# Automated Quality Control by Application of Machine Learning Techniques for Quantitative Liver MRI

Charles Hill

St Hugh's College




University of Oxford

A thesis presented for the degree of

*Doctor of Philosophy*

Michaelmas 2021

Copyright © 2021 by Charles Hill
All Rights Reserved

# Acknowledgements

This thesis wouldn't have been possible without the support of many people, both inside and outside academia. First of all, I would like to thank my three supervisors; Prof Vicente Grau, Dr Luca Biasiolli, and Prof Matthew Robson. Vicente and the group have been invaluable in guiding me into the world of deep learning and being around for a cycle and pint. Luca has somehow been able to support me all the way from Italy while supporting his own young family, and starting a new job. Finally, I am grateful to Matt for probably being the most supportive industry supervisor that existed. Not only did he actually meet me regularly, but he also had some good ideas.

I would also like to thank those who I have lived with, over the years in Oxford. The 50 Windmill Road collective, and the Reliance Way Massive have reminded me that life isn't always about work. Many great memories have been made over the years which I will remember fondly. Thanks also go out to John McGonigle and his team at Perspectum.

Finally I would like to express thanks to family, and my partner for their invaluable support over the past four years. They have been there consistently, through thick and thin. I wouldn't have finished this thesis without them.

# Declaration

I, Charles E. Hill, declare that this thesis is entirely my own work, and except where otherwise stated, describes my own research.

Charles Hill

C Hill 10:21, Feb 13, 2024 UTC

*Signature*

13/02/2024

*Date*

**Abstract**

Quantitative magnetic resonance imaging (qMRI) and multi-parametric MRI are being increasingly used to diagnose and monitor liver diseases such as non-alcoholic fatty liver disease (NAFLD). These acquisitions are comparably more complicated than traditional T1-weighted and T2-weighted MRI scans and are also more prone to image quality issues and artefacts. In order for the output of the qMRI scans to be useable, they must undergo a rigorous and often lengthy quality control (QC). This manual QC is prone to human error and subjective. Additionally, with the development of new qMRI techniques, this leads to the manifestation of new quality issues. This thesis focuses on the development and implementation of automated QC processes for liver qMRI scans, that is where possible tag-free such that the process can be adapted to different imaging techniques. These automated QC processes were implemented using a variety of machine learning (ML) and deep learning (DL) approaches. These methods, developed on T1 mapping in UKBiobank, were designed to output metrics from the MRI scans that could be used to identify a specific quality issue, such as in chapter 3, or give a more general indication of the image quality in chapter 4. Furthermore, it was hypothesised that the introduction of associated meta-data, such as patient factors and scanning parameters, into these deep learning models would increase overall performance. This was explored in chapter 5. Finally, in order to assess the utility of our developed algorithms in a wider setting except for T1 mapping in UKBiobank, we tested it in two settings. Pilot study one assessed the utility of the model in T1 mapping in a separate study (CoverScan). Pilot study two assessed the utility of the model in a different qMRI acquisition; proton density fat fraction (PDFF) acquisitions from UKBiobank.

***Keywords*—** Liver MRI - Deep Learning - Quality Control

# Table of Contents


**1 Introduction** **2**

1.1 Motivation . . . . . 2

1.2 Objectives . . . . . 3

1.3 Thesis Outline . . . . . 5

1.4 Published Work . . . . . 6

1.4.1 Conference Abstracts . . . . . 6

1.4.2 Journal Articles . . . . . 6

1.4.3 Under Review . . . . . 6

**2 Background** **7**

2.1 Introduction . . . . . 7

2.2 Liver Disease and Diagnosis . . . . . 7

2.3 Magnetic Resonance Imaging . . . . . 10

2.3.1 Quantitative MRI . . . . . 12

2.4 Multi-parametric Liver MRI . . . . . 14

2.4.1 ShMOLLI $T_1$ . . . . . 15

2.4.2 Dixon $T_2^*$ for iron correction and PDFF calculation . . . . 16

2.4.3 IDEAL method for PDFF . . . . . 18

2.4.4 Other quantitative MRI techniques . . . . . 19

2.4.5 Curve Fitting . . . . . 20

2.5 Artefacts in MRI . . . . . 20

2.5.1 Tissue related artefacts . . . . . 21

2.5.2 Patient related artefacts . . . . . 22

2.5.3 Scanner Artefacts . . . . . 23

2.6 Image Quality Assessment (IQA) . . . . . 24

2.6.1 FR-IQA . . . 25
2.6.2 RR-IQA . . . 26
2.6.3 NR-IQA . . . 26
2.7 Machine and Deep Learning . . . 27
2.7.1 Machine Learning . . . 28
2.7.2 Deep Learning . . . 30
2.7.2.1 Convolutional Neural Networks (CNNs) . . . 31
2.7.2.2 Region proposal convolutional neural networks (R-CNNs) . . . 32
2.7.2.3 Generative Adversarial Networks (GANs) . . . 32
2.7.2.4 IQA Specific Classifiers . . . 33

**3 Landmark Detection for Slice Location identification 36**
3.1 Introduction . . . 36
3.2 Methods . . . 39
3.2.1 Data and Gold Standard Region Generation . . . 39
3.2.2 Network Architecture and Loss Function . . . 44
3.3 Experiments . . . 47
3.4 Results . . . 48
3.5 Discussion . . . 51
3.6 Conclusion . . . 58

**4 Data imputation for detection of localised artefacts 60**
4.1 Introduction . . . 60
4.2 Materials and Methods . . . 63
4.2.1 Data . . . 63
4.2.2 Network Development . . . 65
4.2.2.1 T1 Mapping network . . . 65

4.2.2.2 Data consistency cGAN . . . . . 66
4.2.2.3 Classification . . . . . 68
4.3 Experimental Setup . . . . . 69
4.3.1 Training . . . . . 69
4.3.1.1 T1 Mapping . . . . . 70
4.3.1.2 cGAN . . . . . 71
4.3.1.3 Classification . . . . . 71
4.3.2 Testing and Evaluation . . . . . 71
4.3.2.1 T1 Mapping . . . . . 71
4.3.2.2 cGAN . . . . . 72
4.3.2.3 Classification . . . . . 75
4.4 Results . . . . . 75
4.4.1 T1 mapping . . . . . 75
4.4.2 cGAN . . . . . 76
4.4.3 Classification . . . . . 78
4.5 Discussion . . . . . 79
4.5.1 Reconstruction Evaluation . . . . . 81
4.5.2 T1 Values . . . . . 82
4.5.3 Classification . . . . . 83
4.5.4 Limitations . . . . . 85
4.6 Conclusions . . . . . 87

**5 Inclusion of Meta Data for Classification of Quality Issues 89**
5.1 Introduction . . . . . 89
5.2 Experimental Setup . . . . . 92
5.2.1 Data . . . . . 92
5.2.2 Random Forests . . . . . 97
5.2.3 Multi-layer Perceptron . . . . . 97

5.2.4 Braided Network . . . . . 98
5.2.5 Diamond Network . . . . . 100
5.2.6 VGG network and random forest combination . . . . . 100
5.3 Results . . . . . 102
5.3.1 Random Forests Vs. MLP . . . . . 102
5.3.2 Braided Net Vs. Diamond Net . . . . . 107
5.3.3 VGG-16 . . . . . 108
5.3.4 VGG-16 with Random Forest addition . . . . . 110
5.4 Discussion . . . . . 111
5.5 Conclusions . . . . . 113

**6 Pilot studies and External validation 116**
6.1 Introduction . . . . . 116
6.2 Study 1: CoverScan . . . . . 117
6.2.1 Data . . . . . 117
6.2.2 Methods . . . . . 119
6.2.3 Results . . . . . 120
6.2.4 Discussion . . . . . 125
6.3 Study 2: QC of the IDEAL Acquisition for PDFF . . . . . 127
6.3.1 Methods . . . . . 127
6.3.2 Results . . . . . 129
6.4 Discussion . . . . . 131
6.5 Conclusion . . . . . 134

**7 Conclusions 136**
7.1 Further Work . . . . . 138

**Bibliography 140**

**Appendix A Appendices** **163**

A.1 Manual Bounding Box Software and Labelling Process . . . . . 163

A.2 LMSQuality - GUI for QC reporting . . . . . 165

# List of Figures


2.1 $T_1$ and $T_2$ recovery . . . . . 13

3.1 fRCNN dataset population statistics . . . . . 39
3.2 Bounding Box Area versus class distribution . . . . . 40
3.3 Bounding box centroid locations per class . . . . . 42
3.4 AP to RL bounding box length ratio . . . . . 43
3.5 Slice positioning on the scanner . . . . . 44
3.6 Faster R-CNN model architecture . . . . . 45
3.7 Training loss for fRCNN backbones . . . . . 47
3.8 PR curves for fRCNN backbones . . . . . 49
3.9 PR curve varying with IoU for mobilenet v3 . . . . . 50
3.10 PR curves by class for mobilenet v3 . . . . . 52
3.11 Normalised volume box plots split by wrong location class label . 53
3.12 Examples of cases which have the correct slice selection and wrong slice selection. . . . . 57

4.1 Conditional generative adversarial network architecture for dataset reconstruction. . . . . 61
4.2 Dataset statistics for the cGAN reconstruction dataset . . . . . 63
4.3 T1 value distribution across the dataset . . . . . 64
4.4 Quality issue classifier for mean absolute error matrices. . . . . 65
4.5 Example mean absolute error matrix . . . . . 66
4.6 T1 map network accuracy regression across T1 ranges . . . . . 70
4.7 Error metric evaluation of reconstruction quality across UKBiobank 72
4.8 Error metric evaluation of T1 map generation post-reconstruction across UKBiobank. . . . . 73

4.9 T1 map accuracy regression across T1 ranges, post-reconstruction 74
4.10 Reconstruction examples for selected image removals/knockouts . 76
4.11 ROC curves for 3D CNN classifier and the normalised mean absolute error across UKBiobank. . . . . 79
4.12 PR curves for 3D CNN classifier and the normalised mean absolute error across UKBiobank. . . . . 79

5.1 **Left:** The distribution of QC classes across the dataset. The majority of cases have no QC issues associated with them. **Right:** The occurrence rate of multi-label cases. . . . . 93
94figure.caption.35
5.3 Quality issue examples in MOLLI T1 mapping dataset . . . . . 96
5.4 Example MLP and random forest architecture . . . . . 98
5.5 Braided network and diamond network architectures . . . . . 99
5.6 VGG-16 architecture . . . . . 100
5.7 Random Forest performance by quality error class . . . . . 101
5.8 MLP performance by quality error class . . . . . 104
5.9 ROC curves by class and data input for random forests . . . . . 105
5.10 ROC curves by class and data input for MLPs . . . . . 106
5.11 ROC curves by class and averaged for braided network . . . . . 107
5.12 ROC curves by class and averaged for diamond network . . . . . 107
5.13 ROC curves by class and averaged for the VGG-16 network . . . 108
5.14 ROC curves for VGG, random forests and combined predictions by class . . . . . 112
5.15 Averaged ROC curves for VGG, random forests and combined model . . . . . 114

6.1 Population statistics for the Coverscan dataset . . . . . 118

6.2 Overall network architecture for CoverScan re-training . . . . . 119
6.3 Mean absolute error comparison between CoverScan and UK-Biobank by instance . . . . . 120
6.4 Normalised root mean squared error comparison between Cover-Scan and UKBiobank by instance . . . . . 121
6.5 Structural similarity comparison between CoverScan and UKBiobank by instance . . . . . 122
6.6 Error metric comparison between CoverScan and UKBiobank when averaged over all instances. . . . . 123
6.7 Example image with a small normalised root mean squared error. . 124
6.8 Example image with a large normalised root mean squared error. . 124
6.9 Mean absolute error for IDEAL dataset after reconstruction . . . . 130
6.10 Normalised root mean squared error for IDEAL dataset after re-construction . . . . . 131
6.11 Structural similarity for IDEAL dataset after reconstruction . . . . 132
6.12 ROC curves for normalised error in predicting quality issue present 133
6.13 Example IDEAL dataset for selected reconstructions . . . . . 134

A.1 User interface for ground truth generation . . . . . 164
A.2 Drawing interface for ground truth generation . . . . . 164
A.3 Drawing interface for ground truth generation, showing bounding boxes . . . . . 165
A.4 The graphical user interface displaying an acquired scan and previous scans. . . . . 166

# List of Tables

4.1 The 3D CNN Architecture used to classify the error maps into the binary accept, reject classes . . . . . 69

5.1 Table depicting the image quality issue types that can occur in the UKBB datasets used for the multi-parametric analysis. . . . . 91

5.2 Table depicting the mean and std. dev. in the AUC, for each data combination, for the random forests . . . . . 103

5.3 Table depicting the mean and std. dev. in the AUC, for each data combination, for the MLPs . . . . . 109

# List of Abbreviations

**MRI** . . . . . Magnetic Resonance Imaging

**MR** . . . . . Magnetic Resonance

**NAFLD** . . . . . Non-Alcoholic Fatty Liver Disease

**BMI** . . . . . Body Mass Index

**NASH** . . . . . Non-Alcoholic Steatohepatitis

**IVC** . . . . . Inferior Vena Cava

**DICOM** . . . . . Digital Imaging and Communication in Medicine

**MOLLI** . . . . . Modified Look-Locker Inversion Recovery

**ShMOLLI** . . . . . Shortened Modified Look-Locker Inversion Recovery

**IDEAL** . . . . . Iterative De-composition of water and fat With Echo Asymmetry and Least-squares estimation

**UKBB** . . . . . UK Biobank

**QC** . . . . . Quality Control

**IQA** . . . . . Image Quality Assessment

**SSIM** . . . . . Structiral Similarity Index

**MAE** . . . . . Mean Absolute Error

**NRMSE** . . . . . Normalised Root Mean Squared Error

**PSNR** . . . . . Peak Signal to Noise Ratio

**ML** . . . . . Machine Learning

**DL** . . . . . Deep Learning

**SVM** . . . . . Support Vector Machine

**MLP** . . . . . Multi-Layer Perceptron

**CNN** . . . . . Convolutional Neural Networks

**R-CNN** . . . . . Region proposal Convolutional Neural Network

**RPN** . . . . . Region Proposal Network

**GAN** . . . . . Generative Adversarial Network

**cGAN** . . . . . conditional Generative Adversarial Network

**ROI** . . . . . Region Of Interest

**IOU** . . . . . Intersect Over Union

**ROC** . . . . . Receiver Operating Characteristic (curve)

**PR** . . . . . Precision Recall (curve)

**AUC** . . . . . Area Under Curve

**AUROC** . . . . . Area Under Receiver Operating Characteristic (curve)

**AP** . . . . . Average Precision

**mAP** . . . . . mean Average Precision

# 1 | Introduction

## 1.1 Motivation

Quantitative Liver Magnetic Resonance Imaging has been shown to be a strong indicator of multiple types of liver disease, from non-alcoholic fatty liver disease (NAFLD) in adults to stratification of auto immune disease in children.(Pavlides et al., 2017; Janowski et al., 2020) These techniques are non-invasive, and allow for early stratification and identification of disease types.(Banerjee et al., 2014) If used clinically, they could save the part of the £6.1 billion cost of obesity, due to the early identification of NAFLD.(Public Health England, 2017a)

These quantitative techniques can be very complex and involve the acquisition of multiple images, potentially at multiple locations. This means the artefacts that have traditionally affected single $T_1$- or $T_2$-weighted images, can affect multiple slices.(Erasmus, Hurter, Naude, Kritzinger, & Acho, 2004) This leads to a time consuming quality control (QC) process, which is open to additional human error. Furthermore, human visual QC is often subjective and, more importantly, based on the experience of the person performing the QC.(Gardner et al., 1995) An experienced radiographer will be more likely to identify any crucial QC issues and re-acquire the dataset immediately, compared to a radiographer earlier in their career development. Implementation of an objective and accurate QC method, would help address the gap between radiographer experience, increasing the baseline of dataset quality.

Additionally, in order to generate quantitative metrics from these acquisitions, the datasets are sent off-site for data analysis and production of outcome reports. If an issue with the acquisitions is found at this stage, then the patient who was scanned

will have to be recalled for further scans. Therefore, there is a need for quality issues to be identified early and accurately. Development of a technique which can be performed within a reasonable time frame of the end of the scans, such that any scans that need to be re-acquired are acquired immediately, would address this issue.

Machine Learning approaches have the potential to address these issues, given a large enough dataset with associated labels. However, these techniques are flawed on two accounts. Firstly, the majority of machine learning algorithms, especially deep learning based algorithms, fall into the category of a "black box". This means there is little understanding in what properties of the input are determining the outcome. In order to promote uptake of these techniques in the clinic and medical research, the reasoning behind classification would be desirable.

Secondly, access to a large dataset with all potential QC issues identified is extremely rare. New QC issues could also develop as the quantitative MRI techniques are further developed. A model that could be easily adapted to new quantitative techniques and QC issues, would have an inherent longevity to it, and so help solve the problem of low quality data in the long-term as well as for specific MRI acquisitions.

## 1.2 Objectives

The objective of this work was to develop and validate a robust, quick and explainable QC algorithm for feedback of quality issues to radiographers, such that a re-acquisition could be performed. The specific objectives are as follows:

- Develop a set of deep learning models and associated framework that are explainable and accurately identify QC issues in the 2D axial shortened-

Modified Look-Locker Inversion Recovery (shMOLLI) T1-mapping sequence on Siemens MRI scanners, in a healthy subject cohort (UKBiobank). We define explainable as the end user of the algorithm would be able to check the outcome against their own knowledge, and the provided information to validate the outcome. For example this could be highlighting the location of an artefact in a scan, when a re-acquisition of the scan has been requested.

- Explore the use of 2D image data, image derived data, and original meta-data in determining the quality of the acquisitions. The 2D data would include the original DICOM data. The image derived data would include the output of any image processing models or algorithms used on the original DICOM data, e.g. segmentation maps or classification task outputs. The original meta-data would include anything non-image based that is present in the DICOM data or can reasonably be expected to be acquired alongside the image. This would include MRI physics parameters, such as Inversion Times, and patient characteristics, such as height. The aim of this is two-fold. Firstly, to determine the optimal framework for combining multiple data types collected during an MRI examination and derived afterwards, and secondly, to explore whether the combination improves prediction accuracy for QC algorithms.
- To show the utility of the developed models in a 2D axial IDEAL MRI acquisition on Siemens' scanners, and in the subject population acquired through the COVERSCAN study. Thus showing the utility of the developed models across multiple quantitative MRI acquisitions and subject populations.

## 1.3 Thesis Outline

Chapter 2 summarises the background of NAFLD, and it's development from biopsy diagnosis to non-invasive MRI diagnosis. Secondly, we summarise the theory behind MRI and in-particular the quantitative protocol performed to help stratify NAFLD, and the introduction of artefacts into MR images. Finally, we explore the use of machine and deep learning algorithms in Liver MRI, and there use in solving image quality assessment.

Chapter 3 covers our initial attempt of developing an explainable deep learning method for identification of a single QC issue, slice placement, using a region detection based algorithm.

Chapter 4 focuses on the development of an explainable deep learning method that can be used to assess the overall quality of an image, without access to ground truth labels. This involves the use of generative adversarial networks and and an additional novel regression network.

Chapter 5 explores the use of multiple types of data into a single QC classification network. We study the optimal combination of 2D image data and 1D meta-data for identification of quality issues.

Chapter 6 tests out the framework from 4 in two new datasets, which are suitably different to the initial dataset that the network was developed on. Firstly, we studied its utility on an up-to-date quantitative T1 dataset, and secondly, we applied the framework to a different type of quantitative MRI.

Chapter 7 summarises the output of this thesis and explores the future directions of this work.

## 1.4 Published Work

### 1.4.1 Conference Abstracts

[1] **Charles E Hill**, Ferenc Emil Mozes, Liam AJ Young, Luca Biasiolli, Matthew D Robson, and Vicente Grau. Deep learning based multi-organ detection: a step closer to liver T1 map quality assurance. In *Proc. Intl. Soc. Mag. Reson. Med. 28 (2020) 2598*

### 1.4.2 Journal Articles

[1] **Hill CE**, Biasiolli L, Robson MD, Grau V, Pavlides M. Emerging artificial intelligence applications in liver magnetic resonance imaging. World J Gastroenterol. 2021 Oct 28;27(40):6825-6843. doi: 10.3748/wjg.v27.i40.6825. PMID: 34790009; PMCID: PMC8567471.

### 1.4.3 Under Review

[1] **Charles E Hill**, Luca Biasiolli, Matthew D Robson, and Vicente Grau. Semi-supervised deep learning model, with domain specific loss, for MRI data consistency check and quality stratification. *Medical Image Analysis*

# 2 | Background

## 2.1 Introduction

This chapter provides background theory and a literature review in the relevant areas for this thesis. In section 2.2, we initially touch on the disease statistics behind the need to develop non-invasive diagnoses. In section 2.3, we then focus on the theory and development of multi-parametric techniques for solving the aforementioned problem. Section 2.5 covers the artefacts than can generally occur in MRI techniques with focus on those that occur in quantitative sequential techniques. The rest of the chapter focuses on the methods that can be used to address the identification and rectification of quality control issues. In section 2.6, we look at current methods for quality controlling images of any form, and those methods which are MR specific. In the following three sections we explore the non-machine learning, traditional machine learning (ML) and deep learning (DL) methods for this in more depth, with background theory to their implementation and utility. In section 2.7, we explore the use of the latter two algorithms in medical imaging and finally focus on their use in quality control.

## 2.2 Liver Disease and Diagnosis

With as many as 1 in 3 people in western populations now obese, there is an ever increasing prevalence of obesity related diseases contributing to the strain on nations medical services. Public Health England estimated that the NHS spent £6.1 billion on overweight and obesity related ill-health in 2014-15.(Public Health England, 2017b) More recently it was estimated that over a million NHS hospital admissions had obesity as a contributing factor, a rise of 17 % on 2018/19.(Thandi

& NHS, 2021) In the United States, the annual cost of healthcare for those who were obese was $2,505 higher than those in a healthy weight category, leading to an aggregate medical bill of $260.6 billion across the USA.(Cawley et al., 2021) This is not just isolated to western populations, with a study in Korea showing that obesity-related liver cancer led to an increased cost of $1.1 trillion, between 2002 and 2015, on the healthcare system.(Lee et al., 2019) Though most of the costs cited above account for all obesity related diseases such as type 2 diabetes and cardiovascular disease (CVD), a significant proportion of resources are dedicated to the treatment of liver obesity related diseases as shown by the Korean study.

Early detection and reliable prognoses are an essential way of tackling some of these costs. In non-alcoholic fatty liver disease (NAFLD), the implementation of lifestyle changes early on in the disease development can lead to a significant improvement in clinical scores.(Hallsworth & Adams, 2019) However, if these lifestyle measures are not brought in early enough, the stages of the disease become irreversible.(Sheth, 1997)

In 1999, initial attempts to stratify NAFLD involved classification into four types, with liver biopsy playing a central role.(Matteoni et al., 1999) Those with only liver steatosis present or both steatosis and lobular inflammation were classified as type 1 and 2 respectively. Those with hepatic steatosis and ballooning degeneration were classified as type 3, while those with liver steatosis, ballooning degeneration, fibrosis and Mallory-Denk bodies were classified as type 4. Both type 3 and 4 were associated with higher liver mortality rates than type 1 or 2.(Matteoni et al., 1999) This was further developed in 2005 and the current gold standard method, the NAFLD activity score, focuses on stratifying patients based on histological features from a liver biopsy.(Kleiner et al., 2005) Out of the 14 features used in the scoring system, 4 are evaluated semi-qualitatively: steatosis (0-3), lobular inflammation (0-2), hepatocellular ballooning (0-2), and fibrosis (F0-F4).

Among these scores, fibrosis scoring has been shown to be predictive of mortality and progression to severe liver disease, i.e. cirrhosis.(Hagström et al., 2017) In modern clinics, when a liver biopsy is performed, the level of liver steatosis and liver fibrosis are the main markers for NAFLD and non-alcoholic steatohepatitis (NASH) diagnosis.(Yoneda et al., 2018)

This stratification is extremely useful in determining the type and severity of treatment to provide for a patient and satisfies this need for early detection. However, liver biopsy is an unnecessarily invasive technique with only local tissue assessed.(Masood, Rosa, Kraemer, Smotherman, & Mohammadi, 2015) This has led to the development of non-invasive assessments of NAFLD, including imaging, or minimally invasive, such as blood biomarkers. Ito *et al.* performed a meta analysis showing that one such minimally invasive biomarker, *Wisteria floribunda* agglutinin-positive Mac-2-binding protein (WFA+-M2BP), had a sensitivity of 0.818 and specificity of 0.839 in diagnosing cirrhosis (F4).(Ito et al., 2017) However, these are fundamentally global measures which are less effective in cases of mixed pathology. Non-invasive imaging techniques have also been developed to address this short fall by biopsy. Ultrasound is a widely available technique which is frequently used in clinic for the diagnosis of chronic liver disease with a greater sensitivity for mild steatosis than biopsy.(Gerstenmaier & Gibson, 2014) However, these techniques are not successful at differentiating between multiple stages of NAFLD.(Dasarathy et al., 2009) Another technique routinely used in clinical practice is transient elastography (one brand being FibroScan from EchoSense), assesses the liver stiffness using a sheer mechanical wave by use of a transducer placed between the rib cage.(Myers et al., 2012) This in turn allows quantification of steatosis and fibrosis in the liver, with high differentiation between F3 and F4, but with low differentiation between F0 to F2.(Eddowes et al., 2019) In addition to this, it has been estimated that between 20-30% of cases fail due to operator

error or patient factors such as an increased skin to liver capsular distance which is prevalent in NAFLD and obese patients.(Şirli, Sporea, Bota, & Jurchiş, 2013; Eddowes et al., 2019)

An assortment of MR methods have been also been developed for the diagnosis and prognosis of liver related diseases, such as NAFLD. These include multi-phasic contrast enhanced imaging, and quantitative MR imaging among other techniques.(Böttcher et al., 2013; Pavlides et al., 2017) These techniques all allow for both local and global assessment of liver damage and in the case of multiple quantitative measures, the ability to assess liver function.

## 2.3 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) was developed primarily in the 1970s after work from Paul Lauterbur and Sir Peter Mansfield into the use of field gradients and selective slice excitation respectively.(Lauterbur, 1973; Mansfield, 1977) Since then MRI has developed rapidly into one of the gold standard non-invasive disease monitoring methods.

MRI works due to the ubiquity of $^{1}H$ in our body (though other magnetically susceptible atoms can be studied), primarily in the form of $H_2O$. These protons with, spin of $\frac{1}{2}$, are magnetically susceptible and so align with the predominant magnetic field. Usually this is the Earth's magnetic field, with minimal effect due to its magnitude. However, when a magnetic field of considerable strength is applied, i.e. that of an MRI scanner, the protons align to this new predominant field. Furthermore, the spins precess around this axis at a frequency known as the Larmor frequency ($\omega$):

$$\omega = \gamma \mathbf{B_0} \tag{2.1}$$

where $\gamma$ is the gyromagnetic ratio, a fixed constant based on the atom we are observing, $\mathbf{B_0}$ the magnetic field. Though all spins will precess at this frequency, when unexposed to relaxation effects, the net magnetization of the combination of spins will be in the direction of the predominant field, i.e. $\mathbf{B_0}$. This net magnetization is dependent on the energy splitting between up and down spin states, and therefore the thermodynamic equilibrium of each state, i.e. how many protons fall into each state, and the strength of the field:

$$\mathbf{M_0} \propto \frac{\gamma^2 h^2 N_s \mathbf{B_0}}{4kT} \tag{2.2}$$

where $\mathbf{M_0}$ is the net magnetization, $h$ is planck's constant, $N_s$ the number of spins, $k$ the Boltzmann constant and $T$ the absolute temperature. As the net magnetization is in the direction of the predominant field direction $\mathbf{B_0}$ and not changing, it is not measurable. This is because no voltage is induced in the receiver coils, the method by which we collect the data required to formulate an image, as per Faraday's law. In order to acquire an image, we must change the magnetic field in order to induce a voltage in the receiver coils and generate an image. Therefore a secondary magnetic field $\mathbf{B_1}$ is applied orthogonal to $\mathbf{B_0}$. This secondary magnetic field, induced by a radio-frequency (RF) pulse at the Larmor frequency, generates an $\mathbf{M_{xy}}$ component. When the RF pulse is turned off, the spins relax back to thermal equilibrium, and a net magnetization of $\mathbf{M_0}$. It is the changing magnetic field caused by this relaxation, that induces a voltage in the receiver coils, that can be used to formulate an image.

This relaxation of the spins back to thermal equilibrium can be split into two

components, the spin-lattice relaxation and the spin-spin relaxation, commonly known in MRI as the $T_1$ and $T_2$ relaxation.

Firstly, the $T_1$ relaxation, describes the relaxation in the direction of the $\mathbf{B_0}$ field, and the recovery of $\mathbf{M_0}$:

$$\mathbf{M_z}(t) = (1 - e^{-t/T_1})\mathbf{M_0} \tag{2.3}$$

where $\mathbf{M_z}$ is the net magnetization in the z-direction at time $t$. Secondly, the $T_2$ relaxation, describes the relaxation of the $\mathbf{M_{xy}}$ component back to zero:

$$\mathbf{M_{xy}}(t) = \mathbf{M_0}e^{-t/T_2}. \tag{2.4}$$

These two relaxation mechanisms are dependent on localised properties that the spins are exposed to, and therefore vary depending on tissue type. Along with spatial encoding, field gradients in x, y, and z to allow localisation of signals, this allows for images with good tissue contrast to be acquired.

### 2.3.1 Quantitative MRI

Qualitative MRI scans are acquired at a set timepoint during the relaxation of the spins back to equilibrium, leading to images either being $T_1$-weighted ($T_1$-w) or $T_2$-weighted ($T_2$-w) images depending on which is the predominant factor in the relaxation at that time point. However, it is not possible to determine the $T_1$ or $T_2$ value of water in a certain region using these images. In order to determine the $T_1$ value of a particular region of an image, then more complicated techniques must be used.

Figure 2.1 demonstrates the basic difference between a single qualitative $T_1$-w

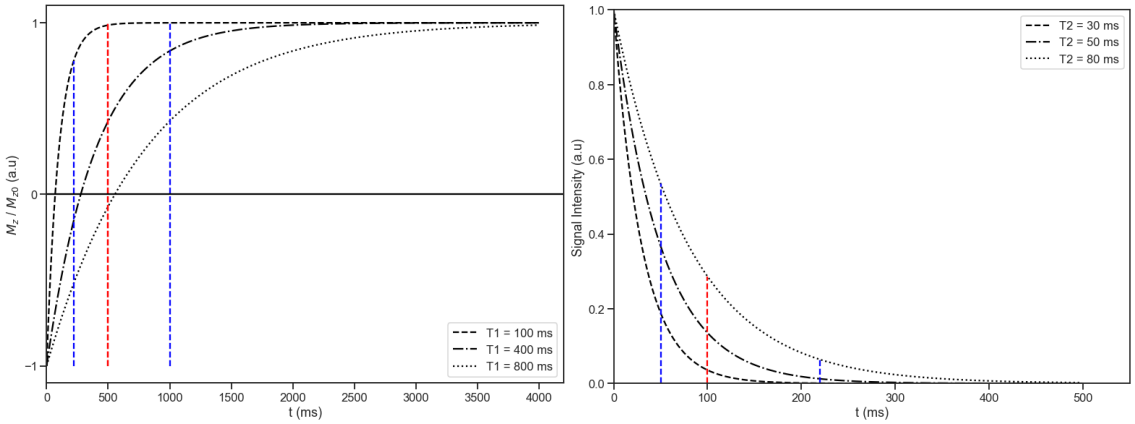



Figure 2.1: Example of $T_1$ and $T_2$ recovery at different values. The red line depicts a single snapshot taken to maximise the contrast between tissues with different $T_1$ and $T_2$ values. The blue lines depict the concept of sampling at different points along these recoveries, to be able to calculate the $T_1$ and $T_2$ values.

image and a quantitative technique. The red line depicts a single snapshot of the $T_1$ recovery. This snapshot will give difference signals to the three tissues with different $T_1$ values, but no more can be read into the images. However, if multiple snapshots at known time points after the inversion of the magnetisation are made, then we begin to sample the $T_1$ recovery curve and can fit against eq. 2.3 for the $T_1$ value. A similar process, can be applied to $T_2$, among other quantitative metrics.

It should be noted that $T_2$ cannot be measured directly, the measured "$T_2$" decays faster than expected. This due to a combination of molecular diffusion, magnetic susceptibility and field inhomogeneities. This quicker measured decay is termed the $T_2^*$ value. This measured value is often combined with the $T_1$ value and other quantitative metrics, to help aid diagnosis and prognosis, in what is known as multi-parametric MRI.

## 2.4 Multi-parametric Liver MRI

Multi-parametric MRI has been used in many anatomical locations for more objective and reliable diagnosis or understanding of the disease/system.(Radtke et al., 2016; B. Zhang et al., 2017) It normally involves the use of $T_1$ and $T_2^*$ mapping combined with additional supplementary MR information like proton density fat fraction (PDFF) or magnetic resonance elastography (MRE).(McDonald et al., 2018) For the case of Liver disease and NAFLD, as previously mentioned, the combination of PDFF values with an iron-corrected $T_1$ ($cT_1$) and $T_2^*$ mapping has correlated well with both liver biopsy and clinical outcomes.(Banerjee et al., 2014) The $T_1$ value is acquired by a Shortened Modified Look-Locker Inversion Recovery (ShMOLLI) and then corrected for the influence of iron by the $T_2^*$ value obtained by the DIXON method for fat-water separation. Finally, an Iterative Decomposition of water and fat With Echo Asymmetry and Least-squares estimation (IDEAL) is used to obtain the PDFF.(Reeder et al., 2005) This multi-parametric method is currently used in clinics and pharmaceutical trials as Liver*MultiScan*$^{TM}$ as developed by Perspectum Diagnostics. It has been shown to have good diagnostic accuracy for NAFLD by comparison to liver biopsy, and potentially diagnosis NASH and cirrhosis.(Pavlides et al., 2017) It has also been shown to have the lowest cost per correct diagnosis of NAFLD versus existing biomarkers in a two centre study.(Eddowes et al., 2018) Recently, a two centre study showed good predictive success for iron accumulation in the liver alongside steatosis and the aforementioned NAFLD.(McDonald et al., 2018) The specific methods and their origin are briefly discussed below.

### 2.4.1 ShMOLLI $T_1$

Initial methods looked to fully sample the $T_1$ relaxation, however in reality these approaches are time consuming and unstable *in vivo*.(Taylor, Salerno, Dharmakumar, & Jerosch-Herold, 2016) The Look-Locker method, is a multi-point method, which first showed that full $T_1$ relaxation was not necessary to acquire a reliable value.(Look & Locker, 1970) With this technique, researchers were able to obtain accurate $T_1$ values in phantoms and other stationary bodies, e.g. the brain, however measurements were highly variable in bodies susceptible to movement over the course of an MR measurement, for example the heart and liver. To address these inadequacies, D.R.Messroghli *et al.* introduced the Modified Look-Locker Inversion Recovery (MOLLI) method for $T_1$ mapping.(Messroghli et al., 2004) This introduced two significant modifications to the original Look-Locker sequence; selective acquisition after each R-wave of the cardiac cycle, and combination of multiple Look-Locker experiments together, with varied inversion times. These multiple Look-Locker acquisitions are carried out with three, three and five single shot readouts respectively, with a minimum of 4 seconds between each Look-Locker inversion recovery experiment. Each Look-Locker inversion recovery experiment begins with a 180°RF pulse, with the aforementioned single shot readouts performed at varying inversion times. The images acquired at these readouts are sorted by ascending inversion time (t), and a three parameter nonlinear curve fitting applied to the signal model:

$$y = A - B \exp \frac{-t}{T_1^*} \tag{2.5}$$

where $\mathbf{y}$ is the signal intensity of each voxel in the images, $\mathbf{t}$ the inversion time for each image and $\mathbf{T_1^*}$ the apparent $\mathbf{T_1}$ due to modification by the Look-Locker ex-

periment. This new sequence model allowed for accurate quantification of Liver $T_1$ values, but still relied on a long breath hold in the region of 20s, which is unsuitable for some patient groups. In 2014, Piechnik *et al.* introduced the shMOLLI, which became the sequence of choice for the UK Biobank study for measuring $T_1$ in the Liver.(Piechnik et al., 2010)

ShMOLLI samples the $T_1$ recovery curve in a similar way to the MOLLI sequence, but with a 5, 1, 1 readout and a shortened recovery period between inversion recovery experiments (1s/1 heartbeat as opposed to the 4s/4 heartbeats in MOLLI). Though this shortened recovery period leads to the $\mathbf{M_z}$ not recovering fully, by using a sub-sample of the readouts for longer $T_1$ values, ShMOLLI is able to achieve comparably accurate $T_1$ maps in a shorter breath-hold, making the method more applicable for clinical use.

The T1 map, as before with the MOLLI acquisition is calculates by performing a non-linear curve fitting to equation 2.5. When applied as part of the LiverMultiScan protocol, this map is then corrected for the iron concentration which is calculated using the DIXON method, to output a corrected $T_1$ ($cT_1$).

### 2.4.2 Dixon $T_2^*$ for iron correction and PDFF calculation

$T_2$ relaxation, also known as spin-spin relaxation, is the process by which the transverse component of spin dephase after being brought into alignment by an Radio Frequency (RF) pulse. Like $T_1$ relaxation, the dephasing of the $T_2$ component is affected by the surrounding environment, but occurs faster than $T_1$ due to additional mechanisms, such as local field inhomogeneities and the size and "tumbling" speed of the molecule upon which the spin resides. As the $T_2$ component is affected by these local inhomogeneities, its dephasing is faster than can be explained by the tissue environment, therefore, the measurements performed

during the DIXON technique are of the observed $T_2$, termed $T_2^*$

The Dixon technique is a multi-echo gradient-echo imaging technique for fat-water separation which is commonly used for sampling $T_2^*$ relaxation.(Dixon, 1984)In the same way as $T_1$, we are able to measure $T_2^*$ by sampling the relaxation curve and fitting the measured data to the signal decay model in eq. 2.4, where t is the echo time of each image (TE).

As mentioned earlier, the Dixon technique is also a method for fat water separation. The Dixon technique utilises the larmor frequency difference between fat and water protons. The difference between these frequencies $\mathbf{\Delta f}$ means that at specific time points, determined by $\mathbf{\Delta f}$, the spins are fully in-phase, and at another specific time point, they are out-of-phase. When the water and fat spins are in-phase, the signal measured can be attributed to the combination of the protons in these two environments. Conversely, when the water and fat spins are out-of-phase, the signal measured can be attributed to the water protons minus the fat protons, as shown in the following equation:

$$\begin{aligned} IP &= W + F \\ OP &= W - F \end{aligned} \tag{2.6}$$

where $\mathbf{IP}$ is the in-phase signal acquired, $\mathbf{OP}$ the out-of-phase acquired, and $\mathbf{F}$ and $\mathbf{W}$ are the contribution of the F and W to the signal respectively. As mentioned earlier the points at which the F and W protons are in-phase is determined by the the $\mathbf{\Delta f}$ between the two environments. At 1.5$\mathbf{T}$, this offset is 220 Hz, leading to the two environments being phase every 4.5 ms. By making an acquisition at an integer multiple of this value, one can obtain the points of the curve described by 2.4. By performing a curve fitting to these points, it is possible to

obtain a $T_2^*$ map. As increased iron concentration leads to increased $T_2^*$ signal decay, the calculated map can be correlated to an iron concentration in the organ imaged. This is done by use of a conversion equation, that has been calculated by comparison with phantom measurements and biopsy from volunteers.(Henninger et al., 2017)As iron concentration has been shown to shorten $T_1$ relaxation as well, the $T_2^*$ map can then be used to calculate a $cT_1$ value by Bloch equation modelling of the ShMOLLI sequence for varying extracellular fluid and iron concentrations.(Tunnicliffe, Banerjee, Pavlides, Neubauer, & Robson, 2017) The multi-parametric liver protocol of interest does just that by acquiring a $T_2^*$ map, in the same plane as the ShMOLLI $T_1$ map.(Banerjee et al., 2014)

Though the Dixon approach has been shown to be versatile in allowing the calculation of different metrics, it has been adapted since its first introduction to deal with the shortcomings of the Dixon sequence.(J. Ma, 2008; Reeder et al., 2004) The iterative decomposition of water and fat with echo asymmetry and least-squares estimation (IDEAL) approach is a common adaptation of the original Dixon approach.

