Technical University Delft

Master Thesis

Recovering Quantitative Maps from Conventional Weighted MR Images using Deep Learning

Bryan Lusse

Summer 2021

TU Delft

Erasmus MC
University Medical Center Rotterdam

Cover Photo: Image showing both a qualitative $T1$-weighted MR image (left) and a quantitative $T1$-map (right) of a healthy brain.

Recovering Quantitative Maps from Conventional Weighted MR Images using Deep Learning

by

Bryan Lusse

to obtain the degree of

Master of Science
in Biomedical Engineering

at the Delft University of Technology,
to be defended publicly on Wednesday August 25th, 2021 at 10:00 AM.

Student number:	4605209	
Project duration:	January 4, 2021 – August 25, 2021	
Thesis committee:	Dr. F. M. Vos,	TU Delft, Chair of the thesis committee
	Prof. Dr. J. A. Hernández-Tamames,	Erasmus MC, supervisor
	ir. S. R. van der Voort,	Erasmus MC, daily supervisor
	Prof. Dr. W. J. Niessen,	Erasmus MC

Abstract

Long acquisition times impede the routine clinical use of quantitative magnetic resonance imaging (qMRI). qMRI quantifies meaningful tissue parameters in T1-, T2-, and PD-maps, as opposed to conventional (qualitative) weighted MRI (wMRI), which only visualises contrast between tissues. Although methods exist that generate synthetic wMRI from qMRI, the inverse problem has not been thoroughly studied yet. A method to generate qMRI from wMRI would be beneficial as it does not change current clinical workflows and enables retrospective quantitative analysis. This thesis investigates to what extent fully convolutional networks are successful in generating qMRI from T1-weighted, T2-weighted, PD-weighted and T2-weighted-FLAIR scans. A set of synthetic wMRI scans from 97 healthy volunteers was split into training, validation and test sets for development of our models. We varied model architectures, loss functions and learning rates during training, in order to find the best performing models. These were able to predict qMRI with median errors of approximately 5% on the test set. Additionally, we determined the amount of information contained in the input scans by training models using different combinations of the input. These results showed that T1-weighted, T2-weighted and PD-weighted scans were the most important. Models trained on synthetic wMRI were tested on an additional dataset of real wMRI. This resulted in higher median errors of 27.4%, 12.0% and 8.7% for T1-, T2- and PD-maps respectively. Furthermore, the same models were tested on a third dataset of synthetic tumour scans and mainly showed errors around the tumour core. These results show that more research is necessary in order to improve the performances of models generating qMRI to a clinical standard.

Contents

List of Figures

List of Tables

List of Abbreviations

MRI	Magnetic Resonance Imaging
qMRI	Quantitative Magnetic Resonance Imaging
wMRI	Weighted (qualitative) Magnetic Resonance Imaging
T1w	T1-weighted MRI
T2w	T2-weighted MRI
PDw	Proton density-weighted MRI
T2w-FLAIR	T2-weighted fluid attenuated inversion recovery MRI
SE	Spin Echo
IR	Inversion Recovery
TE	Echo Time
TR	Repetition Time
TI	Inversion Time
CSF	Cerebrospinal Fluid
WM	White Matter
GM	Grey Matter
AI	Artificial Intelligence
DL	Deep Learning
CNN	Convolutional Neural Network
FCN	Fully Convolutional Network
GAN	Generative Adversarial Network
PSNR	Peak Signal-to-Noise Ratio
MSE	Mean Squared Error
MAE	Mean Absolute Error
SSIM	Structural Similarity Index
RMSE	Root Mean Squared Error

1

Introduction

All things are difficult before they are easy.
- Dr. Thomas Fuller

Medical imaging provides us with essential information about our human bodies. Without it, we would have no non-invasive way of diagnosing or monitoring certain diseases. Magnetic resonance imaging (MRI), X-ray, and computed tomography (CT) are only a few different types of imaging modalities that are available for use in the current day and age. These modalities give different contrast in images and are able to present us with complementary information in different scenario's. CT and X-ray imaging techniques for example, excel at visualising bone, whereas MRI enables us to acquire an image with a higher contrast between soft tissues.

For many applications in neuro-oncology, MR imaging is the modality of choice due to its excellent soft-tissue contrast that enables tumours to be distinguishable. Practically used MRI scans can be separated into *qualitative* or weighted MRI (wMRI), and *quantitative* MRI (qMRI). The former produces images based on relative differences between tissue whereas the latter quantifies tissue parameters in an absolute manner.

Most clinically acquired MR images are qualitative *weighted* scans. There are three tissue-specific parameters that are used to differentiate between tissues: the T1 time, T2 time and the proton density (PD) (explained in Chapter 2). Weightings from all three parameters are always present in every qualitative image, although there generally is one dominant weighting. Each of these mentioned modalities provides a different view of the MR parameters in the body. Examples of brain images made through important qualitative imaging modalities are shown in Figure 1.1.

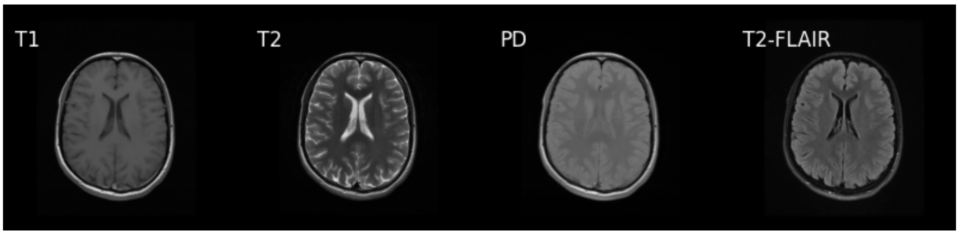

Figure 1.1: Different types of MR images used in practice. From left to right, T1-weighted, T2-weighted, proton density-weighted and T2-weighted-fluid attenuated inversion recovery (FLAIR) images.

These qualitative images are useful because they are relatively easy to acquire and can give appropriate information for many medical applications. General acquisition times of such a brain image fall between 1-3 minutes (1). However, besides tissue-specific parameters there are also system parameters that influence the final image. As a result, qualitative images only visualise *differences* between tissues, i.e. they visualise tissue contrast. Qualitative images with a large contrast between tissues usually give enough information to detect and preliminary diagnose various conditions such as tumours and lesions. However, the problem is that this relative image nature can lead to large signal differences between images of the same patient in different scanners or imaging centres (2). This makes it difficult for radiologists to quantitatively compare images and results, in addition to making it complicated to find subtle pathological differences between subjects, or for the same subject at different time points.

To overcome these limitations, tissue-specific parameters may be quantified. These parameters do not change for a healthy individual (disregarding ageing) and have shown diagnostic value in preliminary studies (3; 4).

qMRI quantifies the tissue-specific parameters that cause the contrast in wMRI. In this way, images can be acquired independent of system parameters, leading to smaller differences for images on a different scanner or at different time points (5; 6) (These errors mainly still occur due to differences in scanners from different vendors, or versions of analysis software). As a result, qMRI potentially enables us to capture small pathological differences between multiple scans of the same patient (Figure 1.2). In this way, tracking the development of tumours may become easier and more consistent. Currently, promising results have been found regarding the difference in T1 and T2 values for malignant and benign tumours (4).

Conventional qMRI methods quantify the tissue-specific parameters by making use of multiple qualitative images. The main downside of acquiring qMRI in this way is a longer acquisition time. Many methods are studied in order to acquire qMRI maps quicker (7–9) or simultaneously for multiple parameters (10; 11). Current acquisition times are around 5-7.5 minutes to gather multiple quantitative maps (12; 13). As such, these methods are slower than qualitative acquisition methods.

To get qMRI in clinical practice scanning times must be at least as quick as wMRI acquisition, also examples should be established in which qMRI is undeniably more beneficial for the patient. Accordingly, a way of generating qMRI from already present wMRI could have extraordinary potential. qMRI studies would benefit from the signif-

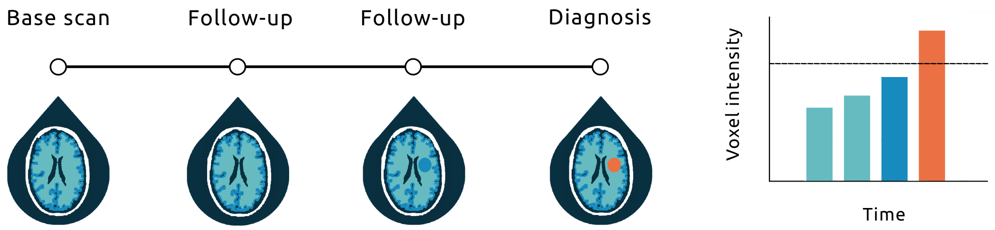

Figure 1.2: Rationale behind quantitative imaging. Multiple visualisations of longitudinal qMRI scans are shown (left). A developing tumour or lesion would become better detectable in these images over time. Similarly, the intensities of the pathological voxels would change significantly over time (right). A diagnosis could be made when a certain diagnostic threshold (dotted line) were to be exceeded.

icantly quicker acquisition, and quantitative maps can also be generated retrospectively from studies where qualitative data was already present in order to enlarge the pool of data that can be used for research.

Deep Learning (DL) is a field in which complex image-related problems are solved. One such way is through image-to-image translation, where MRI scans can be translated into CT scans (14–16), or where one type of MRI scan can be translated into another type (17; 18). This has served as inspiration for our study and others, as already, promising preliminary results for qMRI generation from wMRI have been shown by Wu *et al.* (19) and Moya-Sáez *et al.* (20).

This thesis addresses the problem of generating *quantitative* MRI from different conventional *qualitative* MRI scans (T1-weighted, T2-weighted, PD-weighted and T2-weighted-FLAIR) using DL and employs methodical approaches to find the best performing models to do so.

1.1. Contributions

The contributions of this project are fourfold. Firstly, by studying qMRI generation from synthetic wMRI, we added information to the small scientific literature. Secondly, we investigated the importance of the different wMRI scans for the generation of qMRI. Thirdly, we used synthetic tumour scans to capture the generalisability of our model on pathological scans. Finally, we studied the ability of our method to work in a real-life environment by using real clinical scans.

1.2. Outline

This thesis first goes into the technical background on image formation using MRI, and image processing using artificial intelligence in Chapter 2. In Chapter 3, we will present the methods of our research by explaining network architectures and the incentive behind our experiments. Chapter 4 shows the results of our study and quantifies the performance of our models. Finally, we discuss our results and their impact in Chapter 5.

2

A Background on MRI and AI

In this background chapter we explain important information about the acquisition and analysis of MR images. Section 2.1 describes MRI and qMRI, while Section 2.2 explains DL methods used to analyse images. Ultimately, Section 2.3 concludes by discussing the state of the art in qMRI generation from wMRI.

2.1. Magnetic Resonance Imaging

Magnetic resonance imaging originated in the last century. What started as purely experimental physics gradually turned into an application that would be paramount for healthcare in the years to come. Both Bloch and Purcell laid the basis when observing resonance signals from atomic nuclei immersed in a magnetic field (21; 22). From here it took until 1977 before this knowledge was used by Damadian to produce the first working MRI machine (23). But how do these machines actually work? In this section we will describe the mechanisms behind spin relaxation in the body and the formation of images through an MRI scanner.

2.1.1. Spin Relaxation

An MRI scanner uses the magnetic behaviour of particles in our body to make images. Since hydrogen (H_1), stored in water, is the most abundant particle and allows for easy manipulation of its magnetic moment, it is useful for imaging purposes.

When an H_1 proton is present in a strong magnetic field, its *spin* will align along the direction of this field. Spin is a quantum mechanical property that can be perceived as a vector signifying the magnetic moment of the particle. In short, the proton behaves as a magnet. When looking at many spins, we sum all their magnetic moments together and talk about the net magnetization vector. When aligned along a strong magnetic field (B0-field), the net magnetization vector is in equilibrium. However this equilibrium can be disturbed. A second, weaker magnetic field (B1-field), oriented perpendicular to the main field, can be applied momentarily to the system in the form of a radiofrequency (RF) pulse. By doing this, the orientation of all the spins (and therefore the net magnetization vector) changes, but eventually falls back (or relaxes) to its original alignment (Figure 2.1). The time in which this relaxation happens is mainly determined by the type of tissue that the spins are in but also depends on the strength of the magnetic field. Two types of spin relaxation exist as the used magnetic fields are present in two directions. T1-relaxation (spin-lattice relaxation) is determined as

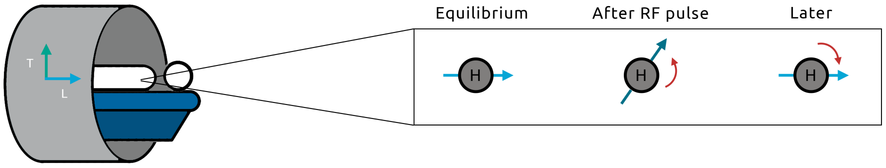

Figure 2.1: Spin relaxation. A patient is shown in a MRI scanner with the longitudinal (blue) and transverse (green) axis indicated (left). The evolution of the spin over time is shown for a single proton (right). Initially it will be aligned along the main magnetic field but this will change due to the RF pulse. Afterwards the magnetisation will fall back to equilibrium again and align as before. This relaxation process happens at different speeds in different tissues.

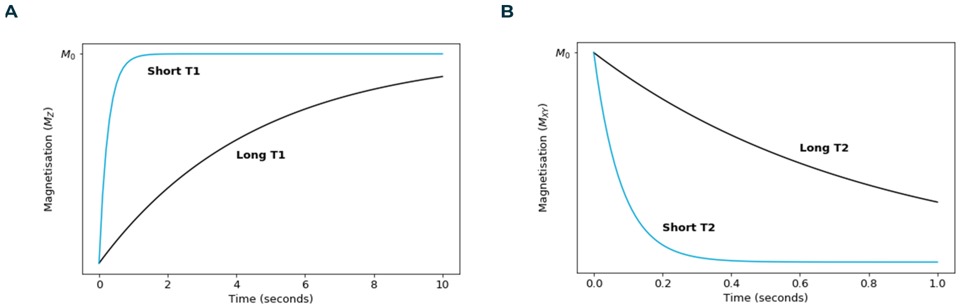

Figure 2.2: T1 and T2 relaxation. **A**: Illustration showing the differences in T1 relaxation for different tissues with a long and short T1 constant. **B**: Illustration showing the differences in T2 relaxation for different tissues with a long and short T2 constant.

the time it takes magnetisation in the longitudinal plane to return to equilibrium. T2-relaxation (spin-spin relaxation) on the other hand, is determined as the time it takes magnetisation to be completely removed from the transverse plane. The time it takes magnetisation in the longitudinal plane to return to equilibrium is not equal to the time it takes magnetisation to be completely removed from the transverse plane.

Relaxation times for both processes differ for various tissues, making it possible to differentiate tissues by comparing relaxation times (Figure 2.2). The number of protons in a unit amount of tissue, or the proton density (PD), also enables us to differentiate between tissues. These three parameters (T1, T2, and PD) are tissue-specific parameters, since they are related to the type of tissue.

2.1.2. Forming an Image

To form an image, we can measure the amount of magnetisation that is present in the transverse plane. In order to also capture T1 effects, RF pulses can be used to flip the magnetisation again to the transverse plane. As the tissue-specific parameters influence the magnitude of this magnetisation, different tissues will give rise to a different signal. If we transform these magnetisation amounts into pixel intensities, a 2D image of a slice of our body is obtained where different tissues have different intensities. When we repeat this process for multiple slices we can reconstruct a 3D volume of the imaged part of the body.

In general, the signal in the transverse plane will be affected by the T1, T2, and the PD of the tissue. Their combined effects do not give us a strong signal contrast between tissues. However, there are multiple methods to influence the signal that is measured in order to modify the contrast between tissues. As a result, dedicated techniques have been devised to acquire signals that are mainly dependent on one tissue-specific parameter. Nevertheless, these methods also introduce imaging-specific (or

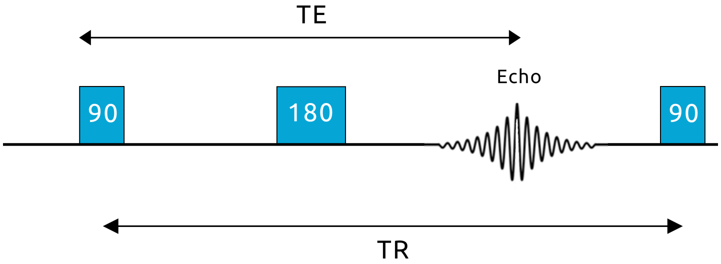

Figure 2.3: Spin echo pulse sequence. A 90° RF pulse is used to change the alignment of the magnetisation. Additionally, the 180° pulse counters signal dephasing. An echo with high signal magnitude arises exactly at time TE. The time in between consecutive 90° RF pulses is the TR.

acquisition) parameters that influence the signal that we measure. In this chapter, we will focus on the acquisition of a T1-weighted (T1w) image and therefore explain the spin echo (SE) and inversion recovery (IR) sequences.

Spin Echo

The SE is the most basic way of acquiring an MRI image (24). An SE consists of the previously mentioned 90° RF pulse with an additional 180° pulse some time later (Figure 2.3). Spins dephase during regular spin relaxation because of irregularities in the strength of the magnetic field and spin-spin interactions. Especially the former results in a faster signal decay over time, and therefore a lower signal and lower tissue contrast that is detectable. The 180° pulse reverses these effects, enabling an echo to be measured. The time between RF pulses, and therefore the amount of magnetisation that is recovered or realigned, is crucial for the appearance of the final image. The time in between the initial RF pulse and the echo is the echo time (TE) and the time in between consecutive 90° RF pulses is the repetition time (TR). The TE and TR are imaging-specific parameters that control the magnitude of the signal that we detect and therefore the contrast visible in an image.

The actual signal that we detect in qualitative MRI is thus dependent on both tissue-specific and imaging-specific parameters. This relation follows,

$$S = PDe^{(-TE/T2)}(1 - e^{(-TR/T1)}) \tag{2.1}$$

where a T1 weighting is acquired by employing a short TE and TR.

