Greenwald, V. Vijayakumar, B. Chang, E. Pao, E. Osterman, W. Graf, and D. Van Valen, "DeepCell Kiosk: scaling deep learning–enabled cellular image analysis with Kubernetes," *Nature Methods 2021 18:1*, vol. 18, pp. 43–45, 1 2021.

[89] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature Methods*, 2020.

[90] N. Eling, N. Damond, T. Hoch, and B. Bodenmiller, "cytomapper: an R/Bioconductor package for visualization of highly multiplexed imaging data," *Bioinformatics (Oxford, England)*, vol. 36, pp. 5706–5708, 12 2021.

[91] M. Sonka, V. Hlavac, and R. Boyle, "Image Processing, Analysis and Machine Vision," *Image Processing, Analysis and Machine Vision*, 1993.

[92] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 3523–3542, 7 2022.

[93] G. Iannizzotto and L. Vita, "Fast and Accurate Edge-Based Segmentation with No Contour Smoothing in 2-D Real Images," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 9, no. 7, 2000.

[94] N. Sharma, V. Jain, and A. Mishra, "An Analysis Of Convolutional Neural Networks For Image Classification," *Procedia Computer Science*, vol. 132, pp. 377–384, 1 2018.

[95] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, pp. 1–21, 5 2021.

[96] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," 7 2022.

[97] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, p. 370, 1972.

[98] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[99] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, Association for Computing Machinery, 2016.

[100] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature 2015 521:7553*, vol. 521, pp. 436–444, 5 2015.

[101] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J Big Data*, vol. 8, p. 53, 2021.

[102] M. Vakalopoulou, S. Christodoulidis, N. Burgos, O. Colliot, and V. Lepetit, "Deep Learning: Basics and Convolutional Neural Networks (CNNs)," *Neuromethods*, vol. 197, pp. 77–115, 2023.

[103] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Networks*, vol. 12, pp. 145–151, 1 1999.

[104] J. Duchi JDUCHI and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization * Elad Hazan," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[105] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[106] G. Hinton, N. Srivastava, and K. Swersky, "Neural Networks for Machine Learning Lecture 6a Overview of mini--batch gradient descent," *Cited on*, vol. 14, p. 2, 2012.

[107] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," *arXiv.org*, 2015.

[108] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[109] S. I. H. Krizhevsky, A., ""Imagenet classification with deep convolutional neural network", in Advances in Neural Information Processing Systems, p. 1097-1105.," *Elsevier Ltd*, 2012.

[110] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey; Deep Facial Expression Recognition: A Survey," 2022.

[111] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep Face Recognition: A Survey," in *Proceedings - 31st Conference on Graphics, Patterns and Images, SIBGRAPI 2018*, 2019.

[112] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI," 2019.

[113] J. Plested and T. Gedeon, "Deep transfer learning for image classification: a survey," *ArXiv*, vol. abs/2205.09904, 2022.

[114] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[115] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, 2015.

[116] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv*, 2015.

[117] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, pp. 3212–3232, 11 2019.

[118] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2015.

[119] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[120] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, "Cell Detection with Star-Convex Polygons," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11071 LNCS, pp. 265–273, Springer Verlag, 2018.

[121] A. Waisman, A. M. Norris, M. Elías Costa, and D. Kopinke, "Automatic and unbiased segmentation and quantification of myofibers in skeletal muscle," *Scientific Reports*, vol. 11, no. 1, p. 11793, 2021.

[122] J. W. Pylvänäinen, E. Gómez-De-Mariscal, R. Henriques, and G. Jacquemet, "Live-cell imaging in the deep learning era," 2023.

[123] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 11 2018.

[124] Z. C. Lipton, "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, pp. 31–57, 6 2018.

[125] R. Marcinkevičs and J. E. Vogt, "Interpretability and Explainability: A Machine Learning Zoo Mini-tour," tech. rep., 2020.

[126] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence 2020 2:1*, vol. 2, pp. 56–67, 1 2020.

[127] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[128] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K. W. Low, S. F. Newman, J. Kim, and S. I. Lee, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering 2018 2:10*, vol. 2, pp. 749–760, 10 2018.

[129] R. Mitchell, E. Frank, and G. Holmes, "GPUTreeShap: Massively Parallel Exact Calculation of SHAP Scores for Tree Ensembles," *PeerJ Computer Science*, vol. 8, 10 2020.