### 2.4.3 IDEAL method for PDFF

The IDEAL method for fat-water separation was developed in 2004 based on a need to be able to accurately decompose the two signals at short echo times (TE).(Reeder et al., 2005) In this case, three or more separate images with varying fat and water relative phase angles are acquired. The initial phase error in each acquired voxel is assumed to be zero such that a least-squares estimation can be obtained. Through iterative means the calculated phase is passed back until the calculated change in the phase is minimal. Once the fitting converges, the fat and water contribution to the voxel signal is used to determine the PDFF in each voxel, as in Eq. 2.7, providing an overall PDFF for the liver.

$$\mathbf{PDFF}(\%) = \frac{\mathbf{FAT}}{\mathbf{WATER} + \mathbf{FAT}} \tag{2.7}$$

As previously mentioned this PDFF is then used in conjunction with the $cT_1$ and $T_2^*$ maps to provide a reliable diagnosis and prognosis for various liver diseases. However, when applied clinically and on a large scale, well known MR artefacts can greatly affect the outcome of the diagnosis.

### 2.4.4 Other quantitative MRI techniques

Though to this point I have focused on the 3 techniques used in combination as part of the LiverMultiScan protocol, other quantitative metrics can be combined together to form multi-parametric techniques. Magnetic Resonance Elastography (MRE) is one such example of a different metric used in combination with other quantitative measures such as $T_1$. MRE is a technique for quantitatively measuring the stiffness of tissue.(Mariappan, Glaser, & Ehman, 2010) The technique has been used to characterise liver tumors showing that malignant tumors had significantly higher stiffness than benign tumors.(Venkatesh et al., 2008) It has also been used to predict advanced fibrosis of the liver in patients with NAFLD.(Loomba et al., 2014)

In addition to combining different quantitative metrics together to form different multi-parametric techniques, it is also possible to use different MRI sequences to measure the same quantitative metrics as used in the LiverMultiScan protocol. The saturation recovery single-shot acquisition (SASHA), is one such alternative MRI sequence for T1 mapping.(Chow et al., 2014) Briefly, it consists of the collection of 10 cardiac triggered single-shot steady-state free procession images. These images are combined using a three-parameter exponential recovery model:

$$Signal = A(1 - \eta_{apparent} \exp^{-TS/T_1}) \tag{2.8}$$

The SASHA sequence has been shown to give accurate $T_1$ measurements *in vivo* that is independent of absolute $T_1$, $_2$ and heart rate among other measures, and provides a viable alternative to shMOLLI T1 mapping.

### 2.4.5 Curve Fitting

As briefly mentioned above, the images acquired for T2-mapping and T1-mapping undergo parameter fitting to a recovery model. This is commonly done on a voxel basis, in order to generate maps, however multiple voxels can be binned into a single voxel to improve to the signal to noise ratio.

In the above signal model for $T_1$ recovery, eq. 2.5, there are three unknown constants to be fitted; $A$, $B$ and $T_1^*$. The most common approach to fit these is to apply a least squares algorithm to minimise the difference between the measured point, i.e. the voxels, and the predicted points by the recovery model.(Y.-C. Kim, Kim, Lee, & Choe, 2021) Multiple implementations of these algorithms exist with variations to improve numerical stability in the curve fitting, such as using the Levenberg-Marquardt technique.(S. Liu et al., 2018) The same approaches are applied in fitting of the Dixon $T_2^*$ values from the echo times and signal on a voxel-wise basis.

## 2.5 Artefacts in MRI

Image artefacts in MRI can be induced in numerous ways, and can broadly be split into three categories; tissue related, subject related and scanner/technique related.

### 2.5.1 Tissue related artefacts

Tissue related artefacts are any artefact that is caused by the properties of the issue and have no relation to the position or motion of the subject within the scanner.

Firstly, a chemical shift artefact can be caused when protons of two different chemical environments, with sufficient difference in resonance frequency (generally lipid protons), exist within a single voxel. As the spatial positioning is assigned along the frequency encoding direction, based on the resonant frequency of the proton, the non-water proton will have the same frequency as those of water molecules in a different voxel. Therefore it will be mapped to the wrong spatial location on the generated image. This appears as a shadow or halo at the boundary between two different tissue types upon the axis of frequency encoding.(Hood, Ho, Smirniotopoulos, & Szumowski, 1999) In echo-planar imaging, these artefacts occur upon the axis of phase encoding.(Jezzard & Balaban, 1995)

Secondly, susceptibility artefacts are caused by the difference in magnetic susceptibility between two tissues, e.g. tissue and air, or the presence of metallic ions or solid metal objects. These paramagnetic objects distort the local magnetic field, varying the frequency of the local protons. This leads to both mis-localisation of the proton and signal losses due to $T_2^*$ dephasing. These artefacts increase with increasing magnetic field as in Eq. 2.9.(Port & Pomper, 2000)

$$\textbf{Susceptibility Artifact Size} \propto \frac{\mathbf{B_0(\Delta Susceptibility)(TE)}}{\textbf{Pixel Bandwidth}} \tag{2.9}$$

Thirdly, dielectric artefacts are caused by the interaction of the Electric field component of RF pulse and the body. When an electromagnetic wave propagates through tissue, reflection/refraction occurs at tissue interfaces and the wavelength decreases. This reduction in wavelength is not an issue at lower fields, however

at fields of 3T and above the RF wavelength becomes shorter than the body. This will induce standing wave currents which constructively and destructively interfere to produce dark and bright regions in an image.(Collins, Liu, Schreiber, Yang, & Smith, 2005) By implementing dual transmit technology, these B1 field inhomogeneities are largely eradicated, however at 7T, dielectric pads are used additionally for impedance matching to reduce the artifact.(Webb, 2011)

### 2.5.2 Patient related artefacts

The main artefact caused by the patient or subject in the scanner are motion artefacts. They appear as either discrete “ghosts" (copies of the body overlaid on the image) or general blurring, dependent on the time-frame of the motion. They mainly occur in the phase encoding direction as the phase sampling interval is on a similar time scale to physiological movements such as swallowing, breathing and the heart beating. In many scans the introduction of these artefacts can lead to analysis of the scans producing unreliable results.(van Dijk, Sabuncu, & Buckner, 2012; Satterthwaite et al., 2012) It has also been shown that by accounting for the production of MRI artefacts, one can increase data quality and reduce costs by 50%.(Dosenbach et al., 2017) One method of reducing the occurrence of motion artefacts is to introduce respiratory and cardiac gating into the sequence, which has become standard practice when imaging the abdomen.(Plathow et al., 2006; Chia, Fischer, Wickline, & Lorenz, 2000) However, motion artefacts can still occur while gating due to erratic breathing or heart rates. Therefore, other techniques for shortening acquisition time to shorter than the physiological time frames are generally implemented such as parallel imaging or compressed sensing.(Deshmane, Gulani, Griswold, & Seiberlich, 2012; Esteban et al., 2017)

### 2.5.3 Scanner Artefacts

Artefacts can also be caused by incorrectly initialising the sequence parameters or the scanner variables. Aliasing is the most common of the scanner related artefacts. This occurs when the sampling rate does not meet the requirements of the Nyquist theorem. In MRI this presents itself as phase wraps, when the field of view (FoV) is smaller than that of the object.(Heiland, 2008) This leads to spatial mismapping of voxels outside the field of view to the opposite side of the image, if the sampling frequency is low.

Partial Volume effects are another common scanner/sequence related artefact. An acquired voxel can cover multiple voxel types, which is particularly problematic when the spatial resolution is low. The signal from this voxel will be a weighted value of the multiple tissue types. This can lead to decreased visibility of fine or low contrast structures.(Heiland, 2008) By increasing the spatial resolution in the slice plane, this can be reduced.

Additionally, you can get artefacts from hardware faults, such as coil malfunctions. Most common acquisitions require the use of multiple coils often contained within scanner company built hardware. The use of multiple coils is used to achieve a good signal to noise (SNR) across the whole body, as for each coil (depending on its size) the SNR drops off with distance.(Hayes & Axel, 1985) If one of the coils malfunctions, you lose a large proportion of the signal that would've previously been acquired close to that specific coil. This will lead to a localised low SNR region compared to the rest of the acquisition. When analysing, you would make the assumption that all the coils are working correctly and so you would miscalculate the properties, be they $T_1$ or $T_2^*$, associated with that region.

The above artefacts are associated with the acquisition in k-space, but artefacts

can also occur in the construction of the image from k-space. A Gibbs artefact occur as we use Fourier transforms to create an image from k-space. By using a finite number of sine waves to estimate a square function, you can't adequately fit high frequency information, i.e. high signal changes in an image at interfaces.(Czervionke, Czervionke, Daniels, & Hauhgton, 1988) These appear in various ways such as artificial edge enhancement or edge widening. Though it is impossible to completely nullify Gibbs artefacts, it is possible to mitigate against them. By increasing the number of phase encoding steps or reducing the field of view, you decrease the rate of signal changes from voxel to voxel and so avoid the fitting of higher frequency data.

All of the above artefacts can often present subtly, and the varying solutions to avoid them is not always an option. Therefore it is important to do an assessment of the image to detect and report back these issues back to the observer.

## 2.6 Image Quality Assessment (IQA)

Image quality assessment can be broadly categorised into two main branches, subjective and objective IQA. Subjective QC usually relies on multiple trained users visually assessing the image against their past experience.(V. et al., 2013) Though adequate for small local data sets, this method has poor reproducibility, which limits its value for multi center comparisons, and is time intensive, which reduces its applicability for large data sets. For accuracy of assessment, it also relies on highly trained personnel and assumes that all quality issues are visible to the user and require no further analysis. Conversely, objective QC allows comparisons within and between data sets, allowing the output of reliable metrics, and is usually an automated or semi-automated process, allowing many images to be quality assessed in an instance. The area of objective QC can be further split into three

areas; Full Reference (FR-), Reduced Reference (RR-) and No Reference (NR-) IQA.

### 2.6.1 FR-IQA

FR-IQA relies on a gold standard version of the image of interest that can be used as a pixelwise reference for the potentially low quality. Wang *et al.* developed the structural similarity index (SSIM) for IQA, one of the first successful objective quality controls in this category, by noting that the human visual system is highly evolved to notice structural similarities between images.(Z. Wang, Bovik, Sheikh, Member, et al., 2004) In this way they were able to obtain a high agreement with both subjective scores and the state of the art objective methods available at the time. Variations of SSIM were introduced in the form of multi-scale SSIM, to account for display resolution and viewing distance among other scale factors, and information content weighted SSIM, to better model for localised quality issues.(Z. Wang, Simoncelli, & Bovik, 2000; Z. Wang & Li, 2011) Since the invention of SSIM, many variants have been produced to account for different quality issues, but all still require an image as reference.(X. Zhang, Feng, Wang, & Xue, 2013; L. Zhang, Zhang, Mou, & Zhang, 2011; L. L. Zhang, Zhang, Mou, & Zhang, 2011).

Other methods of FR-IQA branch away from the structural similarity assumptions, by applying an image statistics-based approach. Charrier *et al.* proposed the Machine Learning-based Image Quality Measure (MLIQM), which passes both the reference image and the quality degraded image to a support vector machine (SVM) for classification into 5 quality classes.(Lebrun et al., 2012) Shnayderman *et al.* use a singular value decomposition approach to reliably deconvolve multiple distortion types applied to an image, thus allowing not only a distortion level, but a classification of the different distortion types.(Shnayderman, Gusev, &

Eskicioglu, 2006) Finally, Sheikh *et al.* proposed an Information Fidelity Criterion (IFC) which is a quality metric derived from natural scene statistics.(Sheikh, 2005) However, as mentioned before all the above metrics and models still rely on a full-reference image to evaluate image of interests quality.

### 2.6.2 RR-IQA

RR-IQA models only need access to a subset of information from the original non-degraded images. They often rely on low-level statistical properties of the image in the transform domain such as a discrete wavelet transform or discrete cosine transform .(Li & Wang, 2009; L. Ma, Wang, Liu, & Ngan, 2016) This has been used in MR image space, with the wavelet transform of a brain MRI scan compared with statistics obtained from the transform of high quality scans.(Jang, Bang, Jang, & Hwang, 2018) Other approaches involve sending a reduced field of view of the original image through the same process, imparting the same degradation on that region. This region is then compared to the whole degraded image to work out a quality score.(Carnec, Le Callet, & Barba, 2005) All approaches in RR-IQA require some reference back to the original image whether through pre-determined statistics, which cannott work on a highly variable image space, or a reduced region, where the degradation needs to be well defined. NR-IQA seeks to solve this issue, by only making quality assumptions based on the image in question.

### 2.6.3 NR-IQA

NR-IQA methods aim to give an objective quality score that has no access to any content from the gold-standard version of the image. Many algorithms in natural image space propose distortion specific solutions based on the compression the image has undergone.(Z. Wang, Simoncelli, & Bovik, 2003; Yammine, Wige,

& Kaup, 2010) However for artefact and noise detection in MR images, this is rarely useful as it requires knowledge of the distortion that has taken place. In MR image space, Magnotta and Friedman propsed signal to noise ratio (SNR) and contrast to noise ratio (CNR) as NR quality metrics to allow data comparison between and within imaging centres.(Magnotta & Friedman, 2006) A few years later, Mortamet *et al.* proposed the Quality Index 1 ($QI_1$) which had a greater than 85% agreement with subjective quality labels applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset.(Mortamet et al., 2009) More recently, Rosen *et al.* showed good agreement between a range of quantitative NR-IQA metrics, including the three above, and subjective quality scores given by three highly trained raters.(Rosen et al., 2018) These scores, in unison, can provide a good quality assessment of images post acquisition, however they make no indication of the type of artefact or the cause of degradation. We hope that by classifying into artefact classes, we will be able to provide more information on the usability of acquired data.

## 2.7 Machine and Deep Learning

Machine learning (ML) algorithms are being increasingly used to help automate analyses of medical images. Though the application of these algorithms is wide ranging, some of the common use cases include; segmentation of organs or objects of interest to automate quantitative analyses, regression or classification to determine disease class and prognosis, image registration for population level comparisons, and data imputation or synthesis for compressed sensing or data recovery.(Moyano-Cuevas et al., 2014; L. He et al., 2019; Trivizakis et al., 2019; Wei et al., 2019; Islam, 2020; Xia et al., 2021) Specifically in Liver MRI imaging, ML approaches have been used to solve similar problems from segmentation to

tumor classification.(Ivashchenko et al., 2020; X. Liu et al., 2021)

### 2.7.1 Machine Learning

ML is defined as the development of algorithms which improve via experience and the use of data.(Mitchell, 1997) The many different ML algorithms can be broadly split into two classes; supervised and unsupervised algorithms. Unsupervised algorithms seek to minimise a predefined cost function, without knowledge of any underlying classes or target values. Supervised algorithms rely on a explicit target objective that the algorithm is then minimised for. This could be anything from the class of a tumour in the liver to the mean $T_1$ value of a region. A couple of common supervised ML algorithms include, the support vector machine (SVM), random forests.(Cortes & Vapnik, 1995; Breiman, 2001)

The SVM was first developed by Cortes et al. as a new two-group classification model. They work by taking n input features which will define an n-dimensional space. They then construct a hyperplane in this space, such that the separation between class 0 and class 1 is maximised. This hyperplane is learnt from training features, where the margin, i.e. the distance between the two classes in n-dimensional space, is maximised. Then when any new cases with the same features are input into the model, they are plotted in this multidimensional space. Where they fall in relation to the hyperplane, determines their assigned class.

SVMs have been used in liver MRI in a number of classification tasks. Liu *et al.* extracted features from tumours manually segmented from contrast-enhanced (Gd-EOB-DTPA) liver MR images.(X. Liu et al., 2021) These features were input into two SVMs, with the first classifying into two independent classes of tumours (cHCC-CC vs. non-cHCC-CC) and the second classifying a different set of independent classes of tumours (HCC vs. non-HCC). They achieved an average accu-

racy of 77 % in their test set. He *et al* utilised an SVM to classify patient groups into MRE liver stiffness measurement of $\leq$ 3 kPa and $\geq$ 3 kPa as surrogates of low and high fibrosis burden respectively,(L. He et al., 2019) They combine features derived from $T_2$-weighted images, with clinical data such as blood scores, BMI and their medical history to achieve an accuacy of 81.8 %. Schawkat*et al.* sought to quantify the liver fibrosis from $T_1$- or $T_2$-weighted MR images, using an SVM.(Schawkat et al., 2020) The SVM was trained to output whether the patient had a high fibrosis score, 3-4, or low fibrosis score, 0-2. They achieved an accuracy of 85.7 %. Importantly, the three papers referenced were all published in the last 3 years, showing that the SVM still has utility in novel ML in medical imaging.

Random forests is an ensemble learning method for classification, like the SVM, or regression. It was first developed in 1995 by Tin Kam Ho, to address the over-fitting effect of a single decision tree.(Kam Ho, 1995) Since then it has been further developed by Leo Breiman in 2001, by the introduction of bootstrap aggregation.(Breiman, 2001) It consists of n-decision trees which each output a predicted class or value. Each decision tree has access to a random sample size from the training data, the bootstrap introduced by Brieman. The decision tree develops to a defined depth, where at each node in the decision tree, only a random subset of features are selected for the decision, as first developed by Kam Ho. Each decision tree then output a classification score or regression value with the accuracy maximised for the each individual decision tree. When random forests is applied to a test set, the outputs of each decision tree are combined into a single prediction.

In liver MRI, random forests have also been used extensively in classification and regression based problems. Kim *et al.* used random forests to predict the post-operative reoccurence time of HCC with a Harrell c-index of 0.716.(S. Kim et al.,

2019) Wang *et al.* similarly used a random forest model to predict the 5-year survival of patients with HCC with an AUROC of 0.758.(Y. Wang et al., 2020) Finally Lin *et al.* used a random forest classifier to predict whether a hepatobiliary phase liver MRI scan was "adequate" or "suboptimal".(Lin, Hasenstab, Moura Cunha, & Schwartzman, 2020) They achieved a highest AUROC of 0.94 when handcrafted features were input into an random forests alongside features extracted from a convolutional neural network (CNN). This last paper especially shows that random forests are still applicable, and also applicable in image quality assessment due to their explainability and ease of use. However, the majority of the field of machine learning in Liver MRI, now focuses on deep learning, a subsection of ML

### 2.7.2 Deep Learning

The previous networks, all required set features which are often processed forms of the underlying data. This often requires a high level of understanding of the dataset and field specific knowledge. Deep learning sought to address this gap by being applied to the raw data and crafting these features itself.(Lecun, Bengio, & Hinton, 2015)

Early attempts into deep learning started with the creation of the perceptron and subsequently, the multi-layer perceptron (MLP).(Rosenblatt, 1961) These networks allowed for the statistical determination of features instead of having to hand craft these features from background knowledge. Later developments into back-propagation in the 1980s using stochastic gradient descent (SGD),(Rumelhart, Hinton, & Williams, 1986) It allowed the gradients computed at the end of the MLP to be successively worked backwards through the modules. However, until the invention of the graphical processing unit, the computing power required to compute these gradients wasn't easily accessible. It is the introduction of GPUs

and mass access to the computational power to compute the gradients which has led to the large uptake of deep learning models in comparison to traditional ML techniques, especially in 2D image space in the case of convolutional neural networks (CNNs). The next three sections describe three types of deep learning models that are used throughout this thesis; the convolutional neural network, region proposal convolutional neural networks (R-CNNs), and generative adversarial neural networks (GANs)

#### 2.7.2.1 Convolutional Neural Networks (CNNs)

In 1989, Yann LeCun *et al.*, were the first group to train a CNN utilising backpropogation.(Lecun et al., 2015) However, it wasn't until twenty years later, that CNNs outperformed traditional ML techniques. AlexNet was the first deep convolutional neural network to achieve top prize at the ImageNet Visual Recognition Challenge, beating its next opponent by over 10% in top 5 error rate with a winning score of 15.6% (the percentage chance that the model does not output the correct class in its top 5 predictions).(A. Krizhevsky, Sutskever, & Hinton, 2017) This performance improvement, installing CNNs as the go to classifier model, was brought about by the introduction of data augmentation and the introduction of dropout layers, to reduce over fitting. These concepts are still the most commonly used to tackle this problem. Due to the unrivalled success of the CNN developed by Krizhevsky *et al.*, subsequent challenges contained many more submissions in this field. A year later, ZF Net tweaked the AlexNet architecture to obtain a top 5 error rate of 11.2%.(Zeiler & Fergus, 2012) Though nothing major was different between the two models, Zeiler *et al.* introduced the use of deconvolutional networks as a means for visualing feature maps in the model, thus allowing a greater level of understanding about what the network was doing. Multiple examples of CNN success in classification followed, which overall reinforced the idea that in general

deeper models are more accurate but also more likely to suffer from overfitting, and that sequential models are not necessarily optimal.(Simonyan & Zisserman, 2014; Zeng et al., 2016; K. He, Zhang, Ren, & Sun, 2016)

#### 2.7.2.2 Region proposal convolutional neural networks (R-CNNs)

Alongside the development of CNNs, Girshick *et al.* proposed the combination of these high performing networks with extracted regions.(Girshick, Donahue, Darrell, & Malik, 2014) This allowed use of CNNs which had been trained for an auxiliary task, akin to transfer learning, to classify specific regions of the image. This increased the performance of the classification by selecting only the essential information needed for the classification task, and also allowed a rough segmentation of the object of interest. Though having a high accuracy, the original R-CNN was computationally intense, taking in the region of 1 minute per image. The introduction of Fast R-CNNs greatly reduced training and classifying time, by sharing information between the convolutions in the CNN and the region selection.(Girshick, 2015) In this approach, the region of interest (ROI) is projected onto the feature map, instead of previously warping the ROI and inputting it as the image into the CNN. Later improvements to R-CNNs flipped the network and had a region selection step from the later feature maps, which were then passed for pooling and classification, in what the authors called *Faster R-CNN*.(Ren, He, Girshick, & Sun, 2017)

#### 2.7.2.3 Generative Adversarial Networks (GANs)

Goodfellow *et al.* revolutionised the field of generative models with the development of the GAN.(Goodfellow et al., n.d.) A GAN seeks to train a more accurate generative model by training it by an adversarial process. This involves use of a discriminator network which seeks to classify whether the output of the generative

network is real or fake, in comparison to real images. The GAN has since been used extensively in medical imaging for tasks such as data generation to augment datasets and address class imbalance.(Qasim et al., 2020)

Briefly, a GAN consists of a generative network which takes an input vector sampled from a gaussian distribution. This convolutional decoder network then outputs an image. The image is input into the discriminator, a classification network, which outputs whether the image is real [1] or fake [0]. The error in the prediction is back-propagated through the discriminator and the generator, updating the weights only in the generator. Additionally a "real" image is also input into the discriminator network, with the error in the prediction back-propagated through the discriminator network only.

GANs have been further adapted, with conditional GANs (cGANs) a common used network in medical imaging, including Liver MRI. cGANS are very similar to GANs but have a conditioned input, This could be the input vector sampled from a particular solution, or an encoder network taking in an image. Zhao *et al.* utilised a cGAN to output contrast enhanced Liver MRIs from non contrast enhanced equivalents.(Zhao et al., 2020) Jiang *et al.* used a cGAN to generate liver MRIs from CT scans.(Jiang et al., 2020) Finally, Liu *et al.* used a cGAN to perform the reverse synthesis of CT scans of the Liver from the MRI.(F. Liu, Feng, & Kijowski, 2019)

#### 2.7.2.4 IQA Specific Classifiers

With the increasing pressure on the healthcare system, there has been an onus on researchers to focus on improving the quality of MRI acquisitions, in the hope of decreasing the number of re-scans and increasing productivity. Deep learning based approaches, including some of the networks mentioned above, have been

used to solve this problem. A common approach in previous work has been the implementation of frameworks for radiographers or analysts to use when assessing the scans.(Esteban et al., 2017) These can often be complicated, often outputting multiple metrics of quality at once and expecting the user to be focused on it at all time, therefore introducing human error into the framework/scanning or done retrospectively, therefore not designed for the reduction of re-scans. Another approach introduces a rating from the general public through a downloadable rating app, termed "Braindr", combined with a subset of gold standard labels from neuro experts.(Keshavan, Yeatman, & Rokem, 2019) They achieved good results (AUC = 0.99) in classifying (VGG16) brain images into pass or fail by expanding their training data in this way. However, this only focused on one class without a solution and that class consisted mainly of blurring/motion artefacts and not local artefacts. Another study looked into multiple different classifiers, including Naïves Bayes and Decision Trees.(Ding, Suffren, Bellec, & Lodygensky, 2019) They achieved reasonable results classifying brains into pass and fail, but not comparable to the previously mentioned paper. Kustner *et al.* have published two recent papers on quality assessment in recent years. Firstly, they looked into classifying motion artefacts in both brain and abdominal MRIs using a patch based CNN.(Küstner, Gatidis, et al., 2018) By using a patch based technique, they were able to identify regions of the image that contained motion and areas which were devoid of it, by generating probability maps. Though still focusing on a specific artefact and not multiple, they were able to localise the artefacts to a good degree, Secondly, they looked into an active learning framework for obtaining the most relevant data for classification of abdominal images into a 1-5 quality score.(Küstner, Liebgott, et al., 2018) They used an SVM backbone for classifying which fed the borderline cases back to the user. The user then put in a query search, checking for similar images, into the overall database and sent

the results off to a trained radiographer for classifying. By implementing active learning, they were able to an accuracy of 90 % with 1000 samples compared to an accuracy of 90 % with 2000 samples when using random sampling. Finally, Biasiolli *et al.* have recently published a study into the localisation and quality control of the aorta in cine CMR.(Biasiolli et al., 2019) 13 observers manually labelled UK Biobank data into poor quality, major issues, minor issues and good quality, dependent on the slice location and the severity of the artefacts. They then used a random forest, with various spatial, histogram and shape features, to detect the ascending (AA) and proximal descending (PDA) aorta. The detection probabilities were correlated with the initial image quality score and they showed that it is possible to differentiate between large quality differences, but not minor ones. They also made no attempt to identify what the source of the artefact may be.

# 3 | Landmark Detection for Slice Location identification

## 3.1 Introduction

Multi-parametric quantitative MRI scans in the liver have the possibility to provide detailed quantitative reports of patients, and help personalise patient care and aid in disease diagnosis and monitoring progression.(Banerjee et al., 2014) However, the MRI scans used to generate the reports can be excluded for a diverse range of quality issues, as discussed in chapter 2. These issues can range from scanner and patient related effects such as motion and field susceptibility artefacts, to radiographer/technician related issues, such as sub optimal slice selection. There is a need for an objective, automated process to identify these issues, otherwise the low quality issues present will lead to errors in any analysis, rendering the data useless in the clinic.

Multiple algorithms and automated methods have been proposed for the quality control of MRI acquisitions. Since the emergence of AlexNet, the uptake of CNNs to classify occurrences in images has been vast.(B. A. Krizhevsky, Sutskever, & Hinton, 2012) Within the field of medical imaging, these networks have been used extensively. They have been used to classify patients likely to progress from Mild Cognitive Impairment to Alzheimer's Disease, and to classifying tumour type in prostate cancer, among many others.(Payan & Montana, 2015; Le et al., 2017) Recently, CNNs have also been applied to the quality control of MRI scans. Pizzaro *et al.* were able to achieve an accuracy of $\approx$75 % when using an SVM to classify structural brain MRIs into a traffic light system from green, top quality, to red, poor quality.(Pizarro et al., 2016) Kustner *et al.* were able to get an accuracy

of 93.7 % when using an SVM to classify images according to their diagnostic usability.(Küstner, Gatidis, et al., 2018) However, many studies seek to condense these issues down to a singular value of accept or reject and do not identify the cause of image degradation.(Küstner, Gatidis, et al., 2018; Esteban et al., 2017)

Though useful in determining the base outcome of whether a scan should be reacquired or not, most of the existing studies lack reasoning for their decision making, and so do not seek to rectify the underlying issue for the decision. Additionally, there is a hierarchy to issues which lead to data re-acquisition. For example, whether or not a field artefact is present is of no consequence if the slice is in the wrong location. Identifying the reason for rejection, i.e. whether it is radiographer based or scanner based, would allow for implementation of an action plan to address the cause of the problem and contribute to achieving a higher quality output from the particular scanner or radiographer in the future.

Whether the slice selection is optimal, *i.e.* in the wrong location, is one such error that should be identified early on in any QC protocol. Regardless of any other artefacts present, the acquisition will have to be reacquired. Additionally, even if the liver is imaged, not getting sufficient coverage of the liver will lead to a sub-optimal analysis, as many metrics rely either on ROIs placed in multiple different liver regions or a whole liver segmentation.(M. Liu et al., 2020) In large studies in particular, the placement of slices is often automated based on previous acquisitions. This could be anything from finding a landmark and placing the slice at that location, to moving a set distance from the previous acquisition.(UKBiobank, 2015) As humans vary hugely in size, in addition to the variation in organ shape and size, these methods are not robust, leading to some sub optimal slice locations. If we can identify the cases where the slice location is sub optimal and provide a fast solution, such as direction of adjustment, we can provide a more robust solution to this problem.

One could try to predict whether the selected slice location is optimal directly. However, the output of this model is flawed for two reasons. Firstly, this output is likely to be inexplicable. By detecting only whether the slice is optimally placed or not, we cannot give any direction on how to address the issue in the case of an incorrect slice location. Secondly, the data available for developing a model is heavily skewed towards correctly acquired slices, as is the case with most large data studies. Therefore if applied in a setting with a less rigorous QC protocol, then the model will likely fail as it has little data to learn from in order to predict a negative outcome. In order to address these limitations of a direct classification, we can instead extract information which is both explainable, allowing for self quality control, and address the issue of class imbalance by identifying an image property for which the available datasets are less imbalanced.

Object detection, and in this case organ detection, is an approach that could solve both these issues. Firstly, detection of the location and size of the organs, mainly the liver, will provide a direct explanation for any decision made on whether a re-acquisition is required. Secondly, by focusing on a sub-task, we are breaking up the problem to help tackle the class imbalance.

In this chapter, we implement a deep learning object detection network called the *Faster RCNN*, originally developed by Ren *et al.*, We evaluate whether it is possible to detect the body, multiple organs (Liver, Heart, Lungs, Spleen), vessels (Aorta, Inferior Vena Cava (IVC)) and abnormalities (Cysts) in UKBiobank abdominal scans robustly such that this model can become a reliable indicator of a QC issue, such as wrong slice acquisition.(Ren et al., 2017)

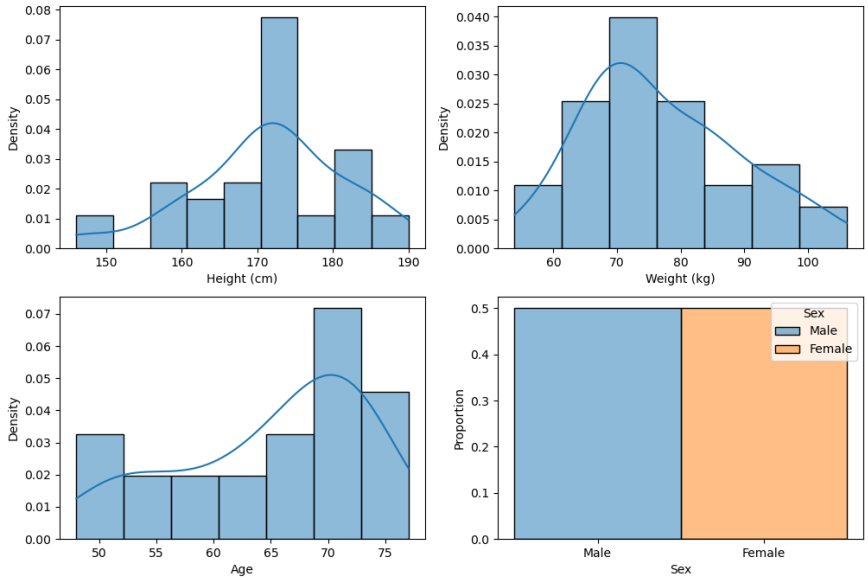



Figure 3.1: Clockwise from top left: The height distribution in cm, the weight distribution in kg, the sex distribution, and the age distribution in years

## 3.2 Methods

### 3.2.1 Data and Gold Standard Region Generation

Bounding box regions for 9 classes (Body, Liver, Lungs, Heart, Spleen, Aorta, IVC, Kidneys and Cysts) were manually cropped and assigned an integer label dependent on the class, using bespoke labelling software as described in appendix A.1. This process was performed on 298 Liver MRI scans in the UK Biobank (ShMOLLI T1 acquisition, 384x288 pixels, 7 instances). The data split by class was as follows; 298 bodies, 256 Livers, 170 Spleens, 121 Lungs, 82 Kidneys, 52 Hearts, 238 Aortas, 150 IVCs, and 48 Cysts. Population distributions are shown in fig. 3.1. The age of subjects included in the dataset is skewed towards the older end of UKBB (mean 65.1 $\pm$ 8.1 yrs, median 67.0 yrs). The distribution

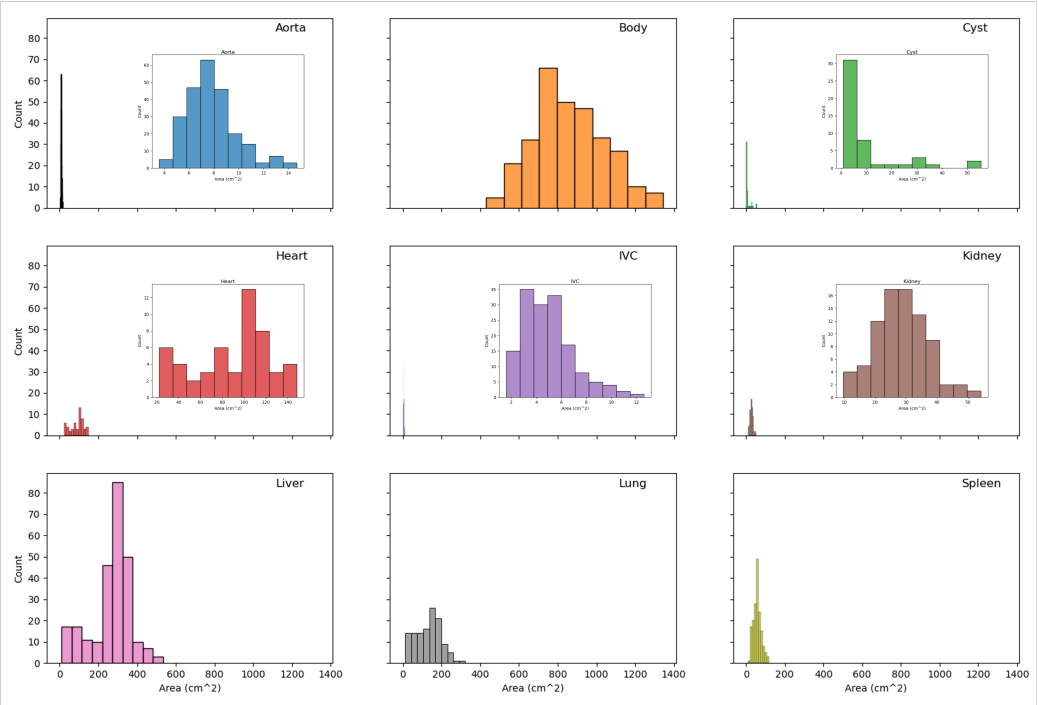



Figure 3.2: The 9 plots show the area distribution across all classes. Inserts are included where the count and/or areas are low

of bounding box areas per class are shown in fig. 3.2. In order to facilitate the comparison of both class occurrence and class areas, the x and y axes are shared across all plots. Inserts are provided where needed to visualise the shape of the distribution. The bounding box centroid location per class, with axes length set to the image size, are shown in fig. 3.3. Organ area distributions are uni-modal, with the exception of Lung and Kidneys which are bimodal. The cyst distribution is the most diffuse distribution across the image. The anterior-posterior (AP) to right-left (RL) bounding box length ratio is shown in fig. 3.4 in order to quantify the symmetry of the bounding boxes in the dataset, where 1.0 is perfectly symmetrical, > 1.0 corresponds with a bounding box which is longer in the AP direction, and < 1.0 corresponds with a bounding box which is longer in the RL direction.

At this point the assumption was made that movement within the time frame of the 7 ShMOLLI acquisitions was minimal and so a bounding box applied to one instance could be applied to the whole sequence. The 1st, 4th and 7th instances of the ShMOLLI acquisition were then passed to the neural network as a three-channel image along with the corresponding bounding box coordinates and class, allowing for information about the T1 values to be passed to the network but without having to adapt the network to take images with more than 3 channels.

The data was randomly split by subject, with no matching of class distribution in the training, validation and test sets. This led to 194 subjects for training, 40 for validation and 64 for testing. Data is available through UKBiobank, and subject IDs are available on request.

Finally, each image had a binary class associated with it. This encapsulated whether the slice had been correctly placed for imaging the liver. During placement of a LMS MOLLI acquisition, the scanner operator is required to position the single slice so that the *porta hepatis* and IVC are within the field of view (as

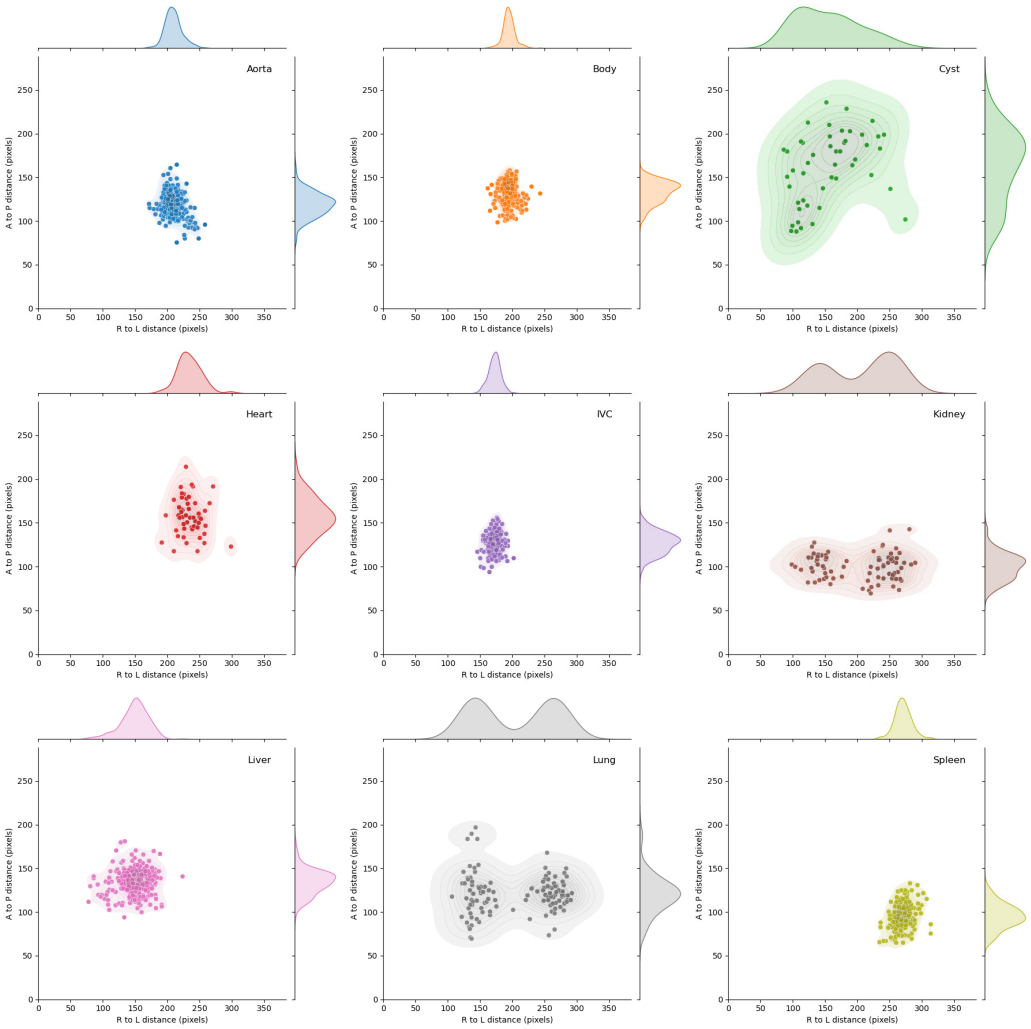



Figure 3.3: The 9 plots show the centroid distribution per class in x and y. The x and y coordinates are set to the size of the input image, allowing for comparison between classes in both distribution location and size.

