DISSERTATION

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Steffen Albert, M. Sc.

born in Karlsruhe, Germany

Oral examination: December 13th, 2023

Prediction of treatment response and outcome in locally advanced rectal cancer using radiomics

Referees: Prof. Dr. Lothar R. Schad
Prof. Dr. Tristan A. Kuder

Vorhersage des Behandlungserfolges und Therapieansprechens von lokal fortgeschrittenen Rektumkarzinomen mithilfe von Radiomics

Mit der zunehmenden Anzahl medizinischer Bilder wird Deep Learning immer häufiger für Radiomics verwendet, leidet aber unter der Verwendung von kleinen und heterogenen Datensätzen. Zu diesem Zweck wurde eine Radiomics-Pipeline für die Vorhersage des Behandlungsergebnisses der neoadjuvanten Therapie bei lokal fortgeschrittenem Rektumkarzinom entwickelt, wobei der Schwerpunkt auf der Entwicklung von Methoden für den Umgang mit kleinen, heterogenen, multizentrischen Datensätzen lag. Für die Normalisierung wurden sechs verschiedene Normalisierungsmethoden (fünf statistische Methoden und eine neuartige Deep-Learning-Methode) in verschiedenen Konfigurationen untersucht: trainiert mit allen Bildern, Bildern aus allen Zentren außer einem und mit Bildern aus einem einzigen Zentrum. Die Auswirkungen der Normalisierung wurden anhand von vier Aufgaben bewertet: Tumorsegmentierung, Vorhersage des Behandlungsergebnisses, Vorhersage des Geschlechts und Vorhersage des Alters. Bei der Segmentierung gab es nur signifikante Unterschiede, wenn mit einem Zentrum trainiert wurde, wobei die Deep-Learning-Methode mit einem DSC von 0.50 ± 0.01 die beste war. Für die Vorhersage von Geschlecht und Behandlungsergebnissen funktioniert die Perzentilmethode in Kombination mit dem Histogramm-Matching in allen Szenarien am besten. Mit mehr Daten sollte sich die Deep-Learning-Methode weiter verbessern. Die Klassifizierungsleistung wurde anhand eines bereits veröffentlichten neuronalen Netzes bewertet. Dieses Netz besteht aus zwei U-Nets, die sich ihre Gewichte teilen, mit Segmentierung als zusätzlicher Aufgabe. Der maximale AUC war 0.75 (95 % CI: 0.52 to 0.92) auf dem Validierungssatz. Dies ist besser als der Zufall, aber nicht besser als die Verwendung eines Klassifikators, der anhand klinischer Merkmale trainiert wurde. Zusammenfassend lässt sich sagen, dass die Normalisierung zur Verallgemeinerbarkeit der neuronalen Netze beigetragen hat, aber es gibt eine Grenze dessen, was korrigiert werden kann.

Prediction of treatment response and outcome in locally advanced rectal cancer using radiomics

With the increasing number of medical images, deep learning is being used more and more in radiomics, but it suffers from small and heterogeneous datasets. To address this, a radiomics pipeline was developed for the prediction of the treatment outcome for neoadjuvant therapy in locally advanced rectal cancer (LARC), focusing on developing methods for dealing with small, heterogeneous multicenter datasets. For normalization, six different normalization methods (five statistical methods and one novel deep learning method) were investigated in multiple configurations: trained on all images, images from all centers except one, and images from a single center. The impact of normalization was evaluated in four tasks: tumor segmentation, prediction of treatment outcome, prediction of sex and prediction of age. For segmentation, there were only significant differences when training on one center, with the deep learning method being the best with a DSC of 0.50 ± 0.01. For the prediction of sex and treatment outcomes, the percentile method combined with histogram matching works best in all scenarios. The classification performance was evaluated using a published neural network. This network consists of two U-Nets sharing their weights, with segmentation as an additional task. The maximum AUC was 0.75 (95 % CI: 0.52 to 0.92) on the validation set. This is better than chance, but not better than using a classifier trained on clinical characteristics. In summary, normalization did help with the generalizability of the neural networks, but there is a limit to what can be corrected.

Acknowledgement

First of all, I would like to thank Prof. Dr. Lothar Schad for giving me the opportunity to start a PhD in his group and for the supervision of this thesis.

My thanks also go to Priv.-Doz. Dr. T. A. Kuder for acting as the second referee of this thesis.

I want to express gratitude to Prof. Dr. Frank Zöllner for the help and guidance with my thesis project and for his guidance through each stage of the process. I especially thank him for his help in proofreading all my publications and my thesis.

I also would like to thank everyone who participated in the DFG SPP radiomics project in Mannheim, which made this thesis possible, and all their help, especially for the publications. My thanks especially go to Barbara Wichtmann, who had the task of segmenting all the tumors and helped me understand the medical aspects of this thesis, and Dr. Angelika Maurer, who also helped with segmentation and contacted all the hospitals and got them to send us their images.

My time at CKM was great, despite some unforseen restrictions, but we made the best of it. I want to thank all current and former members for all the enjoyable coffee and lunch breaks, Tischkicker games, bouldering and gym sessions, and a great ISMRM conference and vacation afterwards.

I am grateful for all your help whenever I had questions or needed help with something or just needed to be distracted from whatever (research) problem I was having at that time. I want to especially thank Simon, Patrick, and Safa for proofreading my thesis and all your suggestions.

Last but not least, I want to thank my family for all the support and encouragement. I am thankful for the support of my girlfriend Simone, both for encouraging me and for proofreading my thesis.

Contents

1 Introduction

The volume of medical data recorded increases each day, with medical imaging being one of the primary tools essential for diagnosing a wide variety of medical conditions. With the advancement of hospital digitalization, medical data is increasingly available for research. Furthermore, having access to the data of many patients opens up a plethora of new possibilities using big data techniques and machine learning to obtain new insights from this data, which can help in the diagnosis and treatment of patients.

While some medical information, such as laboratory reports, is available in a structured format, this does not hold true for medical images. Medical images must be interpreted by an experienced radiologist, which is time-consuming and susceptible to potential bias. This can be a problem with an increasing volume of images, and hinder the proper treatment of patients, which can be exacerbated by the lack of experienced radiologists.

Radiomics aims to solve parts of these problems by extracting quantitative information from images. These so-called features can then be used for predictive or prognostic models. This is especially challenging for Magnetic Resonance Imaging (MRI) images, because the interpretation relies less on image intensity and more on contrast information (Afshar et al., 2019).

The field started with the so-called hand-crafted radiomics, where features defined by experts are extracted from a manually annotated region. Subsequently, meaningful features are then selected and employed to train machine learning models (Afshar et al., 2019).

However, due to advances in deep learning, larger available datasets, and challenges related to the reproducibility of features (Schurink et al., 2022; Michoux et al., 2021; Dreher et al., 2020), deep learning methods are increasingly being favored. In these approaches, features are learned by the network, eliminating the need for expert-defined features. Moreover, the region of interest does not have to be annotated, resulting in substantial time savings and the reduction of bias caused by variations in annotation (Avanzo et al., 2020).

Much of the development of the radiomics field relies heavily on the availability of large datasets. The availability of large public datasets can drive the development of research in a particular field (Varoquaux and Cheplygina, 2022). However, even the largest medical image datasets are small compared to other computer vision datasets.

For instance, The Cancer Imaging Archive (Clark et al., 2013), which is a large repository of medical image data and the largest for cancer, has only ten datasets with images of more than 1000 patients and only one with more than 10,000 individual patients. In contrast, in the field of computer vision, datasets for natural images are considerably more extensive, such as ImageNet (J. Deng et al., 2009), which currently contains approximately 14 million images.

Creating large datasets for medical data is a highly challenging endeavor. Numerous concerns about data privacy and security must be carefully addressed. Additionally, the availability of patients with a specific disease is inherently limited, constraining the pool of potential research subjects. Moreover, the image acquisition and processing should be as standardized as possible, but changing that is not possible for retrospective data. Consequently, the resulting datasets exhibit inherent heterogeneity, a characteristic that introduces many challenges for radiomics.

To investigate these issues locally advanced rectal cancer (LARC) was chosen as an exemplary case. Colorectal cancer is the third most lethal cancer in Europe, with a 5-year survival rate of 68% in Germany (Fitzmaurice et al., 2015). Most of the data was taken from a medical study investigating the order of radiotherapy and radio-chemo-therapy (CRT) for neoadjuvant therapy before surgery (Rödel et al., 2015).

The recommended treatment protocol for LARC is radiotherapy and/or CRT followed by total mesorectal excision (Benson et al., 2015). Radiotherapy or CRT serves as neoadjuvant treatment, aiming to reduce the tumor size and improve operability prior to surgery. In some cases, this neoadjuvant therapy can result in complete remission and no surgery would be needed. Consequently, accurate prediction of treatment response is essential to decide which tumors should be surgically resected and which patients qualify for a watch-and-wait approach. Avoiding an unnecessary surgery can greatly improve the quality of life of patients (Smith and Garcia-Aguilar, 2015).

Predicting the tumor regression grade (TRG) before surgery presents challenges due to the limitations of biopsy and imaging techniques. Biopsies only provide information from a limited number of points within the tumor, making it challenging to assess the overall tumor response. Although imaging allows for a comprehensive tumor analysis, the limited resolution makes it difficult to differentiate between the complete absence of cancer cells and the presence of rare residual cells (Horvat, Carlos Tavares Rocha, et al., 2019).

For the staging process, MRI plays a key role (Horvat, Carlos Tavares Rocha, et al., 2019; Coppola et al., 2021), however, T2-weighted morphological imaging has low sensitivity after neoadjuvant therapy and diffusion-weighted imaging yields controversial results (Horvat, Veeraraghavan, et al., 2018). Nonetheless, emerging machine learning techniques are currently under investigation to address these limitations and improve the accuracy of treatment response prediction.

This problem is well suited for a machine learning approach because ground truth data is available. After surgery, pathologists conduct a histological analysis of the resected tumor, resulting in the pathological TRG. According to Ryan et al., 2015, this analysis is the gold standard for assessing tumor response and can provide reliable ground truth data to train machine learning models.

There are already a few approaches to solving this problem, for example with a hand-crafted radiomics approach in Z. Liu et al., 2017, which performed very well, but all patients were imaged and treated in a single hospital, so it is hard to say how well this approach would generalize.

In a meta-analysis conducted by Jang et al., 2020, six radiomics studies in LARC, using expert-defined features, aimed to predict the TRG. They reported a high specificity of 93.5 % (95 % CI: 91.5 % to 95.1 %), when predicting the complete absence of tumor cells. However, sensitivity was considerably lower, at 32.3 % (95 % CI: 18.2 % to 50.6 %). Sensitivity is of particular importance, as misclassifying a patient as a candidate for a watch-and-wait protocol without achieving complete remission would have severe clinical implications.

A higher sensitivity was achieved through the application of deep learning by Jin et al., 2021. In this work, tumor segmentation was added as an additional task for the deep learning model. Due to the good results and the code being available, the same approach was explored with the new dataset in this thesis (B. D. Wichtmann et al., 2022).

One of the main challenges when predicting the TRG using this dataset is its inherent heterogeneity. In machine learning, it is often an underlying assumption that all data is drawn from the same distribution. However, this assumption does not hold in clinical scenarios. Data originating from a different medical center, scanner, or with a different acquisition protocol belongs to distinct domains. Consequently, it can be challenging for a machine learning model trained in one domain to make accurate predictions when confronted with data from another domain (Guan and M. Liu, 2022; Mårtensson et al., 2020). To mitigate this issue, normalization techniques are commonly employed.

A wide array of normalization techniques is available for MR images, with statistical methods being commonly employed in this context (Reinhold et al., 2019; Shah et al., 2011). However, many of these methods have been developed primarily for brain imaging applications (Carré et al., 2020), which limits their applicability in abdominal imaging. This limitation arises because these methods rely on specific characteristics of the brain, such as for example white stripe normalization, which uses white matter as reference tissue for intensity normalization (Shinohara et al., 2014). Unlike in brain imaging, there is no suitable reference tissue available in the abdomen.

Other methods, such as histogram matching (Nyul et al., 2000), were initially developed for brain images, but can also be adapted for other regions with some modifications. Intriguingly, some methods were originally developed for entirely unrelated purposes. For instance, the ComBat method was originally developed for gene assays to remove batch-dependent effects (Johnson et al., 2007). Nevertheless, it is used in the field of magnetic resonance imaging for addressing scanner-dependent effects (Eshaghzadeh Torbati et al., 2021; Fortin et al., 2017; Mali et al., 2021).