Inversion Recovery

An SE sequence can thus be used to acquire a T1w MRI. However, more often a different method is used, namely the IR sequence (25). This method adds an additional pulse before a conventional SE (Figure 2.4). This additional pulse inverts the

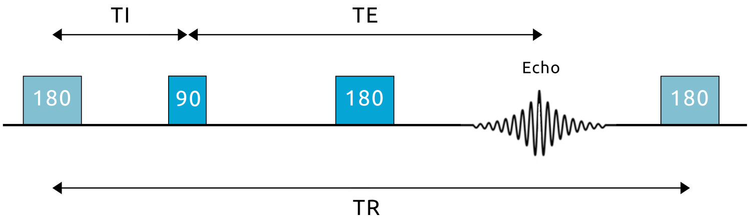

Figure 2.4: Inversion recovery. Due to the extra inversion pulse, the magnetisation starts to recover governed by T1 effects. After a certain TI, a conventional spin echo is used to acquire image signals.

magnetisation by 180°, flipping it from the positive longitudinal axis to the negative longitudinal axis. The IR sequence is particularly useful for T1-weighting as the inversion pulse results in recovery of the magnetisation over a larger signal range compared to an SE, which is governed by the T1. As a result, IR sequences lead to better tissue contrast than regular SE sequences (26). The addition of an extra pulse brings the introduction of an additional imaging-specific parameter with it. This parameter, the inversion time (TI), is the time between the inversion pulse and the 90° pulse of the SE.

When taking the TI into account, the signal evolution from Equation 2.1 changes to

$$S = PDe^{(-TE/T2)}(1 - 2e^{(-TI/T1)} + e^{(-TR/T1)}) \tag{2.2}$$

2.1.3. Quantitative MRI

As mentioned in the previous chapter, qMRI maps are made using multiple conventional MRI scans. There are a multitude of methods for acquiring these maps. These methods either quantify one relaxation parameter or multiple at once. By doing so, the actual underlying tissue-specific constants are found and the influences of imaging-specific parameters on the image are removed. This leads to images that are more reproducible and leave less room for ambiguities.

In T1-mapping for example, the relaxation curve gets sampled by doing measurements at multiple points in time (Figure 2.5). By fitting an exponential function to these data points, the underlying T1 time can be found. The use of multiple images is also performed in T2- and PD-mapping. The possibility of visualising solely tissue-specific parameters makes it easier to evaluate the scans and compare results to other patients or institutes.

2.1.4. Synthetic MRI

The quantification of tissue-specific parameters enables us to do more than just construct T1-, T2- and PD-maps. When we have values for T1, T2 and PD, we can plug these into equations 2.1 & 2.2 to construct wMRI images (27). Additionally we have to decide on imaging-specific parameters to use (values for TE, TR and TI). The resulting images are not exactly the same as conventional wMRI, but they simulate it. Such images are called synthetic MRI. Synthetic wMRI images have shown to reach

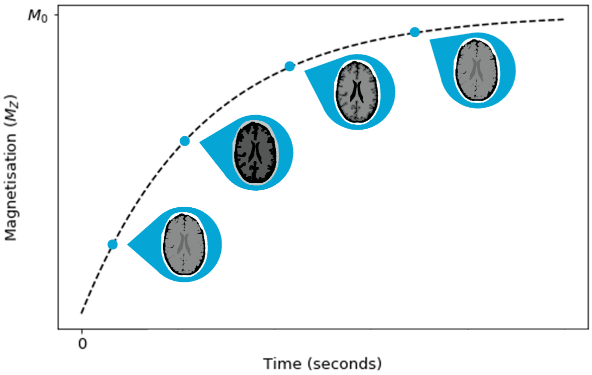

Figure 2.5: Sampling the T1 decay curve. MRI scans (blue dots) need to be acquired at different time points to collect enough information to find the underlying relaxation curves (dotted line) and subsequently the tissue parameter for every voxel (in this case the T1 time).

a similar diagnostic value as real wMRI images, but the contrast and quality of the image can still differ (28; 29). This is mainly due to the fact that synthetic images do not take magnetic field behaviour into account .

2.2. Artificial Intelligence

After acquisition, medical images, such as MRI scans, need to be interpreted correctly to detect pathologies or assess an individual's health. Radiologists are trained to do exactly this. To aid radiologists, artificial intelligence (AI) solutions are studied more and more (30). AI aims to emulate cognitive behaviour in computers to achieve similar or better performances than humans on a variety of tasks. Deep learning (DL) is the most recent and widespread method for doing this. DL employs computer models or 'neural networks' which are inspired by the human brain and try to actively learn associations between an input and the desired output by training and optimising on large amounts of data. These networks learn by means of a *loss function* of which the output signifies to what extent the model is making accurate predictions. The loss is iteratively minimised by tweaking the weights the model gives to the input and the intermediate outputs. Eventually, this aims to produce a model that makes correct predictions. Loss functions are mainly constructed by looking at the differences between the predicted and actual outcomes. In image based solutions for example, loss functions are used based on the difference between pixel values.

Two DL methods that are important for image translation problems are fully convolutional networks (FCNs) and generative adversarial networks (GANs). We present these concepts in the following sections as an introduction to the subsequent chapters. Additionally, the attention mechanism, a method that can improve model performance, is explained.

2.2.1. Fully Convolutional Networks

Convolutional neural networks (CNNs) are widely studied for image analysis. CNNs take an image as input and converge to a single value. This value can signify a binary prediction made from the image (e.g. sick or healthy), a prediction of a certain classification of the image (e.g. car, boat or bicycle) or the actual prediction of a certain value of interest (e.g. distance, length or age). In order to do this, CNNs make use of a central concept in DL, *convolution*. This aims to extract essential information and features from the input image. Convolution is usually followed by a downscaling of the image in order to make the output focused on smaller and smaller image details. CNNs additionally use *dense* layers that ensure that the output prediction is a single value. FCNs are a subclass of CNNs and also make use of convolution. However,

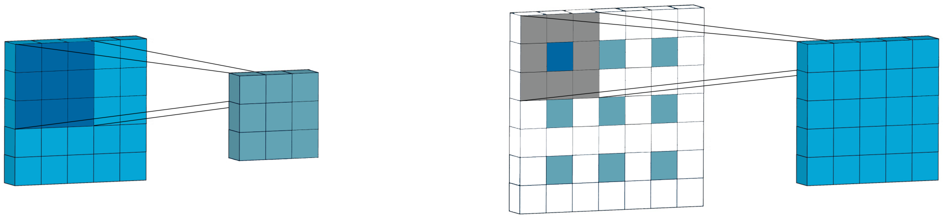

Figure 2.6: The process of convolution. Multiple pixels in the input image get combined to form one output pixel (left). Transposed convolution performs a convolution to end up with an output with the same dimensions as before (right). Here the input image is padded with zero intensity voxels for matching dimensions.

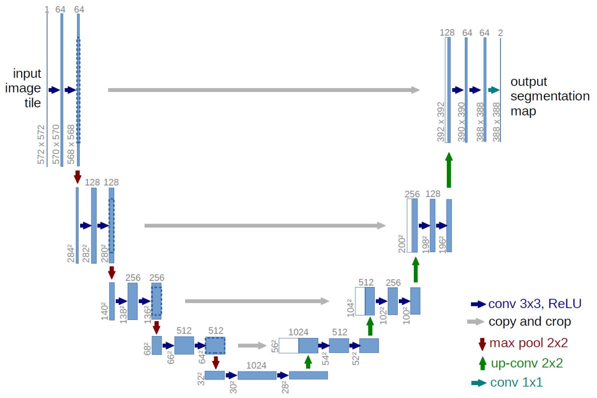

Figure 2.7: U-Net architecture. The input image is downsampled iteratively before being iteratively upsampled to its original size. Long connections between the upsampling and downsampling paths (grey) enable the network to recover information lost during downsampling. Image reproduced from Ronneberger *et al.* (31)

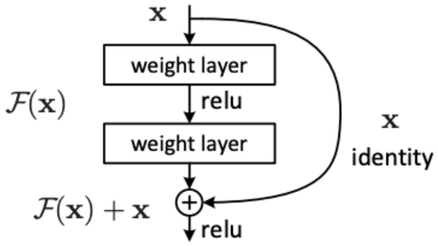
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
$\mathbf{x}$
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2.8: The building block of a ResNet. The input, x, first goes through consecutive convolutional layers to produce the output, $\mathcal{F}(x)$. Additionally, the original input, x, is added to the output. Image reproduced from He *et al.* (32).

as the name implies, FCNs are 'fully convolutional', i.e. they only make use of convolutional layers. The benefit of this is that the outputs can also be an image. FCNs are therefore often used for segmentation and translation tasks. When using FCNs, we require to end up with output images of the same size as the input. Therefore, a second mechanism is used to scale up images, which is called transposed convolution (or deconvolution). In this type of calculation, zero valued pixels are added as padding to increase the size of the input. From there, a normal convolution operation will result in an output with a larger size than the original input. Visualisations of both calculations are shown in Figure 2.6.

Different types of FCNs exist, of which the U-Net model is one of the most well known (31). This model initially downscales an image to identify smaller features before transforming the image back to its original shape (Figure 2.7). Simultaneously, information is extracted during the downscaling path and introduced in the upscaling path through long 'skip connections'. These connections *skip* multiple layers. In this way, information that has been lost during downsampling can still be recovered.

Another important architecture is the ResNet by He *et al.* (32). This model also uses extra skip connections between layers that cause the high performance of the model. These skip connections only skip a single layer (Figure 2.8). Due to these connections, the model can choose to skip the output of a certain layer and instead continue with the output of the previous layer, i.e. performing an identity mapping. As He *et al.* showed, this is beneficial when bigger networks are used, as they prevent such a network from learning redundant information using the extra layers and parameters it has.

The main difference between the connections in a U-Net and the skip connections

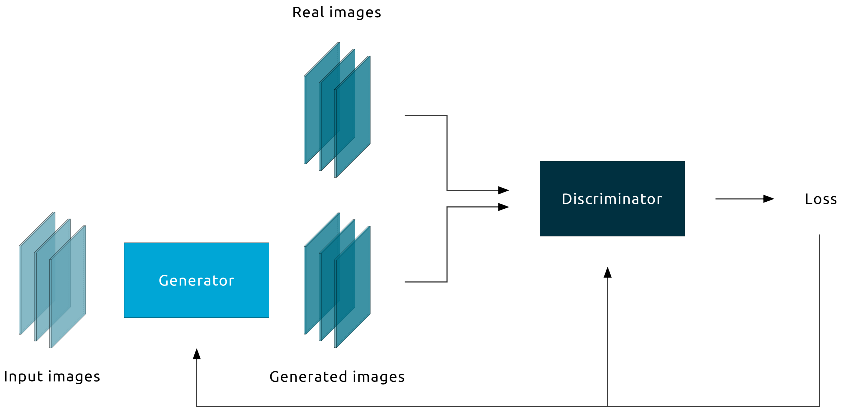

Figure 2.9: Basic GAN architecture. Both the generator and discriminator networks do not have a fixed architecture. The calculated loss influences both the generator and discriminator.

in a ResNet is that ResNet's skip connections are designed to only skip one or two layers, whereas the connections in a U-Net traverse over many layers.

2.2.2. Generative Adversarial Networks

GANs were first introduced by Goodfellow *et al.* (33) as a way to generate images similar to popular computer vision datasets (e.g. handwritten digits and faces). In the hope of achieving better results, GANs do not use one network, but two networks: a generator and a discriminator. Both these networks have opposing goals and compete against each other, enhancing their performance. The generator generates images from the input and the discriminator tries to distinguish the generated images from real images. As opposed to conventional techniques, the loss does not depend on differences between the images, but it depends solely on the output of the discriminator, i.e. to what degree the generated images are indistinguishable from the real images. As both networks optimise, the aim is that the generated images become more and more similar to the real images until at a certain point in time, the discriminator can not distinguish generated from real anymore.

The generator and discriminator in a GAN can essentially be any type of neural network. U-Net- and ResNet-like models are often used as generators, whereas CNN-based models are regularly used as discriminators. A basic visualisation of a GAN that generates multiple images can be seen in Figure 2.9.

A relevant GAN model is the 'pix2pix' model by Isola *et al.* (34). This model uses corresponding (paired) input and output images to learn the mapping between the two.

2.2.3. Attention

A neural network effectively has to learn which parts of the input it should give a high importance to and in what way it should combine them. To help a network with both goals, attention was introduced. Attention is a mechanism by which a neural network can learn the importance of certain parts of the input, in order to arrive at the correct output. Attention is very popular in language-based tasks (35; 36), but has also branched out into image-based tasks (37).

A special subclassification of attention, self-attention, uses only the input data in order to determine important regions in the data. This technique can be used to improve the modelling of relationships between spatial regions in the image. Utilising these relationships more effectively can improve model performance in image classification, segmentation and synthesis. As a result, self-attention has been used in multiple models that generate images (19; 38), where special attention goes to Wang *et al.* (39) and Oktay *et al.* (40) as they implemented self-attention in U-Net models.

Self-attention in image-based networks is calculated by a combination of computations that extract meaningful relations from the input data. The implementation of Zhang *et al.* (38) calculates attention similar to Vaswani *et al.* (36). Here, the input is used to calculate the importance of itself. The input is transformed into three feature spaces $f(x)$, $g(x)$ and $h(x)$ by an additional convolution operation. The attention map is then calculated as

$$\beta_{j,i} = \frac{\exp s_{ij}}{\sum_{i=1}^{N} \exp s_{ij}} \tag{2.3}$$

where $s_{ij} = f(x_i)^T g(x_j)$. This attention map is then multiplied by $h(x)$ and added to the original input.

Oktay *et al.* (40) use a slightly different approach, more suited for U-Nets, as here $f(x)$ and $g(x)$ are the feature spaces of the input of the current layer and the output of the skip connection, respectively. These are then added together, transformed through a convolution and multiplied by the original input of the current layer. As a result, information extracted from a coarser scale is used to focus on salient features.

2.3. Current State of the Art

As mentioned in Chapter 1, models exist that tackle tasks similar to generating qMRI. These models are mainly focused on CT generation from MRI and vice versa. But other interesting research has been done on synthesising T2-weighted (T2w) images from T1w images (18) and contrast enhanced T1w images from regular T1w images (39). Additionally, a great inspiration for our project was the knowledge available on the generation of a missing MRI image out of a standard set of images (17; 41; 42).

When we focus on qMRI generation methods, Wu *et al.* (19) stands out as they generated accurate knee qMRI maps from weighted MRI scans. Interestingly, they succeeded in generating T1-maps directly from a T1w image. Showing that a single weighted images may already contain sufficient information for the generation of a quantitative map. The model used is a U-Net with additional self-attention layers.

Additionally, Moya-Sáez *et al.* (20) came with more evidence that the accurate generation of brain qMRI maps from qualitative images is possible. They generated synthetic weighted scans from qMRI and used these to retrieve the original qMRI maps again. The model they used had a U-Net-inspired encoder-decoder architecture (43) and only used synthetic T1w and T2w images as input in order to generate T1-, T2- and PD-maps.

Other studies either mainly focus on generating qMRI from raw (k-space) MRI signals (44), or try to reduce the amount of weighted scans necessary for conventional qMRI acquisitions (45).

Our research aims to broaden the qMRI generation literature and to add more knowledge to be used in further research. Especially by evaluating the performance of models on real brain scans and scans of brain tumours.

2.3.1. Comparing GANs and FCNs for qMRI synthesis

GANs and FCNs are both used to solve image translation problems. However, the qMRI translation methods that explain their methods all use an FCN (19; 20; 44), which is related to the goal of generating qMRI. For the generation of *weighted* MR images, it is difficult to define a loss function using an FCN. Absolute differences between voxels do not have a considerable meaning since the images are constructed using relative differences. It is therefore hard for a network that learns from absolute differences to make consistent relative predictions. In this scenario, GANs are easier to implement as the discriminator takes care of optimising the performance and no specific loss function has to be devised. For qMRI generation on the other hand, the image values are quantitative, meaning they should be similar on every image. The usage of an FCN is therefore much easier as absolute differences and losses can be employed to enable the model to learn.

3 Methods

This project proposes computational methods based on FCNs to solve the qMRI generation problem. In this chapter, we initially describe our approach and the differences between conventional approaches (Section 3.1). Afterwards, we describe the used data (Section 3.2 & 3.3) and deep learning models (Section 3.4 & 3.5) before presenting the motivation behind different experiments in Section 3.6. All experiments were programmed in Python and used the PrognosAIs software package (version 0.3.5) (46).

3.1. Approach Compared to Conventional Methods

Conventional qMRI generation approaches use multiple weighted scans per tissue-specific parameter in order to arrive at a quantitative map, e.g. multiple T1w scans for a T1-map. Our approach only uses *one* scan per tissue-specific parameter and tries to use differently weighted scans to obtain the same amount of information. Figure 3.1 shows the difference in our approach compared to the general approach in conventional methods of generating qMRI. Our approach is easier for the patient and clinic, and has multiple benefits for the clinical acceptance of qMRI which have been discussed in Section 1.

Noteworthy, our approach is somewhat similar to approaches that aim to generate multiple qMRI maps simultaneously, for example, MR fingerprinting. If we disregard the small differences in acquisition time, there are still benefits to generating qMRI from wMRI. One benefit is that in the end real wMRI scans are present (as opposed to synthetic scans). Secondly, acquiring wMRI is currently still the standard workflow in clinics. When quantitative maps are acquired after wMRI acquisition, the patient spends extra time in the MRI scanner, which is often not considered as pleasant.