[130] E. Winter, "Chapter 53 The shapley value," *Handbook of Game Theory with Economic Applications*, vol. 3, pp. 2025–2054, 1 2002.

[131] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: interpreting, explaining and visualizing deep learning*, vol. 11700. Springer Nature, 2019.

[132] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, and N. C. Bouaynaya, "Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks; Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks," *IEEE Signal Processing Magazine*, vol. 39, 2022.

[133] M. Ivanovs, R. Kadikis, and K. Ozols, "Perturbation-based methods for explaining deep neural networks: A survey," *Pattern Recognition Letters*, vol. 150, pp. 228–234, 10 2021.

[134] P. J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (Un)reliability of Saliency Methods," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11700 LNCS, 2019.

[135] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 618–626, Institute of Electrical and Electronics Engineers Inc., 12 2017.

[136] G. Kwon, M. Prabhushankar, D. Temel, and G. Alregib, "Distorted Representation Space Characterization Through Backpropagated Gradients," *Proceedings - International Conference on Image Processing, ICIP*, vol. 2019-Septe, pp. 2651–2655, 9 2019.

[137] M. D. Zeiler and R. Fergus, "LNCS 8689 - Visualizing and Understanding Convolutional Networks," tech. rep., 2014.

[138] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K. R. Müller, S. Dähne, and P. J. Kindermans, "INNvestigate neural networks!," *Journal of Machine Learning Research*, vol. 20, 2019.

[139] B. D. De Vos, J. M. Wolterink, T. Leiner, and P. A. De Jong, "Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT; Direct Automatic Coronary Calcium Scoring in Cardiac and Chest CT," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, 2019.

[140] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognition*, vol. 65, 2017.

[141] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, 2015.

[142] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification," *Frontiers in Aging Neuroscience*, vol. 10, no. JUL, 2019.

[143] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," 7 2017.

[144] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," 2017.

[145] R. J. Janssen, L. G. Nijtmans, L. P. van den Heuvel, and J. A. Smeitink, "Mitochondrial complex I: Structure, function and pathology," *Journal of Inherited Metabolic Disease*, vol. 29, pp. 499–515, 8 2006.

[146] H. J. Seltman, "Experimental design and analysis," 2012.

[147] A. Molak and A. Jaokar, *Causal Inference and Discovery in Python : Unlock the Secrets of Modern Causal Machine Learning with Dowhy, EconML, Pytorch and More*. Packt Publishing, 2023.

[148] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of Biomedical Informatics*, vol. 35, pp. 352–359, 10 2002.

[149] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J. D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96–110, 2 2014.

[150] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[151] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "LNCS 8693 - Microsoft COCO: Common Objects in Context," 2014.

[152] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," 2023.

[153] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," tech. rep., 2014.

[154] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," 2020.

[155] S. Wales, A. M. C Kiernan DSc, B. C. Cheah MBiostat, J. Burrell MBBS, M. C. Zoing BNurs, M. C. Kiernan, S. Vucic, B. C. Cheah, M. R. Turner, A. Eisen, O. Hardiman, J. R. Burrell, and M. C. Zoing, "Seminar Amyotrophic lateral sclerosis," *Lancet*, vol. 377, pp. 942–55, 2011.

[156] M. Filippi, A. Bar-Or, F. Piehl, P. Preziosa, A. Solari, S. Vukusic, and M. A. Rocca, "Multiple sclerosis," *Nature Reviews Disease Primers*, vol. 4, no. 1, p. 43, 2018.

[157] K. Bushby, R. Finkel, D. J. Birnkrant, L. E. Case, P. R. Clemens, L. Cripe, A. Kaul, K. Kinnett, C. Mcdonald, S. Pandya, J. Poysky, F. Shapiro, J. Tomezsko, and C. Constantin, "Review Diagnosis and management of Duchenne muscular dystrophy, part 1: diagnosis, and pharmacological and psychosocial management," *The Lancet Neurology*, vol. 9, pp. 77–93, 2010.

[158] B. M. Morrison and J. W. Griffin, "Neuromuscular Diseases," *Cerebrospinal Fluid in Clinical Practice*, pp. 121–126, 2009.

[159] S. T. Ahmed, R. W. Taylor, D. M. Turnbull, C. Lawless, and S. J. Pickett, "Quantifying phenotype and genotype distributions in single muscle fibres from patients carrying the pathogenic mtDNA variant m.3243A>G," *medRxiv*, p. 2022.04.04.22272484, 4 2022.