Deep learning can also be used for normalization. However, a significant challenge in this context is the availability of adequate training data. One approach involves capturing images of patients using various scanners and training a neural network to adapt the image style accordingly (Dewey et al., 2019), but this requires a lot of effort, especially when multiple hospitals at different locations are involved.

Thus, it is preferable to use methods that do not require paired training data, which simplifies the creation of a sufficiently large training set. An architecture which has proved to be effective at domain translation is the CycleGAN. It uses adversarial training, which means the generator is trained to translate an image from one domain to another while the discriminator evaluates how well these generated images match real images from the target domain. Additionally, cycle consistency is enforced, meaning that if an image is translated from domain A to domain B and then back from domain B to domain A, it should closely resemble the original image.

In the context of normalization, a CycleGAN can be used to translate the image from the style of one scanner to the style of another scanner, as done by Modanwal et al., 2021 for breast MRIs with scanners from two different manufacturers. For more than two domains, multiple networks must be trained when using this approach (Y. Li et al., 2020). This approach can be extended to the StarGAN, which incorporates multiple encoders and decoders, eliminating the need to train separate models for each pair of domains. Instead, it can translate images into a common latent space and then into any of the domains on which it was trained (Bashyam et al., 2021).

In addition to Cycle- and StarGAN, there are several other multi-source domain adaptation techniques (Guan and M. Liu, 2022). For instance, adversarial training can effectively remove scanner-dependent effects (Ganin et al., 2016). In this technique, domain detection is employed as an adversarial loss.

Adversarial training is a technique also employed by ImUnity, as described in the work by Cackowski et al., 2023. ImUnity utilizes a reference image to specify the style that should be transferred to the image during normalization. An advantage of this approach is that it can be trained unsupervised using random image pairs without the strict delineation of domains.

This overcomes a limitation present in many other approaches. Most existing methods require well-defined domains, such as specific scanners, sites, or protocols. However, this is not the case for the data used in this work. The dataset consists of images from various scanners located at different sites, each featuring slightly different acquisition parameters and post-processing procedures. Consequently, there are no distinct clusters within this data.

Therefore, a new method was developed that does not need defined domains. Unlike ImUnity, this method does not employ unsupervised training. Instead, it incorporates the acquisition parameters as additional target parameters.

Normalization performance is evaluated using downstream deep learning tasks. In addition to predicting the TRG, sex and age, segmentation is employed as an auxiliary task. Apart from its role in evaluation, segmentation plays a crucial role in a hand-crafted radiomics workflow and is integral to treatment planning in radiation therapy.

While some alternative methods for segmentation exist (Joshi et al., 2010; van Heeswijk et al., 2016), most recent publications use neural networks, which have demonstrated remarkable effectiveness for image segmentation. Among these networks, the U-Net stands out as one of the most well-known architectures for segmentation (Ronneberger et al., 2015). It continues to be widely employed, often with certain modifications, for example in the well-regarded implementation by Isensee et al., 2021.

Its usefulness for segmentation of rectal cancer tumors was successfully demonstrated, for instance J. Wang et al., 2018 employed it for this purpose. Furthermore, Huang et al., 2019 extended the U-Net's application to 3D, incorporating the localization of the region of interest as an additional task during training, sharing the encoder. A network with the same encoder-decoder structure as the U-Net is used by J. Lee et al., 2019. While some alternative architectures, such as the dense net (Soomro et al., 2019), have been explored, the U-Net remains the prevailing choice in the field. Consequently, it was selected for use in segmentation in this work.

The primary objective of this thesis is not the development of novel methods or neural network architectures for segmentation and classification. Instead, its central focus lies in the application of existing methods to new data and the enhancement of the dataset itself. There is only limited application to improving performance on a well-known public dataset, which is far from clinical practice.

Within the scope of this thesis, six different normalization techniques were tested, including a novel one based on deep learning. These methods underwent evaluation through various downstream tasks. For the classification, the approach from Jin et al., 2021 was tested and evaluated on the dataset.

The hope is that with improved normalization, it will become feasible to assemble large datasets derived from real-world as found in clinical practice. These datasets can subsequently be used to train existing machine learning models, improving their performance and enabling clinically usable predictions. Ultimately, this can lead to improved patient treatment strategies and improved outcomes for patients.

In summary, this work tries to develop and improve techniques within a radiomics pipeline tailored to the constraints of small and heterogeneous datasets, with a particular emphasis on the normalization of the images and its impact on the segmentation and treatment outcome prediction.

Theoretical Background 2

This chapter starts with Nuclear Magnetic Resonance (NMR) in Section 2.1, which forms the basis of Magnetic Resonance Imaging (MRI), which is presented in Section 2.2. This is followed by Section 2.3, which describes the basics of deep learning. Radiomics is explained in Section 2.4 and the final Section 2.5 explains the medical basics for the diagnosis and treatment of rectal cancer.

2.1 Nuclear Magnetic Resonance

The spin of an atomic nucleus under an external magnetic field B_0 will begin to precess after it is perturbed by an oscillating electromagnetic field. This causes the nucleus to emit a response in the form of characteristic electromagnetic radiation. The frequency depends on the nucleus and the strength of B_0.

Our understanding of this effect started in 1922 with the *Stern-Gerlach* experiment, which demonstrated the quantized nature of the spatial orientation of the angular momentum in silver atoms (Gerlach and Stern, 1922). The first to describe and measure NMR was Rabi in 1938 (Rabi et al., 1938), who observed NMR in a beam of lithium chloride molecules. This discovery earned him the Nobel Prize in 1944.

His work was extended to liquids and solids by Bloch (Bloch, 1946) and Purcell (Purcell et al., 1946) (Nobel Prize 1952), using ^{1}H-nuclei in paraffin. In the 1960s and 1970s, NMR started to be used in chemistry and medicine. This section provides a brief overview of the NMR phenomenon. The explanations in this and the next chapter mostly follow the book by Brown (Brown et al., 2014).

2.1.1 Nuclear Spin

Each elementary particle possesses a spin, which is an intrinsic quantity. The nucleus is made up of protons and neutrons (both of which have a spin of $s = {}^1\!/\!_2$), which are called nucleons. The nuclear spin $\boldsymbol{I}$ is the sum of the spins of the nucleons. If there is an even number of protons and neutrons, all spins cancel each other out. If there is an odd number of both, the spin quantum number I is a positive integer. In case only one of the nucleon types has an odd number, the spin quantum number is a half-integer.

The nuclear spin operator $\boldsymbol{I}$ is a quantum mechanical angular momentum operator and thus has the following commutation properties:

$$\left[\boldsymbol{I}^2, I_a\right] = 0 \qquad [I_a, I_b] = \epsilon_{abc} I_c \qquad a, b, c \in \{x, y, z\} \tag{2.1}$$

where x, y, z are the spatial dimensions and ϵ_{abc} is the Levi-Civita symbol, which is defined as

$$\epsilon_{abc} = \begin{cases} +1 & \text{if } (a, b, c) \text{ is an even permutation of } (x, y, z) \\ -1 & \text{if } (a, b, c) \text{ is an odd permutation of } (x, y, z) \\ 0 & \text{else} \end{cases} . \tag{2.2}$$

Each permutation of two indices results in a sign change. $\boldsymbol{I}^2$ and I_a commute, which means that they can be written in a common eigenbasis with the eigenstates $|I, m_I\rangle$.

Without loss of generality, z is chosen as the quantization axis.

$$\boldsymbol{I}^2 |I, m_I\rangle = I(I+1)\hbar^2 |I, m_I\rangle \qquad I_z |I, m_I\rangle = m_I \hbar |I, m_I\rangle , \tag{2.3}$$

$\hbar$ being the reduced Planck constant $\hbar = {}^{h}/_{2\pi} = 1.05 \times 10^{-34}\,\mathrm{J\,Hz^{-1}}$. m_I can take values between $-I$ and I in integer steps. So, there are $2I + 1$ states. Without an external field, these states are degenerate.

If we only look at ^{1}H, which is the most common hydrogen isotope, the nucleus contains only a single proton. This means $I = {}^1/_2$, so the angular momentum has the two states $m_I = \pm{}^1/_2$. An applied external field interacts with the magnetic moment $\boldsymbol{\mu_I}$ of the nucleus

$$\boldsymbol{\mu_I} = \frac{g_p \mu_p}{\hbar} \boldsymbol{I} = \gamma \boldsymbol{I} \qquad \mu_p = \frac{g_p}{2m_p} \tag{2.4}$$

where $g_p \approx 5.59$ is the gyromagnetic ratio of the proton, $m_p \approx 1.67 \times 10^{-27}\,\mathrm{kg}$ is the mass of the proton. γ is the gyromagnetic ratio of the specific nucleus. For hydrogen in water, the gyromagnetic ratio is $y = 2.68 \times 10^8\,\mathrm{rad\,s^{-1}\,T^{-1}}$, the reduced ratio is $\bar{\gamma} \equiv {}^{\gamma}/_{2\pi} = 42.6\,\mathrm{MHz\,T^{-1}}$ (Brown et al., 2014).

2.1.2 Nuclear Zeeman Interaction

In a magnetic field, the energy levels of quantized magnetic moments (spins) split up, which is called *Zeeman interaction*. There are two types of magnetic moment in an atom, the orbital magnetic moment, and the magnetic moment of the nucleus. In NMR/MRI, we focus on the interaction of the magnetic field with the nucleus.

This means that the Hamiltonian $\mathcal{H}$ in a constant magnetic field $\boldsymbol{B_0} = B_0\boldsymbol{z}$, which is, without loss of generality, applied along the z-axis, is given by

$$\mathcal{H} = -\boldsymbol{\mu_I B_0} = -\gamma \boldsymbol{I B_0} = -\gamma I_z B_0. \tag{2.5}$$

If we use the previous basis $|I, m_I\rangle$ we can calculate the energy contribution of this effect for hydrogen with $I = {}^1\!/_2 \Rightarrow m_I = \pm{}^1\!/_2$

$$\begin{aligned} E_{\pm 1/2} &= \langle I, m_I| \mathcal{H} |I, m_I\rangle = -\gamma B_0 \langle I, m_I| I_z |I, m_I\rangle \\ &= -\gamma m_I \hbar B_0 = \mp\frac{1}{2}\gamma\hbar B_0, \end{aligned} \tag{2.6}$$

so the energy difference between the two states is

$$\Delta E = E_{-1/2} - E_{+1/2} = \gamma\hbar B_0 \tag{2.7}$$

using the Planck relation $E = \hbar\omega$

$$\Delta E = \hbar\omega_L = \gamma\hbar B_0 \quad \Rightarrow \quad \omega_L = \gamma B_0 \tag{2.8}$$

ω_L is also called *Larmor frequency*. The Larmor frequency for a 3 T magnetic field (commonly used in the clinic) is ${}^{\omega_L}\!/_{2\pi} \approx 127.7\,\text{MHz}$.

2.1.3 Precession

Now, if we want to calculate the time evolution of the magnetic moment, we have to solve the Schrödinger equation $i\hbar\frac{\partial}{\partial t}|\Psi\rangle = \mathcal{H}|\Psi\rangle$ for the Hamiltonian given in (2.5). Then, we can calculate the expectation value of the magnetization $\hat{\boldsymbol{\mu}} = \langle\Psi| \boldsymbol{\mu} |\Psi\rangle$. We can write the Hamiltonian by representing the nuclear angular momentum with the Pauli matrices (Pauli, 1927) $I_a = \frac{\hbar}{2}\sigma_a$ as

$$\mathcal{H} = -\gamma I_z B_0 = -\gamma B_0 \frac{\gamma\hbar}{2}\sigma_z = -\hbar\frac{\omega_L}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.9}$$

We use the spinor $|\boldsymbol{\chi}\rangle$, which uses the spin-up $c_+(t)$ and spin-down $c_-(t)$ states as the basis

$$\boldsymbol{\chi} = \begin{pmatrix} c_+(t) \\ c_-(t) \end{pmatrix}. \tag{2.10}$$

In this basis, we get the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\begin{pmatrix} c_+(t) \\ c_-(t) \end{pmatrix} = -\hbar\frac{\omega_L}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} c_+(t) \\ c_-(t) \end{pmatrix}, \tag{2.11}$$

which simplifies to

$$\frac{\partial}{\partial t}c_\pm(t) = \pm i\frac{\omega_L}{2}c_\pm(t). \tag{2.12}$$

The solution is

$$c_\pm(t) = c_\pm(0)e^{\pm i\frac{\omega_L}{2}t}. \tag{2.13}$$

The solution has to be normalized with the condition $\langle\chi|\chi\rangle = 1$, so $|c_+|^2 + |c_-|^2 = 1$. To satisfy this, we choose (without loss of generality) $c_+(0) = \cos\frac{\alpha}{2}$ and $c_-(0) = \sin\frac{\alpha}{2}$. This enables us to calculate the expectation values of the magnetization $\boldsymbol{\mu} = \gamma\boldsymbol{I} = \frac{\gamma\hbar}{2}\boldsymbol{\sigma}$

$$\langle\boldsymbol{\mu}\rangle = \frac{\gamma\hbar}{2}\langle\chi|\boldsymbol{\sigma}|\chi\rangle = \frac{\gamma\hbar}{2}\begin{pmatrix} \sin(\alpha)\cos(\omega_L t) \\ -\sin(\alpha)\sin(\omega_L t) \\ \cos(\alpha) \end{pmatrix}. \tag{2.14}$$

This means that the expectation value of the nuclear magnetization is a vector of magnitude $\gamma\hbar/2$, which spins clockwise around the axis of the magnetic field at the angle α between the magnetization vector and the z-axis, while the rotation is in the x–y plane. The spin is still quantized, so the measured states are just spin-up or spin-down.