3.2. Data Acquisition

3.2.1. Synthetic Healthy Volunteer Data

A cohort of brain MRI's from 97 healthy volunteers was acquired under the HARPS (Harmonization of Resonators based on Physiological Signature) project. For each individual, qMRI maps were acquired and synthetic wMRI images were calculated during post-processing (Figure 3.2). Quantitative maps were obtained using a multiple-dynamic multiple-echo MR sequence. Subsequently, SyMRI (Synthetic MR, Sweden,

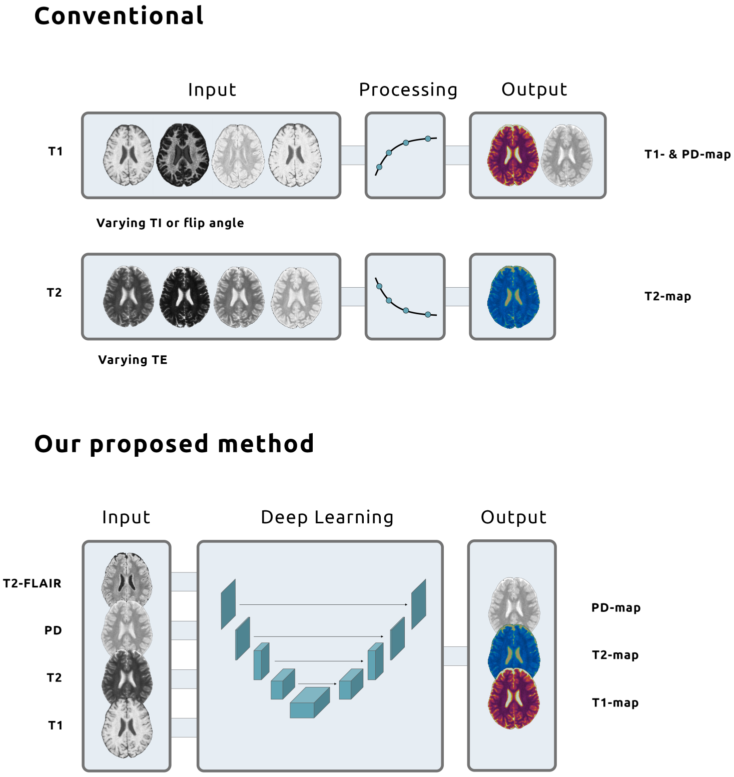

Figure 3.1: Proposed approach of generating qMRI. Conventional methods (left) generally acquire multiple scans where a single imaging-specific parameter is varied. These scans then give enough information to calculate a quantitative map. In our method (right), we aim to use scans that differ in multiple imaging-specific parameters in order to calculate multiple quantitative maps simultaneously. This involves less additional acquisition time.

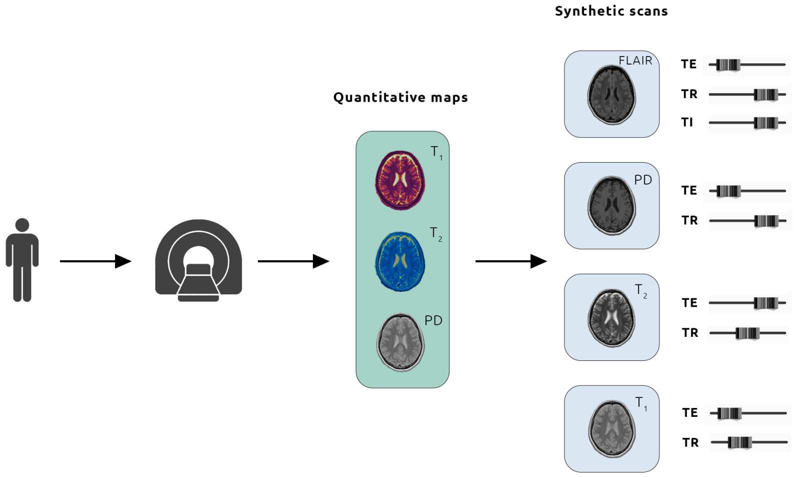

Figure 3.2: Visualisation of the data acquisition process. Healthy volunteers were scanned in the MRI scanner. qMRI maps were acquired directly (middle) and the corresponding qualitative images were synthetically calculated at a later stage (right). Tweaking of the imaging-specific parameters determines the contrast visible on these qualitative image.

Table 3.1: MRI imaging-specific parameters for synthetic healthy volunteer data. Parameter values are in milliseconds (ms).

MRI type	TR	TE	TI
T1-weighted	500	10	N/A
T2-weighted	4500	100	N/A
T2-weighted-FLAIR	15000	100	3000
PD-weighted	8000	10	N/A

Table 3.2: MRI imaging-specific parameters for real healthy volunteer data. Parameter values are in milliseconds (ms). Small differences between real and synthetic acquisitions arise due to the nature of the software.

Data	MRI type	TR	TE	TI
Real	T1-weighted	750	9	N/A
	T2-weighted	7288	104	N/A
	T2-weighted-FLAIR	8500	117	2418
	PD-weighted	3851	9	N/A
Synthetic	T1-weighted	750	9	N/A
	T2-weighted	7280	104	N/A
	T2-weighted-FLAIR	8503	116	2418
	PD-weighted	3860	9	N/A

version 0.45.27) software was used to generate synthetic qualitative images. imaging-specific parameters for all synthetic scans are shown in Table 3.1.
Images were all made on the same 1.5 T MRI system (Signa Artist, GE Healthcare, Milwaukee, WI, USA). All images had a voxel resolution of 0.61 x 0.61 x 5 mm^3.

3.2.2. Tumour Patient Data

Part of the data from the RIGEL (Radiotherapy in IDH mutated Glioma: Evaluation of Late outcomes) study (Nederlands Trial Register, NL7993) was used as a second dataset. The used data consisted of scans from 7 glioma patients which were scanned in the exact same manner as the dataset of healthy volunteers. The data of each patient consisted of quantitative maps and synthetic qualitative images that were acquired with the same parameters as mentioned in Table 3.1.

3.2.3. Real Healthy Volunteer Data

An additional dataset was acquired from two healty volunteers under the HARPS project. For each volunteer, we acquired real wMRI scans and calculated two sets of synthetic wMRI scans. The two sets of synthetic scans differed in the choice of imaging-specific parameters. One set had the same parameters as the real scans, while the other set had the same parameters as Table 3.1. The imaging-specific parameters for the real scans and synthetic scans with the same parameters are shown in Table 3.2. For the real scans, Phased array Uniformity Enhancement (PURE) was used in order to do a bias field correction.

3.3. Data Processing

The raw DICOM data files were first converted to the NIfTI filetype using dcm2niix (version 1.0.20210317) for easier analysis. Subsequently, brain masks were generated using the HD-BET software (47). For consistency, the mask of the PD-weighted MRI was used to mask all scans for every subject.

For the scans with tumours, we created tumour masks using an algorithm by Van der Voort *et al.* (48). We used synthetically generated T2w, PD-weighted (PDw) and T2-weighted-FLAIR (T2w-FLAIR) scans as input to generate tumour masks. The algorithm also expected a post-contrast T1w scan but as this was not available for our dataset, a pre-contrast T1w scan was substituted in its place. During evaluation, we evaluated the model on the full brain and on the tumour mask.

After masking, all scans were cropped to the dimensions of the largest brain mask in the dataset. As a result, we ended up with scans of the same size with a minimal amount of background pixels.

Masked and cropped weighted MR images were normalised using Prognosais. We rescaled the image intensity range from 0.01 to 1. For 2D models, inputs were made by splitting the processed 3D NIfTI files per slice and removing slices where the brain masks had a largest connected area of less than 400 mm^2. This essentially discarded empty slices and slices containing tiny segmentations from the data (A threshold of 400 mm^2 was chosen to ensure that every image contained a reasonable amount of brain tissue). This also allowed every 2D slice to be a data instance for the model. Figure A.1 shows processed T1w slices for an example patient.

For 3D models, after doing the same discarding step, not all processed 3D images had the exact same amount of slices. Images with less slices than the image with the maximum number of slices were padded with zeros to ensure they had the same size. In this fashion, for the dataset of healthy volunteers, we arrived at 97 data instances for 3D models, of which 74 were used for training, 14 were used for validation and 9 were used for testing. For 2D models, scans were divided per patient, which lead to 2607 data instances, of which 1743 were used for training, 653 were used for validation and 211 were used for testing. All scans of the same patient were either in the train, validation, or test set. All other datasets were preprocessed in a similar fashion, resulting in 370 2D slices for the synthetic weighted scans of brain tumours and 60 2D slices for the real weighted scans.

FSL-FAST (FMRIB's Automated Segmentation Tool, version 5.0) (49) was used in order to segment white matter, gray matter, and cerebrospinal fluid to be used in evaluation.

3.4. Model Architectures

Multiple model architectures were studied. We focused on FCNs as these are easy to implement and have shown to lead to satisfactory results in qMRI generation (20; 44). All models used a dense layer as the final prediction layer, shaping our problem into a regression problem.

3.4.1. Regular U-Net

The U-Net by Ronneberger *et al.* (31), is a model architecture that has shown to perform well on many image-related tasks. More information about the U-Net can be found in Section 2.2.1. Our U-Net model is shown in Figure 3.3.

3.4.2. U-ResNet

A second model architecture we investigated was a U-Net model with additional residual layers, as made famous in He *et al.* (50). This can result in more accurate predictions as the model is able to learn identity mappings (Section 2.2.1). Our ResNet model is shown in Figure 3.4.

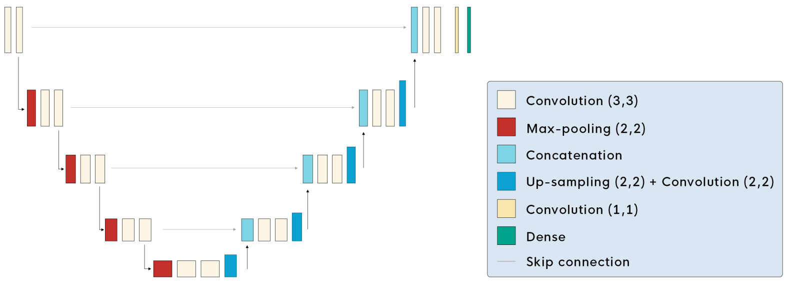

Figure 3.3: U-Net architecture similar to Ronneberger *et al.* (31). The model consists of downsampling and upsampling paths with added skip connections. Dropout was used after the convolutional layers in the downsampling path and before convolutional layers in the upsampling path.

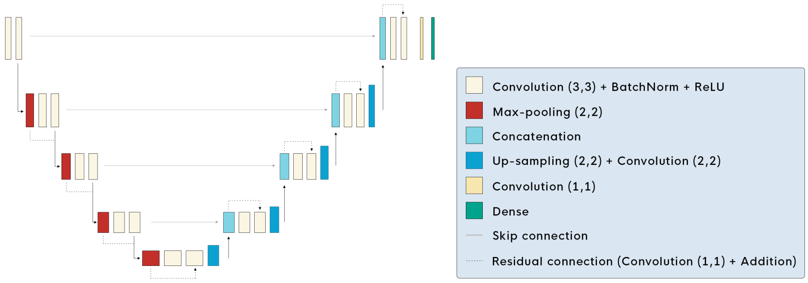
Convolution (3,3) + BatchNorm + ReLU
Max-pooling (2,2)
Concatenation
Up-sampling (2,2) + Convolution (2,2)
Convolution (1,1)
Dense
Skip connection
Residual connection (Convolution (1,1) + Addition)

Figure 3.4: U-ResNet architecture. U-Net with additional residual layers as used in He *et al.* (50). Dropout was used after the convolutional layers in the downsampling path and before convolutional layers in the upsampling path.

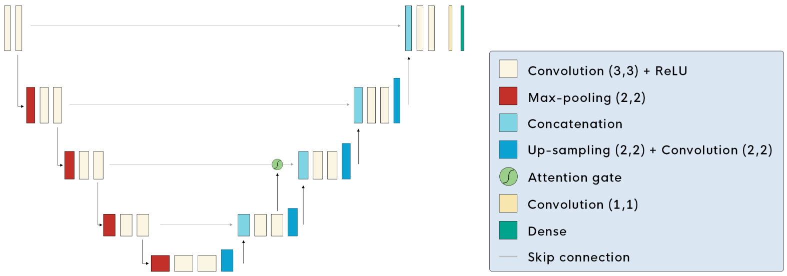

Figure 3.5: U-AttenNet architecture. U-Net architecture with attention gating. The attention gate was inspired by Oktay *et al.* (40) and Wang *et al.* (39). Dropout was used after the convolutional layers in the downsampling path and before convolutional layers in the upsampling path.

3.4.3. U-AttenNet

A third model architecture we investigated was a similar U-Net model with an additional attention gate. We took inspiration from Oktay *et al.* (40) and Wang *et al.* (39) in order to construct a model with a single additive attention gate. This can achieve improved performance due to the fact that the attention gate aims to force the network to use only relevant information from the skip connection (Section 2.2.3). This then makes it easier for the model to learn the accurate representation of the input-output relationship and predict correct outputs. Our attention model is shown in Figure 3.5.

3.5. Model Implementation

3.5.1. Loss Functions

Multiple loss functions were deemed to be promising for accurate qMRI generation. The mean squared error (MSE) and the mean absolute error (MAE) are given by

$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2 \tag{3.1}$$

$$MAE = \frac{1}{N}\sum_{i}^{N}|(y_i - \hat{y}_i)| \tag{3.2}$$

where N is the amount of voxels in the sample, y_i is the ground-truth value in voxel i, and $\hat{y}_i$ is the predicted value in voxel i. In addition, a combination of both MSE and MAE was used. To broaden the amount of loss functions, normalised versions of the three previous loss functions were also studied. As an example, the normalised MAE (nMAE) is given by,

$$nMAE = \frac{1}{N}\sum_{i}^{N}\frac{|(y_i - \hat{y}_i)|}{y_i} \tag{3.3}$$

All loss functions were implemented as ’masked’ loss functions, meaning that only the part of the image representing the brain was used for the loss calculations. Errors on background pixels were disregarded since these did not contribute to the model learning meaningful information.

3.5.2. Evaluation

Evaluation of the model predictions was done by using the peak signal-to-noise ratio (PSNR), root-mean-square error (RMSE) and the structural similarity index (SSIM).

The PSNR is given by,

$$PSNR = 10\log_{10}\left(\frac{I_{max}^2}{MSE}\right) \tag{3.4}$$

where I_{max} is the maximum voxel value in the sample. Generally, a higher PSNR means a higher quality of the generated image.

The RMSE is given by,

$$RMSE = \sqrt{MSE} \tag{3.5}$$

The PSNR and RMSE both focus on the differences in voxel values. Distinctively, the SSIM tries to decompose the luminance, contrast, and structure in an image. These get compared between two images (51). In this fashion, the SSIM tries to quantify differences between images in a way a human would perceive them. The SSIM is given by,

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.6}$$

where x and y are the two images to be compared, μ_x and μ_y are the mean voxel intensities of x and y, σ_x^2 and σ_y^2 are the variances of x and y, and σ_{xy} is the covariance of x and y. C_1 and C_2 are small constants that are added in order to account for situations where $\mu_x^2 + \mu_y^2$ or $\sigma_x^2 + \sigma_y^2$ are very close to zero.

The PSNR and RMSE were customised to only take into account the voxel values inside the brain masks of the patient and not take into account the background pixels. Due to the difficulty in customising the SSIM algorithm for this purpose, we decided to use the whole image (brain and background) for the calculation of the SSIM.

3.6. Experimental Approaches

Our research was developed to investigate the feasibility of qMRI generation using DL. Since this is an elusive problem, initial experiments need to take place in a controlled environment where the chances of success are highest. Afterwards, if the initial experiments are positive, we can expose our model to more real-world-like situations to get an estimate of the actual performance and try to understand the workings of the model.

Considering this, we proposed the following experiments;

1. **Optimisation**

   In a controlled environment (using synthetic scans), multiple model architectures and loss functions will be explored to find the model that is best at generating qMRI. Each experiment will be repeated three times and the best result will be used to compare with other models.

2. **Gaining Knowledge**

   When we have a well-performing model, we can start to examine the fundamentals that our approach is based on. Experiments are:

   - Investigating the amount of information in the input images by studying the performance when reducing inputs. Here we repeat every experiment five times.
   - Investigating if there is a difference in performance between 2D and 3D models.

3. **Real-world Situations** In the clinic, healthy volunteers are not very important, but the attention needs to go to people with diseases or conditions. Patients with brain tumours are especially valuable to identify. As we only have access to a small amount of scans of tumour patients, we have decided to use these scans only for the testing, and not the training, of the model. For the training of the model we use the synthetic scans of healthy volunteers.

   Additionally, the perfect magnetic field that is used in the calculations of synthetic scans is not always valid. On real MRI scanners, imperfections in the magnetic field lead to slightly different weighted images. Validating a qMRI generation method on actual acquired weighted MRI scans is therefore vital. As we also only have access to a small amount of real wMRI scans, we will follow a similar approach as the previous experiments. Therefore, we train our models on synthetic scans of healthy volunteers and test them on real scans of healthy volunteers.

4 Results

In this chapter we present the findings of the different experiments. In the subsequent sections, we show the best performing models, how much information is contained in the input images by training models with different combinations of wMRI scans, and to what extent there is a difference between models using 2D and 3D inputs. Finally, Section 4.4 & 4.5 show the performances of the best models on synthetic wMRI scans of tumour patients and real wMRI scans, respectively.

4.1. Best Performing Models and Parameters

In initial experiments we constructed models that used synthetic T1w, T2w, PDw and T2w-FLAIR images as input, and only predicted a T1-map as output. While doing so, we varied the model architecture, loss function and the learning rate. All models were trained with a batch size of 8 for 75 epochs, which took roughly 2,5 hours on a RTX 2080 Ti GPU.