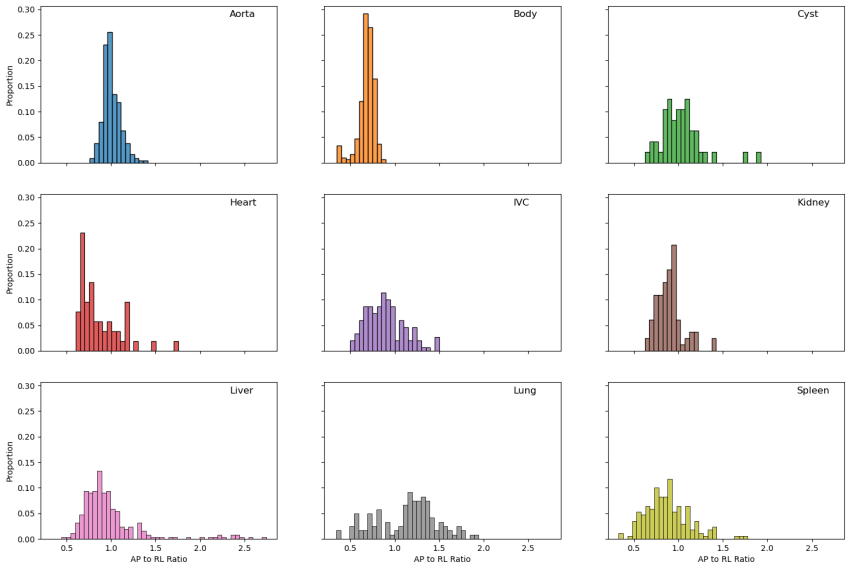



Figure 3.4: The 9 plots show the anterior-posterior length to right-left length ratio across all classes. This is a measure of the symmetry of the bounding boxes.

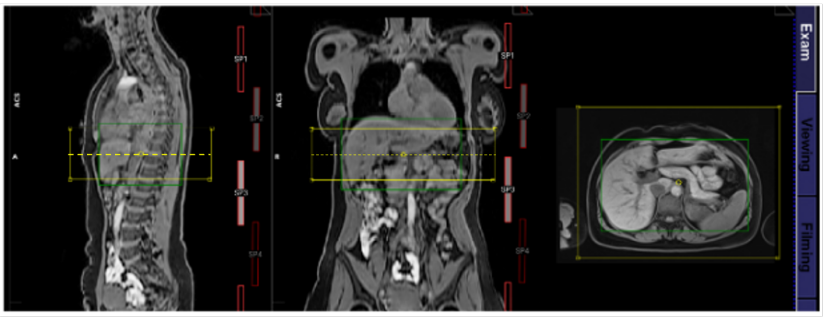



Figure 3.5: The view presented to the scanner operator on a Siemens scanner. The correct scan placement is shown with good coverage of the liver and the IVC, and *porta hepatis* in the field of view

shown in fig. 3.5). Deviations from this position are deemed a none optimal slice placement if it is within the liver, but it is still quantifiable. Any cases where the image is positioned far away from this point, the portion of liver imaged decreases, and is sometimes not present at all, and so a reliable analysis of the slice can't be performed.

### 3.2.2 Network Architecture and Loss Function

An fRCNN consists of a backbone network (normally based off a readily available classifier) and a region proposal network (RPN), which are trained end to end. Five different backbones: Mobilenet v2, Mobilenet v3, Resnet 34, DesneNet 121, and VGG 11 were tested. All backbones were pretrained on ImageNet. Depending on the backbone, different modifications were made to the final layers, as in the original fRCNN paper.(Ren et al., 2017) For the VGG 11 and Resnet 34 networks, the final two layers were removed, leading to an output from the backbone of 512 features. For the mobilenet v2 and v3, and the densenet121, the network was kept the same with numbers of output features of 1280, 960 and 1024 respectively.

The RPN is a small convolutional network which outputs region proposals in the feature maps which have been output from the fRCNN backbone. This involves

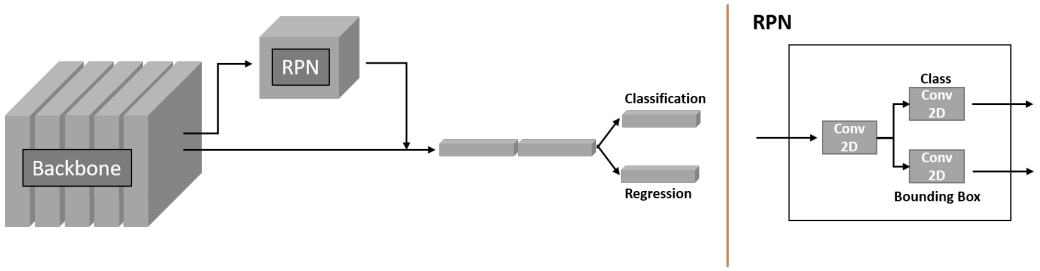



Figure 3.6: The overall architecture for the Faster R-CNN. The backbone is replaced by the appropriate model.

a sliding window passing over the feature maps, outputting k proposals corresponding to a predefined number of aspect ratios and size of region proposals at each location, termed anchors. These anchors are combined via spatial pyramid pooling, as developed by He *et al.*, which outputs feature maps of equal size irrespective of anchor size. These features are input into a classifier and regressor which output the class and box coordinates respectively. The RPN maintains the same architecture regardless of backbone.

A region proposal network was chosen in these circumstances for two main factors. Firstly, the speed at which training data can be acquired is significantly faster for bounding boxes than full segmentations, especially with use of the software developed. Secondly, the process of organ segmentation involves expert domain level knowledge of the boundaries between the organ of interest and the rest of the body. A bounding box, however, only requires that the user has a knowledge of the location and size of the organ needing to be identified. As the purpose of this work is to identify for QC purposes and not perform automated quantification of the dataset, then we believe that bounding boxes suffice for this purpose.

The overall network is optimised by applying a binary class label to each anchor, or region proposed within the features maps, 1/positive for an object or 0/negative for no object. It is assigned positive if there is a greater than 0.7 intersection over

union (IoU) with a ground truth box. If the condition is not satisfied by any anchor, then the highest IoU anchor is given a positive label. If the IoU is less than 0.3 the region is deemed negative and all other anchors are discarded. These anchors are then passed into the multi-task loss function as first described in Fast R-CNN, as seen in Eq. 3.1.(Girshick, 2015)

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{3.1}$$

where $p_i$ is the probability that anchor $i$ is an object, $t_i$ represents parameterised coordinates of the bounding box, and $*$ denotes the gold standard version. $L_{cls}$ is a log loss and $L_{reg}$ is a robust $L_1$ loss, as seen in Eq. 3.2, which contributes only in the case of positive ground truth anchors. Both losses are normalised: $N_{cls}$ by the batch size and $N_{reg}$ by the number of anchors, and the regression loss is weighted by $\lambda$

$$L_{reg}(t^u, u) = \sum_{i\in\{x,y,w,h\}} \text{smooth}_{L_1}(t_i^u - v_i) \tag{3.2}$$

where $x$,$y$,$w$ and $h$ are the coordinates, width and height of the bounding box. $t$ is the predicted box for class $u$ and $v$ is the gold standard. The $\text{smooth}_{L_1}$ loss is defined as follows:

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if}|x| < 1 \\ |x| - 0.5, & \text{otherwise,} \end{cases} \tag{3.3}$$

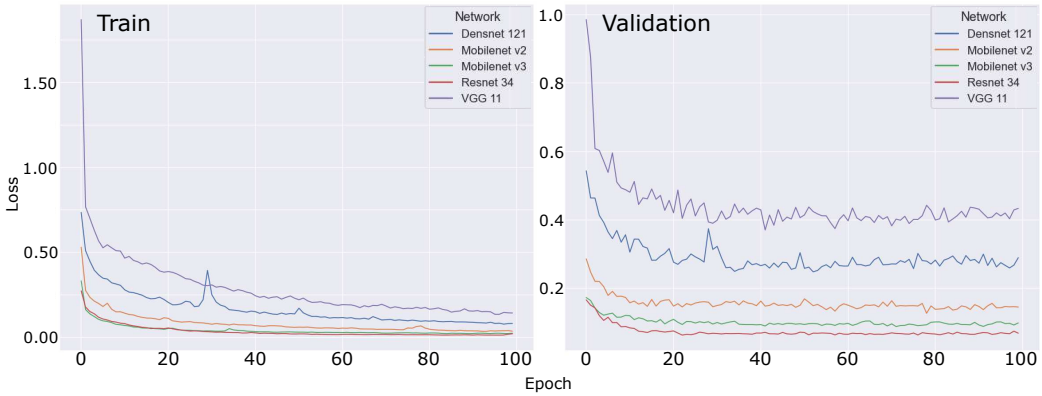



Figure 3.7: Training Loss (Left) and Validation Loss (Right) for the 5 trained backbones

## 3.3 Experiments

The networks were trained for 100 epochs on the aforementioned data on a GTX 1060 6GB (GeForce, Nvidia plc., USA). The loss, Eq. 3.1, was minimised using the Adam optimiser.(Kingma & Ba, 2014) The learning rate was fixed at 1e-3 with a batch size of 2. The training loss and validation loss are shown in Fig. 3.7.Implementation of the networks can be found at: `www.github.com/hillce/fRCNN`.

The trained network was tested on 64 cases, and the bounding box coordinates, class prediction and probability recorded. The Intersection over Union (IoU) was calculated for each box coordinate. Where there were two or more ground truth bounding boxes present of a particular class, the maximum IoU between any prediction and the 1st ground truth box was first calculated and recorded. The ground truth and predicted boxes for that case were then removed. This process was repeated until no ground truth boxes remained. This was to ensure the correct ground truth bounding box corresponded with the correct prediction. Precision recall curves were generated for each network by selecting an IoU cutoff of 0.7.

The curve was generated irrespective of class to determine the best performing network. Additional PR curves were plotted for this network, including at different IoUs, and for each class.

In order to assess the performance of the model in determining whether an incorrect slice location was acquired, the area of each bounding box was normalised to the size of the bounding box of the body, where 1 is the area of the body bounding box in that slice. This was designed to account for larger subjects having larger organs, and not just flag smaller subjects with a smaller absolute area . Tags classifying each subject into whether it was correctly located, were then correlated with these normalised volumes for each organ. This was plotted against whether an incorrect slice was acquired to determine whether the normalised volume of an organ would be an indicator of whether re-acquisition needed to be performed.

## 3.4 Results

The PR curve depicting the performance of the different backbones, shown in Fig. 3.8, shows that the best performing network is the mobilenet v3 backbone with a mean Average Precision (mAP) of 0.822. This is in contrast to the validation loss, which indicates that the ResNet backbone has the best performance. The worst performing network is that with the densenet 121 backbone, with a mAP of 0.594.

Focusing on the mobilenet v3 backbone, Fig. 3.9 shows the performance of the model at different IoU cutoffs. There is a small difference between the mAP from an IoU of 0.5 to 0.7 (0.901, 0.872, 0.822) with a more pronounced drop off for IoU of 0.8 and 0.9 (0.637, 0.210). On a class by class basis, as shown in Fig. 3.10, the average precision (AP) at an IoU of 0.7 is largest for the body, liver and Aorta at 1.000, 0.982, and 0.927 respectively. The worst three classes are Cyst, IVC, and Lungs with APs of 0.180, 0.476 and 0.675 respectively.

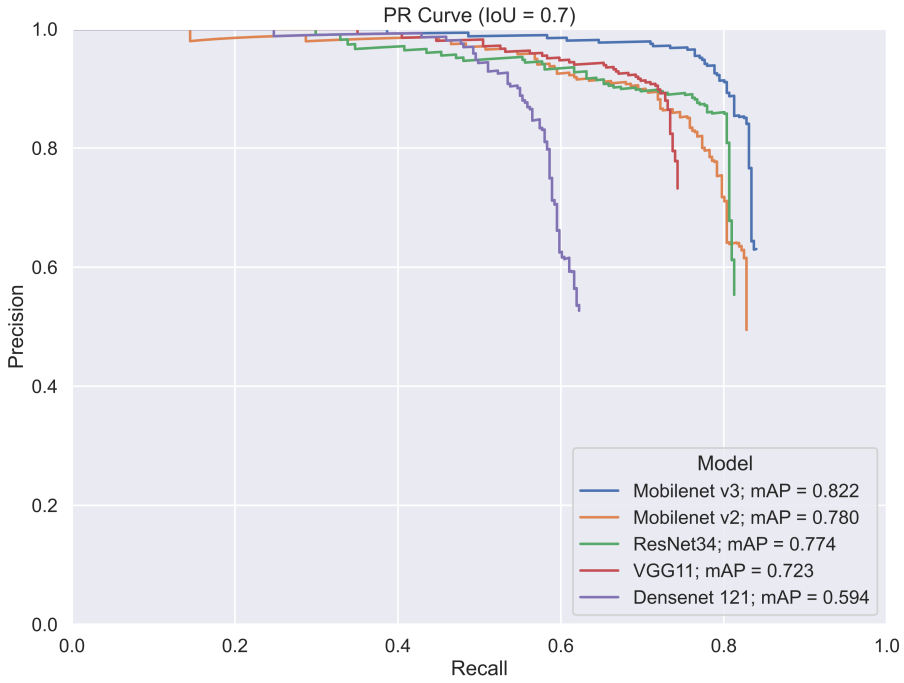



Figure 3.8: The PR curves for the 5 tested backbones at an IoU of 0.7. The best performing model at this IoU is the fRCNN with the mobilenet v3 backbone.

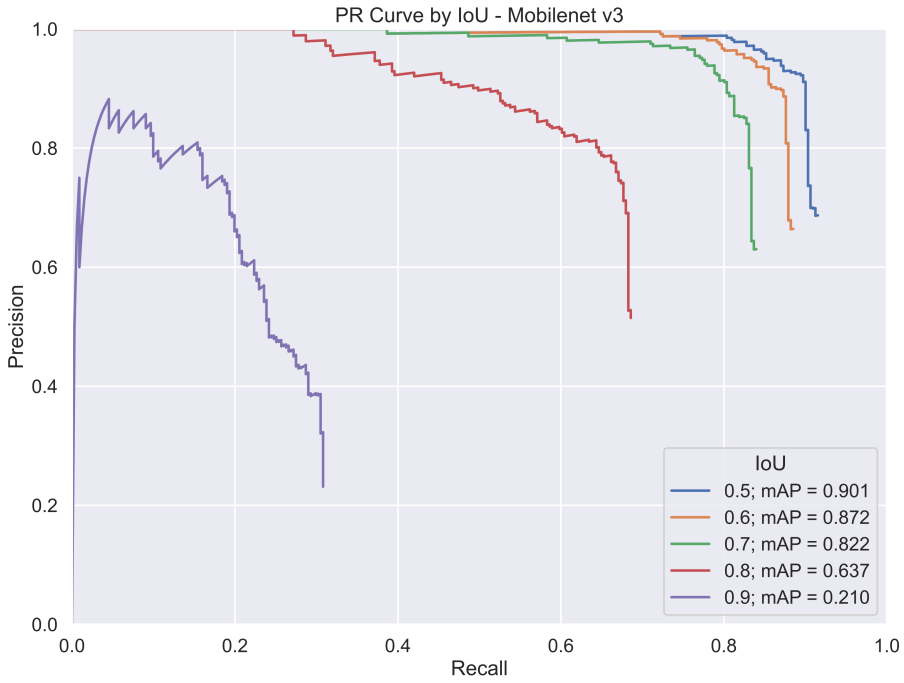



Figure 3.9: The PR curves across multiple IoU cutoff points, for the mobilenet v3 backbone.

Figure 3.11 depicts the normalised volume by organ. The normalised volume for the Liver, Lung, Heart and Spleen show significant differences with a p-value of at least 0.01 using an independent t-test. There was no significant difference between the normalised volume in the Kidney, IVC and Aorta.

A selection of examples for correct slice locations and incorrect slice locations are shown in Fig. 3.12. The Liver is detected in all the correct slice locations shown, along with a selection of other organs of interest. The Liver bounding box is also of considerable size in relation to that of the body in all cases. Additionally, the Aorta is detected in the majority of correct slice location cases, though it also appears in the majority of the incorrect slice location cases. The kidney is detected in one of the correct slices (C), only on the anatomical right side of the body, whereas in the case of the incorrect slice location (G), with kidneys detected, they appear on both sides of the body. Errors have occurred mainly in cyst detection, where the spinal cord (G) and part of the heart (E & F) have been incorrectly classified as cysts. In C, the cyst in the kidney has been missed.

## 3.5 Discussion

Both the training and validation loss curves, as shown in Fig. 3.7, highlight three main candidates mobilenet v2, mobilenet v3 and ResNet 34, as the optimal backbone networks. In all cases, the validation loss plateaus after 40 epochs, indicating that any improvement in the training loss at this point is over-fitting to the training data, though there is no sign of this over-fitting leading to a loss of accuracy in the validation data.

The assessment of the performance of the different backbones at an IoU of 0.7 further supports these three networks as the standout options. Though ResNet 34 previously outperformed mobilenet v2 in the validation set at an IoU cutoff of 0.7,

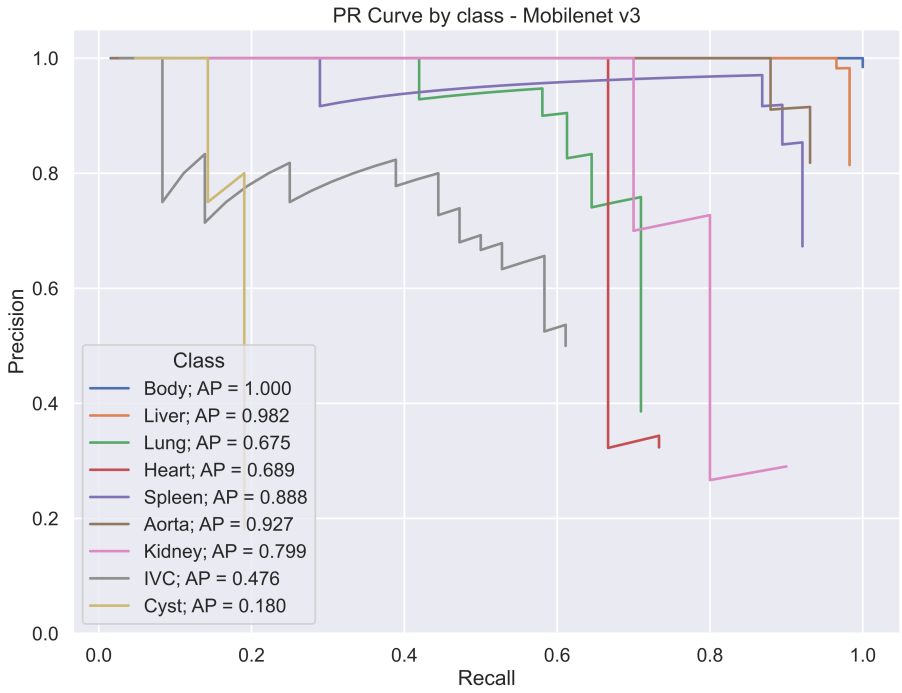



Figure 3.10: The PR curves by class for the backbone mobilenet v3, at an IoU of 0.7.

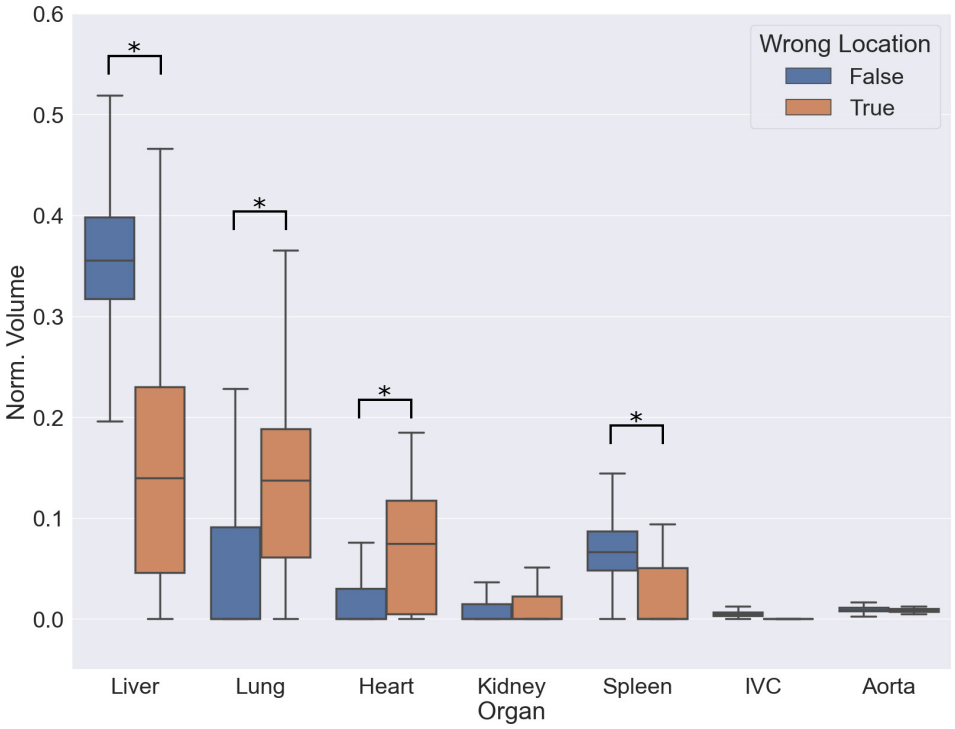



Figure 3.11: Boxplots depicting the volume of the bounding box for each class, normalised with respect to the size of the body, and whether they were identified as being the wrong location or not.

they perform very similarly with very close AP values.

The PR curves depicting change in performance versus IoU show that until under a value of 0.7 the mAP remains high. This justifies 0.7 as a cutoff point for determining whether the prediction is correct or not. This is higher than the standard literature value of 0.5, however, as performance is still high at 0.7, we believe this is more beneficial in a medical imaging setting and for determining whether the slice location is correct. The goal wasn't just detection but quantification, and this allows for more accurate volume calculations. Additionally, the high AP for the body and liver, at an IoU of 0.7, help support that the higher IoU allows for good detection.

For both cases of the liver and the body, the distribution of the centroid locations of the bounding boxes are tightly packed, as seen in fig. 3.3. Additionally, the AP to RL ratio of the bounding boxes are also tightly packed for the body and the liver, with the exception of a few outliers in the liver, as shown in 3.4. We believe this is what drives the performance in both cases due to their highly consistent location and aspect ratio. With focus on the liver, the scenarios where the AP to RL length ratio diverge from the distribution shown in fig 3.4, i.e. the AP length greatly exceeding the RL length, are where the bounding boxes are at the most inferior point of the liver. However, the centroid of these cases still lie on the anatomical right of the body; as seen in fig. 3.12, case G. We believe it is for this reason that it likely still performs well in these slices. Though a more in-depth evaluation of liver centroids, aspect ratios and areas in these most inferior slices would be needed to confirm this.

The two lowest APs occur in the IVC and in cysts. Cysts can occur within any organ, but predominantly feature in the liver, with only a few in the kidney, in our dataset. Even with the localisation in our dataset to a single organ, the location

of the cysts are distributed across the entirety of the liver, as shown in fig. 3.3. This could contribute to the difficulty in localising the cysts in the test dataset, as in comparison to the other organ locations, the cysts has the largest spacing across the whole image. Further to this, the cysts are also varied in size more considerably than other organs. Figure 3.2, shows that the area of the cysts are predominantly small, but there are outliers that are not visible in the other features of their size, the aorta and IVC. This observation combined with the fact that the cysts are the smallest class (48 samples) in our dataset, helps explain why their is poor performance in this class. In order to rectify this, more cyst data could be labelled to fully cover the distribution of sizes and positions in the body. The IVC, however, is localised to a particular anatomical location, with a dense clustering around the centre of the image, as shown in fig. 3.3. Additionally, there is no noticeable difference in the area distribution of the bounding boxes compared to the aorta, the other small vessel, which vastly outperforms the IVC in detection performance (0.476 vs. 0.927). We hypothesise that the reduced performance in the IVC could be down to the IoU cutoff imposed. The IoU cutoff of 0.7 was proposed as localisation of anatomical landmarks could be improved, for only a small reduction in the overall mAP. However, the reduction in performance seen in fig 3.9 could be localised in a a few classes, potentially including the IVC. This is because any slight shift in the centroid of the small predicted bounding box, would result a larger drop off in the IoU in comparison to the liver and body. In order to assess this further, a more in-depth look into the PR curves per class across multiple IoUs might show this to be true.

Example A in fig. 4.10, shows a case where the the spleen, though it is one of the better performing classes (AP = 0.888), has been mis-classified as a lung and a cyst within the liver has been mis-classified as a spleen. By visual inspection, the actual spleen has a large AP to RL ratio, and a centroid further up the body than

seen in most other cases. This puts it on the periphery of both distributions in fig. 3.4, and fig. 3.3 respectively. The AP to RL ratio and the centroid would place it more likely within the lung class. By comparison to H in the same figure, it is possible that a significant proportion of the lung bounding boxes contain other organs such as the spleen, heart and liver, and it is these organs which are similar in appearance to the case in A. A more in depth look into the histogram of lung and spleen classes would be needed to assess this. However, this doesn't explain the prediction of a spleen within the liver. The spleen data is localised to the anatomical left of the body and as such we would never expect to see a prediction on the right hand side. We believe that it is the combination of the area, AP to RL ratio, and contrast of the cyst that is driving the prediction on this side. Future work, would include removal of anatomically unlikely scenarios such as this.

The shape of the liver, means that the area of liver is often maximal just above the centre of the liver near the *porta hepatis*, as shown in fig. 3.5. Either side of this point it decreases in size until reaching the periphery of the organ. As the body doesn't significantly change shape and size over this region, it was hypothesised that where the normalised volume (i.e. the ratio of the area of the liver bounding box to the body bounding box) is large it would be in the region where the correct slice position should be placed, if it is small, or non-existent then it will be at the periphery of the liver and so an incorrect slice location. We additionally wanted to explore if this relation existed in the other classes

The larger organs and those that have a more definitive location are the ones which have a significant difference in the normalised volume. The Liver volume is the stand out in this case with the largest difference. This is arguably the best discriminator as a lack of liver coverage will require re-acquisition. The presence of the spleen and its good coverage is also a good indicator that the slice location is optimal, due to the proximity of the two organs. Conversely, the higher the nor-

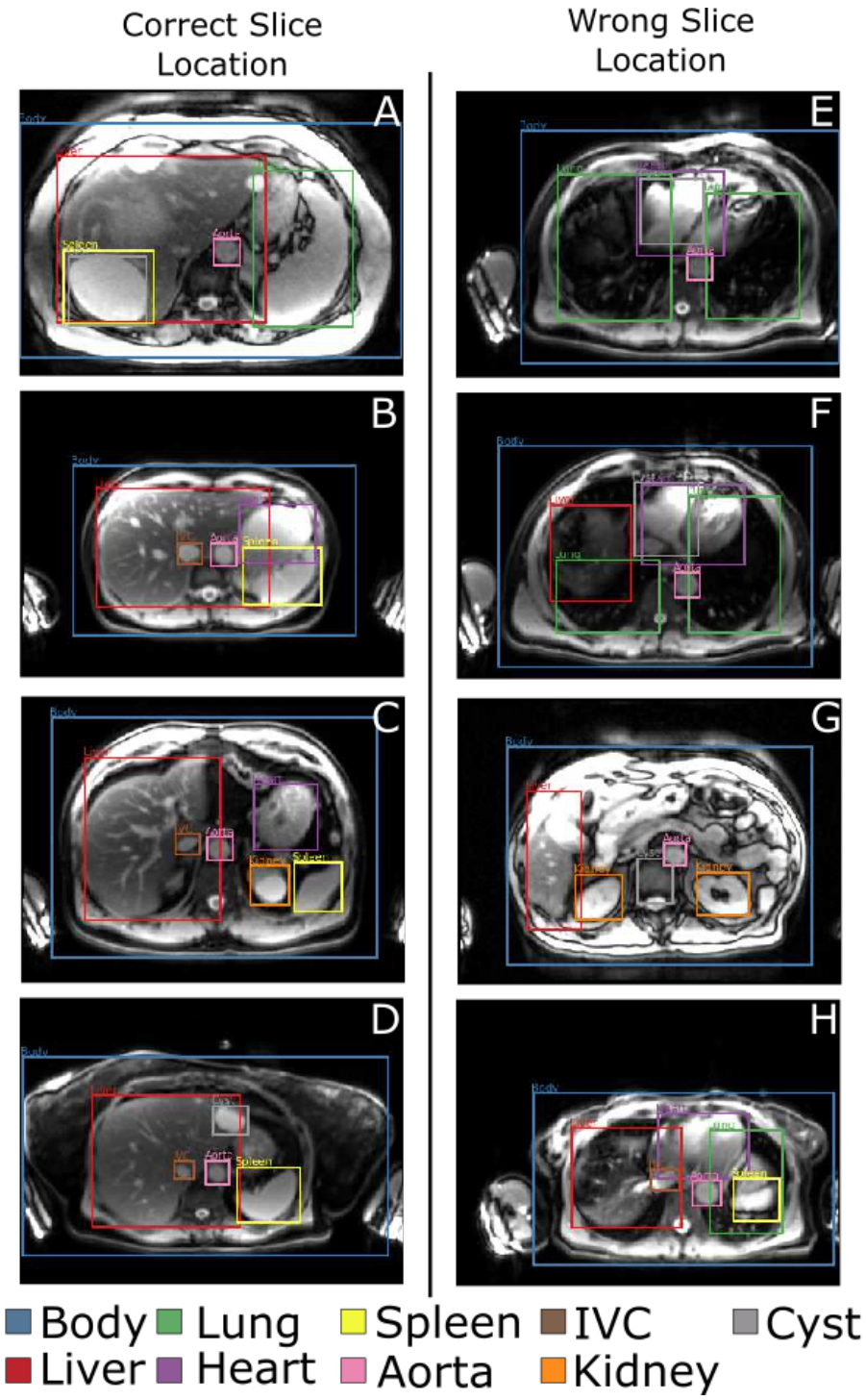



Figure 3.12: Examples of cases which have the correct slice selection and wrong slice selection.

malised volume in the lungs and heart, if present, the more likely re-acquisition is necessary. However, the presence of these two organs alone is not indicative of whether re-acquisition is necessary. The lungs can often be present either side of the liver with still good liver coverage in the slice. Similarly, the apex of the heart can appear in a slice without the need for re-acquisition. The only organ which does not show a significant difference is the kidneys. In an optimal slice, the top of a kidney is often present on the anatomical right side of the body adjacent to the spleen. However, this kidney could also be present without the presence of the liver in the slice. Additionally, an optimal slice placement does not require the presence of a kidney, and an incorrect slice location above the liver, will not contain a kidney. Further post processing of the kidney bounding box predictions could help. For example, if there is a predicted kidney on the anatomical left side of the body, e.g. where the liver is located, then this is likely to be more predictive of an issue. Whereas the anatomical right side kidney should be noted, but will have less of an impact on whether the scan is incorrectly acquired.

## 3.6 Conclusion

We showed that an object detection deep learning model, in the form of Faster RCNN, could be used to accurately bounding box the body, liver, kidney and spleen, present in a 2D axial shMOLLI acquisition. We were unable to match this performance in sub-organ size objects with the same accuracy as the main organs, with the exception of the aorta, due to a combination of training size, size variation and location variation. We went on to show that when the area of the liver, the lungs and the heart, are normalised to that of the body, there is a significant difference between cases with the correct slice placement and those without. Future work, could use a combination of these metrics to feed back

whether a re-positioning and re-acquisition of a slice is required. This would likely involve a threshold approach on the organ volumes, combined with binary classification using the presence or lack of particular organs. This feed back would help stop poor quality issues getting through to analysis. Furthermore, this model could form part of a wider collective model in identifying quality issues beyond the slice location, helping improve radiographers output.

# 4 | Data imputation for detection of localised artefacts

## 4.1 Introduction

Chapter 3 was focused on identifying a singular artefact among many. However, in order for an automated quality control (QC) protocol to have use in an MRI imaging centre, it must encompass most, if not all commonly occurring artefacts. Otherwise, we would be adding additional time pressures to the workload of radiographers with minimal gain in image quality. However, the lack of a large annotated dataset for QC poses a challenge in developing a model to identify particular cases which would require re-acquisition. We therefore explored methods that would be indicative of quality of the image without the need for explicit labels.

In such scenarios, unsupervised and self-supervised methods are often employed to solve the target problem. These include clustering techniques where the number of output labels can be defined and the algorithm seeks to separate the dataset into that number of classes. However, these models are hard to train and may lead to clustering based on undesired metrics, such as patient size, and not based on image quality without further guidance. Self-supervised methods allow the use of information inherent to the data to be used to guide the model towards an accurate solution (Jing & Tian, 2019). In sequential datasets, quality issues often present themselves in an individual image, due, for example, to a cough or ecotopic heart beat. This means the data is often of sufficient quality with the exception of a single image within the sequence. This information can be used to train a self-supervised data interpolation model to generate an idealised version

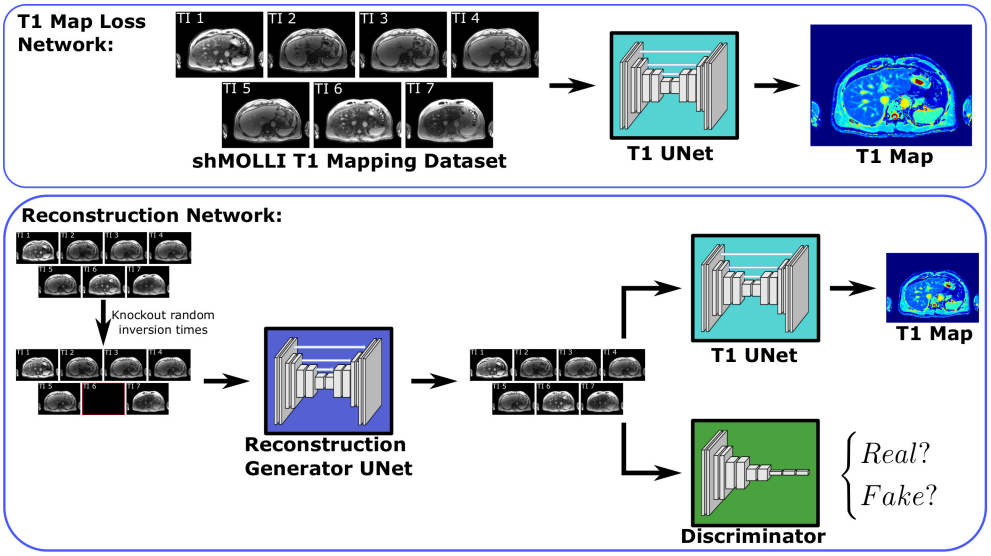



Figure 4.1: The conditional Generative Adversarial Network (cGAN) proposed to reconstruct the MOLLI datasets. Firstly, a UNet is trained to output the T1 map from a fully sampled MOLLI dataset. Secondly, a UNet based generator takes an input of a partially zero filled dataset and outputs a reconstructed dataset. A VGG-16 classifier is used as a discriminator for the adversarial loss. A smooth L1 loss is applied between the output of the generator and the original dataset. An additional smooth L1 loss is applied between the T1 map generated from the original dataset and the reconstructed dataset, to constrain the generator output to be consistent with the T1 map

of the missing image from the remaining high quality data. If the model was developed solely on a good quality dataset, then any artefacts will not appear in the idealised version and a large error between this and the original will highlight them. As the majority of acquisitions in a large population study, such as UKBB, are high quality, an assumption can be made that any data interpolation model will learn to generate high quality data and not fit to the minimal amount of low quality data available.

Various interpolation techniques for missing MR slices or images have previously been proposed to fill in missing data based on the remaining information. A simple model would be to apply a polynomial intensity interpolation between images,

however, due to motion between images in the series, this can lead to interpolation between the wrong anatomical features. Several methods rely on registration, which allows identification of corresponding voxels in a time or slice series so that a more accurate interpolation technique can be applied. This could still be intensity based (Leng, Xu, & Zhang, 2013; Horváth, Pezold, Weigel, Parmar, & Cattin, 2017), or could involve additional processing (Frakes et al., 2008). However, these methods are highly dependent on the input data with no access to population level statistics. More modern methods have utilised neural networks to synthesise missing data points, and have been shown to outperform the registration interpolation methods (Xia et al., 2021).

In this chapter, we present a novel method to quantify the quality of T1 mapping studies, and allow accurate T1 measurements from acquisitions affected by artifacts. The method quantifies data consistency across T1 scans, by reconstructing complete datasets from partial ones. This involves the introduction of a novel T1 map calculation network, which combines with a cGAN to ensure reconstructions are consistent with the T1 inversion recovery model. By testing the methods on a large population study, we test whether the error in this reconstruction is predictive of the presence of artifacts. We also test whether a 3D CNN, when applied to a combination of the reconstruction error across multiple permutations of partial datasets, will lead to a more accurate quality stratification compared to the mean reconstruction error in a dataset.

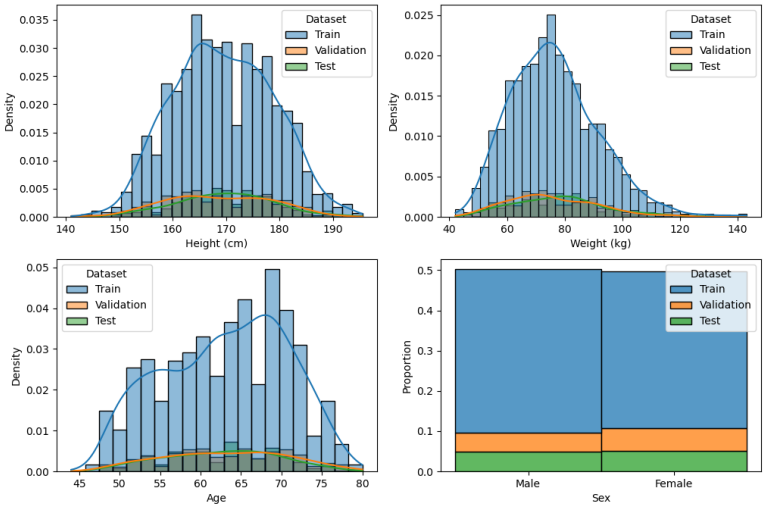



Figure 4.2: The distribution of the age, height, weight and sex in the 24,000 MOLLI cases.

## 4.2 Materials and Methods

### 4.2.1 Data

The T1 U-Net (T) for T1 map generation and the data consistency conditional Generative Adversarial Network (cGAN) were developed on 24,000 MOLLI T1 mapping acquisitions from the UKBB study (50.2 % male, 63.0 $\pm$ 7.5 years, 169.4 $\pm$ 9.3 cm, 76.6 $\pm$ 15.3 kg). The full dataset population distributions are shown in fig. 4.2.