2.1.4 Signal Excitation

Now, in addition to $\boldsymbol{B}_0$, we apply a circular polarized radio-frequency (RF) field $\boldsymbol{B}_1(t)$ with the precession frequency ω perpendicular to $\boldsymbol{B}_0$

$$\boldsymbol{B}(t) = B_0\boldsymbol{z} + B_1\left(\cos(\omega t)\boldsymbol{x} - \sin(\omega t)\boldsymbol{y}\right). \tag{2.15}$$

This changes the Hamiltonian to

$$\mathcal{H}(t) = -\frac{\hbar}{2}\boldsymbol{B}(t)\boldsymbol{\sigma} = -\frac{\gamma\hbar}{2}\begin{pmatrix} \omega_L & \omega_1 e^{i\omega t} \\ \omega_1 e^{-i\omega t} & -\omega_L \end{pmatrix}, \tag{2.16}$$

with $\omega_1 = \gamma B_1$. We assume that the excitation frequency ω is the same as the Larmor frequency ω_L. We transform the wave function into a rotating frame of reference, which is similar to the solution of (2.13)

$$c'_\pm(t) = e^{\pm i\omega_L t/2} c_\pm(t) \tag{2.17}$$

and get the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}\begin{pmatrix} c'_+(t) \\ c'_-(t) \end{pmatrix} = -\hbar\frac{\omega_1}{2}\begin{pmatrix} c'_-(t) \\ c'_+(t) \end{pmatrix}. \tag{2.18}$$

We can use the Ansatz $c'_+(t) = A\cos\left(\frac{\omega_1 t}{2}\right) + iB\sin\left(\frac{\omega_1 t}{2}\right)$. It gives

$$c'_-(t) = iA\sin\left(\frac{\omega_1 t}{2}\right) + B\cos\left(\frac{\omega_1 t}{2}\right). \tag{2.19}$$

The normalization condition is $|c'_+|^2 + |c'_-|^2 = 1$, which results in $|A|^2 + |B|^2 = 1$, so we choose $A = \cos\left(\frac{\theta}{2}\right)$ and $B = \sin\left(\frac{\theta}{2}\right)e^{-i\phi}$ (the total phase is irrelevant; only the phase difference ϕ is important). Now we can use this to calculate the expectation value of the magnetization in the rotating frame

$$\langle\boldsymbol{\mu}\rangle = \frac{\gamma\hbar}{2}\langle c'|\boldsymbol{\sigma}|c'\rangle = \frac{\gamma\hbar}{2}\begin{pmatrix} \sin(\theta)\cos(\phi) \\ \cos(\theta)\sin(\omega_1 t) - \sin(\theta)\sin(\phi)\cos(\omega_1 t) \\ \cos(\theta)\cos(\omega_1 t) + \sin(\theta)\sin(\phi)\sin(\omega_1 t) \end{pmatrix}. \tag{2.20}$$

This means that when applying an external magnetic field, oscillating with the Larmor frequency, the magnetization oscillates around the x' axis (in the rotating frame of reference) with the angular frequency $\omega_1 = \gamma B_1$. θ and ϕ determine the initial angle of the magnetic moment. This is what enables us to change the orientation of the spins with an RF field. If we apply a field for the duration of time τ, the magnetic moment is rotated by the angle α

$$\alpha = \int_0^\tau \gamma B_1(t')\,\mathrm{d}t'. \tag{2.21}$$

If the RF field oscillates with constant frequency and magnitude, the spin changes its orientation by the angle $\alpha = \gamma B_1 \tau$. The RF field used to change the angle of the spin is called *RF pulse*, because it is usually applied only for a very short duration.

2.1.5 Macroscopic magnetization

For MRI, we do not investigate an individual nucleus, but rather a volume with a huge number of individual nuclei. So, we only look at the average over many nuclei. In this case, the energy difference between the spin states (2.7) leads to a different number of nuclei in each state. There is a tendency for the spins to align themselves with the magnetic field because this state has lower energy, but the nuclei can also gain energy as a result of interaction with their surroundings. After some time, an equilibrium arises in which the distribution of energies follows a Boltzmann distribution. The probability of an atom being in an eigenstate with energy E_m in an environment with temperature T is

$$p(E_m) = \frac{1}{Z} e^{-\frac{E_m}{k_B T}}, \tag{2.22}$$

with the partition function

$$Z = \sum_m e^{-\frac{E_m}{k_B T}} \tag{2.23}$$

and the Boltzmann constant $k_B \approx 1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$. In the case of hydrogen, we have two eigenstates $m = 1/2 \equiv m_+$ and $m = -1/2 \equiv m_-$. The ratio of the number of atoms ($N_\pm$) in the eigenstates is proportional to the ratio of the probabilities

$$\frac{N_+}{N_-} = \frac{p_+}{p_-} = \frac{\frac{1}{Z} e^{-\frac{E_+}{k_B T}}}{\frac{1}{Z} e^{-\frac{E_-}{k_B T}}} = e^{-\frac{E_+ - E_-}{k_B T}} = e^{\frac{\Delta E}{k_B T}}. \tag{2.24}$$

In the medical field, measurements are made at body temperature $T = 310\,\mathrm{K}$. Magnetic fields are on the order of $\mathcal{O}(1\,\mathrm{T})$, we assume $B_0 = 3\,\mathrm{T}$. According to (2.7), the energy splitting is $\Delta E \approx 530\,\mathrm{\mu eV}$. This means that the thermal energy $k_B T \approx 27\,\mathrm{meV}$ is orders of magnitude higher. So, we can assume

$$\frac{\Delta E}{k_B T} \ll 1 \tag{2.25}$$

and with a Taylor expansion, we get

$$\frac{N_+}{N_-} \approx 1 + \frac{\Delta E}{k_B T} \approx 1 + 2 \times 10^{-5}. \tag{2.26}$$

This means that only 20 ppm more atoms are in the lower energy parallel state (m_+) compared to the higher energy antiparallel state (m_-). *Voxels* (volume pixels) have a volume of the order of $1\,\mathrm{mm}^3$. There are 6.69×10^{19} hydrogen atoms in this volume of water, so we can still assume that we are averaging over many atoms.

We can now calculate the magnetization along the magnetic field $\boldsymbol{M}_0 = M_0\boldsymbol{z}$ with $\boldsymbol{z}$ being the unit vector in z direction

$$M_0 = \rho_0 \sum_{m=-I}^{I} p\left(E_m\right)\mu_z = -\frac{\rho_0}{Z}\sum_{m=-1/2}^{1/2} \gamma\hbar m e^{-\frac{m\Delta E}{k_B T}} \tag{2.27}$$

with the definition of $\tanh$ as

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{2.28}$$

we can write this as

$$M_0 = \rho_0 \frac{\Delta E}{2} \tanh\left(\frac{\Delta E}{2k_B T}\right) \approx \rho_0 \frac{\gamma^2\hbar^2}{4k_B}\frac{B_0}{T} \tag{2.29}$$

with the density $\rho_0 = {}^{N}\!/_{V}$ and using (2.25) for a Taylor expansion of $\tanh$. This means that magnetization increases with the external magnetic field and decreases with temperature, and is proportional to the spin density ρ_0 and the square of the gyromagnetic ratio γ^2.

This is also called Curie's law and determines the signal strength in MR. This makes hydrogen the nucleus with the highest signal in medical MRI because it has a high gyromagnetic ratio γ and a high spin density ρ_0 because hydrogen is the most common element in the human body.

2.1.6 Relaxation

For small voxel sizes, we can assume that the magnetic fields are constant. So, instead of looking at individual atoms, we consider the average magnetization $\boldsymbol{M}$. The initial magnetization is $\boldsymbol{M}_0$.

So far, we have only looked at the interaction of the nuclear spin with the magnetic field. So, without considering any other effects, after applying an RF pulse to rotate the initial magnetization, the spins would continue to precces at that angle.

This does not happen. Due to thermal interaction, the electromagnetic field created by the electrons, and interaction between neighboring nuclei, the magnetization will return to the initial value. The energy gained from the RF pulse is disseminated in the volume and the magnetization returns to $\boldsymbol{M}_0$. This is called relaxation.

This relaxation was first described formally by Bloch (Bloch, 1946). Because we are using the rotating frame of reference, we split the magnetization into a component along the magnetic field $\boldsymbol{M}_{||} = M_z\boldsymbol{z}$ and perpendicular to it $\boldsymbol{M}_{\perp} = M_x\boldsymbol{x} + M_y\boldsymbol{y}$.

T_1 Relaxation

Due to the thermal and electromagnetic interaction, $M_{||}$ will decay with the following differential equation:

$$\frac{\mathrm{d}M_{||}}{\mathrm{d}t} = -\frac{M_{||} - M_0}{T_1}. \tag{2.30}$$

T_1 is called the *longitudinal relaxation* time because it acts on the component of the magnetization parallel to the magnetic field. The process is also called spin-lattice relaxation because early NMR experiments were performed in solids, with a strong interaction with the crystalline lattice. The name is still used in MR experiments of biological tissue, despite the fact that most tissue does not have a crystalline lattice structure.

The solution is an exponential decay of $M_{||}(t)$ toward the equilibrium value M_0

$$M_{||}(t) = M_{||}(0)e^{-\frac{t}{T_1}} + M_0\left(1 - e^{-\frac{t}{T_1}}\right). \tag{2.31}$$

T_1 is an empirical constant and depends on the type of tissue. It is usually quite long and can be multiple seconds. The relaxation time depends on the mobility of the lattice. It is high for fluids and water-based tissue and low for fat (see Tab. 2.1). Paramagnetic substances can decrease greatly T_1 and are therefore used as a contrast agent.

T_2 Relaxation

There are multiple effects that cause the *transversal relaxation*, the relaxation of the magnetization perpendicular to the magnet field $M_{\perp}$. Most of the relaxation is due to a loss of coherence in the spin ensemble due to random phase shifts. The dominant mechanism is the dipole-dipole interaction; this is why it is also called the spin-spin relaxation. There are also higher-order interactions, such as J-coupling or quadrupole interactions, which have less of an effect.

Another effect is the chemical shift anisotropy. A chemical shift can occur when clouds of electrons partly shield the magnetic field, which reduces the Larmor

Tab. 2.1: Representative approximate values of T_1 and T_2 relaxation times for hydrogen components of different tissue at $B_0 = 1.5\,\mathrm{T}$ and $T = 37\,°\mathrm{C}$, taken from Bojorquez et al., 2017 and Brown et al., 2014. There is a wide range of values being reported in the literature for most organs, which can change depending on the patient cohort, the scanner, and the measurement parameters.

Tissue	T_1 (ms)	T_2 (ms)
Cerebrospinal fluid	3800 – 4200	2200
Prostate	1400 – 1700	80
Blood	1200	100 – 200
Muscle	900 – 1500	50
Fat	380 – 450	40 – 140

frequency. This effect depends on the orientation relative to $\boldsymbol{B}_0$ and is stronger for molecules, which cannot rotate freely.

Phase shifts can also occur due to molecular translation, for example caused by diffusion or blood flow. This can change the precession frequency because of magnetic field inhomogeneities, which can be caused, for example, by differences of susceptibility in tissue, magnetic ions or metallic implants. This change in the Larmor frequency leads to phase shift compared to the surrounding nuclei.

All interactions that result in T_1 relaxation also lead to T_2 relaxation. For example, thermal interaction leads to the spin randomly flipping in a different direction. For this reason, T_2 is always shorter than or equal to T_1.