4.1.1. Single-Output Models

Table 4.1 shows the results for the top 10 models that generated a T1-map. Here we see that the U-AttenNet performed best, if accompanied by a MSE loss and a learning rate of 0.001. The resulting RMSE was 91.6 ± 32.9 ms. Comparisons of the model prediction of the best model and the groundtruth T1-map can be seen for an example slice in Figure 4.1. In the difference map (Figure 4.1C), it can be seen that the model

Table 4.1: Evaluation metrics for models predicting a T1-map from synthetic input scans. The 10 best performing models are shown with their loss functions and learning rates, together with the evaluation metrics on the test set. Best model indicated in bold. Arrows indicate if a metric is desired to be high or low.

Model	**Loss**	**LR**	**PSNR (dB) ↑**	**SSIM ↑**	**RMSE (ms) ↓**
U-Net	MAE	0.01	26.667 ± 2.905	0.982 ± 0.012	210.928 ± 77.146
U-Net	MAE	0.001	28.079 ± 3.351	0.989 ± 0.008	183.413 ± 81.031
U-Net	MAE	0.0001	23.103 ± 2.914	0.984 ± 0.010	319.361 ± 127.254
U-Net	MAE+MSE	0.0001	23.241 ± 2.693	0.983 ± 0.011	311.444 ± 113.079
U-Net	nMAE+nMSE	0.001	27.832 ± 3.206	0.986 ± 0.009	187.445 ± 78.734
U-ResNet	MAE	0.01	26.855 ± 2.453	0.990 ± 0.006	203.182 ± 62.634
U-ResNet	MAE+MSE	0.01	27.030 ± 2.159	0.982 ± 0.012	197.256 ± 53.506
U-AttenNet	**MSE**	**0.001**	**33.837 ± 2.598**	**0.996 ± 0.003**	**91.645 ± 32.863**
U-AttenNet	nMAE	0.01	26.325 ± 3.017	0.979 ± 0.013	220.667 ± 85.593
U-AttenNet	nMAE	0.001	26.326 ± 2.902	0.983 ± 0.011	219.701 ± 80.59

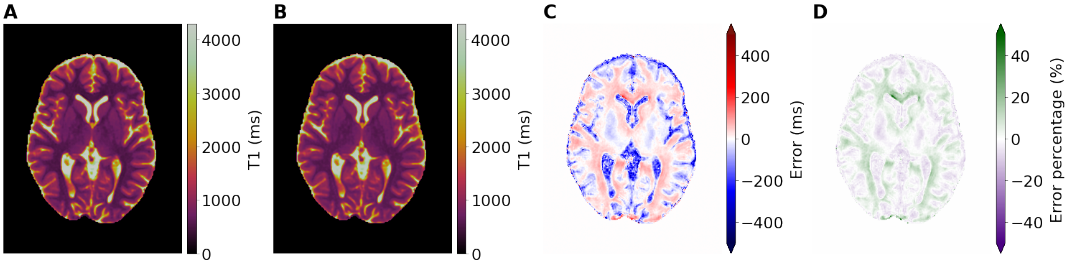

Figure 4.1: Visual performance comparison for model predicting T1-map. Difference values are cut-off at ±500 ms for improved visualisation. Error percentages are cut-off at 50% for improved visualisation.**A**: Groundtruth. **B**: Model prediction. **C**: Difference in T1 between prediction and groundtruth. **D**: Error percentage map showing the percentual error relative to the true T1 values.

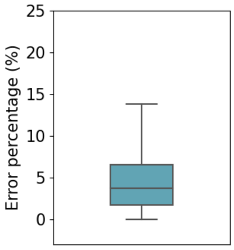

Figure 4.2: Boxplot of error percentages for best model generating a T1-map. The image shows the distribution of the error percentages that we acquire after comparing the model prediction and groundtruth of all images in the test set.

overpredicts (red regions) and underpredicts (blue regions) different regions of the T1-map. These regions seem to coincide with different tissues in the brain. Mainly regions with high T1 values like the CSF are underpredicted while white matter is overpredicted. These former errors are less pronounced when examining the relative error (Figure 4.1D), due to the large T1 value of CSF.

When we use the best model to calculate the percentual error (Figure 4.1D) of every sample in the test-set and create a boxplot showing the distribution of all error percentages, we find that the median error of the best performing model is 3.75% (Figure 4.2).

4.1.2. Multi-Output Models

Subsequently, we investigated models that generated all three quantitative mappings (T1, T2 and PD) from synthetic T1w, T2w, PDw and T2w-FLAIR images, as this was our original goal. Table 4.2 shows the performances for the best three models. The results show that there is no single model that can achieve the best performance for all three quantitative maps. The lowest error on the T1-map was similar to the performance of models with only a T1-map output. We also see that the U-AttenNet model with MAE loss and a learning rate of 0.001 seems to perform the best overall at generating multiple quantitative maps.

Additionally, we investigated the visual performances of the models predicting multiple quantitative maps. Figure 4.3 shows the comparisons of the best overall model's predictions and the groundtruth quantitative maps for a representative slice.

Figure 4.4 shows the distribution of the percentual error for all samples in the test set. Here we again see the differences between the error of different quantitative maps for the same model.

Table 4.2: Evaluation metrics for models predicting multiple quantitative maps from synthetic input scans. The 3 best performing models are shown with their loss functions and learning rates, together with the evaluation metrics on the test set. Best results indicated in bold. Arrows indicate if a metric is desired to be high or low. [1]PD values not in ms but in a.u.

Model	Loss	LR	Map	PSNR (dB) ↑	SSIM ↑	RMSE (ms)[1] ↓
U-Net	nMAE	0.0001	T1	20.571 ± 2.411	0.956 ± 0.026	418.956 ± 132.547
			T2	22.795 ± 3.442	0.975 ± 0.016	27.403 ± 12.102
			PD	**28.710 ± 3.748**	**0.980 ± 0.018**	**5.454 ± 2.800**
U-AttenNet	MAE	0.001	T1	32.660 ± 2.781	0.995 ± 0.004	105.635 ± 40.695
			T2	**25.788 ± 3.011**	**0.973 ± 0.020**	**18.978 ± 7.284**
			PD	26.620 ± 2.998	0.957 ± 0.032	6.672 ± 2.373
U-AttenNet	MSE	0.001	T1	**34.242 ± 2.784**	**0.996 ± 0.003**	**88.194 ± 34.686**
			T2	20.750 ± 2.659	0.938 ± 0.040	33.528 ± 11.657
			PD	23.192 ± 2.623	0.928 ± 0.049	9.779 ± 2.832

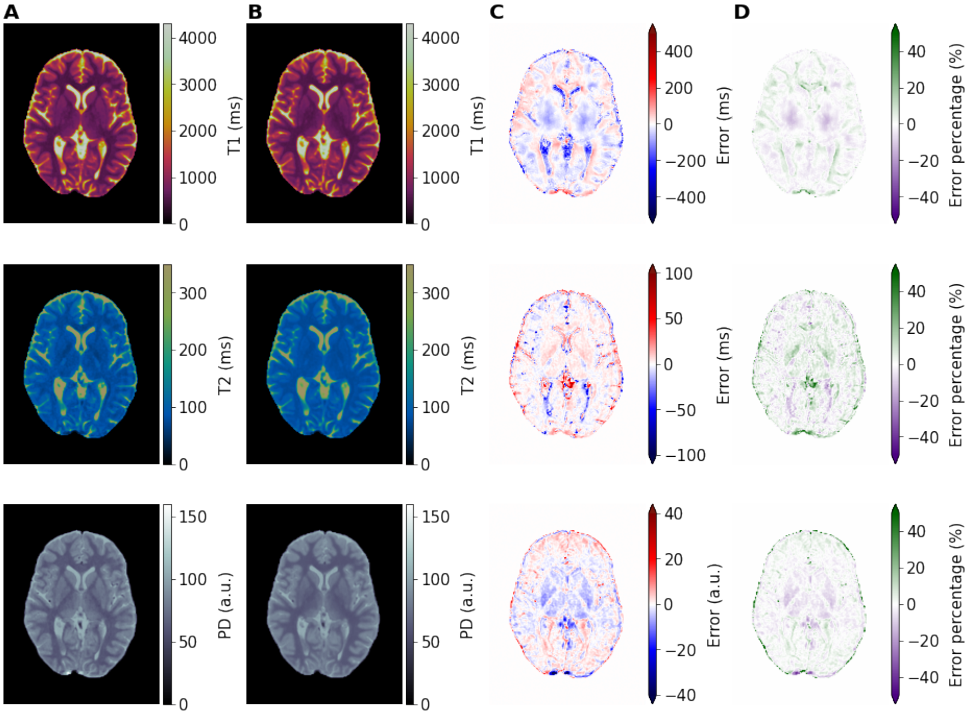

Figure 4.3: Visual performance comparison for models predicting multiple quantitative maps. From top to bottom, T1-, T2- and PD-map predictions. Difference values are cut-off at different values for improved visualisation. Error percentages are cut-off at 50% for improved visualisation. **A**: Groundtruth quantitative map. **B**: Model prediction. **C**: Difference between prediction and groundtruth. **D**: Error percentage map showing the percentual error relative to the true quantitative values.

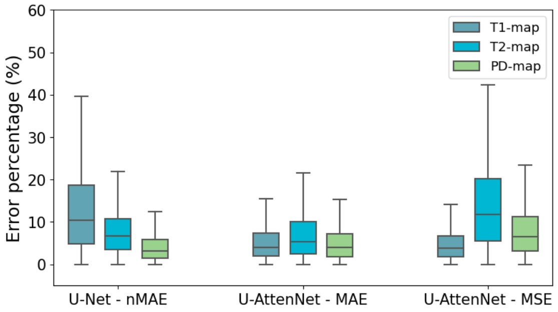

Figure 4.4: Boxplots of error percentages for best models generating multiple quantitative maps. The image shows the distribution of the error percentages that we acquire after comparing the model prediction and groundtruth of all images in the test set for the three best models.

4.2. Effect of Varying the Amount of Input Scans

Next, we varied the amount of input images for the best performing models of the previous experiment. All models here were trained for 200 epochs, which took roughly 5,5 hours on a RTX 2080 Ti GPU.

4.2.1. Single-Output Models

Figure 4.5 shows the output of models predicting a T1-map, compared to the groundtruth for an example slice. Models that did not have access to all weighted images still show reasonable predictions, however, performances seem to get better when more input images are used. We also see that different models underpredict and overpredict the T1 values on different locations. This can be due to the different input data that the models have, or due to a different optimisation path that the models took.

Table 4.3 gives the evaluation metrics for the same models, showing that a model with all weighted images as input can perform better than models with less weighted images. Nevertheless, the difference between the two best performing models is small.

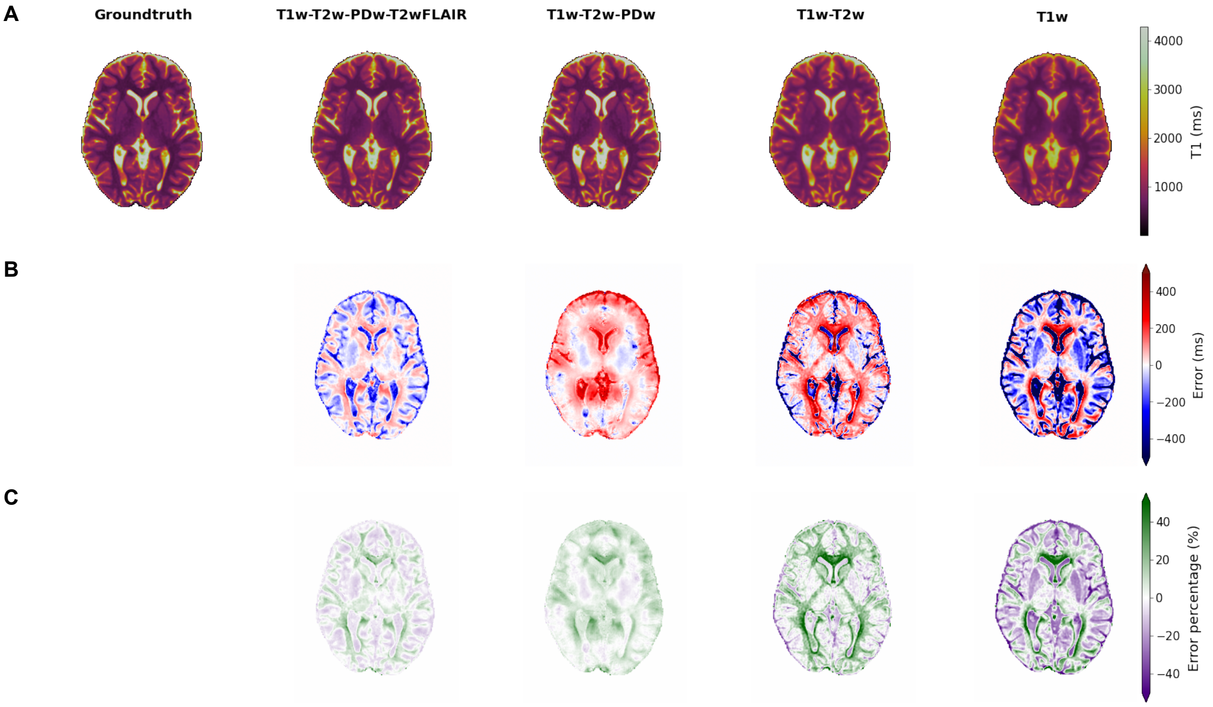

Figure 4.5: Effect of input images on performance for the generation of a T1-map. **A**: Model prediction of T1-map for different amounts of model inputs. **B**: Error maps showing the difference in T1 between prediction and groundtruth for different amounts of model inputs. Error values are cut-off at ±500 ms for improved visualisation. **C**: Error percentage maps showing the difference in T1 between prediction and groundtruth as a percentage of the groundtruth T1 value for different combinations of model inputs. Error percentages are cut-off at ±50% for improved visualisation.

Table 4.3: Evaluation metrics of models predicting a T1-map using different inputs. Best performances are indicated in bold. Arrows indicate if a metric is desired to be high or low.

Input	PSNR (dB) ↑	SSIM ↑	RMSE (ms)
All	**30.233 ± 2.693**	0.992 ± 0.005	**139.376 ± 51.085**
T1w, T2w, PDw	29.859 ± 3.306	**0.995 ± 0.004**	149.046 ± 64.785
T1w, T2w	23.734 ± 2.446	0.970 ± 0.019	290.583 ± 85.656
T1w	16.049 ± 2.192	0.899 ± 0.064	699.075 ± 188.851

4.2.2. Multi-Output Models

The same experiments were done for models generating multiple quantitative maps. Figure 4.6 shows the model predictions compared to the groundtruth. Difference maps and error percentage maps can be seen in Figures A.3 & A.4. Evaluation metrics are shown in Table 4.4. Here we see that, again, overall model performances increase when using more input data.

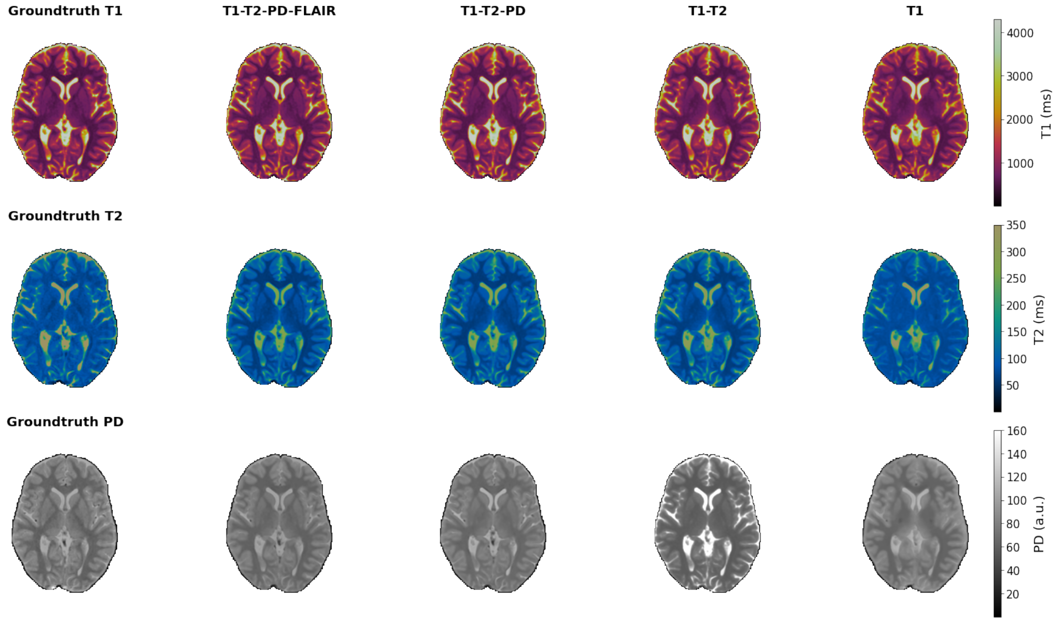

Figure 4.6: Effect of input images on performance for the generation of multiple quantitative maps. The figure shows model prediction of T1-, T2-, and PD-maps for models with access to different combinations of the input data (left to right).

Table 4.4: Evaluation metrics of models predicting multiple quantitative maps using different inputs. Input images and evaluation metrics on the test set are shown. Best results indicated in bold. Arrows indicate if a metric is desired to be high or low. [1]PD values not in ms but in a.u.