The MOLLI T1 mapping acquisition is a sequence consisting of three inversion recovery experiments, conducted in one breath-hold, with seven magnitude and seven phase images being output from the scanner, which are subsequently used to generate a T1 map (Messroghli et al., 2004). Each acquired dataset was acquired on a Siemens 1.5T scanner at one of four imaging centres across the UK

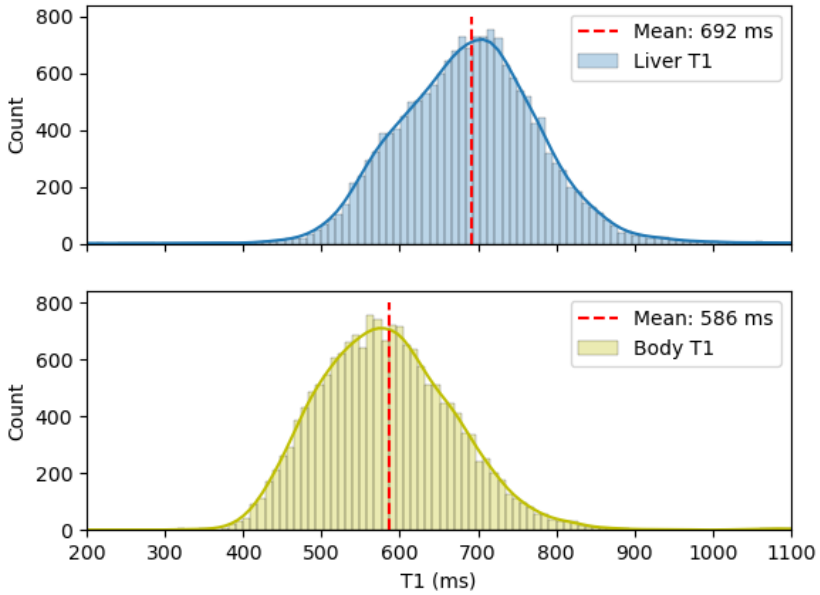



Figure 4.3: The distribution of the mean T1 values within the liver (upper plot) and within the body (lower plot) across the dataset.

(Littlejohns et al., 2020). Minimal pre-processing consisting of removal of corrupt data and extraction of the images in acquisition order was performed on the data. Only the seven magnitude images extracted were then used for network training. T1 maps were generated for all cases in MATLAB using bespoke fitting software. The distribution of T1 values in the dataset are shown in fig. 4.3. An additional 29,751 MOLLI datasets were used to develop the classification network, with overlap only with the test set from the original data. The classification data consisted of a label reflecting whether that dataset had passed QC checks and been deemed acceptable for inclusion in a large population study, or had been rejected. These QC checks were performed by a single trained reviewer at Perspectum Ltd. with random QC of output performed by an additional reviewer.

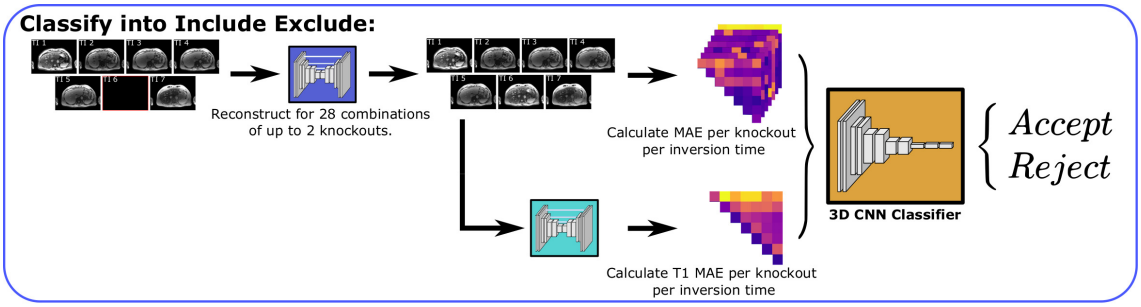



Figure 4.4: A 3D CNN is trained to classify each image into accept (i.e. data of sufficient quality) and reject (i.e. data quality insufficient) using MAE knockout maps for each inversion time and the T1 map.

### 4.2.2 Network Development

The model was developed in three main building blocks, as shown in Fig. 4.1 and Fig. 4.4; the U-Net network for synthesis of T1 maps from full datasets, the cGAN network for reconstruction of full datasets from randomly generated partial datasets, and the data quality classification based on the distribution of errors in the multiple reconstructions.

#### 4.2.2.1 T1 Mapping network

In order to implement a loss function dependent on T1 map accuracy that can be minimised during the training of the cGAN, we developed a T1 mapping network (a 5-layer deep U-Net trained on T1 acquisitions) to output a T1 map. Each layer in the U-Net consists of a 2D convolution, followed by a batch normalisation and a rectified linear unit (ReLU), repeated twice. The output of each layer is downsampled, via a 2x2 max-pooling, for the subsequent layer, and also concatenated to the corresponding layer on the encoding arm. The objective for the training of the model consists solely of a smooth L1 loss, as proposed by Girshick *et al.*, which was chosen over an L2 loss as it is less sensitive to outliers (Girshick, 2015):

$$\mathcal{L}_{recon} = \frac{1}{n}\sum_{i} smooth_{L_1}(z - T(y)) \tag{4.1}$$

in which

$$smooth_{L_1}(x) = \begin{cases} 0.5(x)^2, & \text{if} |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{4.2}$$

where $y$ is the input full MOLLI dataset, $T$ is the T1 mapping U-Net, and $z$ is the gold standard T1 map.

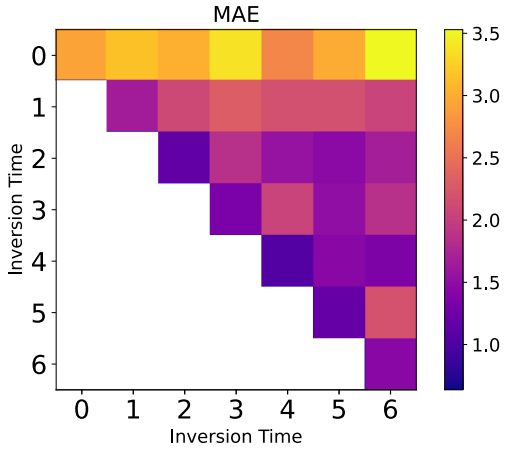



Figure 4.5: An example error array for a single image in a sequential dataset. The error in the reconstruction is dependent on which instances are removed before passing through $G$.

#### 4.2.2.2 Data consistency cGAN

Generative Adversarial Networks (GANs) traditionally consist of two competing networks, the generator ($G$) and the discriminator ($D$). $G$ seeks to generate a plausible data point within the target data distribution, whereas $D$ seeks to maximise the difference between the real data distribution and the cases generated by $G$. cGANs are a variant of these networks where the input to $G$ is no longer a

random vector, but a condition vector or matrix. For the purpose of the work presented here we conditioned a U-Net generator on MOLLI datasets, where up to three of the images have been replaced with zero valued images to create a partial dataset($x$). $G$ is then trained to output the full MOLLI dataset $G(x)$ based off this subset. This reconstruction is then input into the discriminator, which is trained to distinguish between the original dataset $y$ and $G(x)$. They are trained adversarially such that the objective of $G$ is to deceive $D$ and the objective of $D$ is to detect whether the input is a fake output from $G$. This leads to an adversarial loss function as follows (Mirza & Osindero, 2014):

$$\begin{aligned}\mathcal{L}_{cGAN}(G,D) = E_{x,y}[logD(x|y)] \\ + E_x[log(1 - D(y, G(x)))]\end{aligned} \tag{4.3}$$

In addition to the adversarial loss, it has been shown previously that there is benefit to computing the reconstruction loss, $\mathcal{L}_{recon}(G)$ across the network (Isola, Zhu, Zhou, & Efros, 2017). Therefore we evaluate the smooth L1 loss, as with the T1 mapping network previously, between $G(x)$ and $y$.

We implement a final loss term based on the T1 map U-Net. This enforces the condition that all outputs of G should be able to be used to calculate a plausible T1 map and values. The original dataset and the generated dataset are both input into the network, so the smooth L1 loss between the original datasets T1 map, $T(y)$, and the generated dataset T1 map, $T(G(x)$, can be used as an additional objective $\mathcal{L}_{T1}(G,T)$. The total loss function for the network consists of a combination of the above losses, such that:

$$\mathcal{L} = \mathcal{L}_{cGAN}(G,D) + \mathcal{L}_{recon}(G) + \mathcal{L}_{T1}(G,T) \tag{4.4}$$

The above training function was used to train $G$, a 5-layer deep U-Net with the same architecture as $T$, and $D$, a VGG-16 based classifier.The conditional GAN architecture was chosen over other reconstruction networks, such as variational auto encoders, as they have been shown to perform well in fine detail construction.(Tavakkoli, Kamran, Hossain, & Zuckerbrod, 2020) Furthermore, we wanted to include the generation of a $T_1$ map into our model on a pixel-wise basis. The established infrastructure of the cGAN, allowed us to incorporate this $T_1$ map information as a pseudo-discrimnator into out network.

#### 4.2.2.3 Classification

Prior to this point, the model had no access to any labels and so was self-supervised. The mean absolute error (MAE), see eq. 4.5, is calculated between the output of the cGAN and the original full dataset, for all possible permutations of two image removals, as shown in Fig. 4.5, which we term knockouts. This depicts the MAE across all seven channels for the corresponding removals, where the diagonal corresponds to the knockout of a single instance/inversion time, and the off-diagonal corresponds to two knockouts. The MAE is also calculated separately for each individual image in the series, leading to a MAE matrix of 7x7x7 corresponding to the 1st knockout, the 2nd knockout and the channel. Additionally the error between the original dataset T1 map and the T1 map after a knockout is performed is concatenated to this matrix, giving a 7x7x8 error matrix. This error matrix allows for qualitative assessments to be made to determine which images might not be consistent with the rest of the series. This would be visible where there is a local spike in this error matrix, for example if someone was breathing towards the end of an acquisition, you would expect to see a large error in the last image and the (0,6) knockout.

| Layers | Output Size | Kernel Size | Output Features |
|---|---|---|---|
| Conv Block (1) | 8 x 7 x 7 | 3 | 16 |
| Conv Block (2) | 8 x 7 x 7 | 3 | 32 |
| Conv Block (3) | 8 x 7 x 7 | 3 | 8 |
| Linear Layer | 1 x 1 | - | 16 |
| Classification Layer | - | - | 1 |

Table 4.1: The 3D CNN Architecture used to classify the error maps into the binary accept, reject classes

$$MAE = \frac{\sum_i^n |y_i - \hat{y}|}{n} \tag{4.5}$$

In order to automate the feedback of quality assessment decisions, we used the error matrix to train a small 3D CNN Classifier consisting of three 3D Conv-BN-ReLU layers, which classifies into a binary score of whether the data has been accepted or rejected for use in a large population analysis of UKBB data. The objective loss function is a binary cross entropy loss:

$$\mathcal{L}_{cls}(z) = -\frac{1}{N}\sum_i z_i \log(p(z_i)) + (1 - z_i) log(1 - p(z_i)) \tag{4.6}$$

where $z_i$ is the class, and $p(z_i)$ is the probability of the input fitting into class $z_i$.

## 4.3 Experimental Setup

### 4.3.1 Training

$T$ and $G$ were both trained on NVIDIA Tesla P4 GPUs for 50 epochs, by which point the respective losses had stabilised, with the classification network trained on a NVIDIA GTX 1060 Ti GPU for 200 epochs. All networks were designed and implemented in PyTorch, and the implementation can be found at: http://

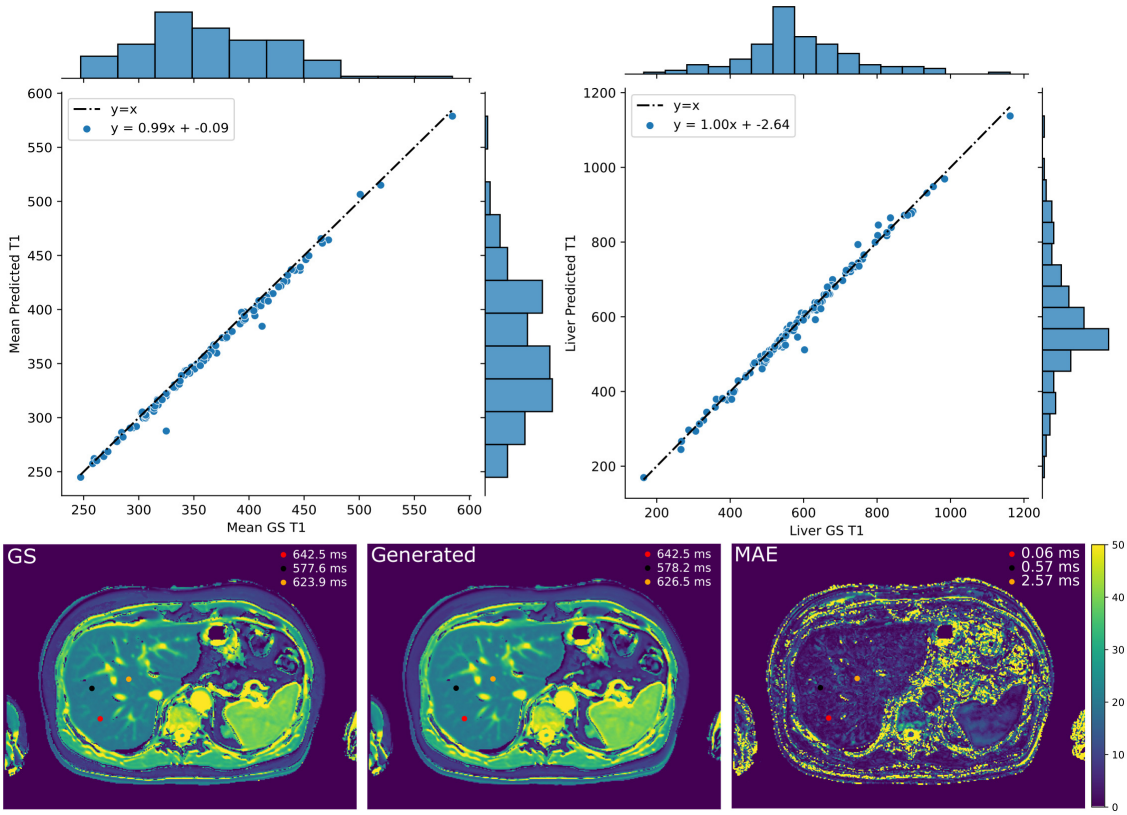



Figure 4.6: Above, two graphs show the predictive ability of the T1 loss network demonstrating that the model is both accurate and unbiased to mean T1 both over the whole body and within the liver. Below is an example case where the network has produced an near identical result to the GS map, with little error in the liver in the MAE map.

www.github.com/hillce/GAN_T1_OUT.

#### 4.3.1.1 T1 Mapping

A set of 20,000 MOLLI acquisitions were used for training $T$, with 2000 reserved for validation and 2000 cases reserved for testing. The batch size used for training was four, and Adam was used as the optimiser.(Kingma & Ba, 2014) A learning rate of $1 \times 10^{-3}$ was used throughout training.

#### 4.3.1.2 cGAN

The same training, validation and testing sets were used for $G$. During training, part of the input was replaced with an image of zeros, what we term knockout, in-situ with equal probability across all possible combinations of up to three images from seven. The batch size used for training was four, and Adam was used as the optimiser for both generator and discriminator. A learning rate of $1 \times 10^{-5}$ was used throughout training.

#### 4.3.1.3 Classification

Once the cGAN and T1 mapping network had been trained and validated, the MAE error maps were calculated for the aforementioned additional MOLLI datasets which were split into a training set of 23,800, a validation set of 2975 and a test set of 2976. The batch size used for training was 40, and Adam was used as the optimiser. A learning rate of $1 \times 10^{-3}$ was used throughout training.

### 4.3.2 Testing and Evaluation

#### 4.3.2.1 T1 Mapping

Once training was complete, the network was run on the testing set, outputting the predicted T1 maps. The T1 map was calculated for each case, and the mean T1 value from the predicted map was compared to the gold standard (GS) mean T1 value. Additionally, we randomly sampled and manually annotated 136 datasets from the test set with 3 ROIs each. The T1 value was calculated for these ROIs and averaged to give a T1 value for the liver, as is commonly done in the clinic. A regression plot of the gold standard liver T1 values, calculated using bespoke software in MATLAB, versus the reconstructed liver T1 values was generated and is shown in fig. 4.6 with an example gold standard and generated T1 map. This

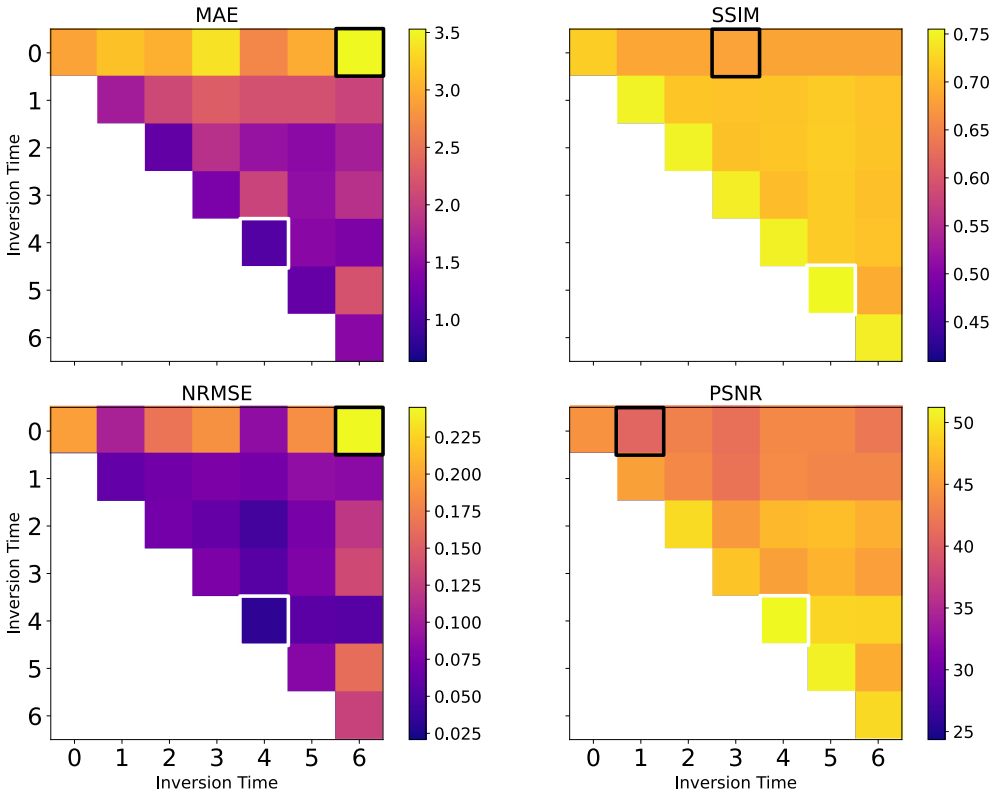



Figure 4.7: Above are four quantitative assessments of the generator's reconstruction ability. They all depict the score for each permutably invariant combination of up to 2 image removals (knockouts). The diagonal corresponds to the removal of 1 image, where the off-diagonal correspond to the removal of 2 images. These are the mean values across the entirety of the test set

was performed prior to training of the cGAN.

#### 4.3.2.2 cGAN

During training, we randomly knocked out up to 3 of the input images and replaced them with blank images. During testing we generated every permutably invariant combination of 1 and 2 replacements. Multiple metrics to determine the performance of the overall model were calculated. Alongside the MAE, we calculated the normalised root mean squared error (NRMSE)

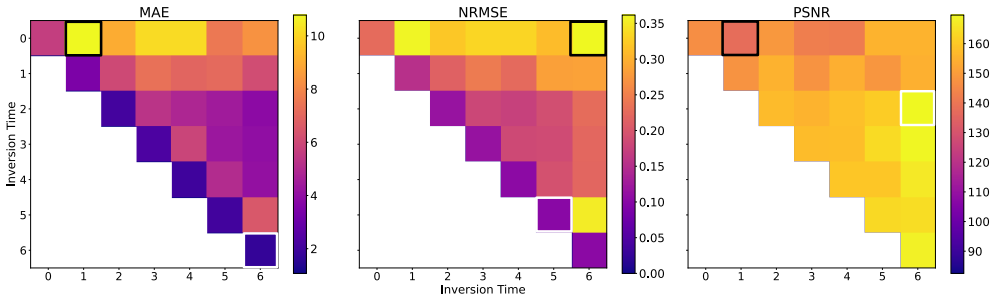



Figure 4.8: Assessment of the generators ability to reconstruct datasets which are consistent with an accurate value within the liver. As before, the diagonal corresponds to the removal of 1 image, where the off-diagonal corresponds to the removal of 2 images prior to reconstruction. These are the mean values across the entirety of the test set

$$RMSE = \sqrt{\frac{\sum_i^n (y_i - \hat{y})}{n}};$$
$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \tag{4.7}$$

where $y_i$ is the output value, and $\hat{y}$ is the desired value. $y_max$ and $y_min$ are the maximum and minimum value in the output image respectively. This measure allows to compare performance across multiple inversion times more easily. We also calculated the peak-signal-to-noise ratio

$$PSNR = 10 \log(\frac{MAX_i^2}{MSE}) \tag{4.8}$$

where $MAX_i$ is the max pixel value in each image, and $MSE$ is the mean squared error across each image. Finally, we calculated the structural similarity index (SSIM)

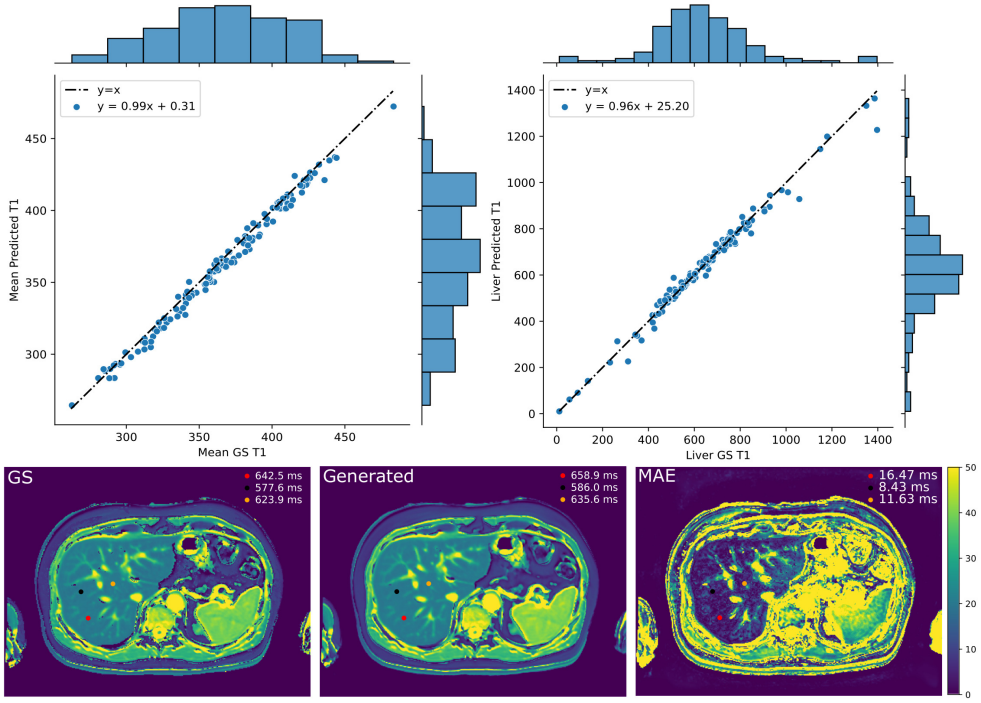



Figure 4.9: The two graphs above show regression plots between the original gold standard T1 map and the generated T1 map after reconstruction from a partial dataset. On the left is the regression plot for the T1 value across the whole body and on the right for the ROIs within the liver. Below is an example of a T1 map generated from a reconstructed dataset (knockout of 2nd and 4th image), compared to the original gold standard T1 map.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.9}$$

where $\mu_x$ and $\mu_y$ correspond to the mean value in image x and image y respectively, $\sigma_x$ and $\sigma_y$ the standard deviation in image x and image y respectively, and $\sigma_x^2$ and $\sigma_y^2$ are the variance in x and y respectively.(Z. Wang, Bovik, Sheikh, & Simoncelli, 2004) $C_1$ and $C_2$ are constants introduced to avoid instability when $\mu_x$, $\mu_y$, $\sigma_x$ and $\sigma_y$ are close to zero.

The measures were then averaged across all datasets and are shown in Fig. 4.7.

Additionally we generated the T1 maps for the original dataset using the T1 network and this was compared back to the original GS T1 maps. We generated a regression plot between the gold standard T1 value and the reconstructed T1 value across the whole body using ROIs previously used for validation of the T1 mapping network, as seen in fig. 4.9. Additionally we used along with an example of a gold standard T1 map and a T1 map generated from a reconstructed dataset. We also qualitatively assessed the error in the reconstruction across selected examples as shown in fig. 4.10.

#### 4.3.2.3 Classification

The classification network was applied to the test set in order to predict whether each image series was accepted (i.e. deemed of enough quality) or rejected (i.e. artifacts were detected that made the series unusable for analysis) for a large population study aiming at generation of T1 values. The receiver operating characteristic (ROC) curve (fig. 4.11), and precision recall (PR) curve (fig. 4.12), were calculated for each case. A comparison was made to the predictive accuracy of the error matrix without a classification network, by taking the mean value of the matrix and normalising it between 0 and 1 across the test set. This normalised error model was then treated as a quality prediction probability.

## 4.4 Results

### 4.4.1 T1 mapping

The regression plot on Fig. 4.6 shows the mean T1 value of the generated T1 map, both in the whole image and within the liver, versus the mean T1 value of the GS T1 maps. The T1 model generates an accurate T1 map which has a mean T1 value consistent with that of the GS, with an $R^2$ score of 0.988. Additionally, it

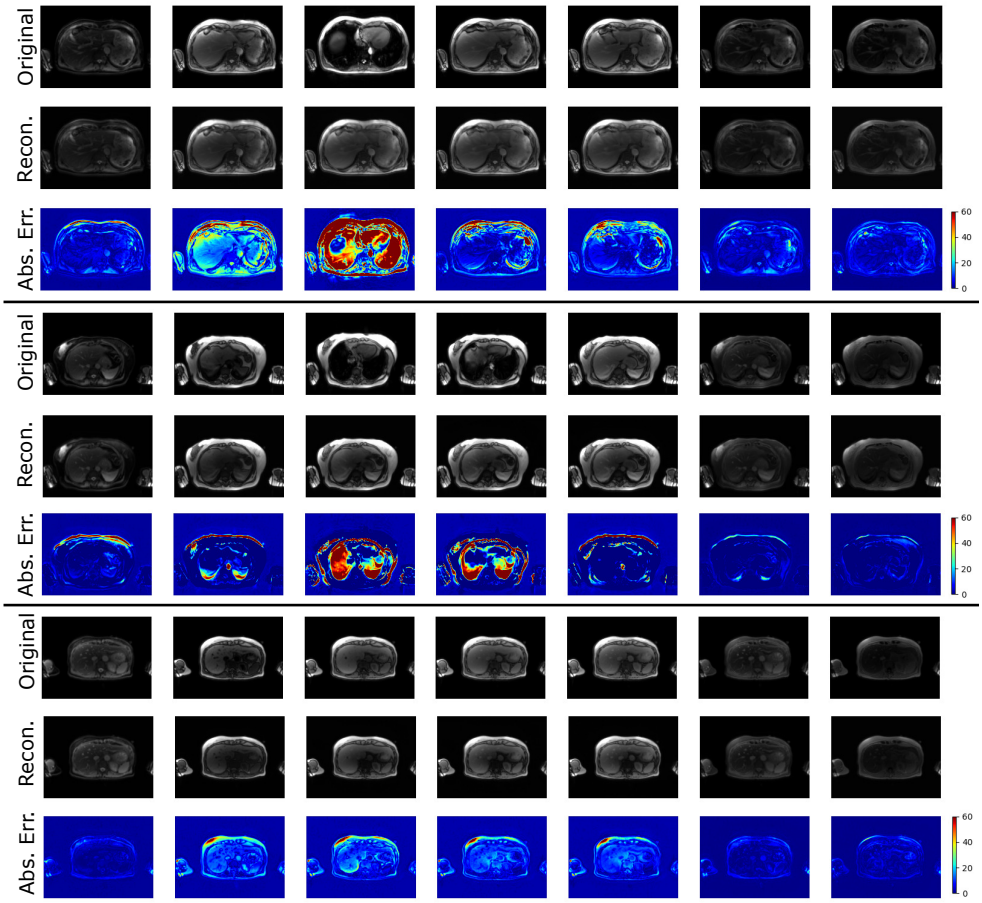



Figure 4.10: Three examples of the the original dataset, the corresponding reconstructed dataset, and the absolute error between the two. Red corresponds to a high error in the reconstruction.

generates an accurate T1 value within the liver, as shown by the second regression plot, which has an $R^2$ score of 0.991. The presented examples show that there is minimal error in the T1 map within the liver, with the error in the T1 map occurring mainly on the periphery of organs in the body.

### 4.4.2 cGAN

The peak performance of the cGAN occurred when knocking out the 4th image in the series, when assessed by MAE, NRMSE and PSNR, with values of 1.061,

0.035, and 51.229 respectively, indicated by a white box on Fig. 4.7. The mean signal intensity in the 4th image of the series across all datasets is 51.99, and so we have a mean shift of 2.04 % in the signal intensity. When assessed by SSIM, the peak performance occurred when knocking out the 5th image in the sequence, with a value of 0.755. The network performs worst when knocking out the 0th and the 6th image in the sequence, when assessed by the MAE and NRMSE with values of 3.529 and 0.245 respectively, indicated by a black box on Fig. 4.7. The SSIM is lowest when knocking out the 0th and 3rd at 0.681, and the PSNR is lowest when knocking out the 0th and the 1st image at 40.587. The performance metrics all show that the network performs better when only one image is knocked out, and the network performs poorly when one removes the 0th image combined with any other image in the sequence. We also calculated the mean absolute error across all possible knockouts, for each of the 7 images in the series, with the following values: $2.42 \pm 2.63, 2.06 \pm 1.65, 2.38 \pm 1.36, 1.66 \pm 1.47, 2.10 \pm 1.31, 2.17 \pm 1.31, 1.60 \pm 1.18$ This corresponds to a mean shift in the signal intensity of 10.44 %, 4.66 %, 4.77 %, 3.22 %, 4.04 %, 11.93 %, 9.12 % for each image in the series.

The network generates a reconstruction that is consistent with the T1 map as shown by Fig. 4.8. The network predominantly performs better when only one image from the sequence is removed, i.e. on the diagonal of the error maps. The PSNR is the exception, where the maximum value occurs when the 2nd and 6th images are removed. The network underperforms when the 0th image is removed across the board, with the worst performing cases indicated by the black boxes. The worst MAE out of all the knockouts is when the 0th and 1st image from the sequence are removed, with a value of 10.780 ms. This value is within the range of repeatability previously determined in a reproducibility and repeatability study of T1 mapping in the liver, which calculated limits of agreements of -53.6 ms to

38.5 ms.(Bachtiar et al., 2019) Fig. 4.9 further supports this agreement between the GS T1 and the reconstructed T1. The linear regression fit shows a good correlation in both the body and the liver and only marginally worse than taking the original unaffected dataset, as shown in fig. 4.6.

When qualitatively assessing the reconstruction, minimal differences between the original images and the reconstructed images were observed in the case of good quality data. However, when there was an image distorted by motion, i.e. a change in the anatomy, there was a large increase in the visual differences between the frames, and this was highlighted in the absolute error maps. In Fig. 4.10 the first two cases show two different participants breathing midway through an acquisition, whereas the final case shows an acceptable quality acquisition. The network reconstructs the dataset with a liver, which would've been present without the motion, where the lung was present. The maximal error in the dataset is present here, due to the large difference between the original and proposed images. It was also observed that the main source of error (apart from anatomical differences) was in the visceral fat, which is demonstrated in all three cases.

### 4.4.3 Classification

The ROC curve in Fig. 4.11, shows the performance of the small 3D CNN model in predicting whether the dataset was accepted or rejected for inclusion in the aforementioned large population study. It also shows the predictive quality of the MAE in the reconstruction if it is averaged across all knockouts and normalised. The normalised error model has an area under the curve (AUC) of 0.867, showing a high predictive quality without the need for training a further model. The 3D CNN model has an AUC of 0.965, outperforming the basic model. This is replicated in the PR curve in Fig. 4.12. The average precision score is greater for the 3D CNN model at 0.881 compared to 0.674 for the normalised error model.

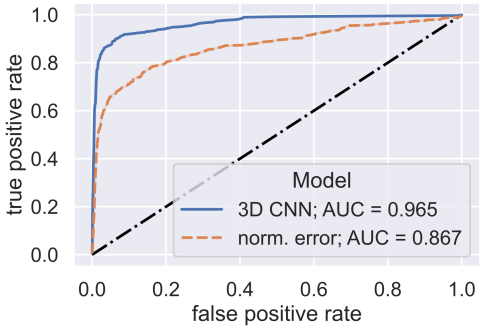



Figure 4.11: The receiver operating characteristic (ROC) curve for the 3D CNN classifier and the normalised mean absolute error.

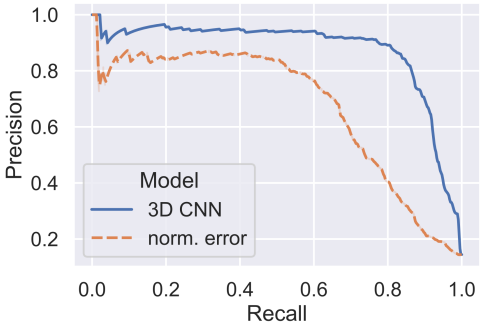



Figure 4.12: The precision recall (PR) curve for the 3D CNN classifier and the normalised mean absolute error.

## 4.5 Discussion

In this chapter, we have presented a cGAN for evaluating the data consistency within sequential datasets, and applied to a T1 mapping dataset. We showed that the use of this network gives accurate reconstructions, when up to 2 of the original datasets are removed and there are no artefacts present. With the additional use of the T1 mapping network we were able to improve the reconstruction of the images such that it was consistent with the GS T1 map. The reconstruction is within the range of a previous repeatability study and so could be used to replace image quality degraded images on data which cannot be reacquired.(Bachtiar et

al., 2019)

We then go on to show that this reconstruction can significantly fail when artefacts or other image quality degrading sources are introduced into a singular instance within the dataset. This allows for the visual identification of instances that are inconsistent with the majority of the dataset, and can be used by scanner operators to decide on whether they should re-acquire the dataset. We show that the MAE in the dataset across all the knockouts has a predictive value in determining whether a scan is of an acceptable image quality, showing that this method can be used completely label-free.

Finally, by introduction of a classifier, we have been able to show that though it is possible to give an accurate classification using the mean reconstruction error normalised across UKBB, there is an optimal weighting of each of the knockout errors which can be learnt by a neural network that further improves the classification performance from 0.867 to 0.965. This classification output can be used in tandem with the reconstruction error maps to highlight to a radiographer when re-acquisition should occur.

A previous study by Xia et al. performs a similar task in cardiac imaging.(Xia et al., 2021) They seek to reconstruct missing slices in 2D cine short axis image stacks, so as to save the data from being discarded in large population studies. For the generator they implement a 3D residual network, with an additional minimisation introduced into the network by regressing for a position label of the missing slice. This is analogous to our use of the T1 mapping network to introduce additional information into the generator. Similarly, as we use the T1 mapping network to determine reconstruction quality during testing, they encode the slice they wish to reconstruct into the network during application.

The work presented here differs from previous studies for two reasons. Firstly, it

can function without any form of labelling or identification of bad data prior to training. Xia *et al.* perform QC prior to training their imputation network in order to generate good quality data.(Xia et al., 2021) Instead we make the assumption that data imputation based on a large mostly high quality dataset will be optimised for the majority of high quality data within the dataset. We believe this holds true when training on UKBB, as the radiographers are repeating the same set of acquisitions from day to day with a consistent imaging protocol. This will create an environment where the radiographers are well trained in the mapping techniques they are using, thus leading to a large reduction of operator and scanner related artefacts. Additionally, the UKBB cohort, consisting of healthy volunteers, is less prone to patient generated artefacts, such as motion. Secondly, previous inpainting and interpolation studies have stopped upon generation of an accurate reconstruction (Xia et al., 2021). In contrast, we use our reconstruction to further highlight artefacts present in the original image(s) which were knocked out. This shows the potential for this network to highlight non-obvious flaws in the scans to a radiographer, helping improve their QC workflow, and allows reasoning to be made when the classification network reports that the dataset should be reacquired. This would be done by comparison back to the error maps, which highlight regions of acquisitions which are vastly different to what is expected as shown by Fig. 4.10. This would hopefully build trust with these techniques and allow QC of the networks to be performed in the minority of cases when it fails.

### 4.5.1 Reconstruction Evaluation

The quantitative evaluation of the network based on the test set showed that the network can reconstruct the dataset accurately. As expected, the error in the reconstruction is at its lowest in the cases when only a single instance is knocked out for the reconstruction, likely due to more ground truth information being available.

We also observe that the case when the reconstruction error is at its highest, when knocking out the 0th and 6th instance, the inversion time is at its lowest in the T1 mapping sequence, at $\sim$170 and $\sim$220 ms respectively (with all other inversion times longer than 250 ms). The rate of change in the T1 signal is largest at low inversion times, and so the measured signal is also at its most variable. Removing the information available in this area makes it difficult to accurately reconstruct much of the dataset, as the trajectory of the T1 recovery is now too undetermined. By contrast, when removing information from instances in the middle of the acquisition, i.e. high inversion time images, the error in the reconstruction is lower, likely due to the smaller variation in the signal as more and more water relaxes to equilibrium. This is corroborated by the mean percentage change in the signal intensity, which shows that the low inversion time data has the highest percentage shift.

This suggests that the model could be used with confidence to replace any artefacts that are present in these high inversion time images. However care should be taken when artefacts are identified in the low inversion images due to their importance in generating the T1 map.

### 4.5.2 T1 Values

The model's ability to reconstruct a particular instance is supported by the T1 map generated. When the new reconstructed instance is substituted in to the original dataset, there is little difference between the original and reconstructed T1 values for that dataset. It suggests that in a study where it is not possible to re-scan the subject, and corruption is limited to a single instance, one could generate accurate T1 maps based off the remaining instances and the reconstructed instance. By inclusion of the additional T1 map generation network, in a similar way to the discriminator loss, we were able to drive the generator towards a reconstruction

with a minimal error in the T1 map generated. This is seen in all knockout cases where the MAE is consistently less than 11 ms when applied to good quality data. In a repeatability study, the repeatability coefficient was calculated as 46.1 ms in the iron corrected T1 map.(Bachtiar et al., 2019) Our reconstruction method is well within this value, and so would be equivalent to reacquiring the dataset, so can be used retrospectively to correct data if needed.

We see this method having the potential to also assess the relative importance of images in sequential datasets, in the quantitative output. When applied to T1 mapping, we have seen that the first two acquisitions ( 170 and 1000 ms) have a large affect on the error in the subsequent T1 map. This could inform further T1 mapping acquisitions to sample more in the low inversion time region. If the same approach was applied to other sequential MRI techniques, it could similarly direct and inform on sequence development and improvement.