The relaxation of the transversal magnetization in the rotating frame of reference $M'_\perp$ follows the equation:

$$\frac{\mathrm{d}M'_\perp}{\mathrm{d}t} = -\frac{M'_\perp}{T_2}. \tag{2.32}$$

The solution is an exponential decay with a time constant T_2

$$M'_\perp(t) = M'_\perp(0)e^{-\frac{t}{T_2}}. \tag{2.33}$$

The time constant is high in fluids and in the order of tens of milliseconds in most biological tissue (see Tab. 2.1).

T_2^* Relaxation

There are constant inhomogeneities (compared to the timescale of the measurement) in the magnetic field. These are due to imperfections in the B_0 field and susceptibility-induced field distortions. The observed time constant T_2^* is thus

$$\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'} \tag{2.34}$$

with T_2' being the time constant caused by constant inhomogeneities. The dephasing caused by these inhomogeneities is reversible, in contrast to the dephasing due to the T_2 relaxation. This can be done by using an RF pulse to flip the spins 180° between excitation and measurement, causing them to rephase and form an echo (see Section 2.2.2 on page 23).

Bloch Equations

The two differential equations (2.30) and (2.32) can be combined with the torque acting on the magnetization due to the magnetic field to obtain the *Bloch equation*

$$\frac{\mathrm{d}\boldsymbol{M}}{\mathrm{d}t} = \gamma \boldsymbol{M} \times \boldsymbol{B} - \frac{\boldsymbol{M}_{\parallel} - M_0}{T_1} - \frac{\boldsymbol{M}_{\perp}}{T_2} \tag{2.35}$$

or, split up into its spatial components,

$$\frac{\mathrm{d}M_x}{\mathrm{d}t} = \gamma \left(\boldsymbol{M} \times \boldsymbol{B}\right)_x - \frac{M_x}{T_2} \tag{2.36}$$

$$\frac{\mathrm{d}M_y}{\mathrm{d}t} = \gamma \left(\boldsymbol{M} \times \boldsymbol{B}\right)_y - \frac{M_y}{T_2} \tag{2.37}$$

$$\frac{\mathrm{d}M_z}{\mathrm{d}t} = \gamma \left(\boldsymbol{M} \times \boldsymbol{B}\right)_z - \frac{M_z - M_0}{T_1}. \tag{2.38}$$

The equilibrium state is $M_x = M_y = 0$ and $M_z = M_0$. The solution is the same precession as derived in (2.20) with the additional decay of $M_{\parallel}$ (2.31) and $M_{\perp}$ (2.34)

$$\boldsymbol{M}(t) = \begin{pmatrix} M_{\perp}(0)e^{-\frac{t}{T_2}} \cos\left(\omega_L t - \phi\right) \\ M_{\perp}(0)e^{-\frac{t}{T_2}} \sin\left(\omega_L t - \phi\right) \\ M_{\parallel}(0)e^{-\frac{t}{T_1}} + M_0 \left(1 - e^{-\frac{t}{T_1}}\right) \end{pmatrix} \tag{2.39}$$

with ϕ being the initial phase of $M_{\perp}$. In addition to the precession, the initial transverse magnetization $M_{\perp}(0)$ decays with time-constant T_2. The parallel magne-

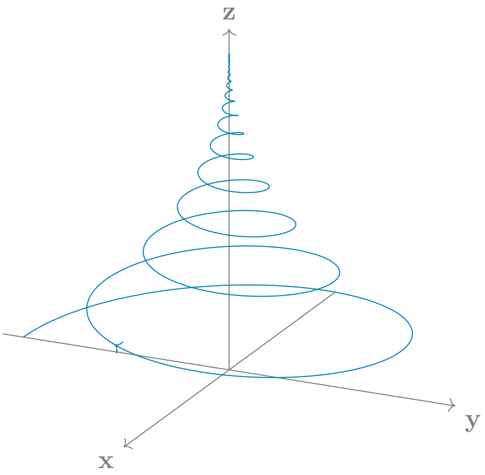


Fig. 2.1: The trajectory of the magnetization over time. The initial magnetization was chosen to be only in the x–y plane ($M_\parallel = 0$). The coordinate system is the static coordinate system. For magnetic fields used in MRI, the cycle duration of the precession is usually a lot shorter than T_1 and T_2, but this was changed for illustrative purposes.

tization returns to the equilibrium magnetization with time-constant T_1. An example of the trajectory can be seen in Fig. 2.1.

2.1.7 Signal acquisition

The precession results in a magnetic field that varies over time. This field will induce a current in the receive coil, which is used to measure the signal. The voltage U_{ind} induced in the coil depends on the change in magnetic flux Φ

$$U_{ind} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t}\int_{S_{coil}} \boldsymbol{B}\,\mathrm{d}\boldsymbol{S}\,. \tag{2.40}$$

Therefore, the induced current depends on the change in magnetic flux through the region enclosed by the coil S_{col}. By introducing the vector potential $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$ and with Stokes theorem, we get the expression

$$\Phi = \int_{S_{coil}} (\boldsymbol{\nabla} \times \boldsymbol{A})\,\mathrm{d}\boldsymbol{S} = \oint \mathrm{d}\boldsymbol{l}\,\boldsymbol{A}. \tag{2.41}$$

We can utilize the principle of reciprocity and express $\boldsymbol{A}$ in terms of the field that the receive coil itself would emit per given unit of current $\boldsymbol{\mathcal{B}} = \boldsymbol{B}_{coil}/I_{coil}$ and get

$$U_{ind} = -\frac{\mathrm{d}}{\mathrm{d}t}\oint \mathrm{d}\boldsymbol{l}\,\boldsymbol{A} = -\frac{\mathrm{d}}{\mathrm{d}t}\int_{V_{Sample}} \boldsymbol{M}(\boldsymbol{r},t)\boldsymbol{\mathcal{B}}_{\boldsymbol{coil}}(\boldsymbol{r})\,\mathrm{d}^3 r\,. \tag{2.42}$$

So we replaced the integral over the surface of the coil with an integral over the volume of the sample V_{Sample} (the volume with non-zero magnetization). For the full derivation, I refer to Brown et al., 2014. The voltage induced in the coil is the opposite of the one driving the current that would result in the same field as the one acting on the coil. This means that the sensitivity of the receive coils depends on the strength of the magnetic field they would emit for each unit of current flowing through them.

This can also cause artifacts in the images. To increase the signal, surface coils are often placed close to the patient. These smaller coils with less surface area have a less uniform $\boldsymbol{B}_{coil}$. Sensitivity (and thus signal-to-noise ratio (SNR)) decreases with depth (Hayes and Axel, 1985). This can lead to the so-called *Surface coil flare*, bright areas on the surface. This can be corrected by measuring the coil sensitivity prior to imaging. Intensity correction can also be performed after imaging (see Section 3.1.3 on page 58).

The signal we measure is proportional to U_{ind}. So for the magnetization given in (2.39) we can measure the following signal:

$$S(t) \propto \int_{V_{Sample}} M_{\perp}(\boldsymbol{r}, 0)\mathcal{B}_{\perp}(\boldsymbol{r})e^{-t/T_2(\boldsymbol{r},t)} \cos\left(\omega_L t + \theta_{\mathcal{B}}(\boldsymbol{r}) - \phi(\boldsymbol{r})\right) \mathrm{d}^3 r \tag{2.43}$$

with $\theta_{\mathcal{B}}(\boldsymbol{r})$ being the angle/phase of the receive coil. This signal is also called the free induction decay (FID). Usually, there are field inhomogeneities present and the signal decays with time constant T_2^* and not T_2. The FID is the response to one excitation without further RF pulses or gradient fields. Only the component perpendicular to the magnetic field can be measured.

If additional gradient fields are used, ω_L depends on the location and time, which can change the frequency and phase of the precession movement. For most measurements, additional RF pulses are also used to change the magnetization. The signal is usually demodulated by multiplying it by a sine and a cosine wave at or near the Larmor frequency ω_L. In this way, the signal can be measured in the rotating frame of reference. The multiplication with a sine and cosine wave results in two channels, which are represented by a complex-valued number measuring the phase and amplitude of the signal. So we get the following complex-valued signal:

$$S(\boldsymbol{r}, t) \propto \int_{V} M_{\perp}(\boldsymbol{r}, 0)\mathcal{B}_{\perp}(\boldsymbol{r})e^{-t/T_2(\boldsymbol{r},t) - i(\theta_{\mathcal{B}}(\boldsymbol{r}) - \phi(\boldsymbol{r}))} \mathrm{d}^3 r \tag{2.44}$$

2.2 Magnetic Resonance Imaging

MRI is a non-invasive imaging technique used in medicine. It is based on NMR. The key to going from NMR to MRI is spatial encoding. In NMR, we only get one signal averaged over the whole volume. In contrast to that, MRI uses gradient fields to change the signal depending on the location, which is used to create 2D and 3D images.

The first published images were generated by Paul Lauterbur in 1973 (Lauterbur, 1973, 1974), who used gradients to spatially encode the signal. He, together with Peter Mansfield, received the Nobel Prize for discoveries concerning MRI in 2003. Mansfield developed the echo-planar imaging (EPI) (Mansfield and Grannell, 1975), which greatly reduced the acquisition times in MRI images. Magnetic resonance imaging has been used in medicine since the 1980s and continues to be used extensively.

2.2.1 Spatial Encoding

A linearly varying field is introduced in addition to the static magnetic field. This is called the gradient field $\boldsymbol{G}$. It is defined by

$$\boldsymbol{G} = \left(\frac{\mathrm{d}B_z}{\mathrm{d}x}, \frac{\mathrm{d}B_z}{\mathrm{d}y}, \frac{\mathrm{d}B_z}{\mathrm{d}z} \right) = (G_x, G_y, G_z) \, . \tag{2.45}$$

The magnetic field still points in the $\boldsymbol{z}$ direction; the gradient fields only change the field strength depending on the location. For spatial encoding, we can use the dependence of the precession frequency on the magnetic field strength

$$\omega = \gamma \boldsymbol{B} = \gamma \left(\boldsymbol{B_0} + \boldsymbol{G r} \right) = \omega_L + \gamma \boldsymbol{G r} = \omega_L + \omega_{\boldsymbol{G}}(\boldsymbol{r}, t). \tag{2.46}$$

These changes in frequency result in a change of phase, which persists even after the gradient field is switched off. The accumulated phase $\phi(\boldsymbol{r}, t)$ when applying a gradient is

$$\phi(\boldsymbol{r}, t) = - \int_0^t \omega_{\boldsymbol{G}}(\boldsymbol{r}, t') \, \mathrm{d}t' = -\gamma \int_0^t \boldsymbol{G}(t') \boldsymbol{r} \, \mathrm{d}t' \, . \tag{2.47}$$

To obtain an image, we have to perform three spatial encoding steps. These are usually slice selection, frequency encoding, and phase encoding. It does not always have to be all three steps; phase encoding can be done in more than one direction, for example, to directly measure a 3D volume instead of multiple 2D slices.