Input	Map	PSNR (dB) ↑	SSIM ↑	RMSE (ms)[1] ↓
All	T1	**33.385 ± 3.338**	**0.996 ± 0.003**	**100.104 ± 48.480**
	T2	20.114 ± 2.878	0.936 ± 0.039	36.471 ± 14.092
	PD	24.243 ± 2.943	0.947 ± 0.038	8.834 ± 3.737
T1w, T2w, PDw	T1	31.146 ± 3.401	0.995 ± 0.004	129.150 ± 58.220
	T2	20.261 ± 2.861	0.936 ± 0.040	35.807 ± 13.627
	PD	**25.196 ± 2.769**	**0.950 ± 0.037**	**7.863 ± 2.654**
T1w, T2w	T1	27.074 ± 3.271	0.985 ± 0.011	204.572 ± 84.482
	T2	**21.690 ± 2.656**	**0.957 ± 0.025**	**30.051 ± 10.078**
	PD	14.959 ± 3.005	0.868 ± 0.081	25.486 ± 8.430
T1w	T1	22.208 ± 2.603	0.971 ± 0.016	348.181 ± 107.922
	T2	19.071 ± 2.709	0.939 ± 0.033	40.871 ± 14.631
	PD	22.475 ± 3.230	0.913 ± 0.056	10.846 ± 3.768

4.3. Effect of Input Dimensionality

Up until now, we have investigated the performances of models that use 2D slices as input. The following experiments consider models that use 3D input image data. For these models we used a batch size of 2. We trained models for 500-1000 epochs, which took 1.5-3 days on a RTX 2080 Ti GPU. We experimented with a 3D version of the best performing 2D model for predicting T1-maps. While doing so, we investigated multiple learning rates in order to find the best performing model. Table 4.5 shows the evaluation metrics of the models. From this we see that the model with the lowest learning rate performed best.

When we compare this performance with the performance of the best 2D model, we find very small differences (Table 4.6). The mean performance of the 2D model seems to be slightly better, but the standard deviation of the results from the 3D models is smaller. This could be due to the fact that we also evaluate on full 3D scans.

Similar results were also found when comparing 2D and 3D models that predicted multiple outputs (Table A.1).

Table 4.5: Evaluation metrics for 3D models predicting a T1-map. Best performing models and evaluation metrics in bold.

Model	**Loss**	**LR**	**PSNR (dB) ↑**	**SSIM ↑**	**RMSE (ms) ↓**
U-AttenNet	MSE	0.01	10.683 ± 1.112	0.588 ± 0.106	1266.985 ± 154.491
U-AttenNet	MSE	0.001	9.185 ± 0.907	0.524 ± 0.116	1501.435 ± 151.328
U-AttenNet	**MSE**	**0.0001**	**33.475 ± 1.528**	**0.997 ± 0.001**	**92.726 ± 19.181**

Table 4.6: Comparison of evaluation metrics for 2D and 3D models predicting a T1-map. Best evaluation metrics shown in bold. Arrows indicate if a metric is desired to be high or low.

	Best 2D Model	Best 3D model
PSNR (dB) ↑	**33.837 ± 2.598**	33.475 ± 1.528
SSIM ↑	0.996 ± 0.003	**0.997 ± 0.001**
RMSE (ms) ↓	**91.645 ± 32.863**	92.726 ± 19.181
Median Error	**3.75%**	4.26%

4.4. Performance on Scans with Pathologies

4.4.1. Single-Output Models

We used the best performing model from Section 4.1.1 (taking synthetic T1w, T2w, PDw and T2w-FLAIR images as input) on the additional dataset of brain tumours to predict T1-maps. Figure 4.7 shows a visual comparison of the model prediction and groundtruth for a single slice. Here, we see that our model is good at predicting the healthy tissue and the tumour core. However, around the tumour core some tissue exists where our model prediction leads to a higher error. Table 4.7 gives performance metrics for the model performance on the whole dataset. It shows that the performance is better when we only evaluate on the tumour. Figure 4.8 shows the distribution of error percentages for the model evaluated on the full brains, healthy brains and on the tumours. Here, minimal differences between evaluating on the healthy or full brain, and tumour are visible.

After the experiments from Section 4.1.1, we tried to improve on the performance of the best model generating a T1-map from synthetic scans of healthy volunteers. We were able to train a model that reached a lower error on the test set than the lowest error shown in Table 4.1, by training this model for a longer time. This model thus performed better on synthetic scans than the model used in the previous paragraph. However, when we used this model to predict the T1-maps from the dataset of synthetic scans of brain tumours, we found a larger error than we reached in Table 4.7. Figure A.5 shows a visual comparison between the model prediction and groundtruth T1-map for an example slice. This lack of generalisability shows that improving the performance on the synthetic scans of healthy volunteers too much can result in overfitting and a worse performance on other datasets.

4.4.2. Multi-Output Models

To predict multiple quantitative maps, we used the overall best model from Section 4.1.2. Results for a single slice can be seen below in Figure 4.9. Here we see that errors are, again, mainly present around the tumour core. The absolute and relative errors are high for predictions made on the T2-map. Table 4.8 shows the evaluation metrics on the full dataset. Here we see an odd behaviour of the SSIM increasing for T1- and T2-maps when only considering the tumour. This can be explained by the fact that the SSIM was not calculated on only the predicted pixels, but also takes the background pixels into account. More background pixels (in the case of selecting only the tumour) results in a higher value for the SSIM. In this scenario the SSIM loses its objectivity. Finally, Figure 4.10 shows the distribution of the errors on the predictions.

Table 4.7: Evaluation metrics for T1-map predictions on tumour scans. Arrows indicate if a metric is desired to be high or low. Best metrics shown in bold.

	PSNR (dB) ↑	SSIM ↑	RMSE (ms) ↓
Full brain	30.566 ± 2.353	**0.992 ± 0.004**	132.449 ± 43.036
Tumour	**33.261 ± 4.526**	0.991 ± 0.004	**106.949 ± 59.533**
Healthy brain	30.584 ± 2.335	**0.992 ± 0.004**	132.110 ± 42.799

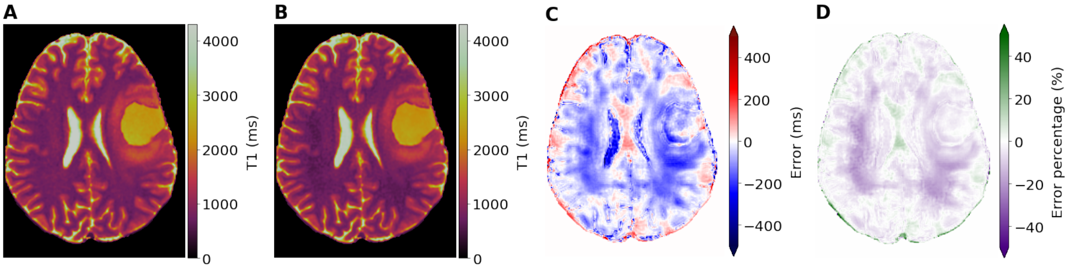

Figure 4.7: Model prediction on tumour data for T1-map. Difference values are cut-off at ±500 ms for improved visualisation. Error percentages are cut-off at 50% for improved visualisation. **A**: Groundtruth T1-map. **B**: Model prediction of T1-map. **C**: Difference in T1 between prediction and groundtruth. **D**: Error percentage map showing the percentual error relative to the true T1 values.

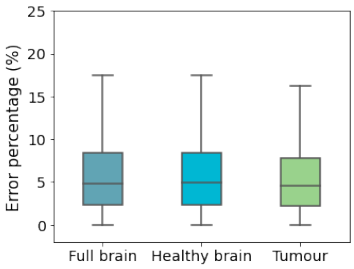

Figure 4.8: Boxplot showing the distribution of error percentages when predicing a T1-map for tumour scans. Error percentages are shown when evaluating on the whole brain, the healthy part of the brain and on the tumour.

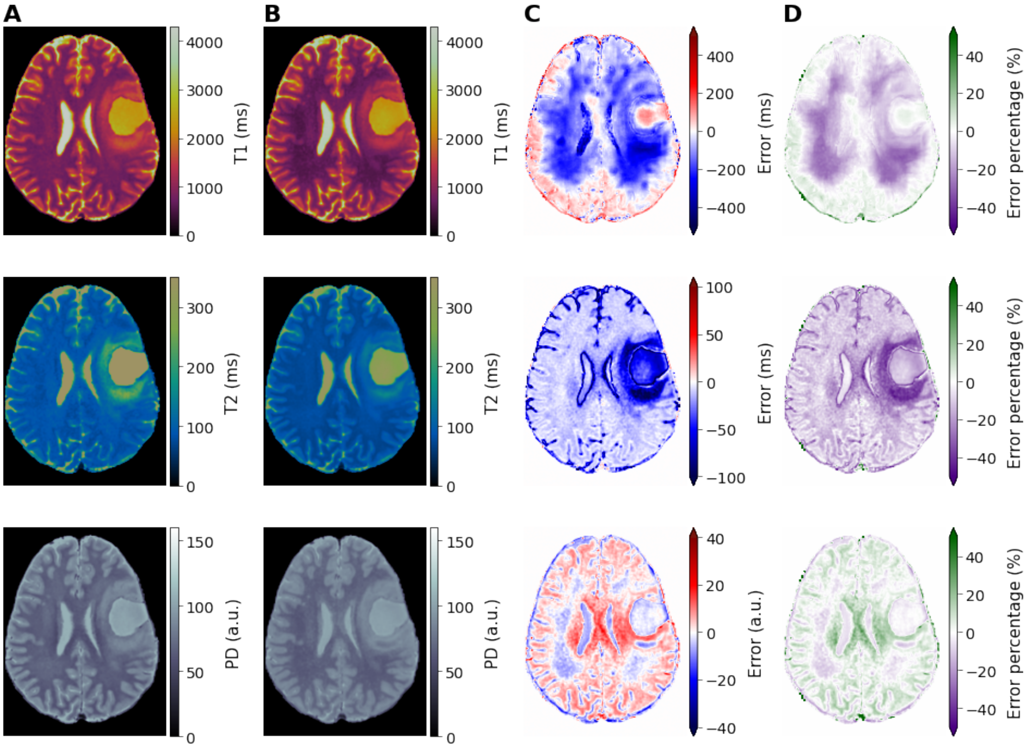

Figure 4.9: Model prediction on tumour data for multiple outputs. Difference values are cut-off at different values for improved visualisation. Error percentages are cut-off at 50% for improved visualisation. From top to bottom: T1-, T2-, and PD-maps. **A**: Groundtruth quantitative map. **B**: Model prediction of the quantitative map. **C**: Difference between prediction and groundtruth. **D**: Error percentage map showing the percentual error relative to the true quantitative values.

Table 4.8: Evaluation metrics for multiple output predictions on tumour scans. Arrows indicate if a metric is desired to be high or low. Best metrics shown in bold. [1]PD values not in ms but in a.u.

	Map	PSNR (dB) ↑	SSIM ↑	RMSE (ms)[1] ↓
Full brain	T1	29.386 ± 2.343	0.991 ± 0.005	151.565 ± 47.389
	T2	22.379 ± 2.018	0.954 ± 0.025	27.359 ± 6.666
	PD	**24.733 ± 2.212**	0.940 ± 0.032	7.678 ± 1.786
Tumour	T1	27.666 ± 4.526	**0.999 ± 0.001**	160.149 ± 93.390
	T2	20.152 ± 3.255	**0.994 ± 0.006**	32.368 ± 13.820
	PD	24.693 ± 2.400	**0.995 ± 0.005**	**2.010 ± 5.397**
Healthy brain	T1	**29.534 ± 2.304**	0.991 ± 0.005	**148.882 ± 46.531**
	T2	**22.573 ± 2.000**	0.954 ± 0.025	**26.750 ± 6.587**
	PD	24.704 ± 2.195	0.940 ± 0.032	7.701 ± 1.775

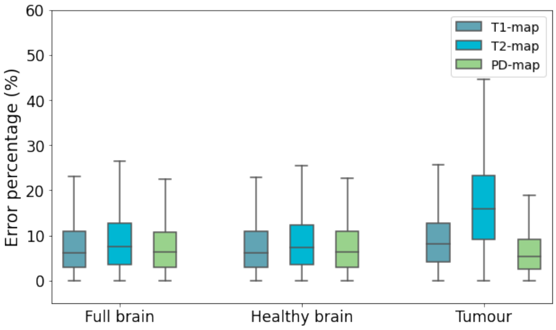

Figure 4.10: Boxplot showing the distribution of error percentages when predicting multiple quantitative maps for tumour scans. Error percentages are shown when evaluating on the whole brain, the healthy part of the brain and on the tumour.

4.5. Performance on Real Scans

Due to a limited amount of real weighted scans, we used models that were trained on the synthetic data and tested these on real data. The differences between real and synthetic data are summarised in Figure A.6.

4.5.1. Single-Output Models

We use the same model as in the previous section to predict quantitative maps. Figure 4.11 compares the model performance for real and synthetic scans. We notice that our model yields poor predictions when predicting high T1-values, as those seen in CSF (Figure 4.11C & D). Figure 4.12 does the same as it shows the distribution of the different error percentages of the predictions. Here, we again see that our model performs poorly on real input scans, whereas the performance is good for synthetic input scans.

The difference in performance between models using synthetic scans with the same imaging-specific parameters as the training images and models using synthetic scans with the same imaging-specific parameters as real data are very small. All evaluation metrics are shown in Table A.2.

Subsequently, we performed more experiments where we used different models optimised for different combinations of the input images. Figure 4.13 shows the results of these experiments. From this we see that models that use T1w, T2w and

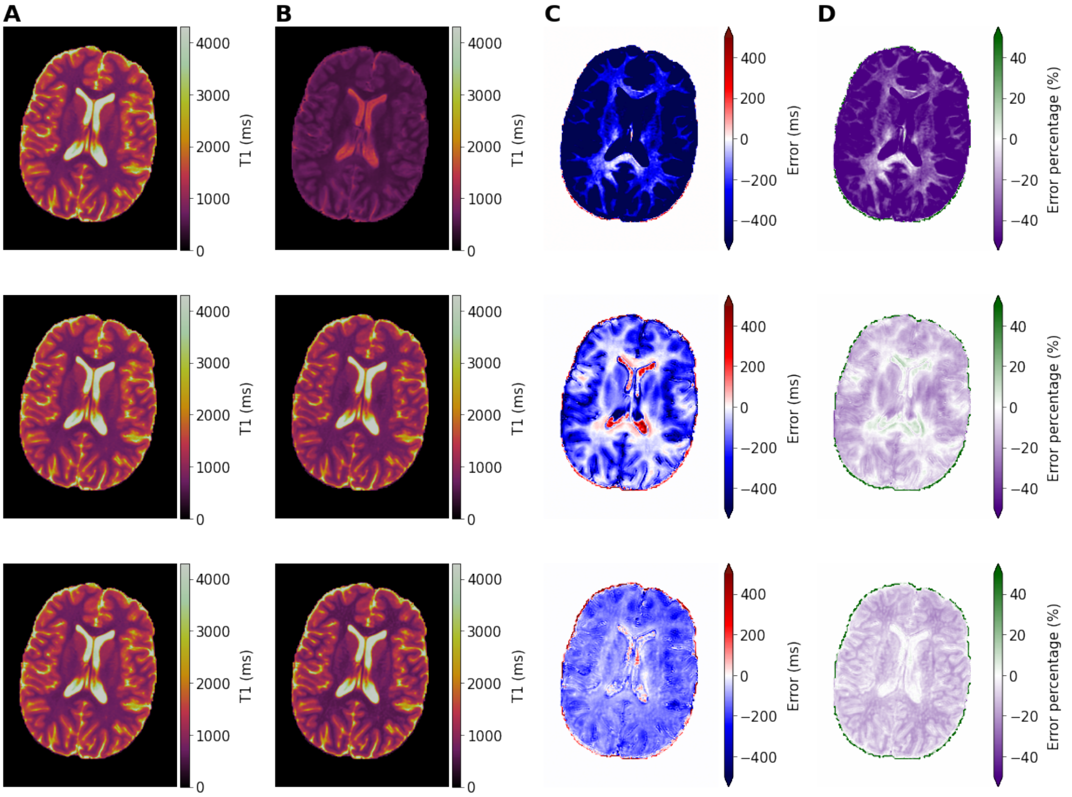

Figure 4.11: Model prediction on real and synthetic input scans for a T1-map. From top to bottom: Real data, synthetic data with the same imaging-specific parameters as real scans and synthetic data with the same imaging-specific parameters as images used for training. Difference values are cut-off at ±500 ms for improved visualisation. Error percentages are cut-off at 50% for improved visualisation. **A**: Groundtruth T1-map. **B**: Model prediction of the T1-map. **C**: Difference between prediction and groundtruth T1 values. **D**: Error percentage map showing the percentual error relative to the true T1 values.

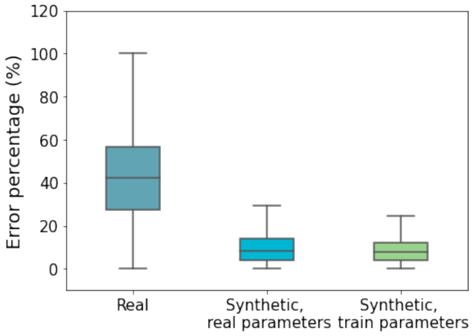

Figure 4.12: Boxplots showing the distribution of error percentages when predicting a T1-map from real and synthetic input scans. Synthetic scans either have the same imaging-specific parameters as real scans or the same imaging-specific parameters as scans used during training.

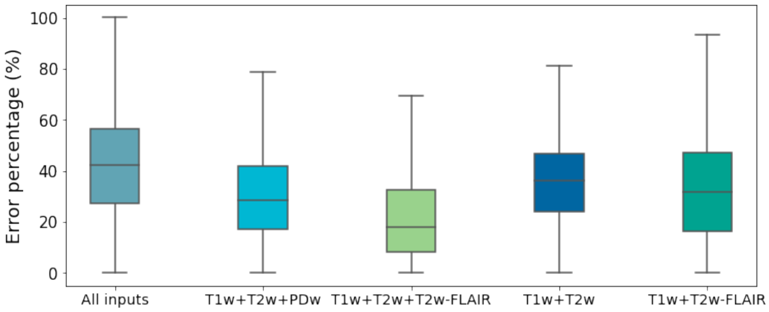

Figure 4.13: Boxplots showing the distribution of error percentages when predicting a T1-map from different combinations of real input scans.

T2w-FLAIR images as input actually perform better than models that use all the input data. This model reaches a median error of 17.9%. However, not all of these models reached the same performance during training, meaning that there are more variables influencing the results besides the amount of input data. Therefore, comparing the importance of the input images is difficult. Nevertheless, no models showed better results on synthetic data than the model using all inputs. Therefore there is definitely an increase in performance noticeable on real input data when removing inputs.