### 4.5.3 Classification

At the start of this chapter, we stressed the need for a QC technique which is label-free. By testing against a labelled dataset, we were able to show that our method has good predictive accuracy in determining the quality of the dataset. However, as we have also shown, we know that if labels are available, then the predictive quality of the model can be increased considerably. Though it could be said at this point that it should be possible to predict whether the images would be accepted or rejected in a population analysis study, we believe that our method has the benefit of being explainable compared to a traditional direct classification technique. By producing the idealised reconstruction of a subset of a dataset, we provide the difference between our idealised case and the original dataset. As shown in Fig. 4.10 the higher error reconstructions correlate with the incorrectly acquired images. In the top two cases, the anatomy present in the image changes over the

course of the first few inversion times acquired. During a shMOLLI acquisition, a breath-hold is required for 8 heartbeats. Before the acquisition is started, breathing instructions are provided to the subject to breathe in, out, and then hold their breath. Occasionally, if the subject is struggling to hold their breath for the required time, they are instructed to do two breathing cycles, before holding their breath in the exhale position. As the images with the high reconstruction error are in the first few inversion times acquired, but the last few inversion times are consistent, it is likely that the acquisition has begun before the subject has completed their breathing instructions, whether that is because they tried to complete two cycles, or were late starting a single breathing cycle. We cannot say whether they were trying to hold their breath overall, but we would be able to flag this case in-situ for re-acquisition. In the third case, the error is diffuse among multiple images. The only localised large error is in the fat tissue in the top of the body, however this is also visible in the other two acquisitions. It is also present across the whole dataset and so is not indicative of a time-localised artefact.

Future work would seek to explore whether error in the first few scans (as shown in the top two cases of fig. 4.10) can be automatically linked to the breathing instructions, so that the scanner operator can be provided with this information and take steps to make sure the subject understands what to do. This would help reduce the number of false positive cases. We would see this model being used in such a way that the radiographer has access to both the classification prediction and the reconstruction error for the worst case. This could be superimposed over the original dataset so that a quick QC could be performed drawing the user's attention to the low quality areas.

### 4.5.4 Limitations

The presented work has some limitations. We assumed that UKBB was suitably large and high quality. Though this is on the whole true, it is known that certain acquisitions suffer from quality degrading artefacts. If time was taken to QC a large enough portion of UKBB data, such that re-training of the network could be performed on this tagged data, then we envisage that the network would be able to differentiate low and high quality data further. UKBB data also only contains data from healthy volunteers. This could limit the application of this network in patient populations. Figure 4.3, shows that the cases trained on are almost all within the healthy range (< 854 ms) for liver $T_1$ values.(Mojtahed et al., 2019). Application in patients with more fibrosis, could bring their $T_1$ maps into a normal range. Given access to a suitable patient dataset, we see no reason why the model would not achieve similar results. However, there still remains the possibility of the network not being able to reconstruct unseen patient pathology, by smoothing the reconstruction to the population average. As the T1 maps are often used to indicate the level of liver fibrosis, and not to detect tumours and different anatomical bodies, we do not believe this would be an important issue. In any case, retraining on a suitable pathological dataset would probably solve this potential issue.

We would not be able to apply this network to any pulse sequence which does not acquire the entirety of k space in one shot, e.g. DIXON T2* mapping. This is because artefacts which occur in a single reading will likely appear across all sampled images. Therefore any attempt to reconstruct by knocking out a single image in the sequence would not remove information about the presence of an artefact. Instead one would have to remove the portion of k space acquired at a single time point. Though we can not plug and play our network into these sequences, we do believe by similarly removing the portion of k space acquired

at a particular time point and re-training the network, we would achieve a similar level of performance as we have by removing a single instance in the T1 mapping sequence.

We have only used magnitude data in this study. By the inclusion of the phase information, or patient meta-data, such as their height and weight, and scanner meta data, such as the inversion times of each image, we could potentially see further improvement. Some artefacts, such as those caused by scanner distortion correction algorithms, can manifest in the phase information. Therefore inclusion of this information could give a more complete picture of potential issues with the dataset. The patient information, on the other hand, would likely give a better reconstruction if included due to correlation of certain factors such as obesity with liver T1. Finally, if we were to introduce scanner meta-data such as the inversion times, we would likely reduce some of the variation in the low inversion time images, as we can fully pinpoint their position on the T1 recovery curve.

In this study we only explored the use of a UNet architecture, as the generator in the cGAN architecture. More recent developments in the field have shown that vision transformers perform well in tasks such as semantic image segmentation, and denoising.(Strudel, Garcia, Laptev, & Schmid, 2021; Chen et al., 2021) However, these networks need vast quantities of data to outperform other networks, with Chen *et al.* using 10M patches to train their denoiser. In the task performed here, due to the semi-supervised approach, we are not limited by data. Future work would look into implementing a transformer network to boost reconstruction performance in good quality cases, such as the third case presented in 4.10

## 4.6 Conclusions

In this chapter, we present our data consistency GAN, which can reconstruct multiple images within a sequential dataset to a high quality standard. In the process of reconstruction, any information which is not present in the remaining images, i.e. wrong anatomy due to motion, will not be present in the idealised reconstruction. This difference between actual acquired data and idealised data can be used to generate an error map which highlights artefacts in the image. To demonstrate the effectiveness of this approach, we qualitatively assessed the error in the reconstruction in previously identified low quality datasets. The error maps clearly display the potential of the network to identify corrupted instances by reconstructing a realistic instance based on the available data. We go further and use our reconstruction error maps to identify datasets which should be re-acquired. We do this with a high predictive ability (AUC of ROC: 0.867) without the use of labelled data at any stage, and only using the normalised error arrays. By using an additional classifier and labels, we can achieve an even higher accuracy in quality prediction (AUC of ROC: 0.965) showing that the structure of these error arrays is important.

Our network could be applied *in-situ* for any healthy volunteer based study to identify some of the poorly acquired data, and inform the scanner operator or radiographer of the need for a re-scan.However, this network is not currently applicable to patient populations. Retraining in a suitable population would rectify this. This network could also be used in an offsite setting to remedy low quality datasets and replace any corrupted instances in the generation of meaningful T1 maps. By applying the network at both stages in a study, this could both lead to the early detection of low quality datasets before the participant leaves the scanner, and salvage any low quality datasets that have slipped through to analysis,

thus minimising loss of time and money on re-scanning.

# 5 | Inclusion of Meta Data for Classification of Quality Issues

## 5.1 Introduction

Up to this point, all work has been focused entirely on what artefacts or issues can be identified only from image data. Though many artefacts do manifest in the images, they are quite often derived from some property of the acquisition parameters. This could be a different inversion time in the case of T1 mapping, or an incorrectly placed shimming volume. These variables are all stored inside the DICOM header information for every scan and so should be used to aid the determination of whether an artefact is present or not.

One method for combination of image and meta information is to bring the image into meta-data space. This could be via extraction of hand-crafted features(Simpson et al., 2015), which requires an in depth knowledge of the image space to maximise the effectiveness of the features extracted. One could also flatten the image space into a linear vector, which doesn't allow for the benefits of 2D convolutions to be realised.

Another method for fusion of meta and image data has included late concatenation of meta data onto the dense layer of a CNN.(Gessert, Nielsen, Shaikh, Werner, & Schlaefer, 2020) This approach has the benefit of treating both data sources independently until the final step, and thus developing the optimal route for the different data formats. However, it does not allow for the interaction of the two data types to benefit the overall outcome. Reed *et al.* developed a similar approach for their discriminator, but allowed for association between the two different data types, by concatenating the data into a late convolutional layer.(Reed et al., 2016)

Though now allowing for combination of the meta data with spatial features, the late concatenation means meta information is only available at low resolutions. Earlier inclusion would allow for the introduction of meta-data in earlier features. Furthermore, in the case of Reed *et al.*, the meta-data was replicated until it was the same spatial dimensions as the feature layer it was to be concatenated to. This is unlikely to be an optimal combination, as different meta values will correspond with different image regions.

Mostapha *et al.* sought to rectify this with inclusion of meta-data throughout the network in the form of a "braided-block".(Mostapha et al., 2020) This braided block consists of a meta-data fully connected arm and an image convolutional arm. There are cross connections between the two arms with image metrics such as the mean and standard deviation concatenated onto the meta-data pathway and a fully connected layer on the meta-data pathway used to scale the output of the image convolution. This inclusion of meta-data at the start of the network allows for the meta-data to be used at different resolutions, and joint optimisation if one wished to output both image and linear information. However, there is no basis for the image statistics to be used as the cross-over and this could be sub optimal.

In this chapter, we seek to evaluate the performance of each method for meta-data and image data fusion by classifying T1 shMOLLI images on whether they were accepted or rejected in a previous analysis, as in chapter 4, and on classifying into the underlying classes which make up the inclusion and exclusion super-classes. We analyse the effect of only using meta-data or only using image data to assess the relative importance of each on classification accuracy. We then proceed to evaluate the classification accuracy of each combined method, and finally compare this to a majority voting process of two separate classifiers for meta-data and image data respectively.

| **Rejection Tag:** | **Description:** |
|---|---|
| Artifacts | Visible artefacts of unknown origin in image. |
| Incorrect Sequence | Incorrect/Outdated sequence has been used for acquisition. |
| Field Artifacts | Visible field artifacts are present in the image, as described in chapter 2. |
| High Iron | Iron levels calculated from DIXON T2* acquisition outside the range of values in which the multi-parametric acquisition has been validated. |
| Outside T1 Range | T1 value falls outside the range in which the multi-parametric acquisitions has been validated. |
| Mask Failure | Over 50 % of the segmentation mask has a poor T1 fitting coefficient. |
| PDFF > 35 % | The PDFF value calculated from the IDEAL acquisition is greater than 35 %; the case falls outside the range of values in which the multi-parametric acquisition has been validated. |
| Rainbow | Banding across the liver leading to "rainbow" colors across the liver in the T1 map. |
| Segmentation Fail | No or poor liver segmentation after passing through semi-automated pipeline. |
| Wrong Location | Image has sub-optimal or no liver coverage. |
| Zero cT1 | T1 value in the liver is 0 after correction for iron from DIXON T2* acquisition. |
| Other Unlisted | Any other reason for exclusion from the large dataset analysis |

Table 5.1: Table depicting the image quality issue types that can occur in the UKBB datasets used for the multi-parametric analysis.

## 5.2 Experimental Setup

Meta-data was extracted from DICOM image data for all shMOLLI liver scans from UKBB. Meta-data parameters were pre-processed by normalisation of all values into a normal distribution, and one-hot encoding of categorical variables. This meta-data was then input into two models without image data; a random forest and a multi-layer perceptron. It was also used alongside image data in two other models; the aforementioned Braided network and our model, a combination of dcGAN and VGG-16, which is explained in more detail below. Finally the output of the random forest and VGG-16 were combined using majority voting and this was compared to using VGG-16 only on image data.

All weights in convolutional networks were initialised with Xavier initialisation.(Glorot & Bengio, 2010) Implementation of the experiments can be found at: `www.github.com/hillce/Meta_T1`.

### 5.2.1 Data

In chapter 4, 31,199 cases of the UKBiobank shMOLLI data had been used to develop a 3D CNN classifier with tags of accept or reject depending on whether it made it through QC for use in a large analysis. Each subject in that data also had a reason for defining as accept or reject, for reasons covered in table 5.1.

In total there are 12 artefact classes, with their distribution in UKBiobank depicted in fig. 5.1. The most populous class is Incorrect Sequence. During the initial stages of the UKBiobank imaging stages, the shMOLLI acquisition was incorrectly setup on each site, which would lead to a systematic shift in the cT1 values calculated from these scans. For this reason, they were excluded. This class is analogous with an incorrect sequence being performed at any site.

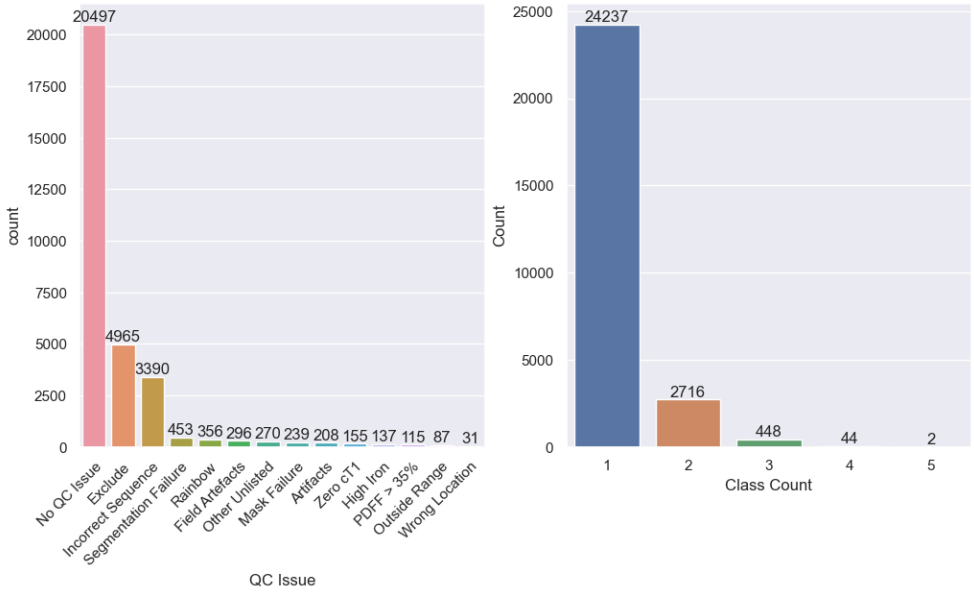



Figure 5.1: **Left:** The distribution of QC classes across the dataset. The majority of cases have no QC issues associated with them. **Right:** The occurrence rate of multi-label cases.

The High Iron, Outside Range, PDFF > 35 % and Zero $cT_1$ classes are specific to the imaging protocols performed during Liver*MultiScan*. All of the classes lead to the Liver*MultiScan* protocol being applied in data outside of the values in which it was validated, and so identification before analysis would be advantageous. They are also linked; a higher iron would lead to a correction of the $cT_1$ which could bring it outside the range in which $cT_1$ is validated.

Mask Failure and Segmentation Fail are semi-automatically labelled pipeline failures. To process the vast quantities of data going through UKBiobank, a U-Net is run on the cases, and then a combination of automated and manual checks are performed on the output. In the case of the Segmentation Fail class, this is where the automatic segmentation has been deemed to be of insufficient quality, which could also be caused by another QC issue, and has to be manually corrected. In the case of Mask Failure, if less than $50\%$ of the pixels within the liver segmen-

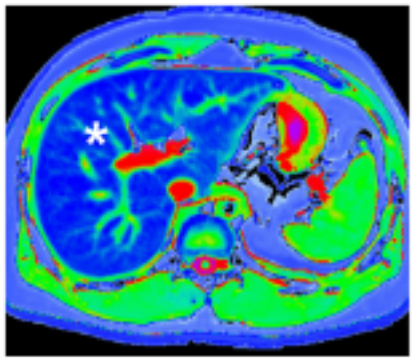


Figure 5.2: An example $cT_1$ colormap.(Banerjee et al., 2014) If a rainbow artefact is present, then bands of different colors appear across the liver.

tation are below a stated goodness of fit threshold, $R^2$, in fitting the T1 recovery model, then this case is flagged as low quality.

The Rainbow class potentially covers artefacts caused by field inhomogeneities, but is only visible in the $cT_1$ map. The name is derived from the presentation of a rainbow of colors across the liver, due to the color mapping, see fig 5.2, used to display the $cT_1$ images in liver*MultiScan*. As the liver is largely homogeneous, variation of multiple different blocks of $cT_1$ values across the organ is deemed a quality issue, and excluded from analysis.

The final artifacts are to differing degrees applicable to any type of MRI image. Firstly the Wrong Location class, as described in chapter 3, occurs when the scanner operator selects the wrong slice for imaging the desired organ, in this case the liver. The Field Artifacts, also known as field susceptibility artifacts, cover any visible artifact that is likely cause by field inhomogeneities. The Artifact class covers any visible artifact that is of unknown origin.

The final class, Other Unlisted, describes any case that has been excluded on any other grounds not covered above, from the population analysis of $cT_1$ values in UKBiobank. The reasons behind each case are unknown.

Examples of the classes are shown in fig. 5.3. Each subject could have multiple labels or none, however the majority of cases have a single class, as shown in fig. 5.1.

Four main types of data were used in assessing the inclusion of meta-data into machine learning models for the classification of these QC issues: meta-data extracted from the DICOM headers, bounding box coordinates for various organs of interest from chapter 3, the error matrix information from chapter 4 and the seven shMOLLI images. These data types can be grouped into meta-data, image derived, i.e. the error matrix and bounding box coordinates, and image data classes.

The meta-data contained multiple data formats including categorical data, such as sex and sequence type, discrete, such as the inversion time, and continuous data, such as weight and imaging frequency.

The categorical data extracted from the DICOM header information included variables such as coils used, and phase encoding direction. Unique Identifiers (UIDs) which reference the sequence used were removed, and any DICOM header that had no variation across the dataset, such as "Imaged Nucleus" which was 1H for all datasets. The remaining categorical data was one hot encoded. The continuous DICOM header data, such as imaging frequency underwent min-max normalisation across the dataset.

The bounding boxes were also normalised, but to the dimensions of the image, such that a value of 288 or 384 in x and y respectively were normalised to one. Input into the model, was either the normalised coordinates of the box (x0,y0,x1,y1) or zeros if the organ was not present.

For the error matrix data, all non-measured values were removed, *i.e.* for the duplicated knockouts of (0,1) and (1,0) only the first value was calculated and

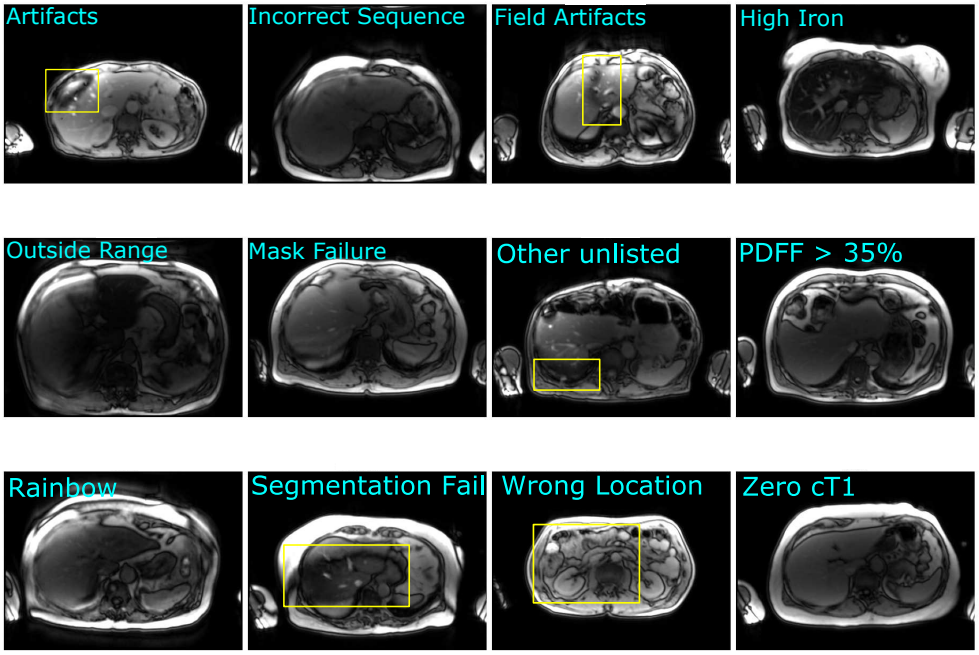



Figure 5.3: Examples of each class. Some artifacts manifest in localised sections of the scan, e.g. field artifacts, others are universal, e.g. wrong location and high iron, and finally others seem not to have any manifestation in the images, e.g. PDFF > 35 %. Yellow boxes highlight regions of interest, whether it be manifestation of the image quality issue, or highlighting lack of liver in the case of wrong location

would be used, and the array was flattened into a singular array with a length of 224. The values were normalised between the max value across UKBB and zero.

The final formatted data had an input feature size of 694 and label size of 11, for each participant. Depending on the experiment, this was combined with the image data (288 x 384 x 7) for each participant, as described in more detail in the data section of 4.

### 5.2.2 Random Forests

Random Forests are an ensemble of multiple decision trees, the general schematic of which is shown in Fig. 5.4, where the output decision is a majority vote of all the decision trees. The number and depth of the decision trees can be controlled, in order to optimise the fitting and return a probabilistic prediction of a class. The random forest returns a binary classification, and as such cannot be used to classify into multiple artefact types. However, using a one versus rest approach, one can use multiple random forests to build a multi-label classifier.

We performed a total of five experiments for each class, based on the inclusion of various data inputs. These were: inclusion of all data; inclusion of only header data and bounding box data; inclusion of only header data; inclusion of error matrix data and bounding box data; and finally inclusion of only the error matrix information.

For all experiments performed, the hyper parameters were set as follows: number of decision trees = 100, maximum decision tree depth = 10, determined empirically by optimising when all data is included. Additionally, we performed repeated stratified 5-fold cross validation for each class with 10 repetitions, in order to provide reliable statistical analysis. An ROC curve was plotted for the best performing random forest for each class, along with the box plot depicting the range of AUROC values across each class.

### 5.2.3 Multi-layer Perceptron

Multi-layer perceptrons (MLP) are a collection of fully connected layers, with N hidden states between input features and output labels. We implemented a MLP with four hidden layers (256,128,32) with dropout and ReLU activation. The network architecture was determined empirically by optimisation when all data

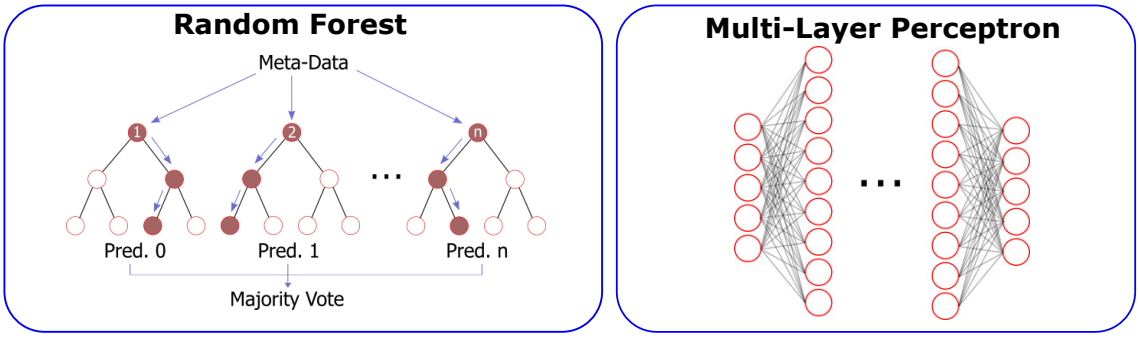



Figure 5.4: Shown are the two meta-data only networks. The random forest, which is an ensemble of decision trees, and the multi-layer perceptron, which is a combination of fully connected layers.

was input. Unlike random forests, MLPs can be used as multi-class, multi-label classifiers, and so we trained it as such. However, in order to facilitate a better comparison between the performance of the MLP and the random forests, we also trained a one versus rest classifier for each class. The splitting of data for multiple experiments was performed as with the random forest experiments (five experiments for each class), plus an additional five experiments for the multi-label, multi-class MLP. Repeated stratified 5-fold cross validation was performed as previously with identical data splits. The hyper parameters for training were set as follows: learning rate = 0.0001, number of epochs = 100, dropout rate = 0.5. We used a binary cross entropy loss function as our objective.

### 5.2.4 Braided Network

The braided network was implemented by ourselves in PyTorch, based off the description in the Mostapha *et al.* paper. The classifier, shown in 5.5, consisted of an initial braided block, a 2x2 max pooling, followed by 4 sets of 2x braided blocks with a 2x2 max pooling. The output of the final braided block was flattened and 4 dense layers were applied, with a final input of 11 classes. Due to time constraints, the experiments that were performed on the random forests and the MLP,

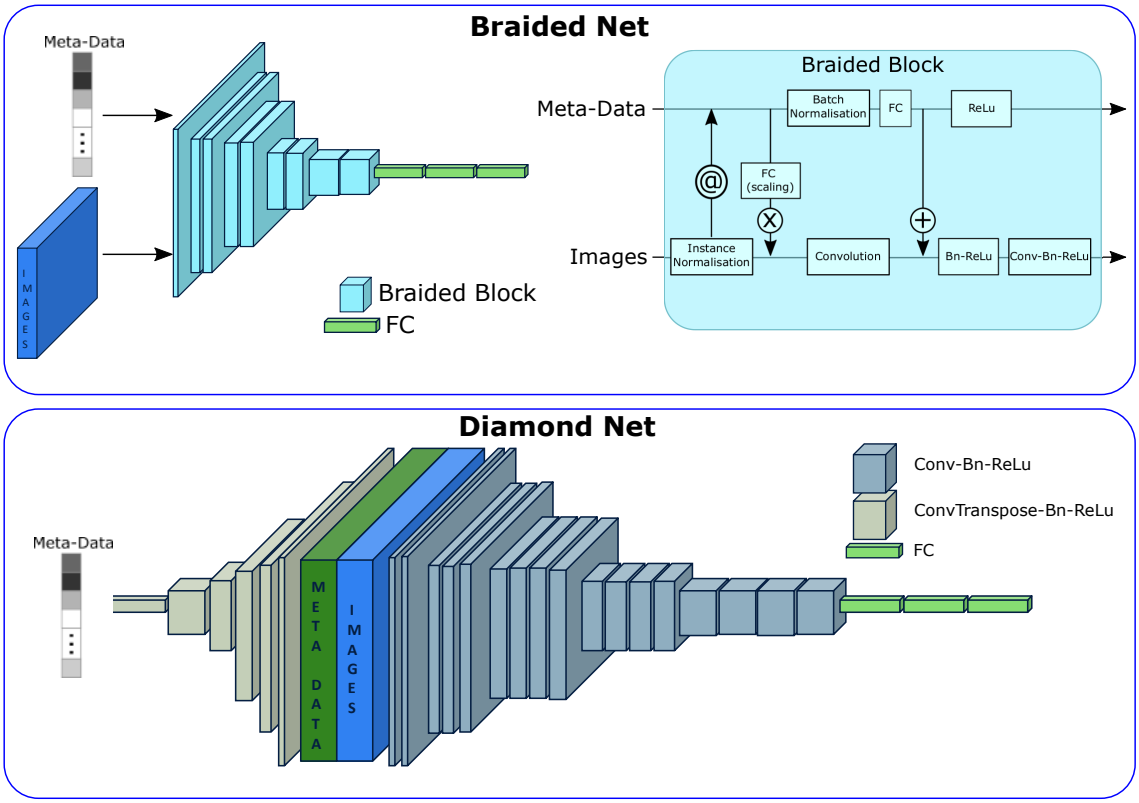



Figure 5.5: Shown are two combined networks which use meta-data and image data in one pass. The braided network, by Mostapha *et al.*, uses the braided block which transfers information from the image data pathway to the meta-data pathway and back. Our diamond network, brings meta-data into image space by use of a network based on DCGAN, followed by classification with a VGG network.

were not applied to the braided network. Instead, we performed one experiment with the inclusion of all meta data (dicom header information and derived features) and image data. For all subsequent networks the same single experiment was performed. The ROC curves for each class were calculated. Additionally, we calculated the micro-averaged and macro-averaged ROC curves. The macro-averaged is the ROC curved calculated across a flattened array of all predictions, and as such will be influenced by a class imbalance. The micro-averaged ROC is the average of the ROC curve for each class, and as such will weight all classes equally.

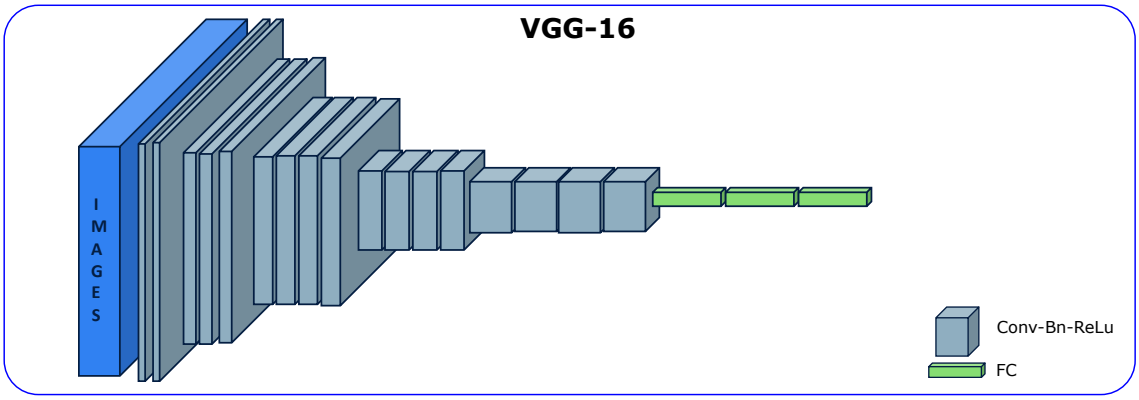



Figure 5.6: Above is shown the standard architecture for a VGG-16 classification network.

### 5.2.5 Diamond Network

The diamond network, our novel model, was also implemented in PyTorch. The architecture was based on DCGAN(?, ?) and VGG-16(Simonyan & Zisserman, 2014). The DCGAN consists of 6 ConvTranspose-Bn-ReLU layers with a final output matched to the size of the image data (i.e. 288 x 384 x 7). This was then concatenated to the image data, leading to an input of 288 x 384 x 14 input into the VGG-16 network, pre-trained on ImageNet, which has a final output of 11 classes. The same statistical analysis which was applied to the braided network was also applied to the diamond network.

### 5.2.6 VGG network and random forest combination

A VGG-16 network, which had been pre-trained on ImageNet, had the initial and final layers adapted from its original 3 input channels and 1000 output classes to 7 input channels and 1 or 12 output classes depending on the experiment. This provided an approach, with no access to the meta-data information data, to compare to. This VGG-16 network was also used in conjunction with the best performing random forest for each class and a majority (mean prediction probability) vote

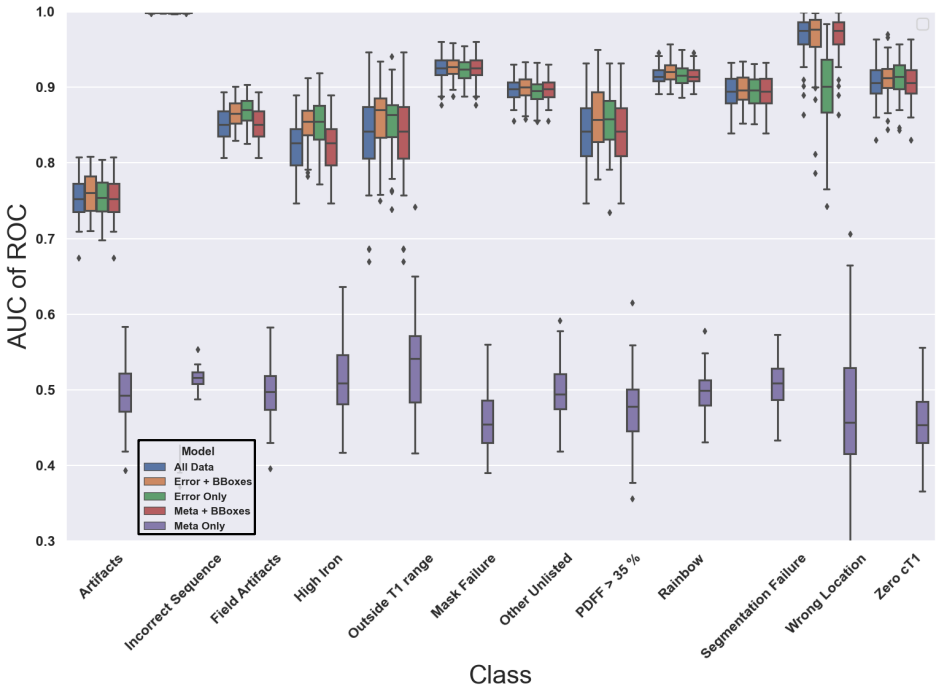



Figure 5.7: Above, boxplots depict the performance of random forests over the repeated 5-fold validation in each class, for multiple different combinations of data input.

output recorded. To test whether any performance increase was due to the combination of image data and meta-data through the two networks or due to the ensemble effect of two predictions, we trained an additional VGG-16 network and combined the two VGG-16 predictions into a majority vote. The same statistical analysis which was applied to the braided network was also applied to these networks.

## 5.3 Results

### 5.3.1 Random Forests Vs. MLP

The results for the random forest experiments are shown in fig. 5.7 and table 5.3.1. The results for the MLP models are depicted in fig. 5.8 and table 5.3.2. The corresponding ROC curves for the best performing models are shown in 5.9 and 5.4 for the random forests and MLPs respectively.

In the random forest experiments, the version with access to the error matrix and the bounding box coordinates obtains the best performance across most classes. The exceptions to this occur in the field artefacts, high iron and wrong location classes. In the field artefact and high iron classes, the highest mean AUC is when including the error only, with mean AUCs of 0.869 and 0.854 respectively. In the wrong location class, the highest mean AUC occurs when only including the meta data (i.e. DICOM headers) and the bounding box coordinates at 0.967, or when including all data with the same AUC. All output a high classification score in "Incorrect Sequence" except for in the case of inclusion of the meta data only. In the case of inclusion of meta-data only, the performance is equivalent to random chance, which is clearly seen in fig. 5.7

In the case of the MLP, there is more variation in which datasets lead to a better performance. Notably, the inclusion of meta-data only is now no longer random chance and in the case of incorrect sequence, field artefacts, high iron and outside T1 range, is the best performing data input for the MLP with mean AUCs of 0.988, 0.752, 0.533 and 0.754 respectively. Additionally, The MLP is outperformed by the random forest models in almost all classes. The random forest models only

Table 5.2: Table depicting the mean and std. dev. in the AUC, for each data combination, for the random forests

| | All Meta | Error + BBoxes | Error Only | Meta + BBoxes | Meta Only |
|---|---|---|---|---|---|
| Artifacts | $0.752 \pm 0.029$ | $\mathbf{0.760 \pm 0.026}$ | $0.755 \pm 0.026$ | $0.752 \pm 0.029$ | $0.494 \pm 0.040$ |
| Incorrect Sequence | $\mathbf{0.999 \pm 0.000}$ | $\mathbf{0.999 \pm 0.000}$ | $0.998 \pm 0.000$ | $\mathbf{0.999 \pm 0.000}$ | $0.516 \pm 0.012$ |
| Field artefacts | $0.853 \pm 0.020$ | $0.866 \pm 0.018$ | $\mathbf{0.869 \pm 0.019}$ | $0.853 \pm 0.020$ | $0.498 \pm 0.040$ |
| High Iron | $0.824 \pm 0.032$ | $0.852 \pm 0.030$ | $\mathbf{0.854 \pm 0.034}$ | $0.824 \pm 0.032$ | $0.515 \pm 0.047$ |
| Outside T1 Range | $0.836 \pm 0.058$ | $\mathbf{0.858 \pm 0.045}$ | $0.856 \pm 0.042$ | $0.836 \pm 0.058$ | $0.530 \pm 0.066$ |
| Mask Failure | $0.924 \pm 0.017$ | $\mathbf{0.928 \pm 0.015}$ | $0.922 \pm 0.016$ | $0.924 \pm 0.017$ | $0.459 \pm 0.039$ |
| PDFF > 35 % | $0.842 \pm 0.043$ | $\mathbf{0.858 \pm 0.042}$ | $0.856 \pm 0.040$ | $0.842 \pm 0.043$ | $0.472 \pm 0.050$ |
| Rainbow | $0.916 \pm 0.012$ | $\mathbf{0.920 \pm 0.013}$ | $0.916 \pm 0.014$ | $0.916 \pm 0.012$ | $0.497 \pm 0.030$ |
| Segmentation Fail | $0.894 \pm 0.022$ | $\mathbf{0.897 \pm 0.020}$ | $0.894 \pm 0.021$ | $0.894 \pm 0.022$ | $0.506 \pm 0.030$ |
| Wrong Location | $\mathbf{0.967 \pm 0.030}$ | $0.960 \pm 0.045$ | $0.897 \pm 0.055$ | $\mathbf{0.967 \pm 0.030}$ | $0.466 \pm 0.095$ |
| Zero cT1 | $0.906 \pm 0.026$ | $\mathbf{0.911 \pm 0.026}$ | $\mathbf{0.911 \pm 0.025}$ | $0.906 \pm 0.026$ | $0.457 \pm 0.038$ |
| Other Unlisted | $0.897 \pm 0.016$ | $\mathbf{0.900 \pm 0.016}$ | $0.894 \pm 0.016$ | $0.897 \pm 0.016$ | $0.497 \pm 0.034$ |

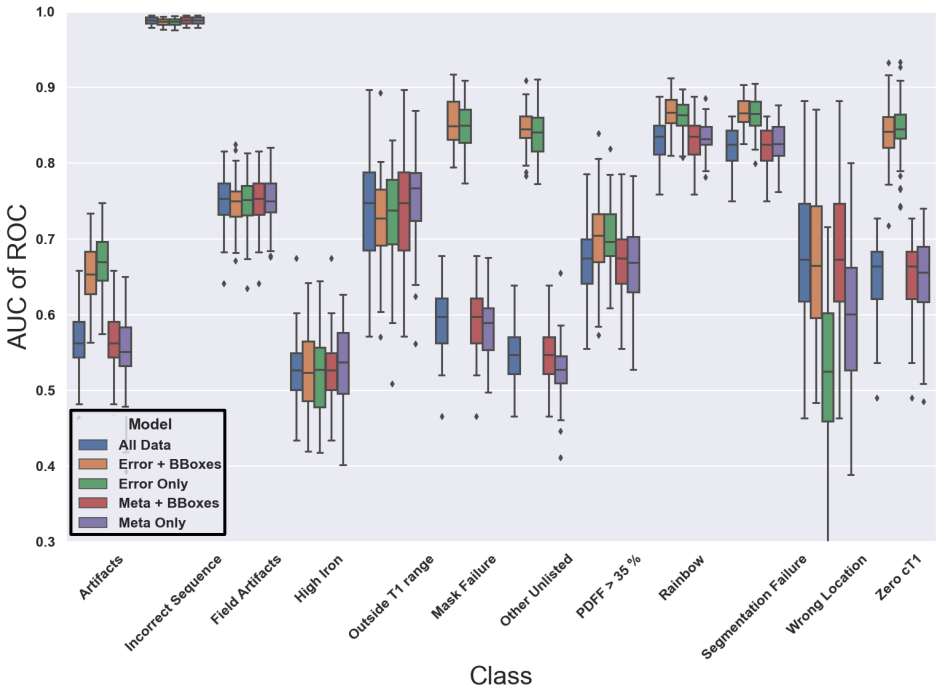



Figure 5.8: Boxplots depict the performance of MLPs over the repeated 5-fold validation in each class, for multiple different combinations of data input.

have one scenario where the best performing data input model combination has an AUC of less than 0.8 (artifacts with error matrix values and bounding boxes). In contrast, only six of the twelve classes for the MLP experiments are above 0.8.

The ROC curves, of the best performing MLP and random forests models depict a similar story to that of the tables and boxplots. The macro-averaged ROC curve for the random forest experiments, fig. 5.9 - F, shows that exclusion of the error matrix information leads to a definitive drop off in classification performance. This is $\sim 0.2$ to meta-data plus bounding boxes from the three data combinations that include the error matrix information. This trend is less pronounced in the case of the MLP models, however the "error and bounding box" and "error only" data combinations are also the best performing classes.