Slice selection

In case of 2D MRI imaging, a slice must first be selected. This can be done by applying a gradient field during the excitation pulse. The gradient is applied along the direction of slice selection (usually z). The thickness of the slice Δz depends on the bandwidth $\Delta\omega_{exc}$ of the RF pulse

$$\Delta z = \frac{\Delta\omega_{exc}}{\gamma G_z}. \tag{2.48}$$

All nuclei in the slice must be excited evenly. To achieve this, a sinc pulse is used. It is an amplitude-modulated RF pulse. The envelope is defined by

$$\mathrm{sinc}\,(\omega t) = \frac{\sin(\omega t)}{\omega t}, \tag{2.49}$$

with $\mathrm{sinc}(0) = 1$. This pulse shape is used because its Fourier transform is a rectangle function.

k-space

We omit the relaxation and assume that the signal is constant during the spatial encoding steps. The receive-coil-dependent phase effects are also omitted. So, we have the signal

$$S(t) = \int_V S_0(\boldsymbol{r}) e^{i\phi(\boldsymbol{r},t)}\,\mathrm{d}^3r = \int_V S_0(\boldsymbol{r}) e^{-i\gamma \int_0^t \boldsymbol{G}(t')\boldsymbol{r}\mathrm{d}t'}\,\mathrm{d}^3r \tag{2.50}$$

after substituting the accumulated phase. If we take (2.50) and use the reduced gyromagnetic ratio $\bar{\gamma}$, we can write the signal as

$$S(t) = \int_V S_0(\boldsymbol{r}) e^{-i2\pi\left(\bar{\gamma}\int_0^t \boldsymbol{G}(t')\boldsymbol{r}\mathrm{d}t'\right)}\,\mathrm{d}^3r\,. \tag{2.51}$$

If we compare this with the definition of the multidimensional Fourier transform of a function $f(\boldsymbol{r})$

$$\mathcal{F}[f(\boldsymbol{r})](\boldsymbol{k}) = \int_{\mathbb{R}^n} f(\boldsymbol{r}) e^{-i2\pi\boldsymbol{k}\boldsymbol{r}}\,\mathrm{d}^n r\,, \tag{2.52}$$

we can see that it is equivalent to $\boldsymbol{k} = \bar{\gamma}\int_0^t \boldsymbol{G}(t')\boldsymbol{r}\,\mathrm{d}t'$. Therefore, the resulting signal is equal to the Fourier transform for a given wave vector $\boldsymbol{k}$

$$S(\boldsymbol{k}) = \int_V S_0(\boldsymbol{r}) e^{-i2\pi\boldsymbol{k}}\,\mathrm{d}^3r = \mathcal{F}[S(t)](\boldsymbol{k}). \tag{2.53}$$

The resulting Fourier space is called the *k-space* in MRI. The measurement is repeated multiple times for different values of $\boldsymbol{k}$ at the time of measurement. In this way, the whole k-space can be sampled. The trajectory through the k-space is defined by the gradients and RF pulses. The magnetization can then be reconstructed with an inverse Fourier transform

$$M(\boldsymbol{r}) \propto \mathcal{F}^{-1}[S(\boldsymbol{k})](\boldsymbol{r}) = \int S(\boldsymbol{k}) e^{2\pi i \boldsymbol{k}\boldsymbol{r}} \mathrm{d}^3\boldsymbol{k}. \tag{2.54}$$

The resulting image (or volume for 3D acquisition) has the lowest frequencies in the center for low absolute values of $\boldsymbol{k}$ and the highest frequencies towards the edges.

Phase Encoding

Phase encoding can be done by applying a gradient between excitation and readout. The simplest method is to apply a gradient $\boldsymbol{G}$ with constant magnitude for time τ_y. So we get if we choose $\boldsymbol{y}$ as the direction of the phase encoding, we get

$$k_y(G_y) = \gamma G_y \tau_y. \tag{2.55}$$

Each nucleus along the y direction will have a different phase. This is repeated N_y times for different combinations of G_y and τ_y. Phase encoding can be done in more than one direction. Generating an image only using phase encoding would be very slow, because only a single point in k-space could be measured at a time, so it is usually combined with frequency encoding.

Frequency Encoding

For the frequency encoding, for example in the x direction, a gradient G_x is applied during signal acquisition following a dephasing gradient. If the acquisition is centered around $t' = 0$, we get for k_x

$$k_x(t') = \gamma G_x t'. \tag{2.56}$$

This looks very similar to phase encoding, but there is a key difference. For frequency encoding, the phase, and wave number k_x changes during acquisition, so multiple points in k-space can be measured in a single data acquisition. This speeds up the process.

In the k-space picture, we select the row we want to measure by selecting a phase k_y. Then we move to the beginning of a line using a dephasing gradient. A line of k-space can then be measured using the frequency encoding. This is only one of many possible trajectories, which can be chosen by varying the phase and frequency encoding. An exemplary sequence is shown in Fig. 2.3.

Resolution and Field of View

The analog-to-digital converter can only sample in a limited time interval δt_x, and only a limited number of phase encoding steps ΔG_y can be chosen. This makes it impossible to continuously sample the k-space. The increments in the k-space are

$$\Delta k_x = \gamma G_x \Delta t \tag{2.57}$$

$$\Delta k_y = \gamma \Delta G_y \tau. \tag{2.58}$$

According to the Nyquist-Shannon theorem, the sampling rate has to be at least twice the highest frequency. So, the resolution in the image space is

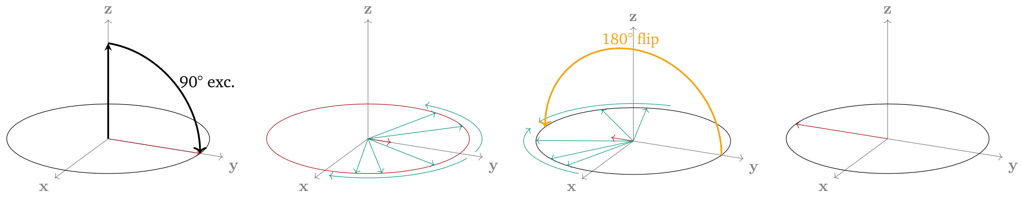
$$\Delta x = \frac{1}{2k_{x,max}} = \frac{1}{N_x \Delta k_x} = \frac{1}{N_x \gamma G_x \Delta t} \tag{2.59}$$

$$\Delta y = \frac{1}{2k_{y,max}} = \frac{1}{N_y \Delta k_y} = \frac{1}{N_y \gamma \Delta G_y \tau}. \tag{2.60}$$

The maximum Field of View (FOV) is thus $\mathrm{FOV}_i = N_i \Delta i = 1/\Delta k_i$. This means that the resolution and the FOV depend on the number of samples in k-space and their bandwidth. There is a trade-off between high resolution and a large field of view.

Artifacts

Spatial encoding can also lead to artifacts. Due to the complicated imaging process, there are many MRI artifacts that must be corrected or taken into account when interpreting MRI images. An example is the chemical shift. The molecular environment can change the magnetic field experienced by the nucleus. In fat, the proton is shielded by the electron clouds of the glycerides, reducing the magnetic field slightly and decreasing the Larmor frequency (by 430 Hz at 3 T). Thus, the fat appears in a position opposite to the frequency-encoding direction for standard sequences. The distance depends on the bandwidth per pixel. This leads to a void of signal on the interface between fatty and non-fatty tissue on one side and an overlap of signal on the other side.



(a) Excitation **(b)** Dephasing **(c)** Rephasing **(d)** Echo

Fig. 2.2: This visualizes the dephasing and rephasing in the spin echo sequence. The first step is the excitation in (a). The spin is flipped by the excitation pulse and starts to precess. Because of inhomogeneities, the spins will start to dephase and the average magnetization (in red) is reduced, as seen in (b). After waiting for $T_E/2$, a 180° pulse is applied to reverse the spins. The phases continue to change in the same direction, as seen in (c). For non or slow-varying inhomogeneities, the dephasing will be reversed. Another $T_E/2$ later, the spins are in phase again and there is a peak in the signal, which is called an echo (d).

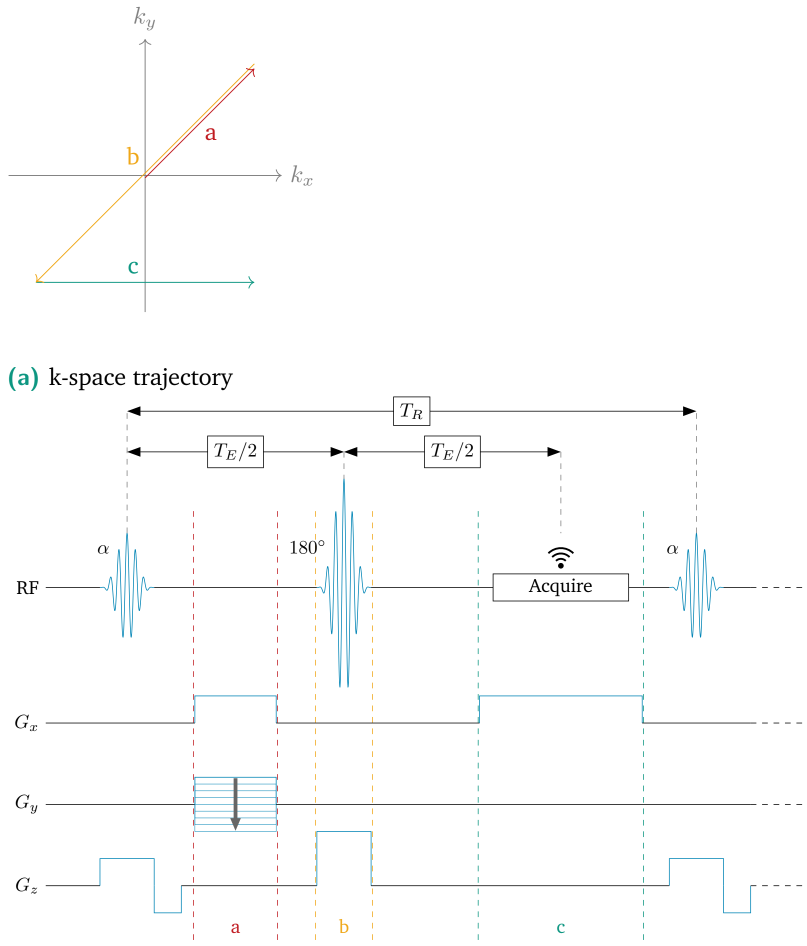
An important group of artifacts are *motion artifacts*. They are caused by either periodic motion, such as the heartbeat and breathing, or by non-periodic motion, such as the relaxation of muscles after laying down for a while. Non-periodic motion mainly causes a loss of signal intensity. Periodic motion can lead to discrete ghosts, because it modulates the signal to have discrete side bands, while non-periodic motion causes diffuse, unfocused artifacts.

2.2.2 T_2 weighted Imaging

There are many sequences for recording MRI images. They define the timing, strength, and phases of gradients, RF pulses, and readout windows. The contrast depends on the sequence used. The sequences shown here are greatly simplified. In reality, there are more effects that have to be corrected for. In addition, many techniques are used to speed up the measurement. For example, multiple slices can be measured at the same time by exciting and measuring other slices while waiting for relaxation in one slice.

Spin Echo

One of the standard sequences is the spin-echo sequence. It can be seen in Fig. 2.3. As described in Section 2.1.6 on page 16, we can only measure T_2^* using a FID signal. T_2^* depends not only on the tissue properties but also on the inhomogeneities in the magnetic field. We want to measure T_2, because it is independent of the measuring device.



(a) k-space trajectory

(b) Sequence Diagram

Fig. 2.3: Image (b) shows the sequence diagram of a spin-echo sequence with arbitrary timescale (adapted from Brown et al., 2014). The second image (a) shows the trajectory in k-space. First, the spins are excited with an RF pulse to flip the magnetization by an angle α while the slice selection gradient G_z is applied along the z-direction.

After the excitation, a rephasing gradient is applied. Then, in a, a dephasing gradient is applied along the x direction and a phase gradient along the y direction. In k-space, we move along the arrow a. Along k_x to the right due to the dephasing gradient G_x and along k_y to the top or bottom, depending on the magnitude of G_y.

Then ($T_E/2$ after excitation, the magnetization is flipped by 180° with another RF pulse. This results in a point reflection, seen in b, with respect to the center in k-space. Therefore, no rephasing pulse is needed, because it will re-phase itself.

Acquisition is T_E after excitation and $T_E/2$ after flip. The frequency encoding gradient G_x has the same direction as the dephasing gradient, due to the 180° flip. Due to the gradient that is applied during data acquisition, we move along the k_x axis in k-space, as seen in c, and a whole line is recorded at once.

Then, T_R after the first excitation, there is another excitation and the next line in k-space is measured. In this way, the whole k-space is measured at a regular interval.

The trick to recover T_2 is to use an RF pulse to flip the magnetization by 180°. The spin-flip results in zero accumulated relative phases due to inhomogeneities at time T_E after excitation, if the spins are flipped by 180° after $T_E/2$. The resulting signal is called an echo, and the time T_E, after which this echo occurs, is called echo time (TE). The principle is visualized in Fig. 2.2.

After the echo, another line in the k-space can be recorded after waiting for the magnetization to return to equilibrium. The time T_R between these two measurements is called repetition time (TR).

It is also possible to record more than one echo by adding more 180° pulses. With a Turbo Spin Echo (TSE)/Fast Spin Echo (FSE) sequence (commercial implementations of the RARE (Hennig et al., 1986) technique), multiple lines of the k-space can be measured in one T_R. The phase is varied between different echos. A T_2 weighted image recorded with a TSE sequence can be seen in Fig. 2.5a.

In the k-space picture (see Fig. 2.3a), we also take the steps a, b and c. But after measuring one line, we do not wait for the next measurement. Instead, we perform another phase encoding to move along k_y to the next line we want to record and acquire the signal in the opposite direction along k_x.