Also, we evaluated performances while using segmentations of white matter (WM), grey matter (GM) and cerebrospinal fluid (CSF). Table 4.9 shows the evaluation metrics per tissue for the model using all inputs and the model using only T1w, T2w and T2w-FLAIR scans as input. Here we see that the error on WM is the lowest, whereas both models have the largest error when predicting CSF.

Table 4.9: Evaluation metrics for different brain regions of models predicting a T1-map from real scans as input. WM = White matter, GM = Grey matter and CSF = Cerebrospinal fluid. Arrows indicate if a metric is desired to be high or low. Best metrics shown in bold.

Inputs	**Region**	**PSNR (dB) ↑**	**SSIM ↑**	**RMSE (ms) ↓**
All inputs	WM	17.362 ± 2.856	0.929 ± 0.044	591.834 ± 206.880
	GM	11.962 ± 2.186	0.909 ± 0.051	1109.795 ± 321.252
	CSF	8.285 ± 1.817	0.929 ± 0.030	1691.565 ± 332.930
T1w, T2w, T2w-FLAIR	WM	**20.391 ± 3.790**	**0.963 ± 0.018**	**426.571 ± 205.047**
	GM	16.194 ± 2.524	0.963 ± 0.021	674.135 ± 194.050
	CSF	11.890 ± 1.434	0.957 ± 0.016	1108.400 ± 174.085

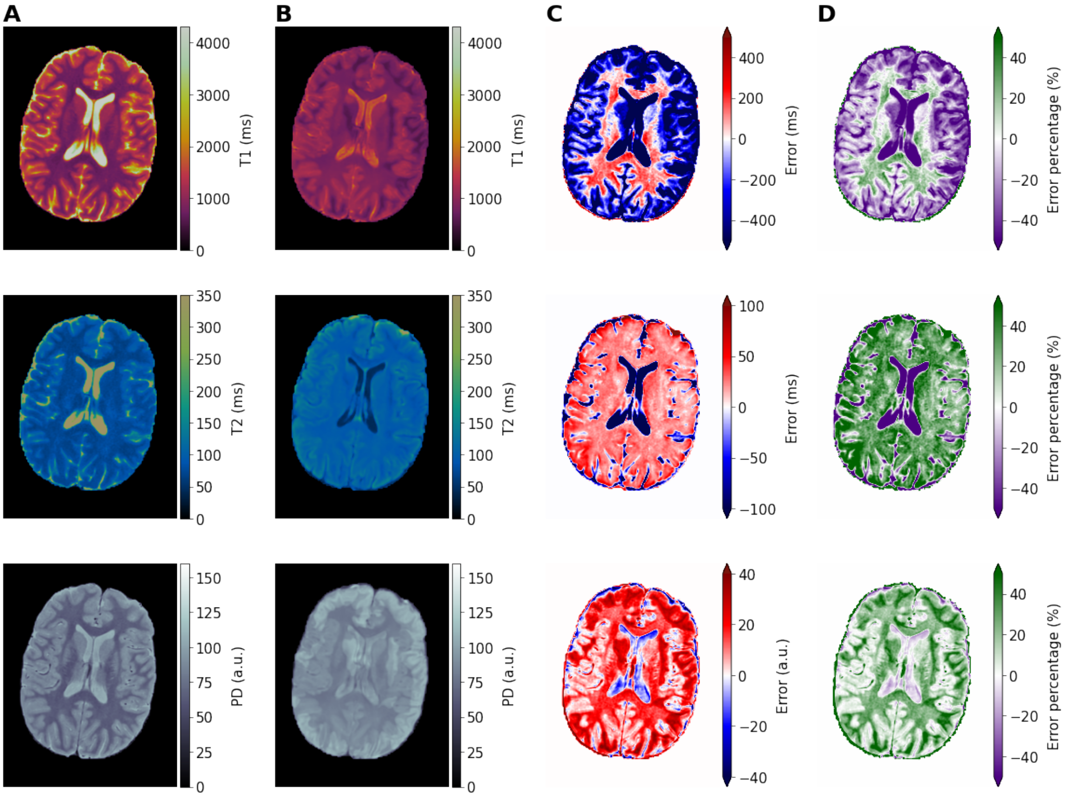

Figure 4.14: Model prediction on real input scans for multiple quantitative maps. From top to bottom: T1, T2- and PD-maps. Difference values are cut-off at different values for improved visualisation. Error percentages are cut-off at 50% for improved visualisation. **A**: Groundtruth quantitative map. **B**: Model prediction of the quantitative map. **C**: Difference between prediction and groundtruth quantitative values. **D**: Error percentage map showing the percentual error relative to the true quantitative values.

4.5.2. Multi-Output Models

Figure 4.14 shows predictions for multiple quantitative maps. Especially high intense regions, like CSF, show a high error on T1- and T2-maps. This is similar to the results of models generating a T1-map. Evaluation metrics are shown in Table A.4. The distributions of error percentages for the prediction of multiple quantitative maps from both synthetic and real inputs are shown in Figure 4.15. A clear difference between performances on real and synthetic scans is again visible, similar to the results of experiments where only a T1-map was predicted.

The combination of T1w, T2w and T2w-FLAIR scans again gave the best performance for T1-maps, but T2-map predictions gave the lowest error when only T1w and T2w scans were used as input (Table A.5). Finally, models that use T1w, T2w and PDw input images reached the best performance when predicting PD-maps.

The model predicting all quantitative maps from T1w, T2w and PDw scans performs the best overall (Figure A.9). Figure A.8 shows the prediction of this model for an examples slice. When we also look at the error per tissue, we find similar results as in the case of models predicting a T1-map, namely that WM has the lowest error and CSF the highest error (Table A.6 & Figure A.10).

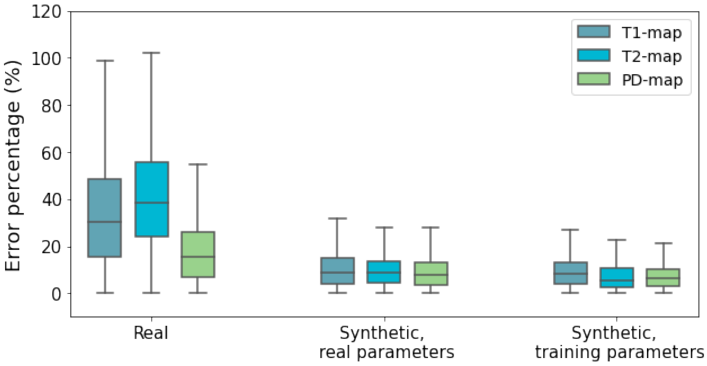

Figure 4.15: Boxplot showing the distribution of error percentages when predicting multiple quantitative maps from real and synthetic input scans. Synthetic scans either have the same imaging-specific parameters as real scans or the same imaging-specific parameters as scans used during training.

5 Discussion

This chapter reflects on the outcomes of our approach. The subsequent sections discuss the general ability of our models to generate qMRI from synthetic wMRI, the performances of models on synthetic tumour scans, and the performances of our models on real wMRI scans. Additionally, the final sections give limitations of our study, recommendations for future research and final conclusions from our results.

5.1. Generating qMRI from Synthetic wMRI

We found that generating qMRI from synthetic wMRI scans achieved a median error of 3.75% when only predicting a T1-map and median errors of 4.07%, 5.31% and 4.03% when simultaneously predicting T1-, T2- and PD-maps respectively (both results using all weighted scans as input). This shows that, i) multiple synthetic wMRI images with different weightings contain sufficient information to reconstruct quantitative maps and that, ii) a DL algorithm is capable of recovering the quantitative information that is stored in these synthetic wMRI scans. However, model performances were not stable and varied between repetitions of the training process.

Additionally, we saw that for models predicting multiple quantitative maps at the same time, the increased performance on the prediction of one quantitative map comes at a cost of a decreased performance on the prediction of another quantitative map, i.e. no single model was able to reach the lowest error on T1-, T2- *and* PD-maps. This means that there is a trade off between the performance and the amount of quantitative maps that can accurately be predicted. We take from this that a multi-output approach may not be the most useful in practice. Higher overall performances might be reachable when constructing a single model for every quantitative mapping and combining them.

For the case of generating solely a T1-map, we saw that T2w-FLAIR images contributed only minimally to the model performance as models without T2w-FLAIR images performed similar to models with access to all scans. Bigger performance differences were visible when PDw and T2w images were removed, strengthening the idea that they (together with T1w images) contain more T1 information and contribute the most to accurate quantitative map generation. A similar performance difference was found in models generating multiple quantitative maps, however, here we again observed that no single model performs the best at generating all three quantitative maps. The errors of the model with access to all data were very close to the lowest errors that we found over all models. Quite logically, we found that the qualitative image corresponding to the quantitative map of interest always needs to be present for the best performance, i.e. to be able to generate the best PD-map, a PD-weighted image has to be used as input.

We found a similar high performance with 3D models as with 2D models, This implies that we had enough data for our 3D model to learn from. Models using 3D data might be able to reach a lower error since they can learn more spatial dependencies. The lack of this improved performance in our results can be explained by the slice thickness of the MRI scans. Our data had a reasonably large slice thickness of 5 millimetres, leading to marked partial voluming in the z-direction. This could impede the model from retrieving correct, additional information from other slices since this information is averaged and lost.

5.2. Differences with Pathologies

After testing the performance of our previous models on synthetic scans with actual tumours, we saw that the prediction error in these cases was higher. We attribute this to tumour tissue behaving differently from healthy tissue in the sense that the relation between quantitative parameters is distorted. The highest errors were made in the tissue around the tumour core, which is a critical area for diagnosis and tumour localisation.

In preliminary research, Meng *et al.* (4) showed that mean T1 values of malignant tumours are 30% higher than mean T1 values of benign tumours. Smaller differences were found between T2 values (12.7%). The error of a model predicting the quantitative parameter values should at least be smaller than these pathological differences. In order for our model to be useful in clinics, errors close to those between conventionally acquired qMRI is desirable. These differences lay around 1-5% on average, when similar software and imaging sequences are used (5; 52).

Currently we reached a median T1 error of 4.5% on *synthetic* tumour scans. When predicting multiple outputs, the errors where much larger (median error of 15.8% on T2 values), showing that predictions need to be better in order to properly distinguish pathological differences in quantitative parameters.

5.3. Real Input Scans

Our experiments showed that the use of real wMRI input scans caused a considerable increase in the error when predicting quantitative maps. These differences are likely caused by assumptions within the calculation of synthetic MRI scans. These calculations assume a perfect magnetic field and do not take into account inhomogeneities in the B0 and B1 field. This leads to differences in contrast between real and synthetic scans. Since DL models are trained on input data that is similar and consistent, high errors can arise even when new inputs have relatively small differences compared to the training data.

Similar performances and errors were found for two different types of synthetic inputs (acquisition parameters as in real images versus acquisition parameters as used in training).

As the differences between the real and synthetic wMRI scans varied per image type, we decided to investigate the performances when using different combinations of wMRI scans as input. What we saw was that the performances increased when using less input images. When generating only a T1-map, the lowest median error of 17.9% was achieved through a model using T1w, T2w and T2w-FLAIR scans as input. This was significantly ($p<0.05$, T-test) less than the median error of 42.5% when using all input scans. Nevertheless, the error is still too large for clinical usage of quantitative map generation. Similar errors were found when generating multiple quantitative maps from T1w, T2w and PDw images; median errors of 27.4%, 12.0% and 8.7% for T1-, T2- and PD-maps respectively.

When we split the error per tissue type, we found that our models were best at predicting white matter, although on some occasions grey matter was predicted almost equally well. Our models produced poor predictions on CSF. This can be explained by the fact that in training this same behaviour was noticed. An additional source of this could be due to that a major part of the differences between real and synthetic scans was present in CSF. Real wMRI scans often *underestimate* the signal contribution in CSF. This happens because, to capture the full magnetic relaxatory behaviour in CSF, one needs to wait very long (CSF has long T1 and T2 times). Conventional MRI sequences do not wait this long and therefore underestimate the signal and image intensities.

Our best models generating a T1-map from real input scans reached median errors of 15.7%, 17.6% and 29.7% on white matter, grey matter and cerebrospinal fluid respectively. When generating multiple quantitative maps we achieved higher errors for T1-maps but lower errors for T2- and PD-maps.

5.4. Limitations of Our Study

The biggest limitation of our study is that we used synthetic wMRI scans as input to train our models. As we have discussed in the previous section, real wMRI scans are very different from synthetic wMRI scans even though the synthetic scans try to emulate their real counterpart. This shows itself in the higher error we have seen when predicting qMRI from real wMRI scans. To counter this problem, new research should focus on real wMRI scans and use these as data to train a model.

Additionally, we have limited our research by only using data from a single MR scanner. Using scans from multiple scanner vendors and locations will likely lead to a more generalisable model, since it has to account for slight differences in qualitative input values.

Furthermore, the MRI scans that we used consisted of very thick slices. With a larger slice thickness, the severity of partial volume effects can be increased as one pixel will reflect a signal being picked up from more types of tissues. This can impair the ability of our model to accurately predict quantitative parameters. This could influence both the 2D and 3D models.

Finally, we ran into problems when training our 3D models due to the memory of our computational sources. This restricted us to train models with a maximal batch size of two. Higher memory limits and batch sizes might facilitate improving the performance of these models.

Different DL models could still be explored in order to reach better performances. A method that has already been named in this thesis is the GAN. Although training of these models is often regarded as difficult, GANs might be able to reach higher performances than FCNs. To date, no studies regarding the ability of GANs to predict qMRI from wMRI have been done.

5.5. Conclusions

We can conclude that we succeeded in creating accurate qMRI generation models using synthetic input scans for both the prediction of a single T1-map and the prediction of T1-, T2- and PD-maps simultaneously.

Additionally, we found that the performances we attained on synthetic wMRI scans of healthy volunteers only partially translated to images of patients with brain tumours. This will give rise to errors that are too large for clinical use.

When predicting qMRI from real wMRI of healthy volunteers, we found increased errors. Therefore, in order to achieve higher performances on real scans and scans of patients with brain tumours, we need to use these scans as training data.

In conclusion, results on synthetic data are encouraging, however, training on real images remains crucial for accurate predictions.

6
Acknowledgements

This has been an extraordinary year. Partly because of the aftermath of the Covid-19 pandemic, but even more so because of the final part of my Masters degree in Biomedical Engineering; my thesis. This thesis marks and end to my academical journey, and to 2 years of further developing myself and learning about interesting things in the world of medical physics.

I would not have been able to finish my research without the help of my supervisors.

Sebastian, thanks for your guidance throughout the project. Being able to use Prognosais has been tremendously useful in setting up my experiments. You always find things that can be improved upon and you always reply quick. I'm glad I have been able to work with such a smart and driven person as you. Thanks for all the interesting discussions we have had about the project and beyond. I'm glad I still got to meet you in real life in the end.

Juan, your vast knowledge was very helpful when trying to interpret our results. Many thanks for acquiring the last few scans for my project so quickly, I think it added a lot. The addition of your humour to the meetings was always a pleasure. Thanks for your supervision.

Frans, thank you for the extra guidance and feedback, and for taking care of the organisational part of things. You were always reachable for a quick chat or update, and you have been throughout my whole Masters, which was very helpful.

Thanks to everyone at BIGR, in particular to the radiomics group where I heard interesting things about new research and also got feedback on my own project.

Lorenzo, Rocher, Elil, Ussama and Virgil. Thanks for your support during my studies. You guys better graduate soon. Let's celebrate with some drinks sometime!

Major thanks to my girlfriend Rebecca for putting up with me talking about MRI's and deep learning for so long. You must have heard about every experiment that failed and every bug I came across. It has been a hard time with Covid, but we are making it through. Thanks for being by my side.

Lastly I want to thank my parents for their unwavering support throughout my thesis and whole academic career.

A

Appendix

A.1. Data

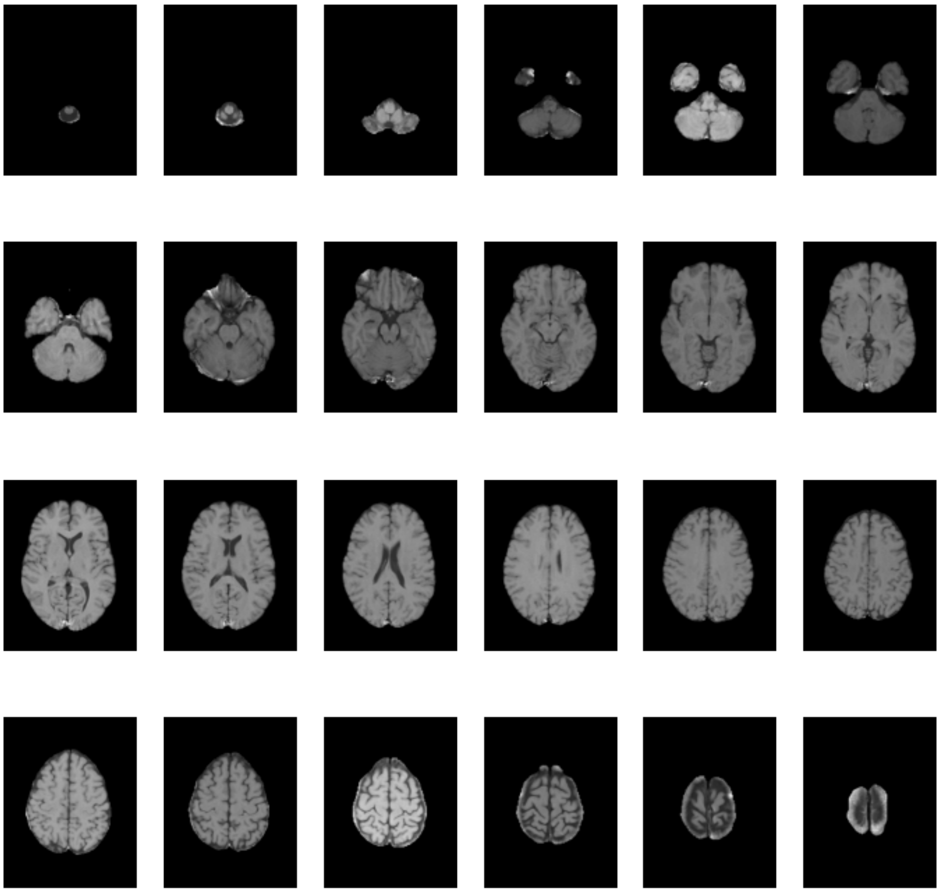

Figure A.1: Representative 2D synthetic T1-weighted scans for an example patient.