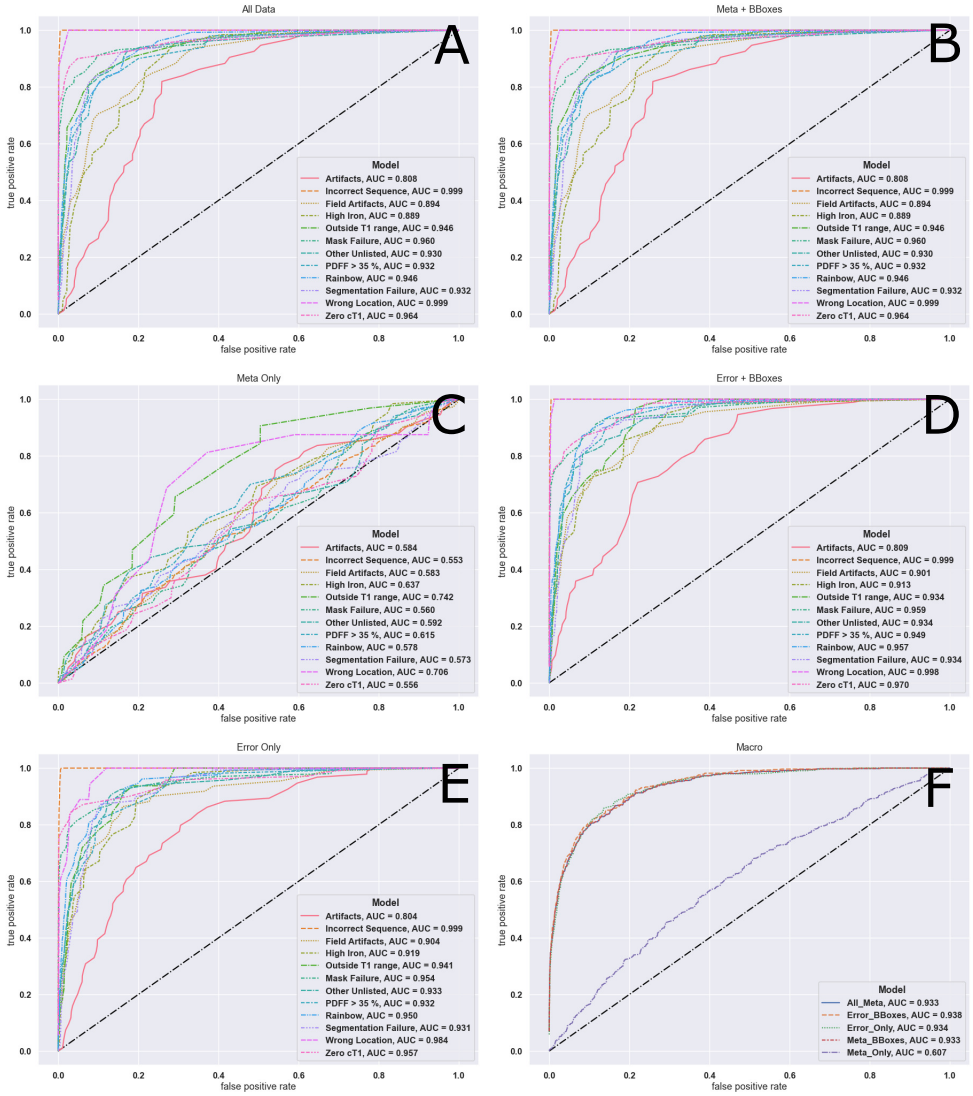



Figure 5.9: Above are shown the ROC curves for the random forest models across multiple classes and data inclusions as follows: inclusion of all data (A), inclusion of dicom header data with bounding box coordinates (B), inclusion of dicom header data only (C), inclusion of error matrix and bounding box coordinates (D), inclusion of error matrix only (E), and finally the macro-average ROC for each data input (F).

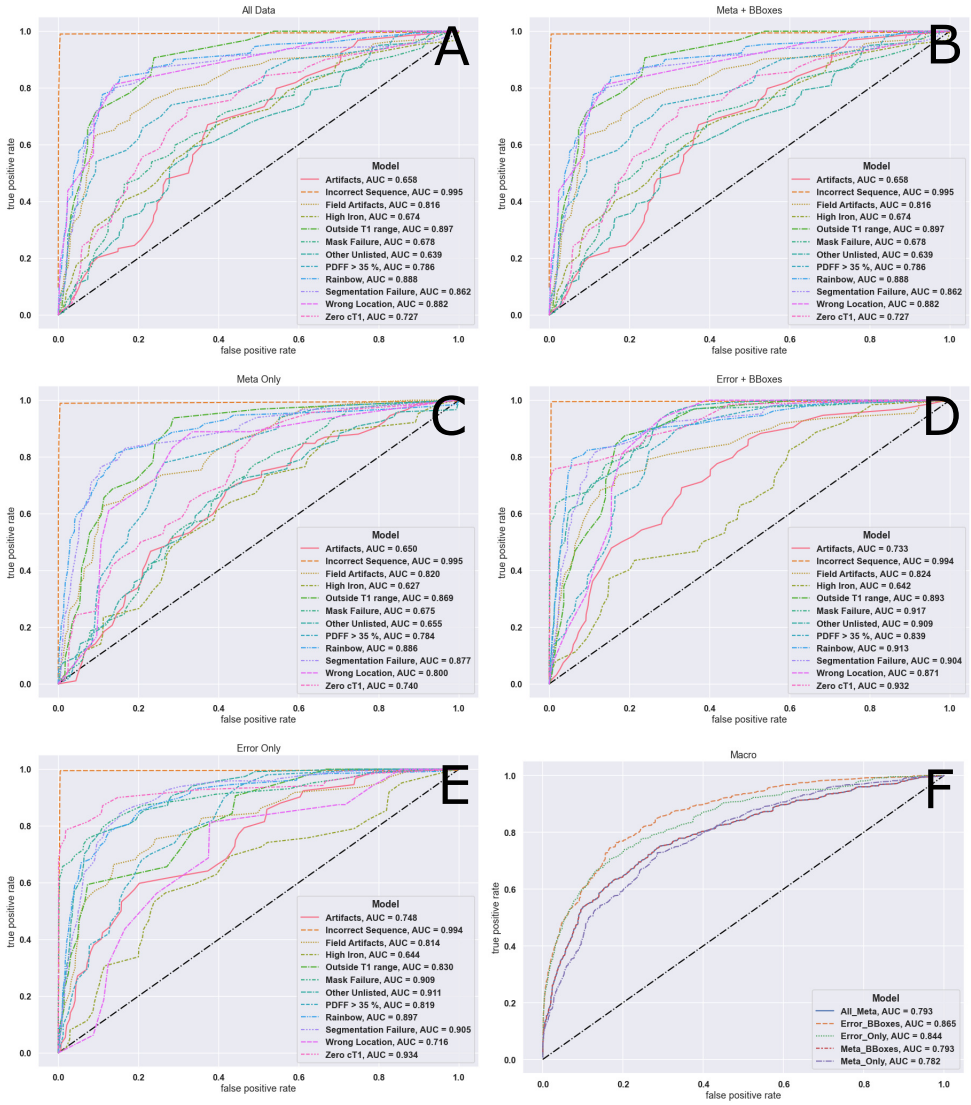



Figure 5.10: ROC curves for the MLP models across multiple classes and data inclusions as follows: inclusion of all data (A), inclusion of dicom header data with bounding box coordinates (B), inclusion of dicom header data only (C), inclusion of error matrix and bounding box coordinates (D), inclusion of error matrix only (E), and finally the macro-average ROC for each data input (F).

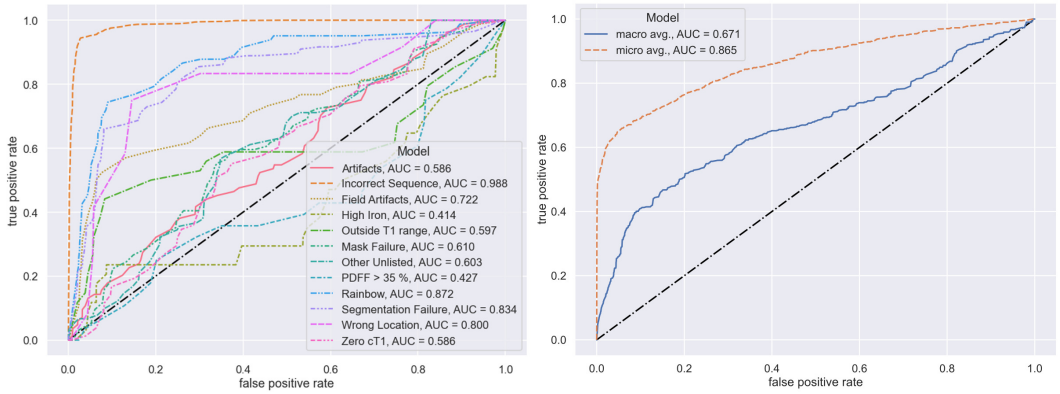



Figure 5.11: Left: The ROC curve for the braided network across each class. Right: The macro- and micro- averaged ROC curves.

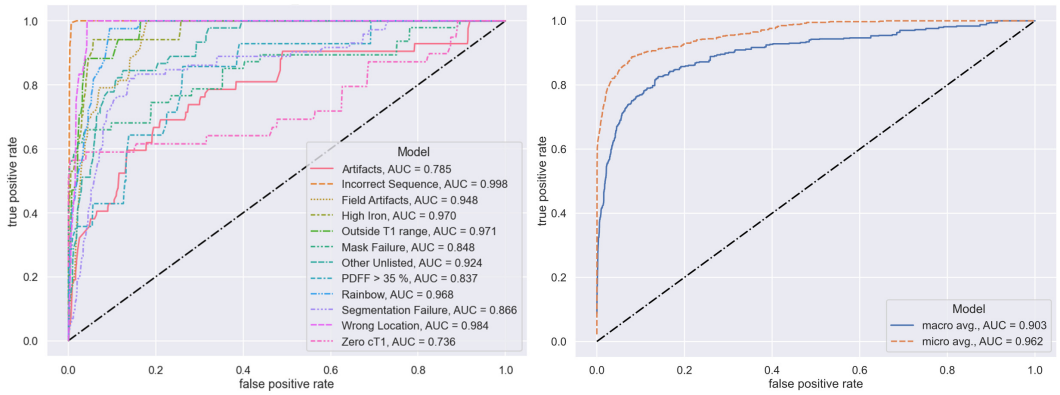



Figure 5.12: Left: The ROC curve for the diamond network across each class. Right: The macro- and micro- averaged ROC curves.

### 5.3.2 Braided Net Vs. Diamond Net

The ROC curves for the braided network and diamond network are depicted in figures 5.11 and 5.12 respectively. The braided network is outperformed in all classes by the diamond network. This is visible in both the class specific and micro and macro averaged ROC curves. The best performances for the braided network occur in the incorrect sequence, at 0.988, the rainbow artefact, at 0.872, and the segmentation fail classes, at 0.834. The worst performances for the braided

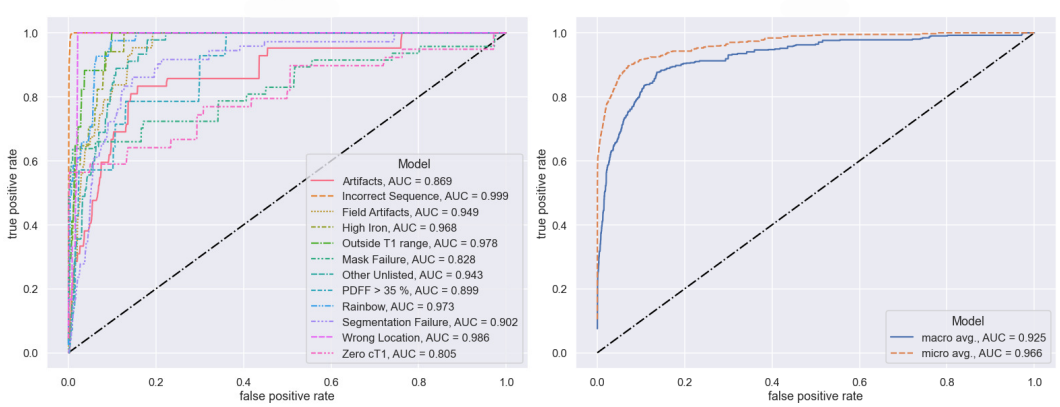



Figure 5.13: Left: The ROC curve for the VGG network across each class. Right: The macro- and micro- averaged ROC curves.

network occur in the PDFF > 35 %, at 0.427, and the high iron class, at 0.413. For the diamond network the best performances, by class, are for the Incorrect Sequence, at 0.998, the wrong location, at 0.984, the outside T1 range, at 0.971, and the high iron, at 0.970. The worst performances for the diamond network occur in the artifacts class, at 0.785, and the zero cT1 class, at 0.736. These are the only two values under 0.8.

When looking at overall model performance, macro-averaged ROC curve for the braided network and the diamond network are 0.671 and 0.903 respectively. The micro-averaged ROC curves show a reduced difference between the braided network and the diamond network with AUCs of 0.865 and 0.962 respectively.

### 5.3.3 VGG-16

The VGG-16 network outperformed the previous two networks in most tasks, as can be appreciated in the ROC curves showing the model performance in figure 5.13. The model's top performance occurs in class incorrect sequence with an

Table 5.3: Table depicting the mean and std. dev. in the AUC, for each data combination, for the MLPs

| | All Meta | Error + BBoxes | Error Only | Meta + BBoxes | Meta Only |
|---|---|---|---|---|---|
| Artifacts | $0.565 \pm 0.039$ | $0.656 \pm 0.039$ | $\mathbf{0.668 \pm 0.039}$ | $0.565 \pm 0.039$ | $0.556 \pm 0.037$ |
| Incorrect Sequence | $\mathbf{0.988 \pm 0.004}$ | $0.986 \pm 0.004$ | $0.987 \pm 0.004$ | $\mathbf{0.988 \pm 0.004}$ | $\mathbf{0.988 \pm 0.004}$ |
| Field artefacts | $0.751 \pm 0.034$ | $0.749 \pm 0.033$ | $0.746 \pm 0.038$ | $0.751 \pm 0.034$ | $\mathbf{0.752 \pm 0.033}$ |
| High Iron | $0.526 \pm 0.045$ | $0.521 \pm 0.055$ | $0.523 \pm 0.054$ | $0.526 \pm 0.045$ | $\mathbf{0.533 \pm 0.053}$ |
| Outside T1 Range | $0.738 \pm 0.070$ | $0.724 \pm 0.058$ | $0.727 \pm 0.067$ | $0.738 \pm 0.070$ | $\mathbf{0.754 \pm 0.060}$ |
| Mask Failure | $0.593 \pm 0.041$ | $\mathbf{0.855 \pm 0.032}$ | $0.848 \pm 0.031$ | $0.593 \pm 0.041$ | $0.585 \pm 0.041$ |
| PDFF > 35 % | $0.674 \pm 0.047$ | $0.702 \pm 0.054$ | $\mathbf{0.703 \pm 0.043}$ | $0.674 \pm 0.047$ | $0.667 \pm 0.057$ |
| Rainbow | $0.832 \pm 0.028$ | $\mathbf{0.866 \pm 0.025}$ | $0.862 \pm 0.023$ | $0.832 \pm 0.028$ | $0.834 \pm 0.022$ |
| Segmentation Fail | $0.822 \pm 0.028$ | $\mathbf{0.867 \pm 0.020}$ | $0.864 \pm 0.023$ | $0.822 \pm 0.028$ | $0.825 \pm 0.028$ |
| Wrong Location | $\mathbf{0.678 \pm 0.104}$ | $0.670 \pm 0.098$ | $0.523 \pm 0.104$ | $\mathbf{0.678 \pm 0.104}$ | $0.597 \pm 0.104$ |
| Zero cT1 | $0.649 \pm 0.052$ | $0.841 \pm 0.040$ | $\mathbf{0.845 \pm 0.041}$ | $0.649 \pm 0.052$ | $0.648 \pm 0.057$ |
| Other Unlisted | $0.547 \pm 0.037$ | $\mathbf{0.845 \pm 0.027}$ | $0.838 \pm 0.029$ | $0.547 \pm 0.037$ | $0.526 \pm 0.039$ |

AUC of 0.999 (Braided Network AUC: 0.988, Diamond Network AUC: 0.998), in class wrong location with an AUC of 0.986 (Braided Network AUC: 0.808, Diamond Network AUC: 0.984), and in class outside T1 range with an AUC of 0.978 (Braided Network AUC: 0.596, Diamond Network AUC: 0.971). The diamond network outperforms the VGG network class mask failure, with an AUC of 0.848 to 0.828, but is outperformed in every other class, if often marginally. The braided network is outperformed in every class.

The micro- and macro- averaged ROC curves for the VGG network have an AUC of 0.966 and 0.925 respectively. These are both better than the braided network and diamond network, with also a smaller difference between the micro and macro AUC values.

### 5.3.4 VGG-16 with Random Forest addition

The performance of the VGG network, the random forest models and the combined model for each class are shown in figure 5.14. The combined model shows similar performance to the better performing model of the two in each case, but rarely exceeds the performance of an individual network, except for in the cases of PDFF > 35 % (combined AUC: 0.919; VGG AUC: 0.899; Random Forest AUC: 0.833), other unlisted (combined AUC: 0.946; VGG AUC: 0.943; Random Forest AUC: 0.891) and wrong location (combined AUC: 0.996; VGG AUC: 0.986; Random Forest AUC: 0.986).

The micro and macro averaged ROC shows that the combined model marginally outperforms the individual models, with an AUC of 0.967 and 0.930 respectively.

## 5.4 Discussion

In this chapter, we have designed a new novel way of combining meta-data and image data in order to aid classification of image quality issues in shMOLLI T1 mapping acquisitions. We have compared this to a wide range of different available networks, with a range of data included into each network. The best performing network or combination was the use of a VGG-16 network to classify from only image data combined with a random forest which classifies based only on meta-data. The classification accuracy across every class is high enough that it has the potential to be used in-situ on scans to advise re-scans, or as a tool to aid analysis at a later date.

We have also showed that the use of image derived metrics can have a great benefit in classification accuracy in particular classes. This is seen especially with the bounding box coordinates in the wrong location class. In both the random forest and the MLP, this is the only class where the use of the error matrix only is the worst performing combination of model and data, whereas any model with inclusion of the bounding boxes performs well.

The better performance of the random forest and MLP models when using the error matrix, further validates the use of this error in informing quality control decision making in a research setting. In almost all cases inclusion of this error was amongst or was the top performer in the data combination experiments, as shown in fig. 5.7 and 5.8. Furthermore, in the case of the mask failure, other unlisted classes, and the zero cT1, exclusion of the error matrix data led to a considerate drop in model performance.

This study has limitations. Firstly, the classes are at times ill defined or automatically generated. The class "other unlisted" will likely overlap with other classes,

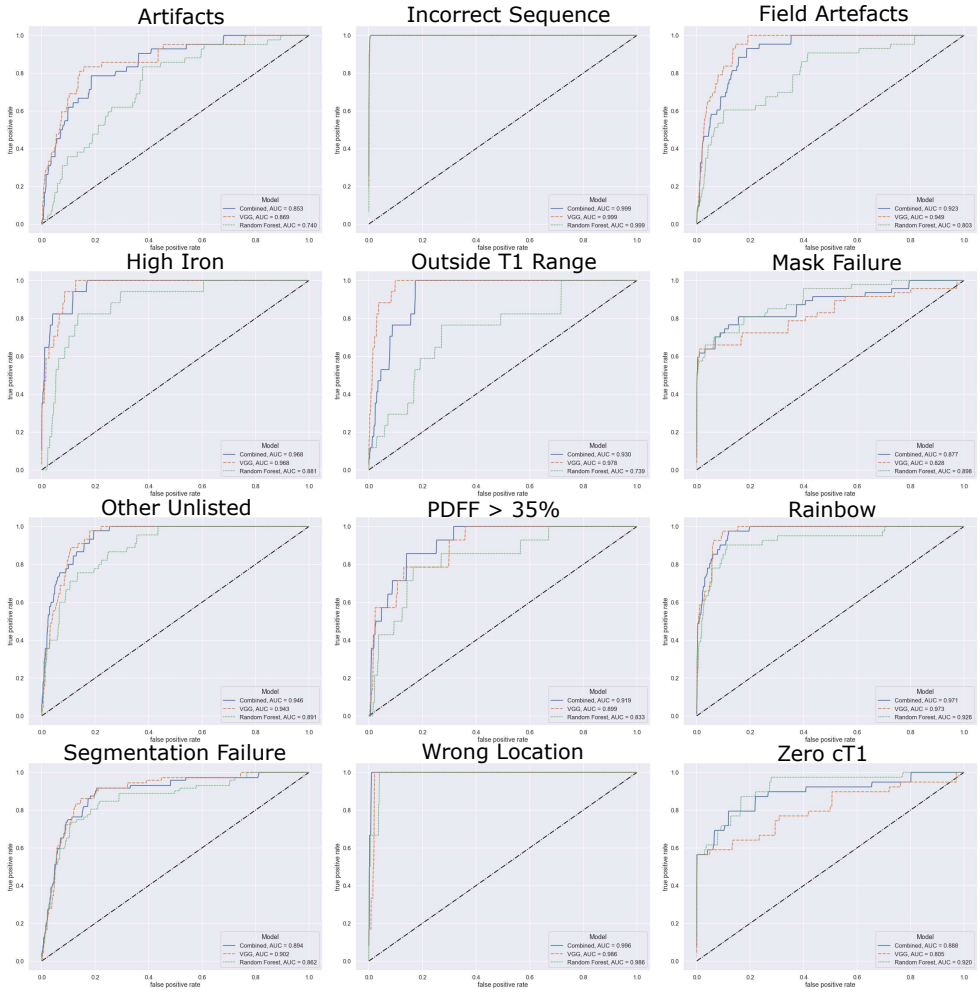



Figure 5.14: Above are shown the ROC curves for the combined VGG and Random Forest combined model across each class.

but due to the ambiguity in the images during manual QC, they were added to this class. It is also likely diverse, due to the very reason that it could be anything which didn't fit into a specific class. The "Incorrect Sequence" class is also classified almost perfectly and is the largest class, which drives the performance in the overall performance metrics. This class was caused by use of an incorrect sequence during the scanning of initial participants in UKBB, and as such classifying into this class will have little to no use in other studies. In cases where a scanning centre uses an incorrect sequence, perhaps because they have two versions of a particular protocol setup, a check of the sequence UIDs would suffice in identifying these cases. In future work, the incorrect sequence class would be removed, allowing both the researcher and the model to focus on the more complex classes. The "segmentation fail" class is auto generated, when the the liver segmentation pipeline fails at any step. This will also mean that this class is highly diverse, and the underlying reason for the segmentation fail is undefined. However, in this case, being able to predict that an acquired case would likely cause this pipeline to crash, would be useful feedback in determining whether the acquisition should be reacquired, or whether the acquisition should be sent directly for manual acquisition. Secondly, the model was again developed on UKBB data which being a large data study is inherently well controlled and includes only "healthy" subjects. Application to patients would not be applicable at this stage, however, we believe this network has merit for use in a research setting.

## 5.5 Conclusions

Inclusion of meta-data and deep learning image derived features in QC classification can benefit the performance of predicting whether an acquisition should be reacquired, but is not always a benefit for all reasons for re-acquisition, and

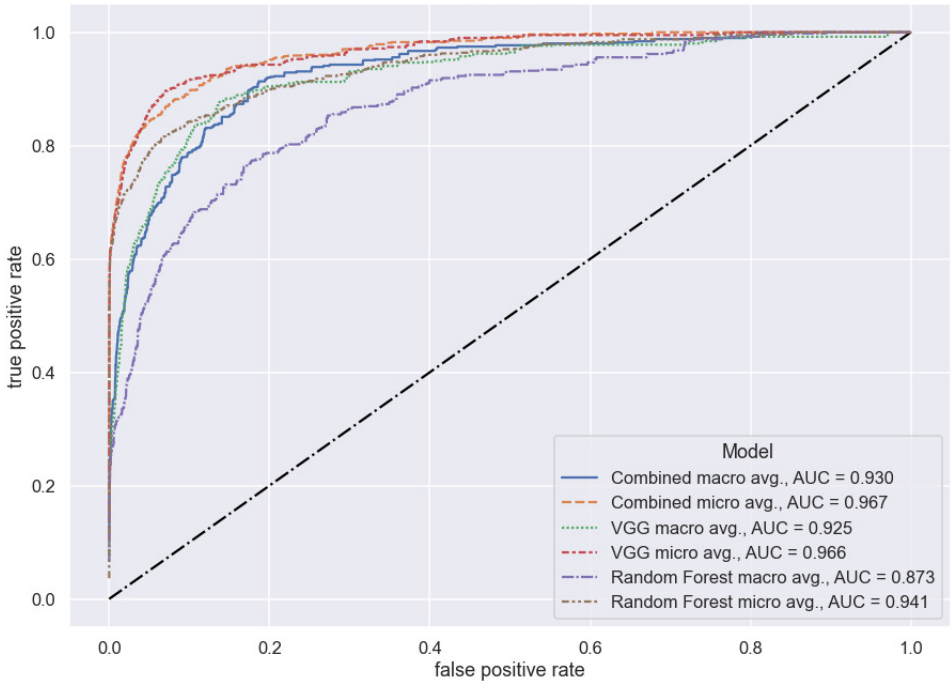



Figure 5.15: Above are shown the micro- and macro-averaged ROC curves for the VGG, random forest, and combined model.

is dependent on the presentation of those quality issues in different data formats. The overall benefit to performance can be seen to benefit different reasons for re-acquisition. Furthermore, the use of well separated, established models for the different data types elucidates the best performance, and as such we would recommend that future work does not dismiss the combination of these more traditional methods. Using this combined model, we believe the accuracy across the classes is high enough that it could be used in a research setting to improve quality output and streamline workflow.

# 6 | Pilot studies and External validation

## 6.1 Introduction

Up to this point, this thesis has focused on the development of novel techniques to detect and localise image quality issues in liver T1 mapping acquisitions. The models have been solely developed on UKBiobank (UKBB) data, and though useful due to its size and availability, the acquisitions are performed as part of a rigid framework with the radiographers repeating the same acquisitions daily. The reality in research institutions and clinics is different. Radiographers often perform a larger range of scan types, with the more novel techniques, such as multi-parametric mapping, being more infrequently used. Therefore in order to assess the viability outside the controlled environment of UKBB, we need to apply the model to datasets more representative of the wider environment that the models will be applied in. We therefore sought to assess the model in two new different settings.

Firstly, we tested our framework on a new T1 mapping MOLLI dataset, which formed part of a COVID-19 study independent of UKBB. T1 mapping has been developed over the years since the first MOLLI acquisition was acquired at UKBB. These scans were also performed in healthy or asymptomatic volunteers. The model has a limited applicability in studies involving symptomatic patients, and is partially restricted in its use. By testing on this new study, we both have access to a new successor sequence to the MOLLI acquisition that is used in UKBB, and to a new patient population which has been exposed to COVID-19 and will potentially have organ damage which will affect the T1 values acquired. It should be noted that though this sequence has no quality tags associated with it, the framework we have developed in this Thesis has the ability to replace images in a sequence

if that image was deemed low quality. Therefore by assessing the ability of the network to reconstruct high quality images accurately, we can validate whether the generated images could be used to replace low quality data.

Secondly, we tested our framework on a different quantitative technique in the liver; the iterative decomposition of water and fat with echo asymmetry and least-squares estimation (IDEAL) technique for calculating the proton density fat fraction (PDFF) of the liver. This sequence allows us to assess the use of our QC framework on the same subject cohort, using the same scanners, and with the same radiographers, and so is controlled for changes only in the sequence. This data has quality tags associated with it, and so we are able to quantify the ability of our model to identify low quality cases. This is done without training of an additional classifier or any knowledge of the quality of the dataset being introduced into the network, as previously done in chapter 4. We could therefore demonstrate that if the framework was applied to a new dataset or sequence, regardless of whether QC information is available, the method will be able to identify cases that need re-acquisition.

## 6.2 Study 1: CoverScan

### 6.2.1 Data

The CoverScan study is a prospective longitudinal study assessing the degree and prevalence of organ injury following a COVID-19 infection, using multi-parametric MRI. The study includes up to 3 visits for MRI scans with over 500 participants expected to be scanned. At the time of this QC study, 104 participants had been acquired, which included 586 MOLLI acquisitions (1.5T Siemens Aera, 384x288 pixels, 7 instances) of the liver among other organs and sequences. The

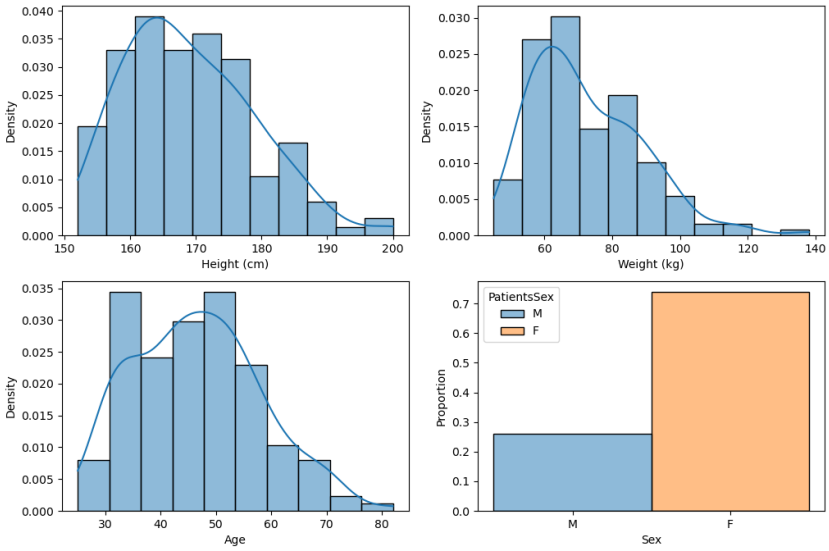



Figure 6.1: The distribution of the age, height, weight and sex in the CoverScan cases

study population statistics are shown in 6.1. In comparison to chapter 4, there is a bias towards females in the dataset, with a corresponding shift in the height and weight distributions as well. Finally, the CoverScan data contains younger subjects in comparison to UKBiobank. The data was randomly split into training, validation and testing sets of 276, 158, and 152 slices respectively ensuring that all scans of a particular participant were restricted to a single data split, i.e ensuring the multiple visits were restricted to a specific split. Access to post acquisition analysis was restricted at this stage and as such we did not have access to the T1 maps generated. Additionally, no quality control tags had been associated with this study at this point.

The CoverScan dataset is not available publicly, for access to the data please get in contact (charles.hill@perspectum.com).

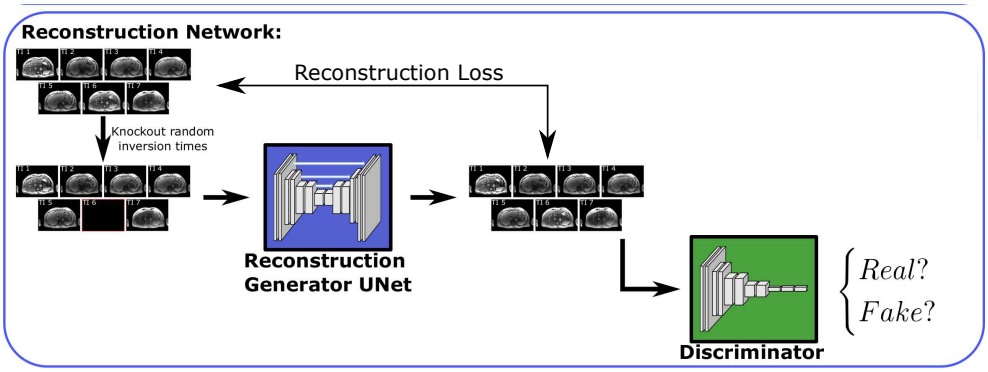



Figure 6.2: The conditional Generative Adversarial Network (cGAN) proposed to reconstruct the MOLLI datasets, as in chapter 4, minus the T1 mapping loss. A UNet based generator takes an input of a partially zero filled dataset and outputs a reconstructed dataset. A VGG-16 classifier is used as a discriminator for the adversarial loss. A smooth L1 loss is applied between the output of the generator and the original dataset.

### 6.2.2 Methods

The generator model, from chapter 4, was used to initialise the U-net, with fine-tuning taking place using the training data split, with the validation set monitored to prevent overfitting. Unlike during training for the method proposed in chapter 4, there was no minimisation for T1, as we have no access to the T1 maps, as analysis of the data had not been completed at this stage. Furthermore the original T1 mapping network was developed on an old MOLLI sequence, and so is not applicable for this dataset. Therefore the network was fine-tuned using a standard cGAN setup, with only the single discriminator, as shown in fig. 6.2.

The original model, i.e. the model only trained on UKBB, and the fine-tuned model were then both applied to the test set and the mean absolute error calculated as in section 4.2.2. Additionally, in order to prevent the background of the image having a large effect on the results, the images were clipped to a region of 100 by 100 pixels around the center of the image.

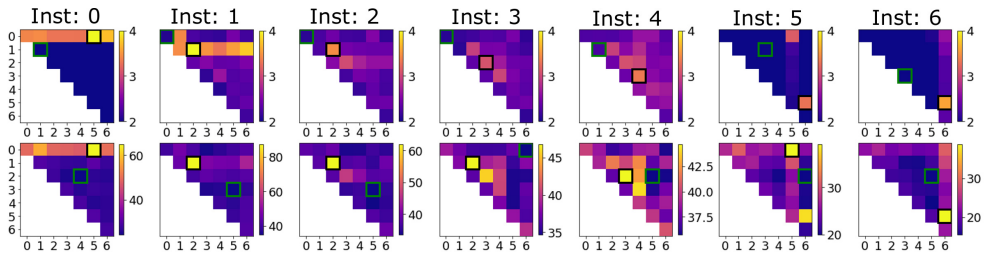



Figure 6.3: A comparison of the mean absolute error in the biobank dataset (above) and CoverScan dataset (below). The "inst: X" refers to the instance in which the error was calculated. The x and y axis refer to the instances that have been knocked out. The highest value in each matrix is highlighted by a black box, with the lowest value highlighted by a green box. In general, the highest value in each matrix occurs when that particular instance has been knocked out.

We assessed the performance of the network by direct comparison to that of UK-Biobank. This involved calculation of the mean absolute error (MAE), the structural similarity (SSIM), and normalised root mean square error (NRMSE) matrices.

The code is a small adaptation of the code used for chapter 4, and is available at: https://github.com/hillce/GAN_PDFF_OUT

### 6.2.3 Results

The MAE follows the same trend in CoverScan as previously seen in the UKBB MOLLI dataset, as depicted in fig. 6.3. The error is greatest in each instance when that image is one of the knocked out images. Additionally, the overall trend is similar with 4 out of the 7 instances having their maximum value with the same instances knocked out (instance 0 (0,5), instance 1 (1,2), instance 2 (1,2), instance 6 (5,6)), and when they don't exactly correspond with high errors in the other dataset. The highest errors are different in instance 3 (UKBB: (2,3), CoverScan: (1,2)), instance 4 (UKBB: (3,4), CoverScan: (2,3)) and instance 5 (UKBB: (5,6), CoverScan: (0,5)). The error is an order of magnitude higher

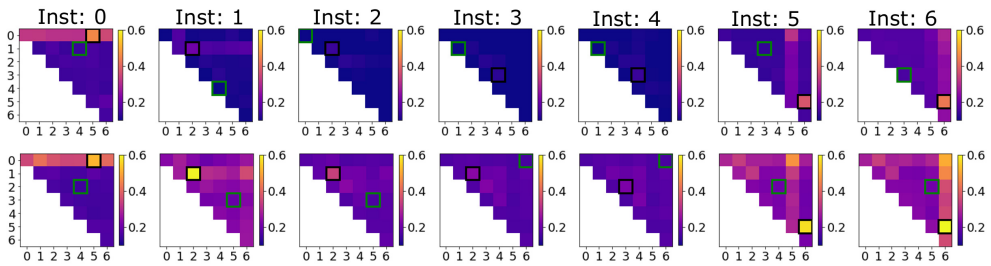



Figure 6.4: A comparison of the normalised root mean squared error in the biobank dataset (above) and CoverScan dataset (below). The "inst: X" refers to the instance in which the error was calculated. The x and y axis refer to the instances that have been knocked out. The highest value in each matrix is highlighted by a black box, with the lowest value highlighted by a green box. In general, the highest value in each matrix occurs when that particular instance has been knocked out.

across the whole dataset, but this is due to the higher signal intensity of the new sequence, mean value of $245.3 \pm 81.7$, compared to the lower signal intensity of the UKBB sequence, with a mean value of $36.6 \pm 15.0$. Due to this difference we would expect an average MAE which is 6.7 times higher. This is supported by the similar ranges of error in the NRMSE matricies.

The NRMSE error agrees with this analysis, with the errors being in the same range, though slightly higher across all instances and knockouts, as shown in fig. 6.4. This is likely due to the considerably smaller dataset upon which the CoverScan model was retrained. The SSIM likewise follows a similar trend across both datasets, with the lowest values occurring in the instance where that instance has been knocked out prior to reconstruction, e.g. the lowest value in 0th instance occurs when the 0th and 5th instance have been knocked out. Similarly to the NRMSE error, the SSIM matrices in CoverScan are slightly lower across all knockouts and instances, again likely due to the smaller training dataset.

When the three error metrics are averaged across all the channels, the similar trends are still visible between UKBB and the CoverScan datasets, with the errors

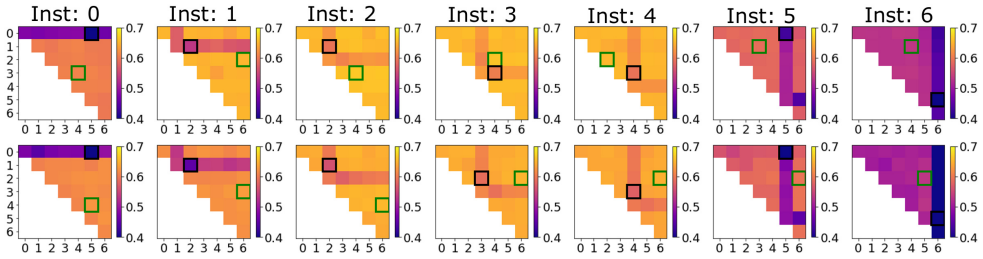



Figure 6.5: A comparison of the structural similarity metric in the UKBiobank dataset (above) and CoverScan dataset (below).

in CoverScan being worse in all cases, even by accounting for the difference in signal intensity as mentioned earlier. The largest MAE value for UKBB and CoverScan are at knockouts (0,5) and knockouts (1,2) with values of $2.876 \pm 1.524$ and $44.697 \pm 25.025$ respectively. The lowest MAE value for UKBB and CoverScan are at knockouts (3,3) and (3,5) with values of $1.957 \pm 0.954$ and $30.039 \pm 7.504$ respectively. The only other deviation between UKBB and CoverScan occurs in the largest NRMSE value where the maximum occurs at (0,5) and (5,6) with values of $0.202 \pm 0.150$ and $0.331 \pm 0.294$ respectively.

Figure 6.7 and figure 6.8 depict an examples of datasets with a low and high normalised error respectively. In the low normalised error case, fig. 6.7, the absolute error in the liver is minimal across all images in the sequence. The high error regions are localised to the spleen, stomach and on the periphery of the body. However when examining the percentage error case, the error across the body, it is almost constant, with the largest error occurring on organ boundaries, as seen around the stomach region again. In the high normalised error case, fig. 6.8, the 1st instance, with an inversion time (TI) of $\sim$ 1150 ms, has a large error in the liver region. However, again when examining the percentage error, this error is minimal and in a similar range to that of the low normalised mean aboslute error, "good quality" case.