Half-Fourier Acquisition Single-shot Turbo spin Echo (HASTE) is an echo-planar fast spin echo sequence, which was trademarked by Siemens. In this sequence, all k-space data are collected using only one shot using a long echo train and only partially sampling the k-space. The advantage is that it is very fast, which can help reduce motion artifacts.

2.2.3 Diffusion-weighted imaging

diffusion-weighted (DW) images can be used to measure the diffusion of nuclei. This can be done using a pulsed gradient spin echo (PGSE) sequence developed by Edward Stejskal and John Tanner and published in 1965 (Stejskal and Tanner, 1965) (see Fig. 2.4). It is very similar to the spin-echo sequence, but two additional gradients are added to dephase and rephase the spins. Nuclei that move during measurement will not return to their original phase and will have a reduced signal.

The advantage is that in this way some information about the cellular structure can be obtained, which cannot be resolved otherwise in the MRI images. This can help identify cancerous tissue (van Heeswijk et al., 2016).

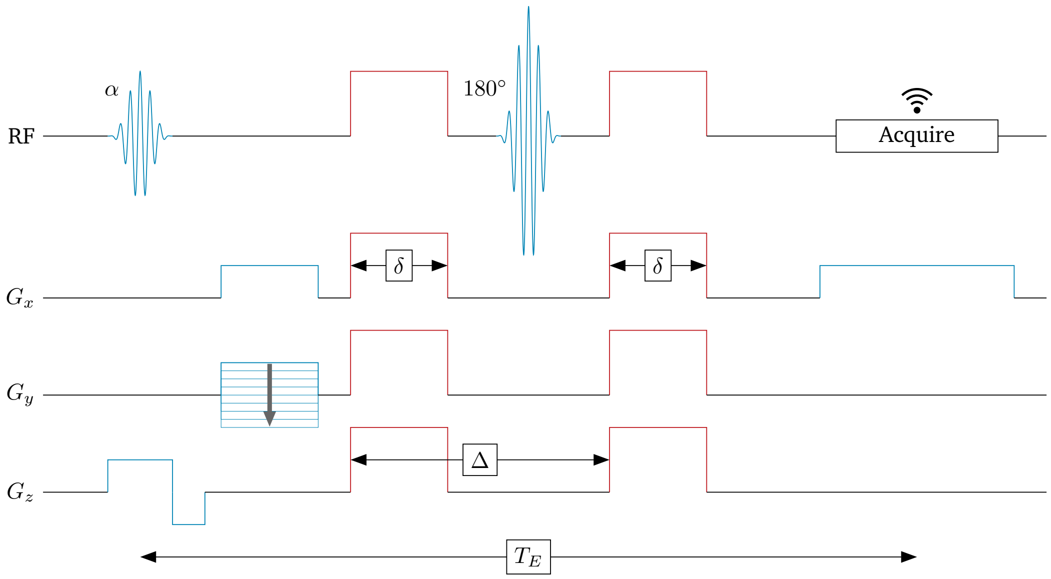


Fig. 2.4: The sequence diagram of the diffusion sequence. It is very similar to the spin echo sequence. The main difference is that additional diffusion gradients are added (in red). They are applied before and after the 180° RF pulse for the time δ and are Δ apart.

Basics of Diffusion

The diffusion process can be modeled as a random walk. In 1D, we assume that the atom jumps to a new position every τ_d with a step size l_d in a random direction $\epsilon_i = \pm 1$. Therefore, the position of a particle starting at $x = 0$ after N steps is

$$x_N = l_d \sum_{i=1}^{N} \epsilon_i. \tag{2.61}$$

There is an equal probability of jumping into the positive and negative direction. This means $\langle \epsilon_i \rangle = 0$, so $\langle x_N \rangle = 0$. The expected squared displacement $\langle x_N^2 \rangle$ is

$$\left\langle x_N^2 \right\rangle = l_d^2 \sum_{k=1}^{N} \sum_{j=1}^{N} \langle \epsilon_k \epsilon_j \rangle \,. \tag{2.62}$$

$\langle \epsilon_i \epsilon_j \rangle = \delta_{ij}$, with δ_{ij} being the Kronecker delta, which is one if $i = j$ and is otherwise zero. This is because the possible values are 1 and -1 with equal probability, so the expectation value is zero. For $i = j$, $\epsilon_i, \epsilon_j = \epsilon_i^2 = (\pm 1)^2 = 1$. So we get:

$$\left\langle x_N^2 \right\rangle = l_d^2 \sum_{k=1}^{N} \sum_{j=1}^{N} \delta_{ij} = l_d^2 \sum_{k=1}^{N} 1 = l_d^2 N. \tag{2.63}$$

Now we can use the diffusion coefficient $D = {}^{l_d^2}/_{2\tau_d}$ as derived by Albert Einstein (Einstein, 1905) and write the expected squared displacement as $\langle x(t)^2 \rangle = \frac{l_d^2}{\tau_d} t = 2Dt$ with time $t = N\tau_d$.

Response to pulse sequence

The probability of finding a particle with displacement x after time t is a normal distribution due to the central limit theorem because we are averaging a vast amount of randomly moving particles. With the mean and variance are the $\langle x \rangle$ and $\langle x(t)^2 \rangle$ we just derived, we get the probability $P(x, t)$ for a displacement x after time t of

$$P(x,t) = \frac{e^{-\frac{1}{2}\frac{x^2}{\langle x^2 \rangle}}}{\sqrt{2\pi \langle x(t)^2 \rangle}} = \frac{e^{-\frac{x^2}{4Dt}}}{\sqrt{4\pi Dt}}. \tag{2.64}$$

Now, we can use (2.47) to calculate the accumulated phase. For the first diffusion gradient of length δ, we get the phase $\phi_1 = -\gamma\delta G x$ assuming that the gradient G is constant over time δ. Due to the 180° pulse before the second diffusion gradient is switched on, the second phase ϕ_2 shift is $\phi_2(x) = \gamma\delta G x$.

With Δ being the time difference between the rising edges of the first and second gradients. If $\delta \ll \Delta$, we can assume that diffusion in time Δ between the two pulses dominates the diffusion process and neglects the diffusion processes during the pulses. The phase shift $\Delta\phi$ is thus

$$\Delta\phi = \phi_1(x_1) + \phi_2(x_2) = \gamma\delta G(x_2 - x_1), \tag{2.65}$$

if the particle moves from position x_1 to x_2 due to diffusion during time Δ. This expression is invariant to translation; only displacement due to diffusion is relevant. This means that we can use the probability $P(x, \Delta)$ as in (2.64) to obtain a displacement $x = x_2 - x_1$ after time Δ. Therefore, each spin has the complex magnetization amplitude

$$M_d = e^{i\Delta\phi} M_0. \tag{2.66}$$

The reduction in signal due to diffusion $\langle {}^{S}/_{S_0} \rangle$ is equal to the reduction in magnetization. Its expectation value is

$$\begin{aligned} \left\langle \frac{S}{S_0} \right\rangle = \left\langle \frac{M_d}{M_0} \right\rangle = \int_{-\infty}^{\infty} e^{i\gamma\delta G x} P(x, \Delta)\, \mathrm{d}x = \frac{1}{\sqrt{4\pi D\Delta}} \int_{-\infty}^{\infty} e^{i\gamma\delta G x} e^{-\frac{x^2}{4Dt}}\, \mathrm{d}x \\ = e^{-\gamma^2\delta^2 G^2 D\Delta}. \end{aligned} \tag{2.67}$$

This is also often written as $\langle {}^{S}/_{S_0} \rangle = e^{-bD}$ with $b = \gamma^2 \delta^2 G^2 \Delta$. b is called the *b-value* and is used to measure the strength and timing of gradients and is an important parameter in diffusion sequences. In modern scanners, the values can range from 0 to as high as 3000 s mm^{-2} (Han et al., 2015).

In reality, it is not possible to fulfill the condition $\delta \ll \Delta$. The expression for arbitrary values of δ is derived by Stejskal and Tanner (Stejskal and Tanner, 1965). The result is a reduced b-value of

$$b = \gamma^2 \delta^2 G^2 \left(\Delta - \frac{1}{3}\delta \right). \tag{2.68}$$

The recording of the positions (in the form of a phase shift) is less sharp, because the phase accumulates over a longer time period, so the signal decreases is reduced. For $\delta \to 0$, we recover our initial expression.

We have assumed, so far, that the particles can move without restriction. This is not the case in biological tissue. The diffusion is limited by obstacles such as cell membranes or macromolecules. This also often makes the diffusion anisotropic. Tissue with high anisotropy on a macroscopic scale is, for example, skeletal muscles or parallel nerve fibers in the white matter of the brain. Therefore, the diffusion coefficient is replaced by the effective or apparent diffusion coefficient (ADC) D^{eff}. The apparent diffusion coefficient depends on the direction.

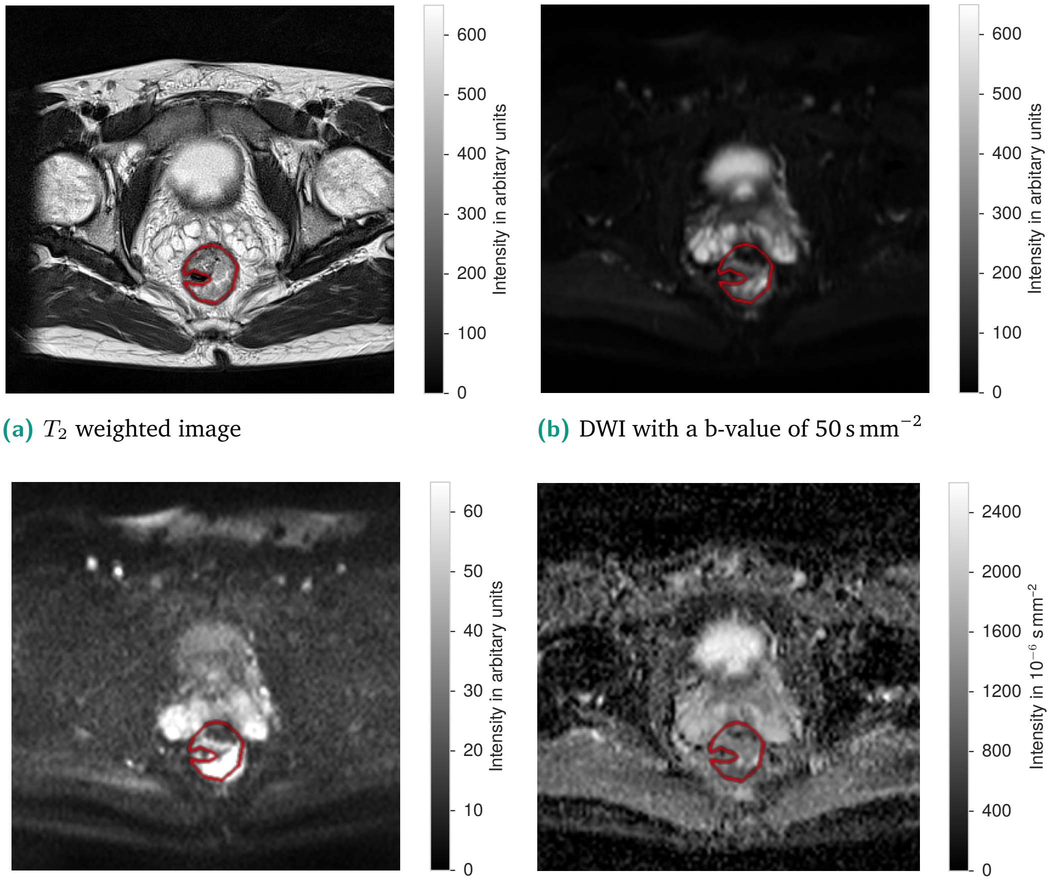
Additional Diffusion Models

Basser et al. developed a formalism (Basser et al., 1994), which extends the approach of Stejskal and Tanner. The signal reduction is given by

$$S/S_0 = \exp\left(-\sum_{i=i}^{i} \sum_{i=j}^{j} B_{ij} D_{ij}^{eff} \right) \tag{2.69}$$

$$B = \begin{bmatrix} B_{xx} & B_{xy} & B_{xz} \\ B_{yx} & B_{yy} & B_{yz} \\ B_{zx} & B_{zy} & B_{zz} \end{bmatrix}, \quad D^{eff} = \begin{bmatrix} D_{xx}^{eff} & D_{xy}^{eff} & D_{xz}^{eff} \\ D_{yx}^{eff} & D_{yy}^{eff} & D_{yz}^{eff} \\ D_{zx}^{eff} & D_{zy}^{eff} & D_{zz}^{eff} \end{bmatrix}.$$

B and D^{eff} are two 3×3 tensors. B_{ij} describes the strength and duration of the two gradient pulses in the i and j directions. If, as in our initial example, the gradients are in the x-direction, only B_{xx} is nonzero and we measure D_{xx}^{eff}. Similarly, all tensor elements can be measured. This is also called diffusion tensor imaging and is used, for example, to measure brain connectivity (Skudlarski et al., 2008).