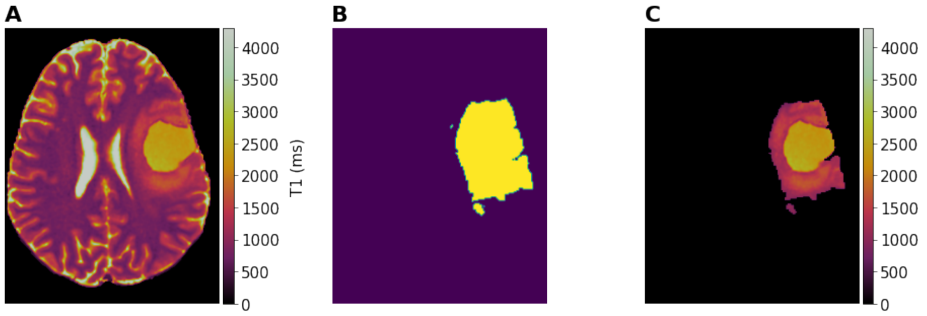

Figure A.2: Quantitative maps and corresponding tumour mask for one slice. **A:** Full T1-map. **B:** Tumour mask. **C:** Masked T1-map.

A.2. Experiments

A.2.1. Different Inputs

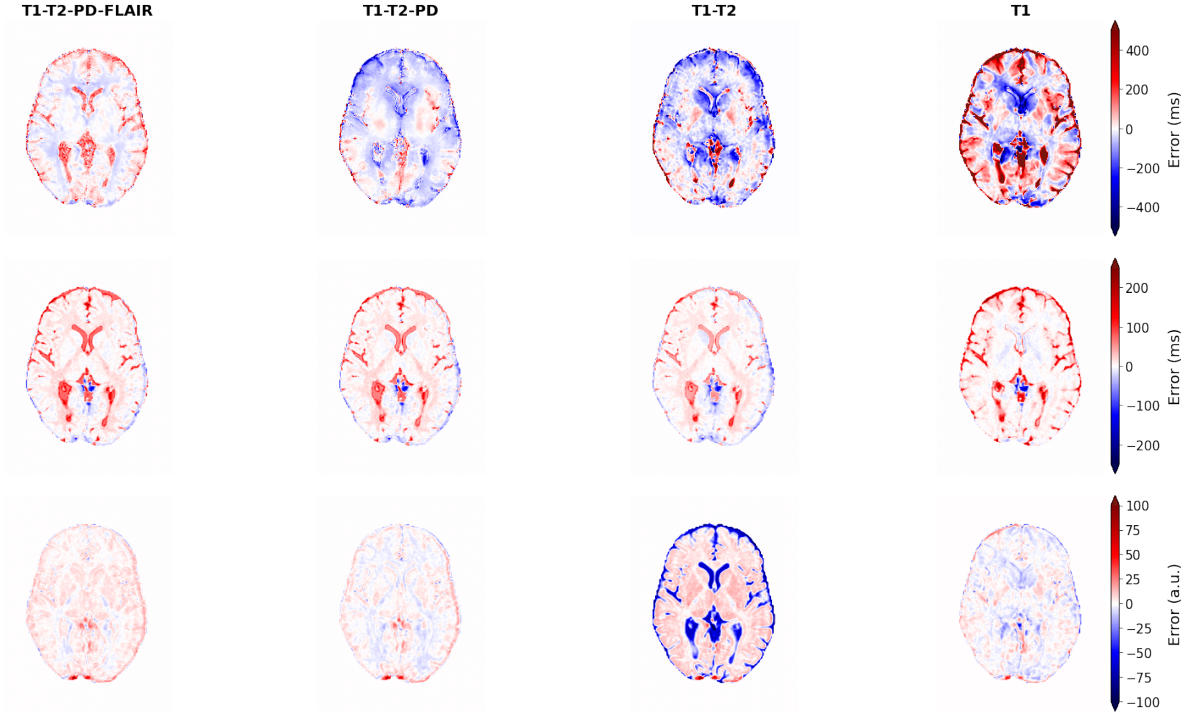

Figure A.3: Difference maps for models generating multiple quantitative maps using different combinations of input scans. From top to bottom: T1-maps, T2-maps and PD-maps. Different models are seen to overpredict (red regions) and underpredict (blue regions) the quantitative values in different regions.

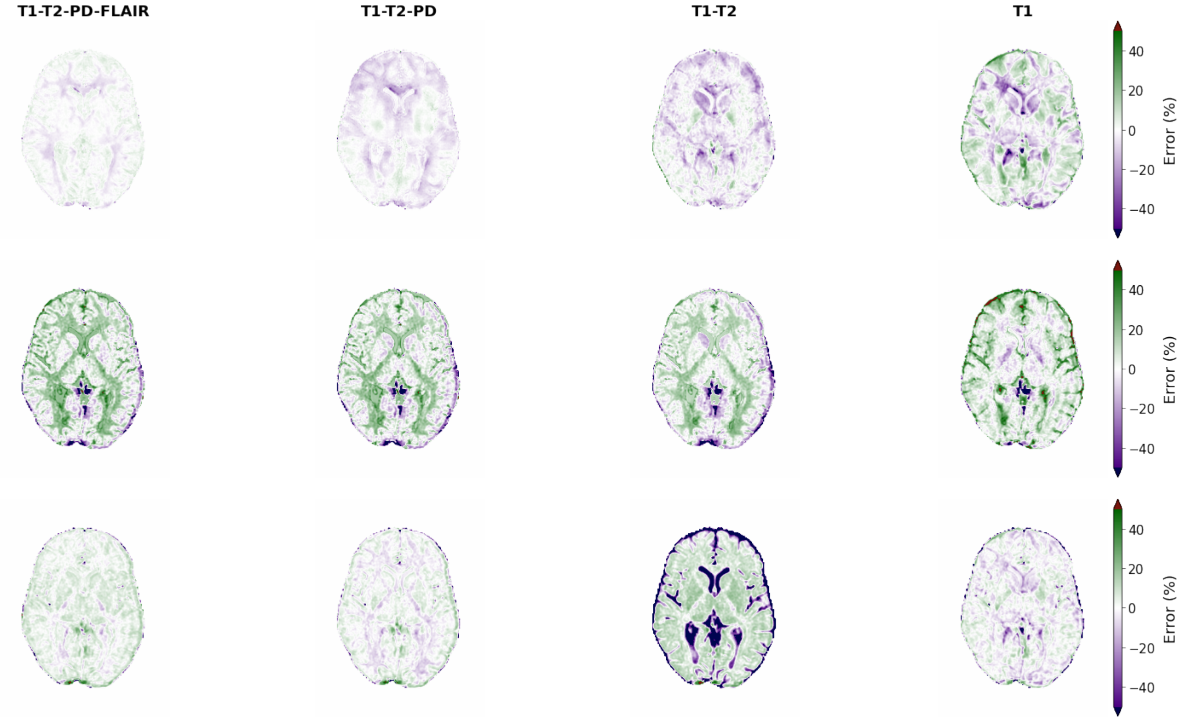

Figure A.4: Error percentage maps for models generating multiple quantitative maps using different combinations of input scans. From top to bottom: T1-maps, T2-maps and PD-maps.

A.2.2. Input Dimensionality

Table A.1: Comparison of evaluation metrics for best models predicting multiple quantitative maps using MSE loss function for both 2D and 3D input scans. [1]PD values not in ms but in a.u.

	Best 2D multi model MSE			**Best 3D multi model MSE**		
	T1	T2	PD	T1	T2	PD
PSNR (dB) ↑	**34.242 ± 2.784**	20.750 ± 2.659	**23.192 ± 2.623**	33.015 ± 1.102	**22.554 ± 0.804**	22.624 ± 0.516
SSIM ↑	**0.996 ± 0.003**	0.938 ± 0.040	**0.928 ± 0.049**	0.996 ± 0.004	**0.952 ± 0.036**	0.923 ± 0.055
RMSE (ms)[1] ↓	**88.194 ± 34.686**	33.528 ± 11.657	**9.779 ± 2.832**	96.871 ± 12.321	**26.196 ± 2.474**	11.849 ± 0.710
Median Error ↓	**3.77%**	11.77%	**6.55%**	5.31%	**7.81%**	7.31%

A.2.3. Tumour Predictions

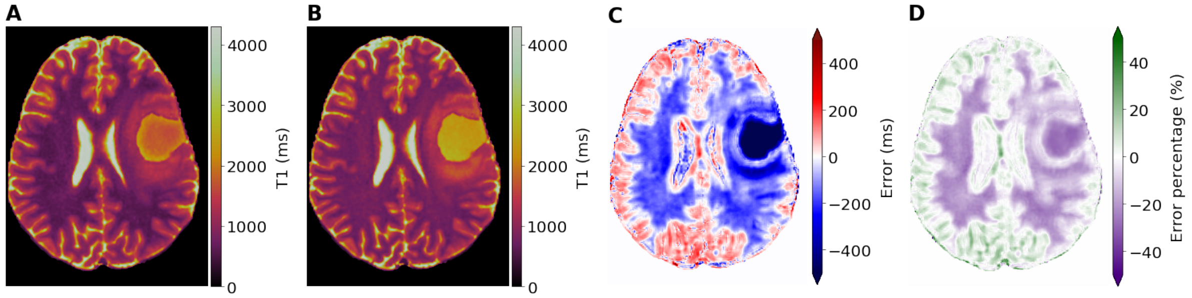

Figure A.5: Model prediction on tumour scans for overtrained model. The model reaches a higher performance on the synthetic healthy scans than the models in Section 4.1.1, but it reaches a higher error when predicting quantitative maps for synthetic tumour scans. This is seen most clearly in the tumour core. Difference values are cut-off at ±500 ms for improved visualisation. Error percentages are cut-off at 50% for improved visualisation. **A**: Groundtruth T1-map. **B**: Model prediction of T1-map. **C**: Difference in T1 between prediction and groundtruth. **D**: Error percentage map showing the percentual error relative to the true T1 values.

A.2.4. Real Data

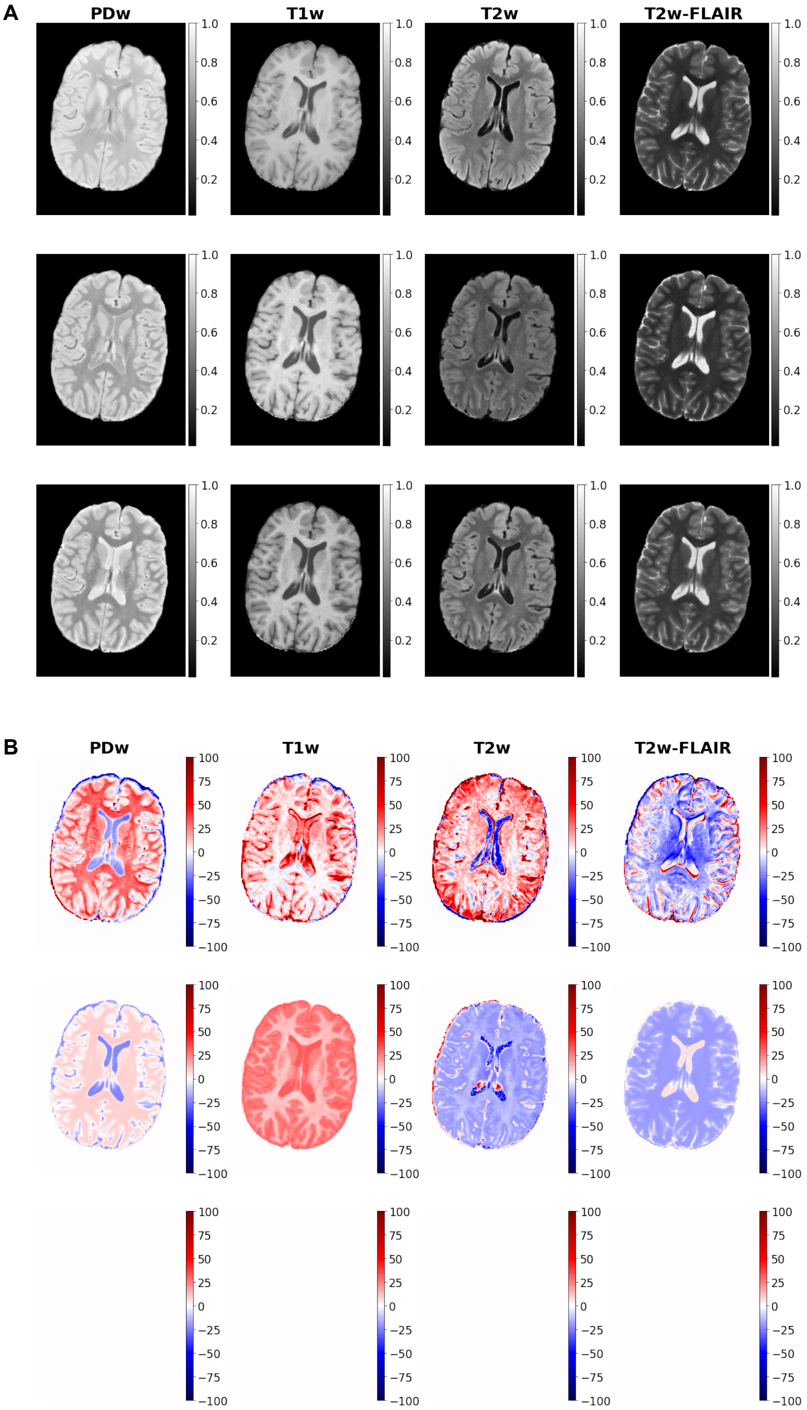

Figure A.6: Comparison of real data and different types of synthetic data. **A:** Plots showing example slices for the different types of data. From top to bottom: Real data, synthetic data with real parameters and synthetic data with training parameters. **B:** Plots showing the percentual difference in pixel intensities compared to the synthetic data with training parameters. From top to bottom: Real data, synthetic data with real parameters and synthetic data with training parameters. As a result, the error on the bottom row is zero.

Table A.2: Evaluation metrics for models generating a T1-map from real and synthetic input scans. Synthetic scans are divided into synthetic scans with the same detection-specific parameters as used in training the models and synthetic scans with the same detection-specific parameters as the real scans. Best metrics are shown in bold.

Input scans	PSNR (dB) ↑	SSIM ↑	RMSE (ms) ↓
Synthetic train	**26.898 ± 2.197**	**0.987 ± 0.004**	**200.857 ± 53.943**
Synthetic real	24.661 ± 2.357	0.980 ± 0.006	261.547 ± 79.373
Real	12.119 ± 2.069	0.821 ± 0.082	1097.932 ± 288.173

Table A.3: Evaluation metrics for models generating a T1-map from different combinations of real input scans. The best metrics shown in bold.

Input scans	PSNR (dB) ↑	SSIM ↑	RMSE (ms) ↓
All inputs	12.118 ± 2.069	0.821 ± 0.082	1097.932 ± 288.173
T1w, T2w, PDw	13.569 ± 2.016	0.840 ± 0.071	928.082 ± 241.748
T1w, T2w, T2w-FLAIR	**16.159 ± 2.132**	**0.868 ± 0.058**	**691.290 ± 191.782**
T1w, T2w	13.908 ± 1.656	0.847 ± 0.069	883.888 ± 183.767
T1w, T2w-FLAIR	12.440 ± 2.416	0.824 ± 0.077	1071.373 ± 346.220

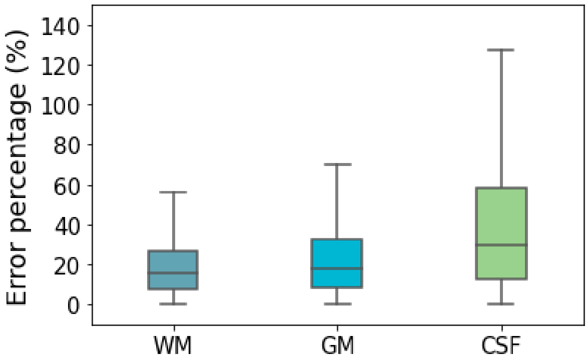

Figure A.7: Boxplot showing the distribution of error percentages per tissue when predicting a T1-map from the best combinations of real input scans. The model uses T1w, T2w, and T2w-FLAIR as input. WM = white matter, GM = grey matter and CSF = cerebrospinal fluid.

Table A.4: Evaluation metrics for models generating multiple quantitative maps from real and synthetic input scans. Synthetic scans are divided into synthetic scans with the same detection-specific parameters as used in training the models and synthetic scans with the same detection-specific parameters as the real scans. The best metrics are shown in bold. ^{1}PD values not in ms but in a.u.