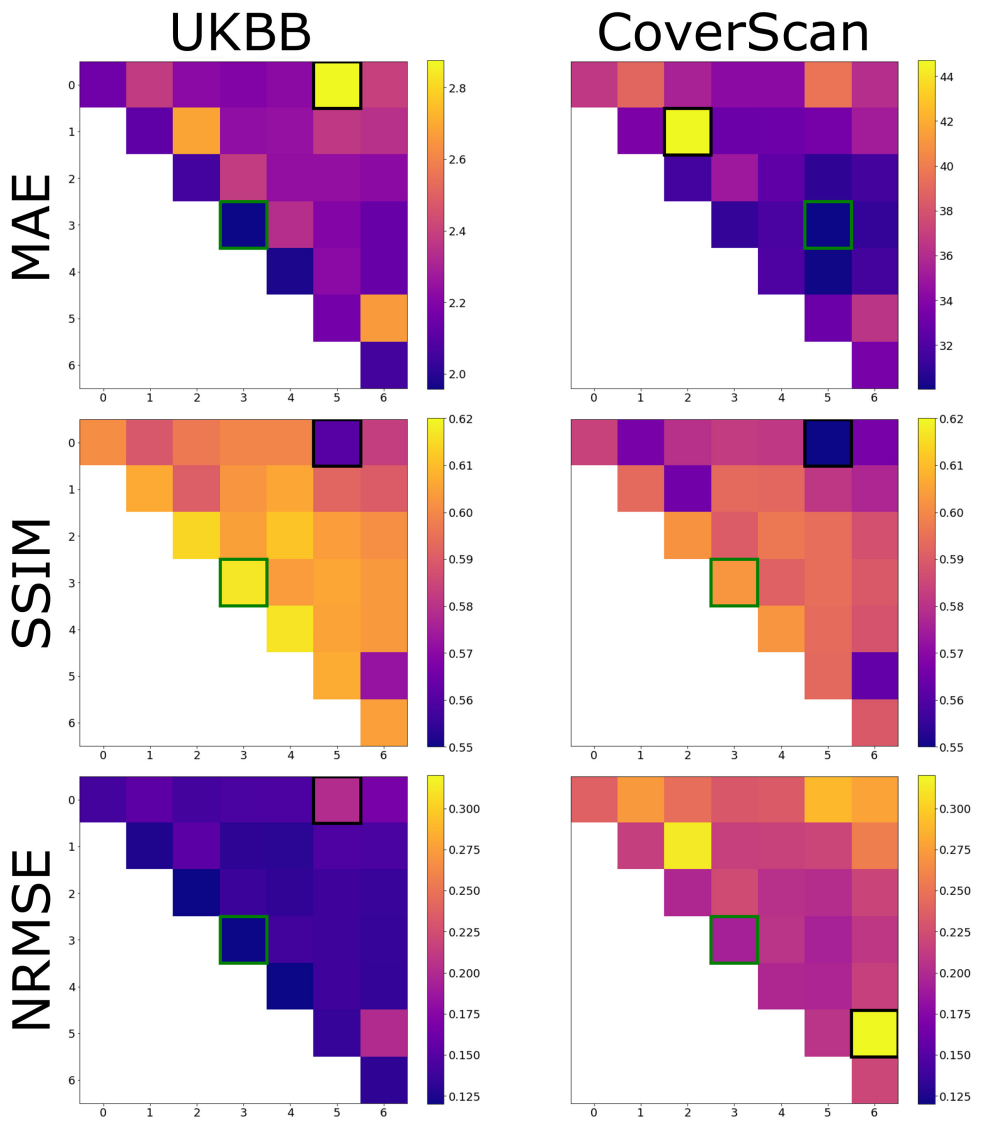



Figure 6.6: Error metric comparison between CoverScan and UKBiobank when averaged over all instances.

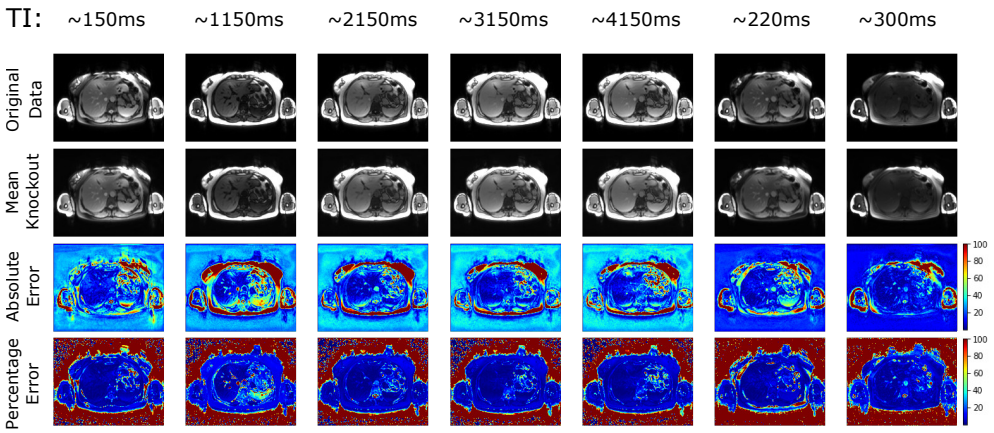



Figure 6.7: An example of an image sampled with a small normalised mean absolute error.

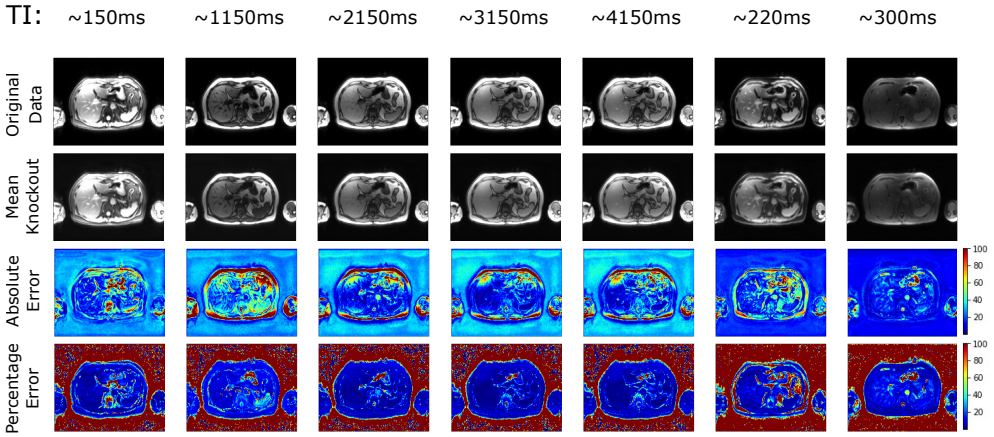



Figure 6.8: An example of an image sampled with a large normalised mean absolute error.

### 6.2.4 Discussion

The fact that the metrics follow almost identical trends between the two different datasets was encouraging and expected as the CoverScan MOLLI acquisition is a successor of the UKBB MOLLI acquisition. As such the inversion times are in a similar range, likely leading to a similar importance of individual instances on the overall reconstruction. Since the development of MOLLI and its use in the UKBB study, improvements have been made on the signal to noise ratio (SNR), and that is likely why we see a higher MAE in the CoverScan dataset. Additionally, the similarities in the trends across the metrics between the two datasets suggests that the framework works independently of disease state, as the confirmed COVID-19 status of all participants has had only a minor effect on the relative importance of particular inversion times on reconstruction quality.

The NRMSE and SSIM address the differences in each sequences SNR, and show a truer representation of the performance of the framework. The SSIM and NRMSE are worse across all points acquired, as expected due to the reduced training set size, as mentioned earlier. When differences occur between maxima and minima of the metrics between the UKBB and CoverScan datasets, this is almost always marginal with the maxima or minima in one study corresponding to a similarly high or low value in the other study.

When the metrics were averaged over all channels, again the same trends occur. However, the minimum in the MAE in CoverScan occurs in an off-diagonal component of the error matrix, as seen in fig. 6.6. This is counteractive as the more information that is knocked out from the dataset, the less information that the model has to reconstruct it. This minimum, at knockout (3,5), corresponds with the 2nd highest inversion time ( 3150 ms) and the 2nd lowest inversion time ( 220 ms). Though when the NRMSE is taken into account, the minimum is back on

the diagonal, i.e. with one knockout only. The model could potentially be over-correcting the dataset based off one of the images and under-correcting based off the other image, however further experiments would have to be performed, such as training a model on those knockouts individually.

When assessing the "high" and "low" quality cases as determined by their reconstruction error, there are only marginal differences between the datasets. Though there are hyper-intensities in "low" quality cases, there is a marginal effect on the liver, our organ of interest. This is likely due to the CoverScan study being acquired by in house Perspectum radiographers, who also aid in development of the new sequences within Perspectum. They are therefore highly qualified in the sequences performed to make assessments on the quality of the scans and perform a re-acquisition if necessary. It is thus expected that the dataset is high quality. We can not use this study to assess if the framework is able to identify quality issues. However, we can use it to assess the ability of the framework to reconstruct an accurate replacement. If an artefact or image quality degrading issue was present, this study shows that a good quality alternative would be reconstructed, in two independent research settings. In future work, the choice of metric for identifying low and high quality cases should be assessed to maximise the difference between the two cases. For example, should the normalised error be calculated only in the liver or any other organ of interest, and should the error in that region be compared to the region across the whole of the image?

One potential limitation of the study is caused by the bias in the dataset towards females. Application in the male population could potentially be poor in regions where anatomy differ such as the chest. However, we don't expect any anatomical differences in the liver and so expect reconstruction to be good in those locations. Retraining of the network with a suitably balanced dataset would rectify this issue.

## 6.3 Study 2: QC of the IDEAL Acquisition for PDFF

We have extensively tested our QC protocol in MOLLI T1 mapping of the Liver, however when it comes to multi-parametric imaging, this only forms a single part of the whole scanning protocol. Furthermore, in order to generate additional meaningful metrics, different quantitative measures are often used. Therefore, regardless of whether the T1 mapping images have been correctly acquired, there is still a non-negligible chance that another acquisition will have a detrimental outcome on the whole protocol.

Currently, our designed technique relies on eliminating artefacts that are acquired in a certain time frame of an acquisition, e.g. we have knocked out the slice corresponding to a particular inversion time in the MOLLI acquisition. In theory, this technique could be applied to any sequence with consistent anatomy, but with an additional varying component across the acquired slices, whether that be time or something else.

In this small study, we assess the feasibility of the framework being applied to the IDEAL GRE acquisition of the liver, which forms another crucial part of the LiverMultiScan multiparametric protocol. The IDEAL acquisition is used to acquire fat and water images, and thus quantify the PDFF of the liver.

### 6.3.1 Methods

The Liver IDEAL acquisition, as covered in chapter 2, consists of six images; three fat images and three water images. The whole dataset consisted of 10,000 subject (UKBB, Siemens 1.5T Magnetom) which were split into training, validation and testing sets of 8000, 1000 and 1000 respectively. Additionally each subject had an accept/reject (9579:421) tag associated with it, as in chapter 4, de-

rived from whether the PDFF map was deemed of high enough quality to use in a large population analysis. As with the MOLLI acquisitions from previous chapters, all subjects were acquired on 1.5T Siemens MAGNETOM scanners. The Dataset is available through the UKBiobank study.

The framework applied to the dataset was a subsection of that used in chapter 4, and is shown in figure 6.2. Briefly, each individual dataset has up to 2 images replaced with a zeroed image and is fed into the UNet generator. The reconstructed images are then fed into a discriminator with the corresponding original data. Unlike in chapter 4, no additional loss function was used as we did not have access to the PDFF map fitting algorithm, making the framework very similar to a standard conditional generative adversarial network (cGAN). The network was trained for 200 epochs, with a learning rate of 1e-3 and a batch size of 12, using an Adam optimiser.

The same metrics of MAE, SSIM, and NRMSE were used to assess the performance of the network. These metrics were averaged across the test dataset to generate a 6x6x6 matrix of values corresponding to the image in which the error was calculated, and the two images that were knocked out. Additionally we averaged the metrics across all the images, to give a 6x6 matrix depicting the error in the whole dataset dependent on the images knocked out. In all cases the diagonal corresponds to a single knockout, where the off-diagonal corresponds to two knockouts.

In order to assess the use of the error matrix for identifying whether the dataset should be accepted or rejected, we normalised the error, between 0 and 1, for each combination of channel and knockouts across the test set. These normalised errors were then combined to give a single prediction between 0 and 1 for each subject and a receiver operating characteristic (ROC) curve calculated for the normalised

error. Additionally we examined the ROC curve for each individual channel and knockouts combination (126 values corresponding to the non-zero values of the 6x6x6 MAE matrix). The top 10 candidates with the highest area under the ROC curve (AUROC), were combined into a single value and the ROC curve also calculated for this combined prediction.

### 6.3.2 Results

The MAE matrices are shown in fig. 6.9. As previously shown in chapter 4 and in the CoverScan dataset, the error was highest in each channel when that channel was knocked out along with any other channel. The highest and lowest MAE are depicted by a black and green box respectively. The highest MAE overall occurs in instance 0, with a value of $2.013 \pm 0.699$, when the 0th and 2nd instances in the series are knocked out. This is also the highest when averaging across all channels, with a value of $0.707 \pm 0.987$. The lowest MAE value occurs in instance 1, with a value of $0.006 \pm 0.007$, when the 5th instance is knocked out. The lowest MAE value when averaging across all channels, $0.167 \pm 0.294$, occurs when the 4th instance in the series is knocked out.

The maximum NRMSE across all channels and knockouts occurs in instance 5 when the 3rd and 5th instance are knocked out, with a value of $0.404 \pm 0.095$. The minimum NRMSE across all channels occurs in instance 1 when the 4th instance is knocked out, with a value of $0.003 \pm 0.004$. When the channels are averaged, the maximum value occurs when the 3rd and 5th instances are knocked out, and the minimum value when the 2nd instance is knocked out, with values of $0.128 \pm 0.174$ and $0.028 \pm 0.043$ respectively.

The SSIM matrices, fig. 6.11, have similar trends as the NRMSE and MAE matrices. The maximum SSIM across all channels and knockouts occurs in instance

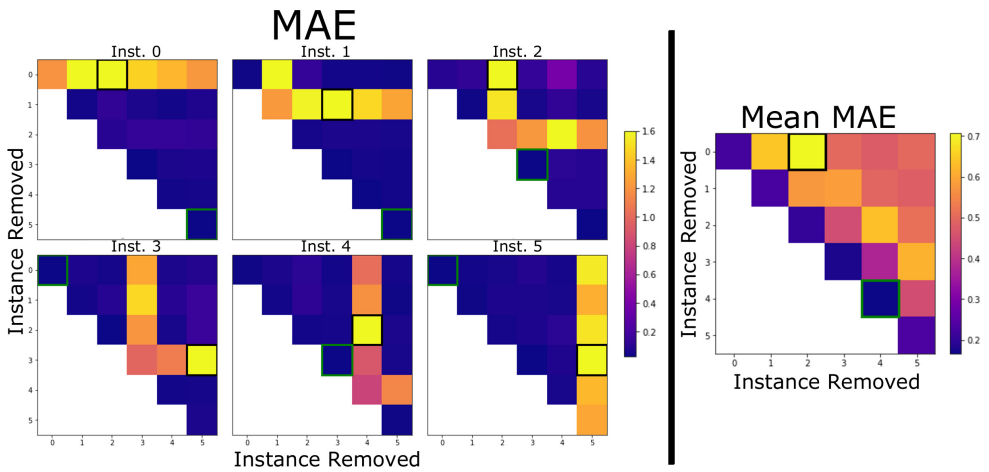



Figure 6.9: The mean absolute error across the IDEAL dataset by instance (left) and the mean absolute error averaged across all instances.

5, after knocking out the 0th channel, with a value of 0.998 $\pm$ 0.273. The minimum SSIM across all channels and knockouts occurs in instance 5, after knocking out the 3rd and 5th channels, with a value of 0.341 $\pm$ 0.033. When averaging across all the channels, the highest and lowest SSIM occur when the 1st channel is knocked out, and when the 3rd and 5th channel and knocked out, with values of 0.897 $\pm$ 0.187 and 0.757 $\pm$ 0.273 respectively. An example of a 3rd and 5th knockout is shown in figure 6.13. In the top row is highlighted the original dataset and the structural information present in the 3rd and 5th instances. In the bottom dataset, this have been translated into a a dark area with a lack of fine detail or a lighter area in the case of the breast tissue. This is shown by a large percentage error in the regions depicted. Conversely, the middle dataset depicts the reconstructed dataset after removal of the 2nd instance. The percentage error in the liver is low, in the region of 4 %.

The ROC curves show the predictive quality of the normalised error when combining all 126 errors, i.e. all permutations of up to 2 knockouts across all 6 instances, with an AUROC of 0.665. By comparison, when we selectively take the most

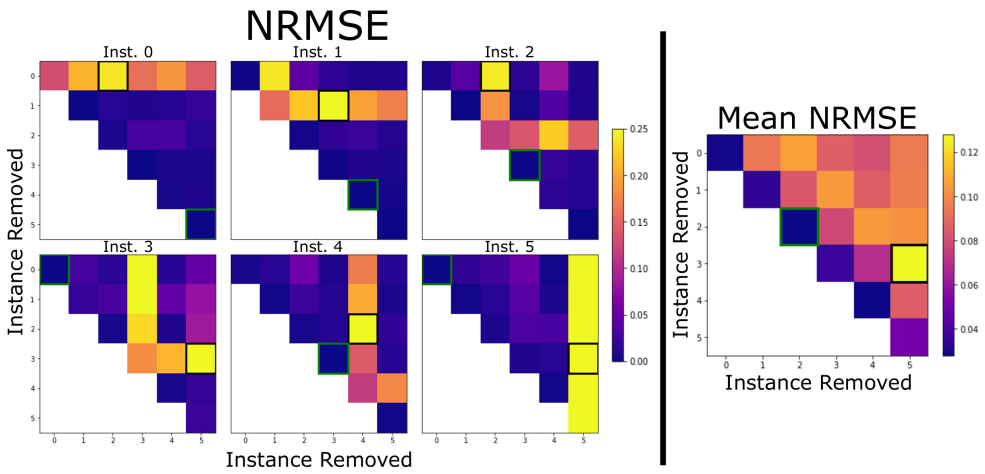



Figure 6.10: The normalised root mean squared error across the IDEAL dataset by instance (left) and the mean normalised root mean squared error across all instances.

influential values, i.e. those with the highest AUROC when individually assessed, we achieve a maximum AUROC of 0.727 in predicting whether the dataset should be excluded from the analysis.

## 6.4 Discussion

As previously seen in the MOLLI datasets, the maximum error occurs in the instance that is knocked out. Additionally the lowest errors occur when only one knockout is knocked out. The NRMSE indicate that the worst reconstruction occurs when the 3rd and 5th instance are knocked out, which corresponds with the echo times which have the most distinguishing features within them, as seen in fig. 6.13. When this information is knocked out, the model's ability to reconstruct the data is highly limited. This would suggest a strategy of focusing on these two images for manual QC as well, in order to ensure that the dataset is correctly acquired. Additionally, unlike in the T1 mapping datasets above, there is a starker contrast between the diagonal (single knockout) and off-diagonal (double knock-

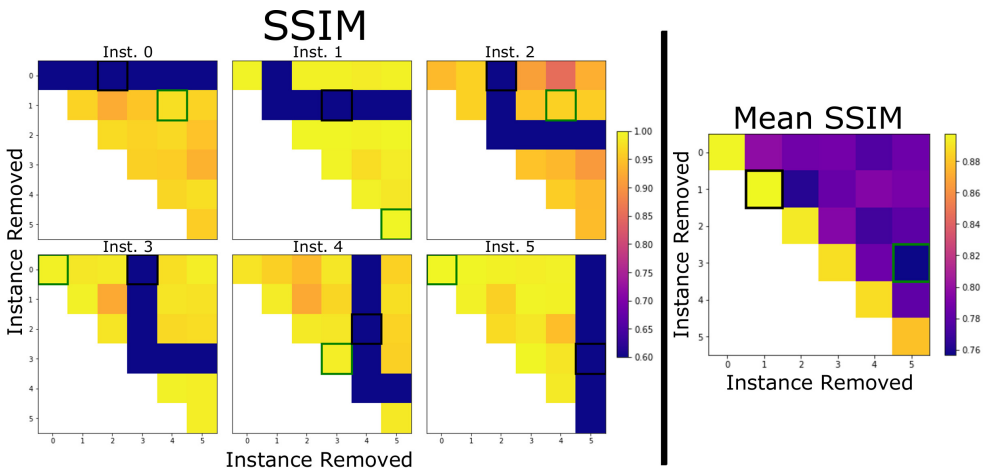



Figure 6.11: The structural similarity across the IDEAL dataset by instance (left) and the mean structural similarity across all instances.

out) component of the matrices. This could suggest that the dataset has a small level of redundancy in it, compared to the MOLLI sequences, as if we can almost perfectly reconstruct the datasets from 5 out of 6 images, then the information must be available in the remaining dataset.

The ROC curves firstly show that there is a correlation between the normalised error of reconstruction and whether the image was accepted or rejected from a large scale study. In chapter 4, the AUROC value for the normalised error was 0.867, which is significantly larger than that of the AUROC for the normalised error here (0.727). This could be down to the training dataset, which is smaller for the PDFF dataset compared to the MOLLI Dataset (8,000 versus 24,000 subjects). Secondly, the acquisition of the MOLLI protocol is a single shot process and so any time dependent artefacts will be present in one slice only. In the case of the IDEAL acquisition, portions of k-space for all constructed magnitude images. Therefore a time-dependent artefact, such as motion, will appear in all images. This would lead to a good reconstruction and therefore less differentiation between high and low quality cases.

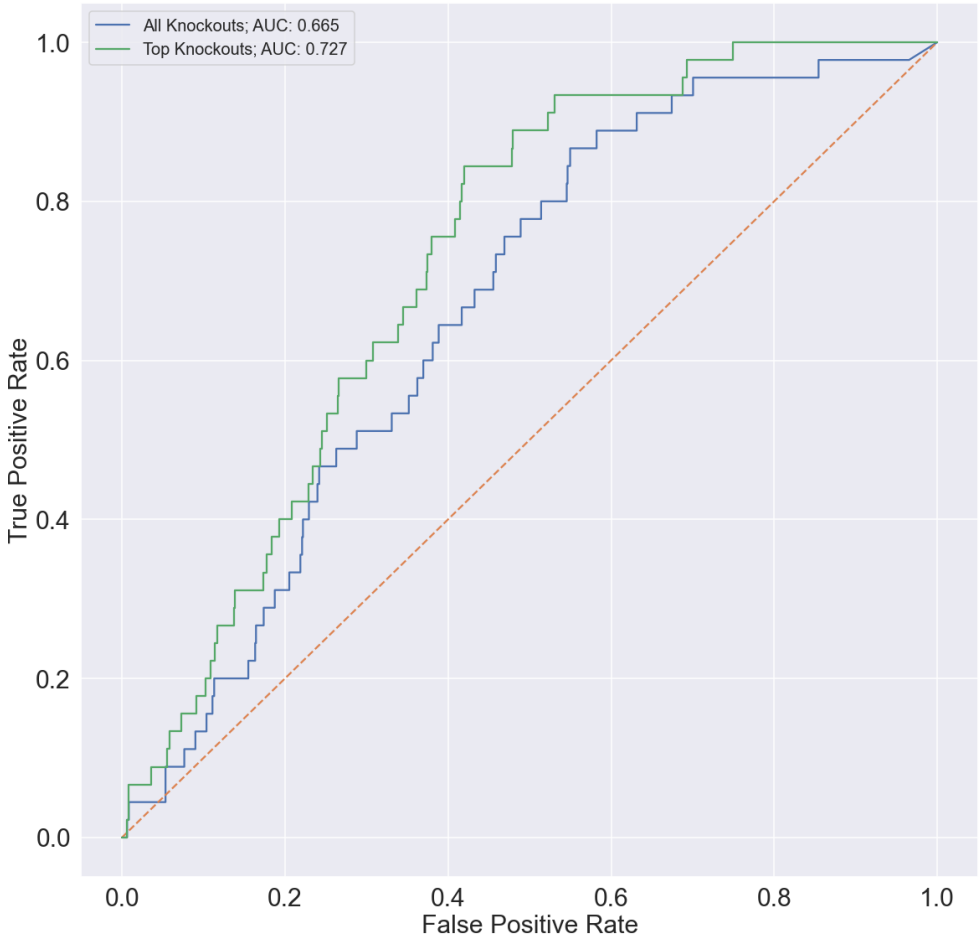



Figure 6.12: The ROC curves above show the performance of the normalised error when we use all possible knockouts (Blue) and when we select for the knockouts which individually are the most predictive of the errors (Green).

The training of the network was performed without any knowledge of the QC tags available. This shows that the technique could be applied to unseen datasets, under the assumption that the majority of the dataset are acceptable quality, and identify those that would need to be reacquired, without having to explicitly label them, which is time consuming. Furthermore, when we selectively use the error metrics which have the highest impact, we can also improve the ability to identify low quality datasets. Though in this case we needed knowledge of the tags to develop

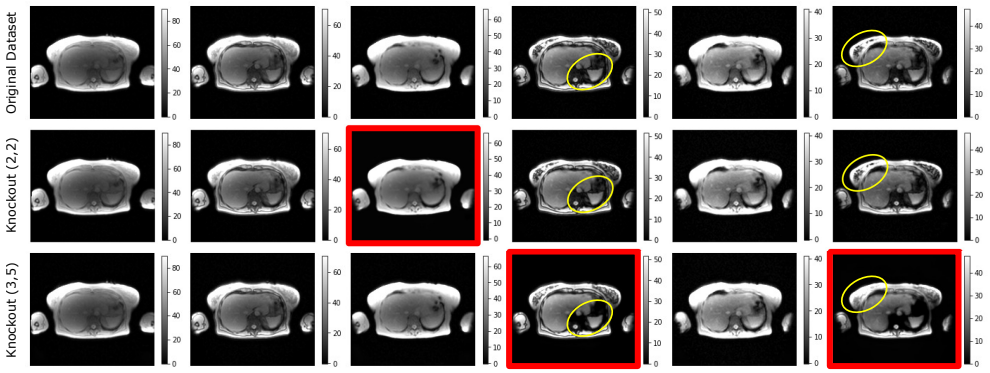



Figure 6.13: An example IDEAL dataset where the 2nd instance had been knocked out, corresponding with the lowest NRMSE, and where the 3rd and 5th instance have been knocked out, corresponding with the highest NRMSE and lowest SSIM.

this, we believe an understanding of the dataset could also be used to improve the detection rate, i.e. knowing which images in the sequence are likely to contain certain artefacts.

## 6.5 Conclusion

In this chapter, we present two pilot studies where we have applied our framework outside of the controlled environment in which it was developed. Firstly, we have applied the framework to a high quality study completely disconnected from UKBB which includes non-healthy subjects. This showed that the framework has the ability to reconstruct data from a more developed sequence and different cohort, to an accurate degree, and as such, could have use in replacing low quality data if that was to arise. Secondly, we applied the model in a quantitative technique different to T1 mapping, in the form of the IDEAL sequence for PDFF quantification. The ROC curves show that the framework has the ability to identify low quality cases with no access to any data quality information. This is an important aspect of the framework as it allows it to be used in settings by training

on data previously acquired, to identify any future low quality data. Furthermore, the model can be used to replace the low quality data and provide reasoning for doing so by presentation of an error map. We believe this process, though more convoluted than direct classification, has many more benefits and so would be more useful in the reality of large studies and clinical practices.

# 7 | Conclusions

The aim of this thesis was to develop a QC algorithm that would identify artefacts that are present in multi-parametric quantitative liver MRI scans. By being able to identify and report back to radiographers on any QC issues identified, we would be able to help limit the number of scans that require a re-acquisition. This would help save costs for healthcare and help facilitate the uptake of these new quantitative techniques to aid in the detection and prognosis of many liver diseases.

Together chapter 3 and chapter 4 present ways of identifying QC issues in liver MRI scans. Unlike a standard classification network, they break the problem down into multiple stages, so that more information beyond the outcome is available to the end user. The fRCNN in chapter 3, allows for identification of organs of interest in the body. The area of these organs of interest was shown to correlate with whether the image had a correct slice placement. Furthermore, the images could be presented back to the user, depicting the organs present and helping them reason and validate the result. This satisfies the explainable objective defined in the introduction. The method was quality label free, but required the use of manually generated bounding boxes. Though this doesn't fully satisfy the objective of being label free, it does show that explicit quality labels aren't needed to identify quality issues.

Chapter 4, was able to show that a fully label free approach is possible. The only data used to determine whether a quality issue was present was the raw dataset itself, and the T1 map output. If one was to apply this model to a new dataset, the output quantitative maps would be available, or could be generated automatically for most quantitative techniques. Therefore this approach still needs no explicit, time consuming quality labelling of the dataset. This was demonstrated in the

final ROC curve showing a good performance with no further models, when the reconstruction error was normalised. Beyond quality issue identification, with introduction of the T1 mapping network, we were able to generate reconstructions which output a T1 value which was equivalent to the case if the original acquisition was good quality. This goes beyond our initial objectives of identification, and would allow the correction of poor quality data at a later date.

Chapter 5 sought to show that the combination of meta-data, in this case the DICOM header information, could be used to improve the performance of the QC framework. We showed that the combination of the output of a random forest, the meta-data network, combined with the output of an image only classification network, the VGG-16, would outperform either of them separately. In the multi-class situation, the performance of one would help supplement the lack of performance of the other and *vice versa*. Additionally, we showed that treating the meta-data separately, as opposed to having them joint in the same network, benefits performance. Though we didn't go onto incorporate this into the QC framework in the pilot studies, it will hopefully inform future development of deep learning networks in MRI.

Chapter 6 was used to show that the framework developed in chapter 4, could be used outside the limited scope of MOLLI T1 mapping in the liver. The first study into CoverScan, was a good quality dataset, with limited ability in identifying QC issues. However, it did show that the reconstruction framework produced accurate outputs, that agreed closely with the input datasets. This implies that given a low quality dataset, which had an artefact in a localised instance, that it would be identified. The second study had a quality labels available, so we could directly assess the ability of the framework to identify low quality acquisitions in a different quantitative sequence. The ROC curve at the end showed that there is a correlation between the reconstruction error and the quality issues, as was the case

in chapter 4. Together these two studies demonstrate the utility of the framework in a range of datasets and across a range of MRI sequences.

## 7.1 Further Work

We believe that this work has set the groundwork for a full QC framework that could be performed *in-situ* and correctly identify the cases that would need to be re-acquired. However, further work needs to be undertaken in order for the process to have an impact in the clinic.

Firstly, this work has been performed and developed retrospectively. We have shown that numerous artefacts and image degrading issues can be identified, and that a heat map can be generated to inform the scanner operator on the positioning of errors. The next step would be to implement this QC protocol into a research study *in-situ*. This could involve, condensing the code into a single package and installing on a local scanner computer to asses the accuracy and speed of the algorithm. Preliminary work was performed in this area, prior to the COVID-19 pandemic, and is present in appendix A.2. Expansion of this work would assess the local applicability of the process. With suitable time points this work could be used to demonstrate whether the framework improves the quality data at a centre overtime.

Secondly, we performed exploratory work into the inclusion of multiple data types in a single joined deep learning model, with some success. More work is needed to study the relationship between the types of data (Bounding box data, DICOM header data and magnitude image data), and the outcome of image quality. We know that the inversion time is important in the construction of a T1 map. Inclusion of this information into the T1 map network in chapter 4, would allow us to study the fusion of these two data types in a more controlled environment

compared to the multi-class problem studied in 5.

Finally, chapter 6 studied the utility of the reconstruction network in the IDEAL dataset. Liver*MultiScan* relies on three different quantitative measures; the T1 value, the T2* values and the PDFF value. We have used the network in two out of the three scenarios. The next step is to apply the same framework to DIXON T2* imaging. If the reconstruction network has similar performance in this dataset, then we can explore the joining of these three QC protocols into a single Liver*MultiScan* QC protocol.

# Bibliography


Bachtiar, V., Kelly, M. D., Wilman, H. R., Jacobs, J., Newbould, R., Kelly, C. J., ... Banerjee, R. (2019). Repeatability and reproducibility of multiparametric magnetic resonance imaging of the liver. *PLoS ONE*, *14*(4), 1–12. doi: 10.1371/journal.pone.0214921

Banerjee, R., Pavlides, M., Tunnicliffe, E. M., Piechnik, S. K., Sarania, N., Philips, R., ... Neubauer, S. (2014). Multiparametric magnetic resonance for the non-invasive diagnosis of liver disease. *Journal of Hepatology*, *60*(1), 69–77. Retrieved from http://dx.doi.org/10.1016/j.jhep.2013.09.002 doi: 10.1016/j.jhep.2013.09.002

Biasiolli, L., Hann, E., Lukaschuk, E., Carapella, V., Paiva, J. M., Aung, N., ... Piechnik, S. K. (2019, 2). Automated localization and quality control of the aorta in cine CMR can significantly accelerate processing of the UK Biobank population data. *PLOS ONE*, *14*(2). doi: 10.1371/journal.pone.0212272

Böttcher, J., Hansch, A., Pfeil, A., Schmidt, P., Malich, A., Schneeweiss, A., ... Renz, D. M. (2013). Detection and classification of different liver lesions: Comparison of Gd-EOB-DTPA-enhanced MRI versus multiphasic spiral CT in a clinical single centre investigation. *European Journal of Radiology*, *82*(11), 1860–1869. Retrieved from http://dx.doi.org/10.1016/j.ejrad.2013.06.013 doi: 10.1016/j.ejrad.2013.06.013

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. Retrieved from https://doi.org/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Carnec, M., Le Callet, P., & Barba, D. (2005). Visual features for image quality assessment with reduced reference. In *Ieee international conference*

*on image processing 2005* (p. I-421). IEEE. Retrieved from http://ieeexplore.ieee.org/document/1529777/ doi: 10.1109/ICIP.2005.1529777

Cawley, J., Biener, A., Meyerhoefer, C., Ding, Y., Zvenyach, T., Gabriel Smolarz, N. B., & Ramasamy, A. (2021, 3). Direct medical costs of obesity in the United States and the most populous states. *Journal of Managed Care and Specialty Pharmacy*, *27*(3), 354–366. doi: 10.18553/JMCP.2021.20410

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., ... Gao, W. (2021). Pre-Trained Image Processing Transformer. In *2021 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 12294–12305). doi: 10.1109/CVPR46437.2021.01212

Chia, J. M., Fischer, S. E., Wickline, S. A., & Lorenz, C. H. (2000). Performance of QRS detection for cardiac magnetic resonance imaging with a novel vectorcardiographic triggering method. *Journal of Magnetic Resonance Imaging*, *12*(5), 678–688. doi: 10.1002/1522-2586(200011)12:5<678::AID-JMRI4>3.0.CO;2-5

Chow, K., Flewitt, J. A., Green, J. D., Pagano, J. J., Friedrich, M. G., & Thompson, R. B. (2014). Saturation recovery single-shot acquisition (SASHA) for myocardial T1 mapping. *Magnetic Resonance in Medicine*, *71*(6), 2082–2095. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.24878 doi: https://doi.org/10.1002/mrm.24878

Collins, C. M., Liu, W., Schreiber, W., Yang, Q. X., & Smith, M. B. (2005). Central brightening due to constructive interference with, without, and despite dielectric resonance. *Journal of Magnetic Resonance Imaging*, *21*(2), 192–196. doi: 10.1002/jmri.20245

Cortes, C., & Vapnik, V. (1995, 9). Support-vector networks. *Machine Learning*,

*20*(3), 273–297. Retrieved from http://ieeexplore.ieee.org/document/163674/http://link.springer.com/10.1007/BF00994018 doi: 10.1007/BF00994018

Czervionke, L. F., Czervionke, J. M., Daniels, D. L., & Hauhgton, V. M. (1988). Characteristic features of MR truncation artifacts. *American Journal of Roentgenology*, *151*(6), 1219–1228. doi: 10.2214/ajr.151.6.1219

Dasarathy, S., Dasarathy, J., Khiyami, A., Joseph, R., Lopez, R., & McCullough, A. J. (2009, 12). Validity of real time ultrasound in the diagnosis of hepatic steatosis: A prospective study. *Journal of Hepatology*, *51*(6), 1061–1067. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0168827809005856 doi: 10.1016/j.jhep.2009.09.001

Deshmane, A., Gulani, V., Griswold, M. A., & Seiberlich, N. (2012). Parallel MR imaging. *Journal of Magnetic Resonance Imaging*, *36*(1), 55–72. doi: 10.1002/jmri.23639

Ding, Y., Suffren, S., Bellec, P., & Lodygensky, G. A. (2019). Supervised machine learning quality control for magnetic resonance artifacts in neonatal data sets. *Human Brain Mapping*, *40*(4), 1290–1297. doi: 10.1002/hbm.24449

Dixon, W. T. (1984, 10). Simple proton spectroscopic imaging. *Radiology*, *153*(1), 189–194. Retrieved from http://pubs.rsna.org/doi/10.1148/radiology.153.1.6089263 doi: 10.1148/radiology.153.1.6089263

Dosenbach, N. U., Koller, J. M., Earl, E. A., Miranda-Dominguez, O., Klein, R. L., Van, A. N., … Fair, D. A. (2017). Real-time motion analytics during brain MRI improve data quality and reduce costs. *NeuroImage*, *161*(August), 80–93. Retrieved from https://doi.org/10.1016/j.neuroimage.2017.08.025 doi: 10.1016/j.neuroimage.2017.08

.025

Eddowes, P. J., McDonald, N., Davies, N., Semple, S. I., Kendall, T. J., Hodson, J., ... Hirschfield, G. M. (2018). Utility and cost evaluation of multiparametric magnetic resonance imaging for the assessment of non-alcoholic fatty liver disease. *Alimentary Pharmacology and Therapeutics*, *47*(5), 631–644. doi: 10.1111/apt.14469

Eddowes, P. J., Sasso, M., Allison, M., Tsochatzis, E., Anstee, Q. M., Sheridan, D., ... Newsome, P. N. (2019, 5). Accuracy of FibroScan Controlled Attenuation Parameter and Liver Stiffness Measurement in Assessing Steatosis and Fibrosis in Patients With Nonalcoholic Fatty Liver Disease. *Gastroenterology*, *156*(6), 1717–1730. doi: 10.1053/j.gastro.2019.01.042

Erasmus, L., Hurter, D., Naude, M., Kritzinger, H., & Acho, S. (2004, 6). A short overview of MRI artefacts. *South African Journal of Radiology*, *8*(2), 13. Retrieved from https://sajr.org.za/index.php/sajr/article/view/127 doi: 10.4102/sajr.v8i2.127

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PloS one*, *12*(9), e0184661. doi: 10.1371/journal.pone.0184661

Frakes, D. H., Dasi, L. P., Pekkan, K., Kitajima, H. D., Sundareswaran, K., Yoganathan, A. P., & Smith, M. J. (2008). A new method for registration-based medical image interpolation. *IEEE Transactions on Medical Imaging*, *27*(3), 370–377. doi: 10.1109/TMI.2007.907324

Gardner, E. A., Ellis, J. H., Hyde, R. J., Aisen, A. M., Quint, D. J., & Carson, P. L. (1995). Detection of degradation of magnetic resonance (MR) images: Comparison of an automated MR image-quality analysis system with trained human observers. *Academic Radiology*, *2*(4), 277–281. doi:

10.1016/S1076-6332(05)80184-9

Gerstenmaier, J. F., & Gibson, R. N. (2014). Ultrasound in chronic liver disease. *Insights into Imaging*, *5*(4), 441–455. doi: 10.1007/s13244-014-0336-2

Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX*, *7*. doi: 10.1016/j.mex.2020.100864

Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 1440–1448. doi: 10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 580–587. doi: 10.1109/CVPR.2014.81

Glorot, X., & Bengio, Y. (2010, 11). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (Vol. 9, pp. 249–256). Chia Laguna Resort, Sardinia, Italy: PMLR. Retrieved from https://proceedings.mlr.press/v9/glorot10a.html

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (n.d.). *Generative Adversarial Nets* (Tech. Rep.). Retrieved from http://www.github.com/goodfeli/adversarial

Hagström, H., Nasr, P., Ekstedt, M., Hammar, U., Stål, P., Hultcrantz, R., & Kechagias, S. (2017, 12). Fibrosis stage but not NASH predicts mortality and time to development of severe liver disease in biopsy-proven NAFLD. *Journal of Hepatology*, *67*(6). doi: 10.1016/j.jhep.2017.07.027

Hallsworth, K., & Adams, L. A. (2019, 12). Lifestyle modification in NAFLD/NASH: Facts and figures. *JHEP Reports*, *1*(6). doi: 10.1016/j.jhepr.2019.10.008

Hayes, C. E., & Axel, L. (1985, 9). Noise performance of surface coils for magnetic resonance imaging at 1.5 T. *Medical Physics*, *12*(5), 604–607. Retrieved from http://doi.wiley.com/10.1118/1.595682 doi: 10.1118/1.595682

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. doi: 10.1109/CVPR.2016.90

He, L., Li, H., Dudley, J. A., Maloney, T. C., Brady, S. L., Somasundaram, E., ... Dillman, J. R. (2019). Machine learning prediction of liver stiffness using clinical and T2-Weighted MRI radiomic data. *American Journal of Roentgenology*, *213*(3), 592–601. doi: 10.2214/AJR.19.21082

Heiland, S. (2008). From A as in aliasing to Z as in zipper: Artifacts in MRI. *Clinical Neuroradiology*, *18*(1), 25–36. doi: 10.1007/s00062-008-8003-y

Henninger, B., Zoller, H., Kannengiesser, S., Zhong, X., Jaschke, W., & Kremser, C. (2017). 3D Multiecho Dixon for the Evaluation of Hepatic Iron and Fat in a Clinical Setting. *Journal of Magnetic Resonance Imaging*, *46*(3), 793–800. doi: 10.1002/jmri.25630

Hood, M. N., Ho, V. B., Smirniotopoulos, J. G., & Szumowski, J. (1999). Chemical Shift: The Artifact and Clinical Tool Revisited 1 Recipient of a Certificate of Merit award for a scientific exhibit at the 1997 RSNA scien- tific assembly. *RadioGraphics*, *19*, 357–371.