600
500
400
300
200
100
0
Intensity in arbitary units
600
500
400
300
200
100
0
Intensity in arbitary units
60
50
40
30
20
10
0
Intensity in arbitary units
2400
2000
1600
1200
800
400
0
Intensity in 10^{-6} s mm^{-2}

(a) T_2 weighted image

(b) DWI with a b-value of 50 s mm^{-2}

(c) DWI with a b-value of 800 s mm^{-2}

(d) ADC image

Fig. 2.5: Example images for the different modalities. (a) shows a T_2 weighted image. A radiologist segmented the tumor in that image (outlined in red). The same patient was imaged with diffusion-weighted imaging.
In (b), the b50 image is visible with the same intensity scale as the T_2 weighted image. It is used instead of a b0 image to reduce the influence of perfusion and suppress the signal of the blood vessels.
In the b800 image (c), the signal is greatly reduced for all tissue (note the reduced scale), but parts of the tumor are visible with a higher intensity compared to the surrounding tissue.
The last image (d) shows a map of the ADC, there, the tumor tissue has a lower value compared to the surrounding tissue because the high intensity in high b-value images correlates with a low ADC. This can also be used to help diagnose rectal cancer (Horvat, Carlos Tavares Rocha, et al., 2019). The bladder has the highest ADC value because there are no obstacles that hinder diffusion there.

The ADC can refer to different things. In addition to D^{eff}, it is also often used to refer to the average of the diagonal elements of D^{eff} and is the same as the average of the eigenvalues (Minati and Wglarz, 2007)

$$\text{ADC} = \frac{\text{Trace}\left(D^{eff}\right)}{3} = \frac{D_{xx}^{eff} + D_{yy}^{eff} + D_{zz}^{eff}}{3}. \tag{2.70}$$

A DW image refers to an image weighted with the apparent diffusion coefficient (see Fig. 2.5). The image is also still T_2 weighted, so in an DW image, we see the combined effect of DW and T_2 weighting.

$$S = S_0 e^{-b\text{ADC}} \tag{2.71}$$

at a certain b-value. S_0 is the signal without any diffusion weighting. This can be measured by taking multiple images with different gradient directions. At least three perpendicular gradients have to be used, but more directions are often used to reduce artifacts and noise.

If at least two DWI images are measured at different b-values, the ADC can be calculated by performing an exponential fit (or a linear fit in log space). In this way, an ADC map can be created.

However, this model is only valid for the medium range of b-values. For lower b-values, there are additional perfusion effects that increase the measured ADC values, due to blood microcirculation (up to approximately $200\,\text{s}\,\text{mm}^{-2}$) (Iima and Le Bihan, 2016) for higher b-values ($> 1000\,\text{s}\,\text{mm}^{-2}$), the ADC is lower than expected from the diffusion model, because obstacles that limit movement and other effects that limit diffusion (Jensen and Helpern, 2010). This is called kurtosis because the distribution of particle displacement differs from the normal distribution.

Imaging Artifacts

Diffusion imaging usually uses EPI, which can be used to image a slice in just one echo (or, for better SNR, a few echos), which is very sensitive to inhomogeneities in the magnetic field.

To achieve high b-values, strong gradients are needed. There are multiple problems with this. Switching high gradients can induce currents in the material in and near the bore. Currents can be induced in conductive surfaces in the scanner, called *eddy currents* (Le Bihan et al., 2006). Eddy currents can even persist after the gradient is turned off, leading to inhomogeneities in the magnetic field. This can lead to geometric distortions. This can be partially compensated for by calibrating the MRI machine and correcting for eddy currents (Rohde et al., 2004). But it is not always possible to completely correct for the eddy currents.

Another major cause of artifacts is the motion of the patient. Due to the long acquisition times, patient motion is a big problem for MRI in general. DWI is especially affected. Even small changes in the location of the patient can cause a significant loss of signal. The spatial shifts due to the motion of the patient can be magnitudes larger than those due to diffusion.

In the abdomen, the main causes are breathing and heartbeat. This is one of the main reasons why EPI is used. If a 2D image can be acquired in one shot (usually 100 ms), there are fewer motion artifacts. The drawback is the reduced spatial resolution (Le Bihan et al., 2006).

2.3 Deep Learning

Deep learning is a part of machine learning. It uses artificial Neural Networks (NNs) to approximate functions by adjusting the weights of the neurons. It is called deep because the networks have multiple layers (up to hundreds or thousands). It is inspired by biological neurons (but works very differently from a biological neuron and is greatly simplified). Deep learning has been successfully used in many fields, such as computer vision, natural language processing, speech recognition, but also protein folding, and high-energy physics.

It is an optimization problem with the goal of finding a mapping between the multidimensional input (such as an image) and the usually low-dimensional output (for example the cancer stage) using NNs. Instead of telling the computer exactly what steps to take, only the structure of the information flow through the neurons (called *Architecture*) is defined, and the weights are adjusted to minimize the cost function. A neuron is visible in Fig. 2.6. The weights are adjusted in incremental steps by a process called *backpropagation*, which is used to calculate the gradients of the weights with respect to the cost function.

The field started in the 19th century, when neurons were first described and Cajal postulated that the neural network is made up of independent cells (López-Muñoz et al., 2006). The dynamics of the action potential was described by Hodgkin and Huxley in 1952 (Hodgkin and Huxley, 1952). The first computational model was developed in 1943 (McCulloch and Pitts, 1943) and implemented by Rosenblatt in

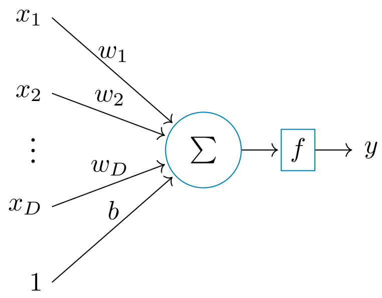


Fig. 2.6: Here, a single artificial neuron is shown. It has inputs x_1 to x_D, which are multiplied by the weights w_1 to W_D and summed up. The bias b is also added. Then, the activation function f is applied to get the output $y = f(\sum_{i=0}^{D} w_i + b)$.

1958 (Rosenblatt, 1958). But at that time, computing power was not sufficient to run large neural networks. There was slow progress in the 80s and 90s, with the first convolutional neural network (CNN) in 1979 (Fukushima, 1980). This was further developed, resulting in LeNet by LeCun (LeCun et al., 1989), which was the starting point for many modern neural networks for computer vision.

However, despite these advances, neural networks still did not outperform other machine learning techniques, such as support vector machines. The breakthrough in neural networks was achieved by a combination of multiple factors at the beginning of the 2010s. Although CNNs and backpropagation had been around for a while (the term backpropagation was made popular in 1986 Rumelhart et al., 1986) and the use of graphical processing units (GPUs) to speed up neural networks was not new (Oh and Jung, 2004), the combination of everything led to significant advances in deep learning. Another factor was larger and larger datasets, which were available for training. A very famous one is ImageNet, which contains many natural images and serves as a benchmark for classification performance (Russakovsky et al., 2015).

CNNs received a lot of attention when AlexNet won the ImageNet challenge with a top-5 error of 15 % (more than 10 % better than the second place) in 2012 (Krizhevsky et al., 2017). At the same time, other advances in speech recognition and computer vision occurred (Ciregan et al., 2012; L. Deng et al., 2013; Dahl et al., 2014), which sparked a resurgence of interest in the field. Since then, many new networks have been proposed and successfully applied in various fields and continue to make amazing progress, such as predicting the structure of proteins (Jumper et al., 2021) or having conversations (W. X. Zhao et al., 2023).

With all the advances, the basic building blocks have remained mostly the same. These are explained in this section. It follows the excellent book by Goodfellow, Bengio and Courville (Goodfellow, Bengio, et al., 2016).

2.3.1 Optimization

The goal of optimization is to reduce the *cost function* $\boldsymbol{J}(\boldsymbol{\theta})$ by changing the parameters $\boldsymbol{\theta}$. $\boldsymbol{\theta}$ refers to the weights and biases of the neurons (and other learned parameters in the network). There are also the *hyperparameters*, which are not learned parameters (for example, the number of layers or the number of neurons per layer), which are not included in $\boldsymbol{\theta}$. The hyperparameters need to be optimized separately. This can be done either by hand or by a (semi-)automatic optimization scheme.

Without regularization, the cost function can be written as an average of the loss function per-example L

$$\boldsymbol{J}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\hat{p}} L(f(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{y}). \tag{2.72}$$

$f(\boldsymbol{x};\boldsymbol{\theta})$ is the predicted output, $\hat{p}$ the empirical distribution and $\boldsymbol{y}$ the output of the NN. We want to reduce the expected generalization error, which is also called risk. The problem is that we do not know the true distribution of the data $\hat{p}$, so we have to reduce empirical risk, using an estimated distribution $\hat{p}_{\text{data}}$.

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\hat{p}_{\text{data}}} L(f(\boldsymbol{x};\boldsymbol{\theta}), \boldsymbol{y}) = \frac{1}{m}\sum_{i=1}^{m} L\left(f\left(\boldsymbol{x}^{(i)};\boldsymbol{\theta}\right), \boldsymbol{y}^{(i)}\right) \tag{2.73}$$

Empirical risk is prone to overfitting because the model could memorize the training set. Often, the NN has more parameters than the number of examples in the training set.

Therefore, a different loss is used, which has to be differentiable, for example the log-likelihood. One difference from conventional optimization is also that the algorithm does not necessarily continue until a (local) minimum is reached. There might not even be a local minimum. Instead, the network is trained for a predetermined number of iterations or until some other predefined condition is reached.

The optimization algorithm usually does not run on the complete dataset; it is applied in *batches*. This works because the loss can be separated into individual terms. The standard error of a sample is given by $\frac{\sigma}{\sqrt{n}}$, so using bigger batches has a limited return compared to the increased computational cost. Batches are selected randomly to avoid correlated samples. The gradient can be estimated using a batch with m examples

$$\hat{\boldsymbol{g}} = \frac{1}{m}\nabla_{\boldsymbol{\theta}}\sum_{i=1}^{m} L\left(f\left(\boldsymbol{x}^{(i)};\boldsymbol{\theta}\right), \boldsymbol{y}^{(i)}\right). \tag{2.74}$$

To calculate the gradient, the derivative of the loss with respect to the network parameters $\nabla_{\boldsymbol{\theta}} \boldsymbol{L}\left(f\left(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}\right), \boldsymbol{y}^{(i)}\right)$ needs to be calculated with the help of a technique called backpropagation.

Backpropagation

When input flows forward through a feedforward NN, it is called *forward propagation*. The computation results in the cost function. The backpropagation algorithm is then used to transmit information about the cost back through the network.

Gradients can then be calculated using the chain rule. For the functions $\boldsymbol{y} = g(\boldsymbol{x})$ and $z = f(g(\boldsymbol{x})) = f(\boldsymbol{y})$, this is simple to calculate.

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_i} \quad \Leftrightarrow \quad \nabla_{\boldsymbol{x}} z = \left(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right)^{\top} \nabla_{\boldsymbol{y}} z \tag{2.75}$$

Therefore, the gradient of the output of a function can be calculated with backpropagation by multiplying the Jacobian matrix by the gradient of the input of the function. Starting from the loss, this calculation is repeated until the gradients of all parameters are calculated. The Jacobian is usually calculated analytically.

The basic idea of the algorithm is fairly simple, although a fast numerical implementation for millions and billions of parameters and operations is challenging. Another problem is the memory required to store all the derivatives.

Stochastic gradient descent (SGD)

Stochastic gradient descent (SGD) is one of the most used optimization algorithms in machine learning. The basic idea was presented by Robbins and Monro (Robbins and Monro, 1951). It is similar to the gradient descent, which was first presented in Cauchy, 1847 and was studied for nonlinear problems by Curry (Curry, 1944). It is called stochastic because the gradient is just an approximation, because only a batch of data is used to estimate it instead of the full dataset.

An estimate of the gradient is taken by averaging the gradient of a batch, as in (2.74). The parameters are then updated by $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \hat{\boldsymbol{g}}$ using the learning rate ϵ. It is crucial to choose the right learning rate. The learning rate is one of the most important hyperparameters. It does not have to be constant.