Input scans	Map	PSNR (dB) ↑	SSIM ↑	RMSE (ms)^1 ↓
Synthetic train	T1	**26.486 ± 2.307**	**0.986 ± 0.005**	**211.507 ± 61.863**
	T2	**24.210 ± 2.623**	**0.965 ± 0.015**	**22.226 ± 7.394**
	PD	**22.970 ± 1.841**	**0.950 ± 0.024**	**8.495 ± 1.948**
Synthetic real	T1	24.286 ± 2.325	0.977 ± 0.007	272.840 ± 82.399
	T2	22.556 ± 3.023	0.960 ± 0.016	27.423 ± 11.431
	PD	21.690 ± 2.023	0.915 ± 0.040	9.774 ± 1.995
Real	T1	12.760 ± 2.220	0.833 ± 0.075	1024.902 ± 294.116
	T2	13.330 ± 2.312	0.825 ± 0.081	77.955 ± 25.369
	PD	6.498 ± 2.136	0.834 ± 0.081	17.734 ± 3.204

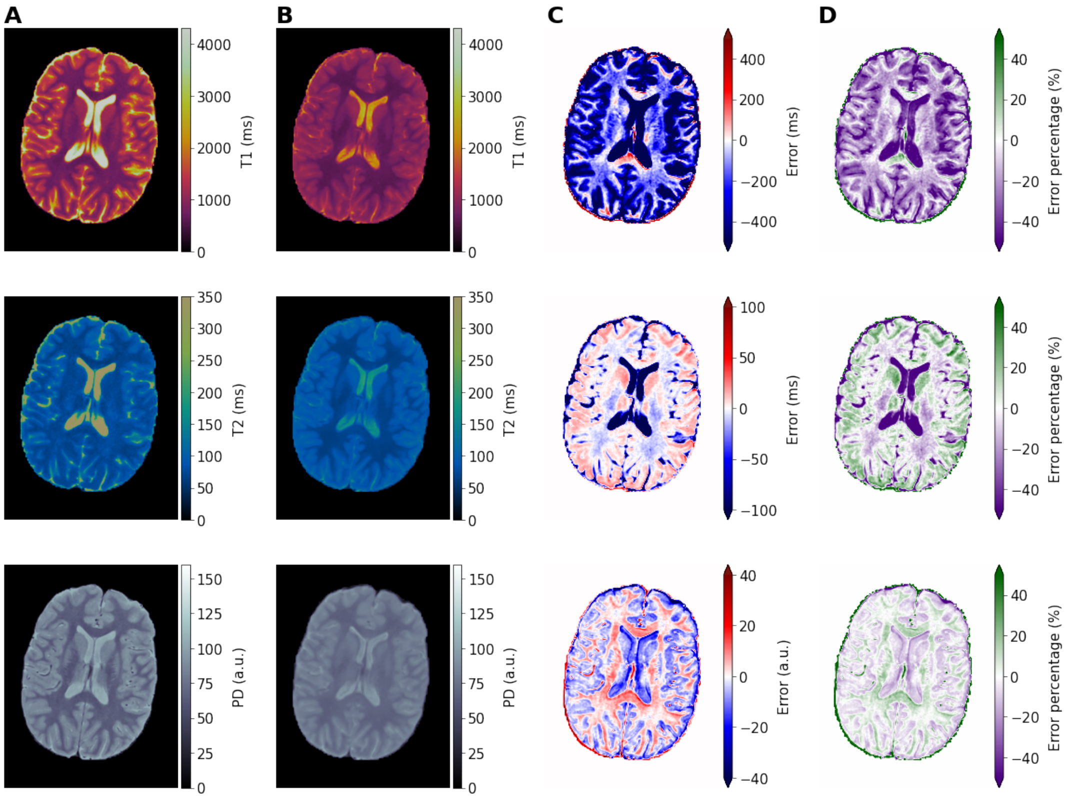

Figure A.8: Predictions of multiple quantitative maps from real T1w, T2w and PDw scans. From top to bottom: T1-, T2-, and PD-maps. Difference values are cut-off at a different value per quantitative map for improved visualisation. Percentages are cut-off at 50% for improved visualisation. **A**: Groundtruth quantitative map. **B**: Model prediction of the quantitative map. **C**: Difference between prediction and groundtruth. **D**: Error percentage map showing the percentual error relative to the true quantitative values.

Table A.5: Evaluation metrics for models predicting multiple quantitative maps from different combinations of real input scans. The best metrics shown in bold. [1]PD values not in ms but in a.u.

Input scans	Map	PSNR (dB) ↑	SSIM ↑	RMSE (ms)[1] ↓
All inputs	T1	12.760 ± 2.220	0.833 ± 0.075	1024.902 ± 294.116
	T2	13.330 ± 2.312	0.825 ± 0.081	77.955 ± 25.369
	PD	16.498 ± 2.136	0.834 ± 0.081	17.734 ± 3.204
T1w, T2w, PDw	T1	13.594 ± 2.054	0.842 ± 0.071	926.508 ± 247.051
	T2	15.770 ± 2.336	0.879 ± 0.054	58.951 ± 18.295
	PD	**18.933 ± 1.873**	**0.865 ± 0.064**	**13.463 ± 2.919**
T1w, T2w, T2w-FLAIR	T1	**16.073 ± 1.964**	**0.876 ± 0.055**	**694.709 ± 177.480**
	T2	16.769 ± 2.393	0.885 ± 0.052	52.383 ± 16.838
	PD	18.775 ± 1.732	0.855 ± 0.068	13.618 ± 2.380
T1w, T2w	T1	14.831 ± 1.665	0.859 ± 0.062	794.939 ± 165.854
	T2	**16.907 ± 1.946**	**0.887 ± 0.050**	**50.847 ± 12.528**
	PD	14.453 ± 2.286	0.831 ± 0.076	22.619 ± 5.425

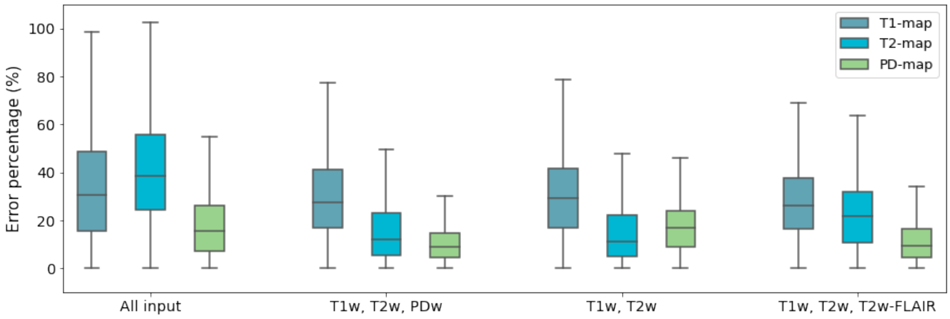

Figure A.9: Boxplot showing the distribution of error percentages when predicting multiple quantitative maps from multiple combinations of real input scans. The model with T1-, T2-, and PD-weighted scans as input reaches the lowest median error overall.

Table A.6: Evaluation metrics for different brain regions of models predicting multiple quantitative maps from different combinations of real input scans. WM = White matter, GM = Grey matter and CSF = Cerebrospinal fluid. Best metrics shown in bold. [1]PD values not in ms but in a.u.

Input scans	Map	Region	PSNR (dB) ↑	SSIM ↑	RMSE (ms)[1] ↓
All inputs	T1	WM	18.765 ± 3.014	0.947 ± 0.031	506.953 ± 193.643
		GM	12.705 ± 2.359	0.926 ± 0.038	1026.086 ± 321.759
		CSF	8.645 ± 1.798	0.936 ± 0.026	1622.312 ± 316.514
	T2	WM	19.485 ± 2.599	0.964 ± 0.022	36.037 ± 9.497
		GM	14.510 ± 2.628	0.945 ± 0.030	67.729 ± 23.058
		CSF	8.591 ± 1.897	0.939 ± 0.025	132.768 ± 30.463
	PD	WM	17.140 ± 2.895	0.967 ± 0.022	16.225 ± 4.420
		GM	17.316 ± 2.330	0.981 ± 0.013	15.744 ± 3.407
		CSF	14.256 ± 1.665	0.973 ± 0.009	22.819 ± 3.731
T1w, T2w, PDw	T1	WM	**19.332 ± 2.962**	**0.955 ± 0.023**	**470.373 ± 181.393**
		GM	13.174 ± 2.029	0.935 ± 0.034	960.661 ± 280.315
		CSF	9.600 ± 1.623	0.944 ± 0.023	1447.885 ± 255.678
	T2	WM	**24.634 ± 3.145**	**0.981 ± 0.011**	**20.636 ± 8.121**
		GM	16.497 ± 2.858	0.965 ± 0.019	55.174 ± 22.071
		CSF	10.790 ± 1.906	0.956 ± 0.018	103.059 ± 22.053
	PD	WM	**22.183 ± 2.263**	0.973 ± 0.018	**9.029 ± 2.705**
		GM	19.315 ± 2.041	**0.986 ± 0.009**	12.505 ± 2.892
		CSF	14.825 ± 1.608	0.976 ± 0.010	21.486 ± 4.253

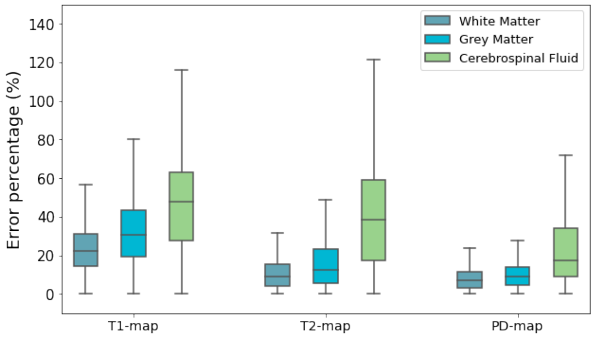

Figure A.10: Boxplot showing the distribution of error percentages per tissue when predicting multiple quantitative maps from the best combinations of real input scans. The model uses T1w, T2w, and PDw scans to predict T1-, T2- and PD-maps.

Bibliography

[1] Prakkamakul S., et al., Ultrafast brain MRI: clinical deployment and comparison to conventional brain MRI at 3T. *Journal of Neuroimaging* **26**, 503–510 (2016).

[2] Biberacher V., et al., Intra-and interscanner variability of magnetic resonance imaging based volumetry in multiple sclerosis. *Neuroimage* **142**, 188–197 (2016).

[3] Fernandes J. L., Rochitte C. E., T1 mapping: technique and applications. *Magnetic Resonance Imaging Clinics* **23**, 25–34 (2015).

[4] Meng T., et al., The diagnostic performance of quantitative mapping in breast cancer patients: a preliminary study using synthetic MRI. *Cancer Imaging* **20**, 1–9 (2020).

[5] Lee Y., Callaghan M. F., Acosta-Cabronero J., Lutti A., Nagy Z., Establishing intra- and inter-vendor reproducibility of T1 relaxation time measurements with 3T MRI. *Magnetic resonance in medicine* **81**, 454–465 (2019).

[6] Gracien R.-M., et al., How stable is quantitative MRI?–Assessment of intra-and inter-scanner-model reproducibility using identical acquisition sequences and data analysis programs. *NeuroImage* **207**, 116364 (2020).

[7] Deoni S. C., Peters T. M., Rutt B. K., High-resolution T1 and T2 mapping of the brain in a clinically acceptable time with DESPOT1 and DESPOT2. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **53**, 237–241 (2005).

[8] Nguyen T. D., et al., Feasibility and reproducibility of whole brain myelin water mapping in 4 minutes using fast acquisition with spiral trajectory and adiabatic T2prep (FAST-T2) at 3T. *Magnetic resonance in medicine* **76**, 456–465 (2016).

[9] Claeser R., Zimmermann M., Shah N. J., Sub-millimeter T1 mapping of rapidly relaxing compartments with gradient delay corrected spiral TAPIR and compressed sensing at 3T. *Magnetic resonance in medicine* **82**, 1288–1300 (2019).

[10] Warntjes J., Leinhard O. D., West J., Lundberg P., Rapid magnetic resonance quantification on the brain: optimization for clinical usage. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **60**, 320–329 (2008).

[11] Ma D., et al., Magnetic resonance fingerprinting. *Nature* **495**, 187–192 (2013).

[12] Liao C., et al., 3d mr fingerprinting with accelerated stack-of-spirals and hybrid sliding-window and grappa reconstruction. *Neuroimage* **162**, 13–22 (2017).

[13] Cao X., et al., Fast 3D brain MR fingerprinting based on multi-axis spiral projection trajectory. *Magnetic resonance in medicine* **82**, 289–301 (2019).

[14] Han X., MR-based synthetic CT generation using a deep convolutional neural network method. *Medical physics* **44**, 1408–1419 (2017).

[15] Wolterink J. M., et al., Deep MR to CT synthesis using unpaired data (in *International workshop on simulation and synthesis in medical imaging*), pp. 14–23 (2017).

[16] Emami H., Dong M., Nejad-Davarani S. P., Glide-Hurst C. K., Generating synthetic CTs from magnetic resonance images using generative adversarial networks. *Medical physics* **45**, 3627–3636 (2018).

[17] Sharma A., Hamarneh G., Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. *IEEE transactions on medical imaging* **39**, 1170–1183 (2019).

[18] Yang Q., et al., MRI Cross-Modality Image-to-Image Translation. *Scientific Reports* **10**, 1–18 (2020).

[19] Wu Y., Ma Y., Du J., Xing L., Deciphering tissue relaxation parameters from a single MR image using deep learning (in *Medical Imaging 2020: Computer-Aided Diagnosis*), Vol. 11314, p. 113140Q (2020).

[20] Moya-Sáez E., Peña-Nogales O., Sanz-Estébanez S., de Luis-Garcia R., Alberola-López C., CNN-based synthesis of T1, T2 and PD parametric maps of the brain with a minimal input feeding (in *ISMRM Proceedings*), (2020).

[21] Bloch F., Nuclear induction. *Physical review* **70**, 460 (1946).

[22] Purcell E. M., Torrey H. C., Pound R. V., Resonance absorption by nuclear magnetic moments in a solid. *Physical review* **69**, 37 (1946).

[23] Damadian R., Goldsmith M., Minkoff L., NMR in cancer: XVI. FONAR image of the live human body. *Physiological chemistry and physics* **9**, 97–100 (1977).

[24] Hahn E. L., Spin echoes. *Physical review* **80**, 580 (1950).

[25] Bydder G., Young I., MR imaging: clinical use of the inversion recovery sequence (1985).

[26] Stehling C., et al., Comparison of a T1-weighted inversion-recovery-, gradient-echo-and spin-echo sequence for imaging of the brain at 3.0 Tesla. *RoFo: Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin* **177**, 536–542 (2005).

[27] Warntjes J., Synthesized Magnetic Resonance Imaging and SyMRI$^{©}$, Theory and Application. A White Paper, (SyntheticMR), White paper (2017).

[28] Blystad I., et al., Synthetic MRI of the brain in a clinical setting. *Acta radiologica* **53**, 1158–1163 (2012).

[29] Tanenbaum L. N., et al., Synthetic MRI for clinical neuroimaging: results of the magnetic resonance image compilation (MAGiC) prospective, multicenter, multireader trial. *American Journal of Neuroradiology* **38**, 1103–1110 (2017).

[30] Hosny A., Parmar C., Quackenbush J., Schwartz L. H., Aerts H. J., Artificial intelligence in radiology. *Nature Reviews Cancer* **18**, 500–510 (2018).

[31] Ronneberger O., Fischer P., Brox T., U-net: Convolutional networks for biomedical image segmentation (in *International Conference on Medical image computing and computer-assisted intervention*), pp. 234–241 (2015).

[32] He K., Zhang X., Ren S., Sun J., Deep residual learning for image recognition (in *Proceedings of the IEEE conference on computer vision and pattern recognition*), pp. 770–778 (2016).

[33] Goodfellow I. J., et al., Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014).

[34] Isola P., Zhu J.-Y., Zhou T., Efros A. A., Image-to-image translation with conditional adversarial networks (in *Proceedings of the IEEE conference on computer vision and pattern recognition*), pp. 1125–1134 (2017).

[35] Bahdanau D., Cho K., Bengio Y., Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[36] Vaswani A., et al., Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[37] Xu K., et al., Show, attend and tell: Neural image caption generation with visual attention (in *International conference on machine learning*), pp. 2048–2057 (2015).

[38] Zhang H., Goodfellow I., Metaxas D., Odena A., Self-attention generative adversarial networks (in *International conference on machine learning*), pp. 7354–7363 (2019).

[39] Wang T., Lei Y., Curran W. J., Liu T., Yang X., Contrast-enhanced MRI synthesis from non-contrast MRI using attention CycleGAN (in *Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging*), Vol. 11600, p. 116001L (2021).

[40] Oktay O., et al., Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* (2018).

[41] Yurt M., Dar S. U. H., Erdem A., Erdem E., Çukur T., mustGAN: Multi-stream generative adversarial networks for MR image synthesis. *arXiv preprint arXiv:1909.11504* (2019).

[42] Dar S. U., et al., Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE transactions on medical imaging* **38**, 2375–2388 (2019).

[43] Chartsias A., Joyce T., Giuffrida M. V., Tsaftaris S. A., Multimodal MR synthesis via modality-invariant latent representation. *IEEE transactions on medical imaging* **37**, 803–814 (2017).

[44] Lyu Q., Wang G., Quantitative MRI: Absolute T1, T2 and Proton Density Parameters from Deep Learning. *arXiv preprint arXiv:1806.07453* (2018).

[45] Wu Y., Ma Y., Du J., Xing L., Accelerating quantitative MR imaging with the incorporation of B1 compensation using deep learning. *Magnetic Resonance Imaging* **72**, 78–86 (2020).

[46] van der Voort S., PrognosAIs (version 0.3.5) (2021).

[47] Isensee F., et al., Automated brain extraction of multisequence MRI using artificial neural networks. *Human brain mapping* **40**, 4952–4964 (2019).

[48] van der Voort S. R., et al., Who 2016 subtyping and automated segmentation of glioma using multi-task deep learning. *arXiv preprint arXiv:2010.04425* (2020).

[49] Zhang Y., Brady M., Smith S., Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imag* (2001).

[50] He K., Zhang X., Ren S., Sun J., Identity mappings in deep residual networks (in *European conference on computer vision*), pp. 630–645 (2016).

[51] Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P., Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**, 600–612 (2004).

[52] Keenan K. E., et al., Multi-site, multi-platform comparison of MRI T1 measurement using the system phantom. *PloS one* **16** (2021).