Horváth, A., Pezold, S., Weigel, M., Parmar, K., & Cattin, P. (2017). High Order Slice Interpolation for Medical Images. In S. A. Tsaftaris, A. Gooya,

A. F. Frangi, & J. L. Prince (Eds.), *Simulation and synthesis in medical imaging* (pp. 69–78). Cham: Springer International Publishing.

Islam, R. (2020). Reconstruction of Compressed Sensing MRI: An Experimental Review among Priors- And Deep Learning- based Algorithms. *2020 2nd International Conference on Sustainable Technologies for Industry 4.0, STI 2020*, *0*, 19–20. doi: 10.1109/STI50764.2020.9350519

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017, 7). Image-To-Image Translation With Conditional Adversarial Networks. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr).*

Ito, K., Murotani, K., Nakade, Y., Inoue, T., Nakao, H., Sumida, Y., ... Yoneda, M. (2017). Serum Wisteria floribunda agglutinin-positive Mac-2-binding protein levels and liver fibrosis: A meta-analysis. *Journal of Gastroenterology and Hepatology (Australia)*, *32*(12), 1922–1930. doi: 10.1111/jgh.13802

Ivashchenko, O. V., Rijkhorst, E. J., ter Beek, L. C., Hoetjes, N. J., Pouw, B., Nijkamp, J., ... Ruers, T. J. (2020). A workflow for automated segmentation of the liver surface, hepatic vasculature and biliary tree anatomy from multiphase MR images. *Magnetic Resonance Imaging*, *68*(May 2019), 53–65. Retrieved from https://doi.org/10.1016/j.mri.2019.12.008 doi: 10.1016/j.mri.2019.12.008

Jang, J., Bang, K., Jang, H., & Hwang, D. (2018). Quality evaluation of no-reference MR images using multidirectional filters and image statistics. *Magnetic Resonance in Medicine*, *80*(3), 914–924. doi: 10.1002/mrm.27084

Janowski, K., Shumbayawonda, E., Dennis, A., Kelly, M., Bachtiar, V., DeBrota, D., ... Socha, P. (2020, 9). Multiparametric MRI as a Non-Invasive Monitoring Tool for Children with Autoimmune Hepatitis. *Jour-*

*nal of Pediatric Gastroenterology & Nutrition*, *Publish Ah*(9), 1689–1699. Retrieved from https://journals.lww.com/10.1097/MPG.0000000000002930 doi: 10.1097/MPG.0000000000002930

Jezzard, P., & Balaban, R. S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. *Magnetic Resonance in Medicine*, *34*(1), 65–73. doi: 10.1002/mrm.1910340111

Jiang, J., Hu, Y. C., Tyagi, N., Rimner, A., Lee, N., Deasy, J. O., ... Veeraraghavan, H. (2020). PSIGAN: Joint Probabilistic Segmentation and Image Distribution Matching for Unpaired Cross-Modality Adaptation-Based MRI Segmentation. *IEEE Transactions on Medical Imaging*, *39*(12), 4071–4084. doi: 10.1109/TMI.2020.3011626

Jing, L., & Tian, Y. (2019, 2). Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8828*(c), 1–1. Retrieved from https://ieeexplore.ieee.org/document/9086055/http://arxiv.org/abs/1902.06162 doi: 10.1109/TPAMI.2020.2992393

Kam Ho, T. (1995). Random decision forests perceptron training. *ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition*, 278–282.

Keshavan, A., Yeatman, J. D., & Rokem, A. (2019). Combining Citizen Science and Deep Learning to Amplify Expertise in Neuroimaging. *Frontiers in Neuroinformatics*, *13*(May), 1–13. doi: 10.3389/fninf.2019.00029

Kim, S., Shin, J., Kim, D. Y., Choi, G. H., Kim, M. J., & Choi, J. Y. (2019). Radiomics on gadoxetic acid-enhanced magnetic resonance imaging for prediction of postoperative early and late recurrence of single hepatocellular carcinoma. *Clinical Cancer Research*, *25*(13), 3847–3855. doi:

10.1158/1078-0432.CCR-18-2861

Kim, Y.-C., Kim, K. R., Lee, H., & Choe, Y. H. (2021). Fast calculation software for modified Look-Locker inversion recovery (MOLLI) T1 mapping. *BMC Medical Imaging*, *21*(1), 26. Retrieved from https://doi.org/10.1186/s12880-021-00558-8 doi: 10.1186/s12880-021-00558-8

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. , 1–15. Retrieved from http://arxiv.org/abs/1412.6980

Kleiner, D. E., Brunt, E. M., Van Natta, M., Behling, C., Contos, M. J., Cummings, O. W., ... Sanyal, A. J. (2005, 6). Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology*, *41*(6). doi: 10.1002/hep.20701

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017, 5). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. doi: 10.1145/3065386

Krizhevsky, B. A., Sutskever, I., & Hinton, G. E. (2012). Cnnå®éè®ç»ç. *Communications of the ACM*, *60*(6), 84–90.

Küstner, T., Gatidis, S., Liebgott, A., Schwartz, M., Mauch, L., Martirosian, P., ... Schick, F. (2018, 11). A machine-learning framework for automatic reference-free quality assessment in MRI. *Magnetic Resonance Imaging*, *53*(November 2017), 134–147. Retrieved from https://doi.org/10.1016/j.mri.2018.07.003 doi: 10.1016/j.mri.2018.07.003

Küstner, T., Liebgott, A., Mauch, L., Martirosian, P., Bamberg, F., Nikolaou, K., ... Gatidis, S. (2018). Automated reference-free detection of motion artifacts in magnetic resonance images. *Magnetic Resonance Materials in Physics, Biology and Medicine*, *31*(2), 243–256. doi: 10.1007/s10334-017-0650-z

Lauterbur, P. C. (1973, 3). Image Formation by Induced Local Interactions:

Examples Employing Nuclear Magnetic Resonance. *Nature*, *242*(5394), 190–191. Retrieved from https://www.nature.com/articles/242190a0http://www.nature.com/articles/242190a0 doi: 10.1038/242190a0

Le, M. H., Devaki, P., Ha, N. B., Jun, D. W., Te, H. S., Cheung, R. C., & Nguyen, M. H. (2017). Prevalence of non-Alcoholic fatty liver disease and risk factors for advanced fibrosis and mortality in the United States. *PLoS ONE*, *12*(3), 1–13. doi: 10.1371/journal.pone.0173499

Lebrun, G., Charrier, C., Le, O., Lézoray, O., Lebrun, G., Charrier, C., ... Lebrun, G. (2012). Machine learning to design full-reference image quality assessment algorithm. *Signal Processing: Image Communication*, *27*(3), 209–219. doi: 10.1016/j.image.2012.01.002

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi: 10.1038/nature14539

Lee, J. E., Nam, C. M., Lee, S. G., Park, S., Kim, T. H., & Park, E. C. (2019, 9). The economic burden of cancer attributable to obesity in Korea: A population-based cohort study. *European Journal of Cancer Care*, *28*(5). doi: 10.1111/ecc.13084

Leng, J., Xu, G., & Zhang, Y. (2013). Medical image interpolation based on multi-resolution registration. *Computers and Mathematics with Applications*, *66*(1), 1–18. Retrieved from http://dx.doi.org/10.1016/j.camwa.2013.04.026 doi: 10.1016/j.camwa.2013.04.026

Li, Q., & Wang, Z. (2009, 4). Reduced-Reference Image Quality Assessment Using Divisive Normalization-Based Image Representation. *IEEE Journal of Selected Topics in Signal Processing*, *3*(2), 202–211. Retrieved from http://ieeexplore.ieee.org/document/4799311/ doi: 10.1109/JSTSP.2009.2014497

Lin, W., Hasenstab, K., Moura Cunha, G., & Schwartzman, A. (2020, 12). Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment. *Scientific Reports*, *10*(1). doi: 10.1038/s41598-020-77264-y

Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., ... Allen, N. E. (2020, 12). The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature Communications*, *11*(1), 2624. Retrieved from http://dx.doi.org/10.1038/s41467-020-15948-9http://www.nature.com/articles/s41467-020-15948-9 doi: 10.1038/s41467-020-15948-9

Liu, F., Feng, L., & Kijowski, R. (2019). MANTIS: Model-Augmented Neural neTwork with Incoherent k-space Sampling for efficient MR parameter mapping. *Magnetic Resonance in Medicine*, *82*(1), 174–188. doi: 10.1002/mrm.27707

Liu, M., Vanguri, R., Mutasa, S., Ha, R., Liu, Y. C., Button, T., & Jambawalikar, S. (2020). Channel width optimized neural networks for liver and vessel segmentation in liver iron quantification. *Computers in Biology and Medicine*, *122*(April), 103798. Retrieved from https://doi.org/10.1016/j.compbiomed.2020.103798 doi: 10.1016/j.compbiomed.2020.103798

Liu, S., Bustin, A., Ferry, P., Codreanu, A., Burschka, D., Menini, A., & Odille, F. (2018). A vectorized Levenberg-Marquardt model fitting algorithm for efficient post-processing of cardiac T1 mapping MRI. *Computers in Biology and Medicine*, *96*, 106–115. Retrieved from https://www.sciencedirect.com/science/article/pii/S001048251830060X doi: https://doi.org/10.1016/j.compbiomed

.2018.03.009

Liu, X., Khalvati, F., Namdar, K., Fischer, S., Lewis, S., Taouli, B., ... Jhaveri, K. S. (2021). Can machine learning radiomics provide pre-operative differentiation of combined hepatocellular cholangiocarcinoma from hepatocellular carcinoma and cholangiocarcinoma to inform optimal treatment planning? *European Radiology*, *31*(1), 244–255. doi: 10.1007/s00330-020-07119-7

Look, D. C., & Locker, D. R. (1970). Time saving in measurement of NMR and EPR relaxation times. *Review of Scientific Instruments*, *41*(2), 250–251. doi: 10.1063/1.1684482

Loomba, R., Wolfson, T., Ang, B., Hooker, J., Behling, C., Peterson, M., ... Sirlin, C. (2014). Magnetic Resonance Elastography Predicts Advanced Fibrosis in Patients With Nonalcoholic Fatty Liver Disease: A Prospective Study.
doi: 10.1002/hep.27362/suppinfo

Ma, J. (2008). Dixon techniques for water and fat imaging. *Journal of Magnetic Resonance Imaging*, *28*(3), 543–558. doi: 10.1002/jmri.21492

Ma, L., Wang, X., Liu, Q., & Ngan, K. N. (2016). Reorganized DCT-based image representation for reduced reference stereoscopic image quality assessment. *Neurocomputing*, *215*, 21–31. Retrieved from http://dx.doi.org/10.1016/j.neucom.2015.06.116 doi: 10.1016/j.neucom.2015.06.116

Magnotta, V. A., & Friedman, L. (2006, 6). Measurement of Signal-to-Noise and Contrast-to-Noise in the fBIRN Multicenter Imaging Study. *Journal of Digital Imaging*, *19*(2), 140–147. Retrieved from http://link.springer.com/10.1007/s10278-006-0264-x doi: 10.1007/s10278-006-0264-x

Mansfield, P. (1977, 2). Multi-planar image formation using NMR spin echoes. *Journal of Physics C: Solid State Physics*, *10*(3), L55-L58. Retrieved from https://iopscience.iop.org/article/10.1088/0022-3719/10/3/004 doi: 10.1088/0022-3719/10/3/004

Mariappan, Y. K., Glaser, K. J., & Ehman, R. L. (2010). Magnetic resonance elastography: A review. *Clinical Anatomy*, *23*(5), 497–511. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/ca.21006 doi: https://doi.org/10.1002/ca.21006

Masood, S., Rosa, M., Kraemer, D. F., Smotherman, C., & Mohammadi, A. (2015). Comparative Cost-Effectiveness of Fine Needle Aspiration Biopsy Versus image-Guided Biopsy , and Open Surgical Biopsy in the Evaluation of Breast Cancer in the Era of Affordable Care Act : A Changing Landscape. , *43*(8), 605–612. doi: 10.1002/dc

Matteoni, C. A., Younossi, Z. M., Gramlich, T., Boparai, N., Yao Chang Liu, & McCullough, A. J. (1999). Nonalcoholic fatty liver disease: A spectrum of clinical and pathological severity. *Gastroenterology*, *116*(6), 1413–1419. doi: 10.1016/S0016-5085(99)70506-8

McDonald, N., Eddowes, P. J., Hodson, J., Semple, S. I. K., Davies, N. P., Kelly, C. J., ... Fallowfield, J. A. (2018). Multiparametric magnetic resonance imaging for quantitation of liver disease: a two-centre cross-sectional observational study. *Scientific Reports*, *8*(1), 9189. Retrieved from http://www.nature.com/articles/s41598-018-27560-5 doi: 10.1038/s41598-018-27560-5

Messroghli, D. R., Radjenovic, A., Kozerke, S., Higgins, D. M., Sivananthan, M. U., & Ridgway, J. P. (2004). Modified look-locker inversion recovery (MOLLI) for high-resolution T 1 mapping of the heart. *Magnetic Resonance in Medicine*, *52*(1), 141–146. doi: 10.1002/mrm.20110

Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets. , 1–7. Retrieved from http://arxiv.org/abs/1411.1784

Mitchell, T. (1997). *Machine Learning* (1st ed.). McGraw Hill.

Mojtahed, A., Kelly, C. J., Herlihy, A. H., Kin, S., Wilman, H. R., McKay, A., ... Harisinghani, M. (2019). Reference range of liver corrected T1 values in a population at low risk for fatty liver diseaseâa UK Biobank sub-study, with an appendix of interesting cases. *Abdominal Radiology*, *44*(1), 72–84. Retrieved from https://doi.org/10.1007/s00261-018-1701-2 doi: 10.1007/s00261-018-1701-2

Mortamet, B., Bernstein, M. A., Jack, C. R., Gunter, J. L., Ward, C., Britson, P. J., ... Krueger, G. (2009, 8). Automatic quality assessment in structural brain magnetic resonance imaging. *Magnetic Resonance in Medicine*, *62*(2), 365–372. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2780021/pdf/nihms119705.pdfhttp://doi.wiley.com/10.1002/mrm.21992 doi: 10.1002/mrm.21992

Mostapha, M., Mailhe, B., Chen, X., Ceccaldi, P., Yoo, Y., & Nadar, M. (2020, 4). Braided Networks for Scan-Aware MRI Brain Tissue Segmentation. In *2020 ieee 17th international symposium on biomedical imaging (isbi).* IEEE. doi: 10.1109/ISBI45749.2020.9098601

Moyano-Cuevas, J. L., Plaza, A., Dopido, I., Pagador, J. B., Sanchez-Margallo, J. A., Sánchez, L. F., & Sánchez-Margallo, F. M. (2014). 3D Segmentation of MRI of the Liver Using Support Vector Machine. In L. M. Roa Romero (Ed.), *Xiii mediterranean conference on medical and biological engineering and computing 2013* (pp. 368–371). Cham: Springer International Publishing.

Myers, R. P., Pomier-Layrargues, G., Kirsch, R., Pollett, A., Duarte-Rojo, A., Wong, D., ... Elkashab, M. (2012). Feasibility and diagnostic performance

of the FibroScan XL probe for liver stiffness measurement in overweight and obese patients. *Hepatology*, *55*(1), 199–208. doi: 10.1002/hep.24624

Pavlides, M., Banerjee, R., Tunnicliffe, E. M., Kelly, C., Collier, J., Wang, L. M., ... Barnes, E. (2017, 7). Multiparametric magnetic resonance imaging for the assessment of non-alcoholic fatty liver disease severity. *Liver international : official journal of the International Association for the Study of the Liver*, *37*(7), 1065–1073. Retrieved from http://doi.wiley.com/10.1111/liv.13284http://www.ncbi.nlm.nih.gov/pubmed/27778429http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5518289 doi: 10.1111/liv.13284

Payan, A., & Montana, G. (2015). Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. , 1–9. Retrieved from http://arxiv.org/abs/1502.02506

Piechnik, S. K., Ferreira, V. M., Dall'Armellina, E., Cochlin, L. E., Greiser, A., Neubauer, S., & Robson, M. D. (2010). Shortened Modified Look-Locker Inversion recovery (ShMOLLI) for clinical myocardial T1-mapping at 1.5 and 3 T within a 9 heartbeat breathhold. *Journal of Cardiovascular Magnetic Resonance*, *12*(1), 69. Retrieved from http://www.jcmr-online.com/content/12/1/69 doi: 10.1186/1532-429X-12-69

Pizarro, R. A., Cheng, X., Barnett, A., Lemaitre, H., Verchinski, B. A., Goldman, A. L., ... Mattay, V. S. (2016). Automated Quality Assessment of Structural Magnetic Resonance Brain Images Based on a Supervised Machine Learning Algorithm. *Frontiers in Neuroinformatics*, *10*(December). doi: 10.3389/fninf.2016.00052

Plathow, C., Ley, S., Zaporozhan, J., Schöbinger, M., Gruenig, E., Puderbach, M., ... Kauczor, H. U. (2006). Assessment of reproducibility and stability

of different breath-hold maneuvres by dynamic MRI: Comparison between healthy adults and patients with pulmonary hypertension. *European Radiology*, *16*(1), 173–179. doi: 10.1007/s00330-005-2795-9

Port, J. D., & Pomper, M. G. (2000). Quantification and minimization of magnetic susceptibility artifacts on GRE images. *Journal of Computer Assisted Tomography*, *24*(6), 958–964. doi: 10.1097/00004728-200011000-00024

Public Health England. (2017a). *Health matters: obesity and the food environment.* Retrieved from https://www.gov.uk/government/publications/health-matters-obesity-and-the-food-environment/health-matters-obesity-and-the-food-environment--2

Public Health England. (2017b, 3). *Health matters: obesity and the food environment.* Retrieved from https://www.gov.uk/government/publications/health-matters-obesity-and-the-food-environment/health-matters-obesity-and-the-food-environment--2

Qasim, A. B., Ezhov, I., Shit, S., Schoppe, O., Paetzold, J. C., Sekuboyina, A., ... Menze, B. (2020, 11). Red-GAN: Attacking class imbalance via conditioned generation. Yet another medical imaging perspective. In T. Arbel, I. Ben Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, & C. Pal (Eds.), *Proceedings of the third conference on medical imaging with deep learning* (Vol. 121, pp. 655–668). PMLR. Retrieved from https://proceedings.mlr.press/v121/qasim20a.html

Radtke, J. P., Schwab, C., Wolf, M. B., Freitag, M. T., Alt, C. D., Kesch, C., ... Hadaschik, B. A. (2016). Multiparametric Magnetic Resonance Imaging (MRI) and MRIâTransrectal Ultrasound Fusion Biopsy for Index Tumor Detection: Correlation with Radical Prostatectomy Specimen. *Euro-*

*pean Urology*, *70*(5), 846–853. Retrieved from http://dx.doi.org/10.1016/j.eururo.2015.12.052 doi: 10.1016/j.eururo.2015.12.052

Reed, S., Akata, Z., Yan, X., Logeswaran REEDSCOT, L., Schiele, B., & Lee SCHIELE, H. (2016). *Generative Adversarial Text to Image Synthesis* (Tech. Rep.).

Reeder, S. B., Pineda, A. R., Wen, Z., Shimakawa, A., Yu, H., Brittain, J. H., ... Pelc, N. J. T. J. (2005). Iterative decomposition of water and fat with echo asymmetry and least-squares estimation (IDEAL): Application with fast spin-echo imaging. *Magnetic Resonance in Medicine*, *54*(3), 636–644. doi: 10.1002/mrm.20624

Reeder, S. B., Wen, Z., Yu, H., Pineda, A. R., Gold, G. E., Markl, M., & Pelc, N. J. (2004). Multicoil Dixon Chemical Species Separation with an Iterative Least-Squares Estimation Method. *Magnetic Resonance in Medicine*, *51*(1), 35–45. doi: 10.1002/mrm.10675

Ren, S., He, K., Girshick, R., & Sun, J. (2017, 6). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Rosen, A. F., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., ... Satterthwaite, T. D. (2018). Quantitative assessment of structural image quality. *NeuroImage*, *169*(October 2017), 407–418. Retrieved from https://doi.org/10.1016/j.neuroimage.2017.12.059 doi: 10.1016/j.neuroimage.2017.12.059

Rosenblatt, F. (1961). *Principles of Neurodynamics*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986, 10). Learning representations by back-propagating errors. *Nature*, *323*(6088). doi: 10.1038/

323533a0

Satterthwaite, T. D., Wolf, D. H., Loughead, J., Ruparel, K., Elliott, M. A., Hakonarson, H., ... Gur, R. C. R. E. (2012). Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth. *NeuroImage*, *60*(1), 623–632. Retrieved from http://dx.doi.org/10.1016/j.neuroimage.2011.12.063 doi: 10.1016/j.neuroimage.2011.12.063

Schawkat, K., Ciritsis, A., von Ulmenstein, S., Honcharova-Biletska, H., Jüngst, C., Weber, A., ... Reiner, C. S. (2020). Diagnostic accuracy of texture analysis and machine learning for quantification of liver fibrosis in MRI: correlation with MR elastography and histopathology. *European Radiology*, *30*(8), 4675–4685. doi: 10.1007/s00330-020-06831-8

Sheikh, H. R. (2005). An Information Fidelity Criterion for Image Quality Assessment. , *14*(12), 2117–2128. Retrieved from http://live.ece.utexas.edu/research/quality/ifcvecrelease.zip.

Sheth, S. G. (1997, 1). Nonalcoholic Steatohepatitis. *Annals of Internal Medicine*, *126*(2). doi: 10.7326/0003-4819-126-2-199701150-00008

Shnayderman, A., Gusev, A., & Eskicioglu, A. M. (2006). An SVD-based grayscale image quality measure for local and global assessment - IEEE Journals & Magazine. , *15*(2), 422–429.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. , 1–14. Retrieved from http://arxiv.org/abs/1409.1556

Simpson, M. S., You, D., Rahman, M. M., Xue, Z., Demner-Fushman, D., Antani, S., & Thoma, G. (2015, 1). Literature-based biomedical image classification and retrieval. *Computerized Medical Imaging and Graphics*, *39*. doi: 10.1016/j.compmedimag.2014.06.006

Şirli, R., Sporea, I., Bota, S., & Jurchiş, A. (2013, 8). Factors influencing reliability of liver stiffness measurements using transient elastography (M-probe) - Monocentric experience. *European Journal of Radiology*, *82*(8), e313-e316. doi: 10.1016/j.ejrad.2013.03.002

Strudel, R., Garcia, R., Laptev, I., & Schmid, C. (2021). Segmenter: Transformer for Semantic Segmentation. In *2021 ieee/cvf international conference on computer vision (iccv)* (pp. 7242–7252). doi: 10.1109/ICCV48922.2021.00717

Tavakkoli, A., Kamran, S. A., Hossain, K. F., & Zuckerbrod, S. L. (2020, 12). A novel deep learning conditional generative adversarial network for producing angiography images from retinal fundus photographs. *Scientific Reports*, *10*(1). doi: 10.1038/s41598-020-78696-2

Taylor, A. J., Salerno, M., Dharmakumar, R., & Jerosch-Herold, M. (2016, 1). T1 Mapping. *JACC: Cardiovascular Imaging*, *9*(1). doi: 10.1016/j.jcmg.2015.11.005

Thandi, S., & NHS. (2021, 5). *Statistics on Obesity, Physical Activity and Diet, England 2021*. Retrieved from https://digital.nhs.uk/data-and-information/publications/statistical/statistics-on-obesity-physical-activity-and-diet/england-2021

Trivizakis, E., Manikis, G. C., Nikiforaki, K., Drevelegas, K., Constantinides, M., Drevelegas, A., & Marias, K. (2019). Extending 2-D Convolutional Neural Networks to 3-D for Advancing Deep Learning Cancer Classification with Application to MRI Liver Tumor Differentiation. *IEEE Journal of Biomedical and Health Informatics*, *23*(3), 923–930. doi: 10.1109/JBHI.2018.2886276

Tunnicliffe, E. M., Banerjee, R., Pavlides, M., Neubauer, S., & Robson, M. D.

(2017). A model for hepatic fibrosis: the competing effects of cell loss and iron on shortened modified Look-Locker inversion recovery T1(shMOLLI-T1) in the liver. *Journal of Magnetic Resonance Imaging*, *45*(2), 450–462. doi: 10.1002/jmri.25392

UKBiobank. (2015). Body (abdominal) MRI scan, Version 1.0. , 1–7. Retrieved from http://www.ukbiobank.ac.uk/

V., K., S., M., N., L., D., L., G., V., P., M., ... J., S. (2013). Quality assessment of cardiovascular magnetic resonance in the setting of the European CMR registry: Description and validation of standardized criteria. *Journal of Cardiovascular Magnetic Resonance*, *15*(1), 1–13. Retrieved from http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L52647628%5Cnhttp://dx.doi.org/10.1186/1532-429X-15-55%5Cnhttp://sfx.library.uu.nl/utrecht?sid=EMBASE&issn=10976647&id=doi:10.1186%2F1532-429X-15-55&atitle=Quality+assessment+of+card

van Dijk, K. R., Sabuncu, M. R., & Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *NeuroImage*, *59*(1), 431–438. Retrieved from http://dx.doi.org/10.1016/j.neuroimage.2011.07.044 doi: 10.1016/j.neuroimage.2011.07.044

Venkatesh, S. K., Yin, M., Glockner, J. F., Takahashi, N., Araoz, P. A., Talwalkar, J. A., & Ehman, R. L. (2008, 6). MR elastography of liver tumors: Preliminary results. *American Journal of Roentgenology*, *190*(6), 1534–1540. doi: 10.2214/AJR.07.3123

Wang, Y., Song, Y., Wang, F., Sun, J., Gao, X., Han, Z., ... Yang, G. (2020). A two-step automated quality assessment for liver MR images

based on convolutional neural network. *European Journal of Radiology*, *124*(January), 108822. Retrieved from https://doi.org/10.1016/j.ejrad.2020.108822 doi: 10.1016/j.ejrad.2020.108822

Wang, Z., Bovik, A. C., Sheikh, H. R., Member, S. S., Simoncelli, E. P., & Member, S. S. (2004). Image quality assessment_from error visibility to structural similarity_IEEE trans on image processing.pdf. , *13*(4), 1–14. doi: 10.1109/TIP.2003.819861

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004, 4). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. doi: 10.1109/TIP.2003.819861

Wang, Z., & Li, Q. (2011). Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, *20*(5), 1185–1198. doi: 10.1109/TIP.2010.2092435

Wang, Z., Simoncelli, E., & Bovik, A. C. (2000). ) P ( M Λ. , *2*(Ki L), 1398–1402.

Wang, Z., Simoncelli, E., & Bovik, A. C. (2003). MULTI-SCALE STRUCTURAL SIMILARITY FOR IMAGE QUALITY ASSESSMENT. *IEEE Asilomar Conference on Signals, Systems and Computers*, 1398–1402.

Webb, A. (2011, 7). Dielectric materials in magnetic resonance. *Concepts in Magnetic Resonance Part A*, *38A*(4), 148–184. Retrieved from http://doi.wiley.com/10.1002/cmr.a.20205http://doi.wiley.com/10.1002/cmr.a.20219 doi: 10.1002/cmr.a.20219

Wei, D., Ahmad, S., Wu, Z., Cao, X., Ren, X., Li, G., … Wang, Q. (2019). Morphological Simplification of Brain MR Images by Deep Learning for Facilitating Deformable Registration. In H.-I. Suk, M. Liu, P. Yan, & C. Lian (Eds.), *Machine learning in medical imaging* (pp. 203–211). Cham: Springer International Publishing.

Xia, Y., Zhang, L., Ravikumar, N., Attar, R., Piechnik, S. K., Neubauer, S., ... Frangi, A. F. (2021). Recovering from missing data in population imaging â Cardiac MR image imputation via conditional generative adversarial nets. *Medical Image Analysis*, *67*, 101812. Retrieved from https://doi.org/10.1016/j.media.2020.101812 doi: 10.1016/j.media.2020.101812

Yammine, G., Wige, E., & Kaup, A. (2010). A no-reference blocking artifacts visibility estimator in images. *Proceedings - International Conference on Image Processing, ICIP*, *2*, 2497–2500. doi: 10.1109/ICIP.2010.5652927

Yoneda, M. M., Imajo, K., Takahashi, H., Ogawa, Y., Eguchi, Y., Sumida, Y., ... Nakajima, A. (2018). Clinical strategy of diagnosing and following patients with nonalcoholic fatty liver disease based on invasive and noninvasive methods. *Journal of Gastroenterology*, *53*(2), 181–196. Retrieved from https://doi.org/10.1007/s00535-017-1414-2 doi: 10.1007/s00535-017-1414-2

Zeiler, M. D., & Fergus, R. (2012). Visualizing and Understanding Convolutional Networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8689 LNCS*(PART 1), 818–833. doi: 10.1007/978-3-319-10590-1{_}53

Zeng, G., He, Y., Yu, Z., Yang, X., Yang, R., & Zhang, L. (2016). InceptionNet/GoogLeNet - Going Deeper with Convolutions. *Cvpr*, *91*(8), 2322–2330. doi: 10.1002/jctb.4820

Zhang, B., Tian, J., Dong, D., Gu, D., Dong, Y., Zhang, L., ... Zhang, S. (2017). Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. *Clinical Cancer Research*, *23*(15), 4259–4269. doi: 10.1158/1078-0432.CCR-16-2910

Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). Correspondence FSIM : A

Feature Similarity Index for Image. *IEEE Transactions on Image Processing*, *20*(8), 2378–2386. doi: 10.1109/TIP.2011.2109730

Zhang, L. L., Zhang, L. L., Mou, X., & Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, *20*(8), 2378–2386. doi: 10.1109/TIP.2011.2109730

Zhang, X., Feng, X., Wang, W., & Xue, W. (2013). Edge strength similarity for image quality assessment. *IEEE Signal Processing Letters*, *20*(4), 319–322. doi: 10.1109/LSP.2013.2244081

Zhao, J., Li, D., Kassam, Z., Howey, J., Chong, J., Chen, B., & Li, S. (2020). Tripartite-GAN: Synthesizing liver contrast-enhanced MRI to improve tumor detection. *Medical Image Analysis*, *63*, 101667. Retrieved from https://doi.org/10.1016/j.media.2020.101667 doi: 10.1016/j.media.2020.101667

# A | Appendices

## A.1 Manual Bounding Box Software and Labelling Process

In order to generate bounding boxes for the organs, vessels and objects of interest as used in 3, bespoke software was written in python using the PyQt5 (https://doc.qt.io/qtforpython/) library. The software loaded in the 14 DICOM images (7 magnitude and 7 phase images) which had been pre-downloaded from UKBiobank. The software then shows these images to the user alongside the DICOM header information, as shown in fig. A.1.

At this point the user can look through all 14 images by use of the up and down keys and/or the user interface. Once the user has found the optimal contrast image that facilitates drawing the bounding box of a particular class, they then select the class they want to label and open the cropping window, shown in fig.A.2.

Once the bounding box has been submitted, the table showing the regions labelled, is updated on the right of the user interface. Additionally the cropped region is shown visually on the central image, as shown in fig. A.3. The user repeats the process for all the classes that are visible in the image, and can repeat labels by deleting the bounding box for a specific class and redrawing. Once the user is happy with all bounding boxes, they can move to the next subject using the left and right arrow keys.

The code for this software can be found here: https://github.com/hillce/Manual_Classification

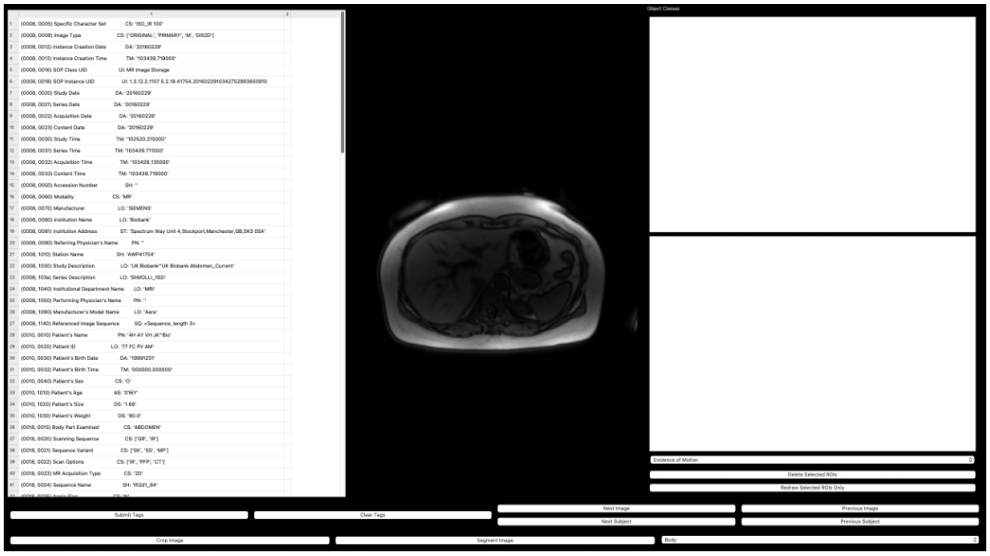


Figure A.1: The user interface presented on opening the application. On the left is the DICOM header information for that subject. In the centre is the DICOM image data. On the right is the tables feeding back the labels and bounding boxes drawn. Finally, at the bottom are the button controls needed to use the GUI. These buttons are mapped to hotkeys to aid in speed of labelling

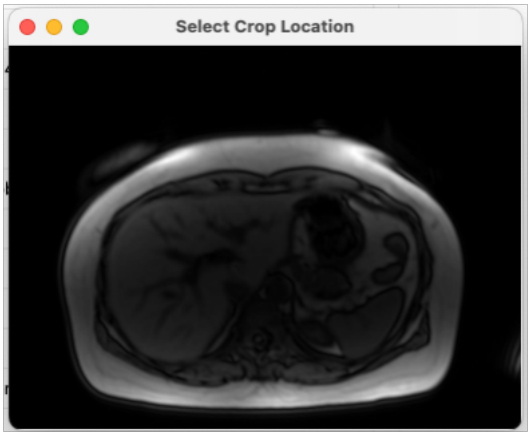



Figure A.2: The actual drawing interface is a canvas object opened as a pop up window from the main interface. It allows the user to quickly drag a box around the class they wish to label and submit it by hitting the space bar

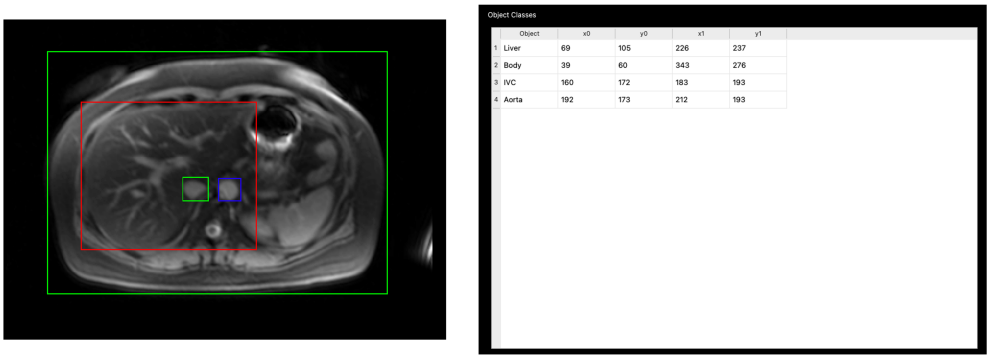



Figure A.3: Drawing interface for ground truth generation, showing bounding boxes

## A.2 LMSQuality - GUI for QC reporting

Preliminary work was performed into the development of a quality control GUI to be run locally on the scanner, allowing us to test instant feedback of quality issues to the scanner operator. As with the manual bounding box user interface, the software was written in python using PyQt5.

The Siemens scanner used for this preliminary study, was setup to automatically dump DICOM into a data directory on a local desktop, as soon as it has been processed. This data directory was monitored by the software. When an acquisition arrived, the software loaded in the DICOM data and formatted it ready for an earlier form of the fRCNN network (No small vessels) to predict which organs were present in the scan. It would then display this information to the scanner operator, as shown in fig. A.4.

In addition to the deep learning checks performed by the fRCNN network, the software also checked for DICOM header information being out of range. One example would be the imaging frequency varying over the course of an acquisition.




Figure A.4: The graphical user interface displaying an acquired scan and previous scans.

The code for this software is currently private, as it is contained within another users git account. If the reader would like a copy of the software, please get in contact with the author.

# Document Information

## Document

| | |
|---|---|
| Title | Thesis |
| ID | aad50472-5455-44a4-81a7-23c348fbf4d6 |
| Status | Signed |
| Sender | Charles Hill (charles.hill@perspectum.com) |
| Dept/business | Perspectum |
| Dept/business ID | perspectum-diagnostics3 |
| PDF Certified | True |

## Signers

| Name | Authentication |
|---|---|
| Charles Hill <charles.hill@perspectum.com> | Email |

## History

| Event | Email | IP address | Date/Time |
|---|---|---|---|
| Thesis created by Charles Hill | charles.hill@perspectum.com | 62.255.31.130 | Tue, 13 Feb 2024 10:19:36 +0000 |
| Thesis emailed to Charles Hill | charles.hill@perspectum.com | | Tue, 13 Feb 2024 10:19:37 +0000 |
| Thesis visited by Charles Hill <charles.hill@perspectum.com> | charles.hill@perspectum.com | 62.255.31.130 | Tue, 13 Feb 2024 10:20:12 +0000 |
| Browser geo-location provided: Lat 51.7291069406044, Long -1.2069640966346389 | charles.hill@perspectum.com | 62.255.31.130 | Tue, 13 Feb 2024 10:20:14 +0000 |
| Signature applied (page 4) | charles.hill@perspectum.com | 62.255.31.130 | Tue, 13 Feb 2024 10:21:02 +0000 |
| Ticked checkbox for statement: I confirm that I am authorised to accept the terms of this document. | charles.hill@perspectum.com | 62.255.31.130 | Tue, 13 Feb 2024 10:21:22 +0000 |
| Ticked checkbox for statement: I understand that accepting the terms of this document creates a legally binding obligation. | charles.hill@perspectum.com | 62.255.31.130 | Tue, 13 Feb 2024 10:21:22 +0000 |
| Signer offered option to accept or reject document | charles.hill@perspectum.com | 62.255.31.130 | Tue, 13 Feb 2024 10:21:22 +0000 |
| Clicked button with text: By clicking this button I hereby ACCEPT the terms of this document | charles.hill@perspectum.com | 62.255.31.130 | Tue, 13 Feb 2024 10:21:22 +0000 |
| Thesis completed by Charles Hill <charles.hill@perspectum.com> | charles.hill@perspectum.com | 62.255.31.130 | Tue, 13 Feb 2024 10:21:22 +0000 |