[160] Connor Lawless, "GitHub - CnrLwlss/quad_immuno: Software for quadruple immunofluorescence histochemistry from Wellcome Centre for Mitochondrial Research, Newcastle, UK," 2019.

[161] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7, pp. 676–682, 2012.

[162] S. T. Ahmed, C. L. Alston, S. Hopton, L. He, I. P. Hargreaves, G. Falkous, M. Oláhová, R. McFarland, D. M. Turnbull, M. C. Rocha, and R. W. Taylor, "Using a quantitative quadruple immunofluorescent assay to diagnose isolated mitochondrial Complex I deficiency," *Scientific Reports*, vol. 7, no. 1, p. 15676, 2017.

[163] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[164] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation; UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," 2020.

[165] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11045 LNCS, pp. 3–11, 2018.

[166] A. Mayeuf-Louchart, D. Hardy, Q. Thorel, P. Roux, L. Gueniot, D. Briand, A. Mazeraud, A. Bouglé, S. L. Shorte, B. Staels, F. Chrétien, H. Duez, and A. Danckaert, "MuscleJ: A high-content analysis method to study skeletal muscle with a new Fiji tool," *Skeletal Muscle*, vol. 8, pp. 1–11, 8 2018.

[167] P. Shrestha, N. Kuang, and J. Yu, "Efficient end-to-end learning for cell segmentation with machine generated weak annotations," *Communications Biology*, vol. 6, no. 1, p. 232, 2023.

[168] S. Qamar, R. Öberg, D. Malyshev, and M. Andersson, "A hybrid CNN-Random Forest algorithm for bacterial spore segmentation and classification in TEM images," *Scientific Reports*, vol. 13, no. 1, p. 18758, 2023.

[169] G. Xu, M. Liu, Z. Jiang, D. Söffker, and W. Shen, "Bearing Fault Diagnosis Method Based on Deep Convolutional Neural Network and Random Forest Ensemble Learning," *Sensors 2019, Vol. 19, Page 1088*, vol. 19, p. 1088, 3 2019.

[170] W. Bakasa and S. Viriri, "VGG16 Feature Extractor with Extreme Gradient Boost Classifier for Pancreas Cancer Prediction.," *Journal of imaging*, vol. 9, 7 2023.

[171] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation," 2 2018.

[172] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. Mcdonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[173] E. Czech, B. A. Aksoy, P. Aksoy, and J. Hammerbacher, "Cytokit: A single-cell analysis toolkit for high dimensional fluorescent microscopy imaging," *BMC Bioinformatics*, vol. 20, pp. 1–13, 9 2019.

[174] F. van Maldegem, K. Valand, M. Cole, H. Patel, M. Angelova, S. Rana, E. Colliver, K. Enfield, N. Bah, G. Kelly, V. S. K. Tsang, E. Mugarza, C. Moore, P. Hobson, D. Levi, M. Molina-Arcas, C. Swanton, and J. Downward, "Characterisation of tumour microenvironment remodelling following oncogene inhibition in preclinical studies with imaging mass cytometry," *Nature Communications 2021 12:1*, vol. 12, pp. 1–14, 10 2021.

[175] J. Eng, E. Bucher, Z. Hu, T. Zheng, S. L. Gibbs, K. Chin, and J. W. Gray, "A framework for multiplex imaging optimization and reproducible analysis," *Communications Biology*, vol. 5, no. 1, p. 438, 2022.

[176] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," 2017.

[177] A. Shrikumar, P. Greenside, A. Y. Shcherbina, and A. Kundaje, "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences," *arXiv*, vol. 1, pp. 0–5, 5 2016.

[178] S Lundberg, "API Reference — SHAP latest documentation," 2017.

[179] N. F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, T. Dougherty, C. C. Fullaway, B. J. McIntosh, K. X. Leow, M. S. Schwartz, C. Pavelchek, S. Cui, I. Camplisson, O. Bar-Tal, J. Singh, M. Fong, G. Chaudhry, Z. Abraham, J. Moseley, S. Warshawsky, E. Soon, S. Greenbaum, T. Risom, T. Hollmann, S. C. Bendall, L. Keren, W. Graf, M. Angelo, and D. Van Valen, "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning," *Nature Biotechnology 2021 40:4*, vol. 40, pp. 555–565, 11 2021.