The stochastic nature of SGD introduces some noise that does not vanish even at a minimum. To guarantee convergence, two conditions must be met.

$$\sum_{k=1}^{\infty} \epsilon_k = \infty \text{ and } \sum_{k=1}^{\infty} \epsilon_k^2 < \infty \tag{2.76}$$

Therefore, it is common to reduce (decay) the learning rate according to a decay schedule. For example, using a constant learning rate at first and decreasing the learning rate each step or when the loss stops decreasing.

SGD can be slow, especially under a lot of noise. A way to make learning faster is to introduce momentum (as suggested in Rumelhart et al., 1986) by accumulating an exponentially decaying moving average of past gradients. If the mass is seen as a unit mass, this is equal to the velocity $\boldsymbol{v}$. The parameters are updated using the velocity

$$\begin{aligned} &\boldsymbol{v} \leftarrow \alpha \boldsymbol{v} - \epsilon \nabla_{\boldsymbol{\theta}} \left(\frac{1}{m} \sum_{i=1}^{m} L\left(f\left(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}\right), \boldsymbol{y}^{(i)}\right)\right) = \alpha \boldsymbol{v} - \epsilon \boldsymbol{g}, \\ &\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{v}. \end{aligned} \tag{2.77}$$

α regulates the speed of decay. A higher value means that the past gradients are taken more into account. Thus, the movement of the parameters through the optimization space is similar to that of an accelerated mass with viscous drag.

A small modification is to use Nesterov momentum (Nesterov, 1983). The gradient is calculated at a slightly different position, with the momentum change already applied to the parameters. This increases the speed of convergence.

$$\boldsymbol{v} \leftarrow \alpha \boldsymbol{v} - \epsilon \nabla_{\boldsymbol{\theta}} \left(\frac{1}{m} \sum_{i=1}^{m} L\left(f\left(\boldsymbol{x}^{(i)}; \boldsymbol{\theta} + \alpha \boldsymbol{v}\right), \boldsymbol{y}^{(i)}\right)\right) \tag{2.78}$$

An algorithm often used is *Adam*, which is derived from “adaptive moments”. It uses Nesterov momentum combined with an adaptive learning rate that scales the learning rate of each parameter with the inverse of the square root of the sum of historical square gradients (Kingma and Ba, 2014). The momentum is incorporated directly into the estimate of the gradient. The momentum is added to the rescaled gradients. A bias correction is also added to account for initialization at the origin. The algorithm can be seen in Algorithm 1.

There are many optimizers, and according to the benchmarks (Schmidt et al., 2020), there is no best optimizer for a variety of different problems, but some optimizers (such as Adam) are consistent in a wide range of different problems.

Algorithm 1: The Adam algorithm. It follows the main idea of SGD, but the learning rate is adapted using the first and second momentum of the gradient. Both moments are initialized with the value 0, which introduces a bias. A step is added for both moments to correct for this bias. Adapted from Goodfellow, Bengio, et al., 2016; Kingma and Ba, 2014.

```
Require: $\epsilon$: Step size
Require: $\rho_1, \rho_2 \in [0, 1)$ decay rates for moment estimates
Require: $\delta$: small constant for numerical stability
Require: $\boldsymbol{\theta}$: parameters
  $t \leftarrow 0$                                                    ▷ timestep
  $s_0 \leftarrow 0$                                                  ▷ 1st moment variable
  $r_0 \leftarrow 0$                                                  ▷ 2nd moment variable
  while stopping criterion not met do
    $t \leftarrow t + 1$
    $\boldsymbol{g}_t \leftarrow \nabla_{\boldsymbol{\theta}} \left( \frac{1}{m} \sum_{i=1}^{m} L \left( f \left( \boldsymbol{x}^{(i)}; \boldsymbol{\theta}_{t-1} \right), \boldsymbol{y}^{(i)} \right) \right)$    ▷ calulate gradient
    $\boldsymbol{s}_t \leftarrow \rho_1 \boldsymbol{s}_{t-1} + (1 - \rho_1) \boldsymbol{g}_t$     ▷ update first moment estimate
    $\boldsymbol{r}_t \leftarrow \rho_2 \boldsymbol{r} + (1 - \rho_2) \boldsymbol{g} \odot \boldsymbol{g}$     ▷ update second moment estimate
    $\hat{\boldsymbol{s}}_t \leftarrow \frac{\boldsymbol{s}_t}{1 - \rho_1^t}$              ▷ correct bias in first moment estimate
    $\hat{\boldsymbol{r}}_t \leftarrow \frac{\boldsymbol{r}_t}{1 - \rho_2^t}$              ▷ correct bias in second moment estimate
    $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \epsilon \cdot \hat{\boldsymbol{s}}_t / \left( \sqrt{\hat{\boldsymbol{r}}_t} + \delta \right)$     ▷ update parameters
  end while
```

Batch Normalization

Batch normalization is an adaptive reparameterization method (Ioffe and Szegedy, 2015). Correction of parameters using a gradient is a linear operation, but the multiplication of multiple weights is non-linear. This makes the effect of a small update unpredictable.

Batch normalization can be used to reparameterize a layer in NNs. If $\boldsymbol{H}$ is a batch of activations of a layer, it is normalized using the variance and mean of the input

$$\boldsymbol{H}' = \frac{\boldsymbol{H} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{2.79}$$

$$\boldsymbol{\mu} = \frac{1}{m} \sum_i \boldsymbol{H}_i; \quad \boldsymbol{\sigma} = \sqrt{\delta + \frac{1}{m} \sum_i (\boldsymbol{H} - \boldsymbol{\mu})_i^2},$$

averaged over one batch. δ is a small numerical constant to prevent undefined gradients. Batch normalization can be included before or after any layer and is backpropagated through when calculating the gradient.

One modification to preserve the expressive power of the network is to us $\gamma \boldsymbol{H}' + \boldsymbol{\beta}$ instead of $\boldsymbol{H}'$ and treat γ and $\boldsymbol{\beta}$ as learned parameters. This means that mean and variance are discrete parameters rather than hidden somewhere in the weights of the layers below, making them much easier to learn.

2.3.2 Loss Function

The exact loss function depends on the problem at hand and the kind of learning that is being performed. In *supervised learning*, there is a *ground truth* for each example that can be compared with the predicted result. A human annotator usually provides the ground truth. For *unsupervised learning*, there is no ground truth. A loss is still needed for optimization. For an autoencoder (see Section 3.2.3), this can be a similarity measure by comparing the reconstructed image with the original image.

Often, the likelihood should be optimized. This can be done by defining the negative logarithmic likelihood as *loss*. This results in *cross-entropy* being used as a loss function for classification. Cross entropy H can be seen as the extra entropy needed when using the wrong probability $\hat{p}$ instead of the true probability p.

$$H(p, \hat{p}) = -\sum_{i} p_i \log \hat{p}_i \tag{2.80}$$

When p and $\hat{p}$ are equal, the cross-entropy is the same as the entropy. The Kullback-Leibler divergence (KL) is the difference between cross-entropy and entropy.

For predicting a value instead of a class, the mean square error can be used as the loss function. For segmentation, the class of each voxel has to be predicted. It would be possible to use the cross-entropy as loss, but there is a major drawback. Most voxels do not belong to the foreground (for example, the tumor), but instead to the background class. So, for example, the Dice similarity coefficient (DSC) (see equation (3.4)) can be used, which takes this into account. In addition to DSC, there is a whole group of loss functions, which can be used to combat class imbalance in medical image segmentation (Yeung et al., 2022).

2.3.3 Dataset

The data should be separated into two sets, the *training set* and the *test set*. Having a separate test set, which is not used during training, prevents overfitting. When the dataset is too small to divide it, *cross-validation* can be used. In this way, the test error can be estimated for all examples, but the computational cost increases. For k-fold cross-validation, the dataset is split into k different non-overlapping samples. The network is then trained k times, with the i^{th} sample as the test set, and the rest of the examples are part of the training set.

The training set should also include a validation set. In this way, the *generalization error* that occurs when generalizing from seen to unseen examples can be estimated and used to adjust the hyperparameters. Due to this optimization, the true generalization error can only be calculated using the test set.

The images should be standardized so that all the values of all pixels lie in the same reasonable range. Images should also be formatted so that they have a similar scale. A NN could adapt to input in any rage, by adjusting the weights and biases of the first layer. But this can cause issues with regularization if the weights in the input layer are very large or very small. It can also make sense to remove variation in the data if it is not relevant to the problem. Preprocessing should be the same for all data.

2.3.4 Architecture

Most neural networks in computer vision are feedforward neural networks, without recurrent connections. This means that there are no output connections that feed into the model itself. This differentiates them from the visual cortex of the brain, which served as inspiration for the structure of neural networks, but has a less strict separation of layers and many recurrent connections (Gilbert and W. Li, 2013). Almost all modern neural networks are structured in layers. First, there is the input layer, then a number of *hidden layers*, which are not visible, and an output layer.

The output of each layer is called a *feature map*, which is a multidimensional tensor. For images, the feature map usually has two or three spatial dimensions and one dimension for the channels. In the input image, the channels can be different colors for natural images or different modalities for MRI images. Each layer takes the feature map of the previous layer as input, processes it, and has another feature map as output. *Skip connections* can be used to allow information from lower layers to flow to higher layers, bypassing layers in between.

A layer accepts an input, which is usually the output of the previous layer $\boldsymbol{h}^{(i-1)}$ and uses a linear transformation with weights $\boldsymbol{W}^{(i)}$ and bias $\boldsymbol{b}^{(i)}$ to produce the output $\boldsymbol{h}^{(i)}$.

After passing the input through a layer, an *activation function* $g^{(i)}$ is then applied. The activation function must be non-linear, so that the network can learn non-linear functions

$$\boldsymbol{h}^{(i)} = g^{(i)}\left({\boldsymbol{W}^{(i)}}^{\top}\boldsymbol{h}^{(i-1)} + \boldsymbol{b}^{(i)}\right). \tag{2.81}$$

Different functions can be used as activation function g, but for optimization, they should be piecewise differentiable, otherwise backpropagation cannot be used.

A very simple and default option is a Rectified linear unit (ReLU), which uses the function $g(z) = \max\{0, z\}$. The gradient is very easy to calculate; it is 0 or 1. It is not differentiable at zero, so a value of zero is usually used.

There are also some improvements to ReLUs, which aim to solve the problem that they do not learn when they are not active. For example, a nonzero slope can be used for $z_i < 0$ with $g(\boldsymbol{z}, \boldsymbol{\alpha})_i = \max(0, z_i) + \alpha_i \min(0, z_i)$. This is also called a *leaky ReLU*.

For the last layer, the activation is different. The simplest method is to use a linear output unit. The output unit needed also depends on the desired output range. Often, probabilities are predicted, so the values should be between 0 and 1.

An example of deciding between two classes (recommended in Goodfellow, Bengio, et al., 2016) is the logistic sigmoid function, defined by

$$\sigma(z) = \frac{\exp(z)}{\exp(z) + 1}. \tag{2.82}$$

This can be extended to multiple classes using the *softmax* function (softened version of the argmax function)

$$\text{softmax}(\boldsymbol{z})_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}. \tag{2.83}$$

The result can be interpreted as a probability because all output values are positive and sum up to one.

2.3.5 Convolutional neural networks (CNNs)

CNNs make excessive use of parameter sharing. They were first developed by Fukushima in 1979 (Fukushima, 1980) and extended by LeCun (LeCun et al., 1989). Parameters can be shared over multiple image locations, and the same features can be calculated at multiple locations. This automatically makes the model invariant to translation and dramatically reduces the number of parameters needed. This can also be seen as introducing a strong prior, and is especially useful when analyzing images.

The other advantage is that, for fully convolutional networks, the input can be of arbitrary size (Shelhamer et al., 2016). Dense layers can be seen as convolutional layers with a kernel the size of a feature map. This is a useful property, for example, for segmentation, because the size of the input images can vary.

The name comes from mathematical convolutions, although they are not necessarily performed. The architecture was inspired by the structure of the visual cortex. Convolutional layers perform convolutions instead of matrix multiplications.

A convolution acts on two functions f and g with real-valued parameters. Convolution is commutative. In a continuous space, an integral gives it. For discrete space, the integral is replaced by a sum

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\,\mathrm{d}\tau = \sum_{n=-\infty}^{\infty} f(n)g(t-n). \tag{2.84}$$

The first function is often referred to as input and the second one as kernel. A convolution can also be defined in multiple dimensions. For CNNs, the kernel is usually a three- or four-dimensional tensor, applied over the spatial dimensions and channels.

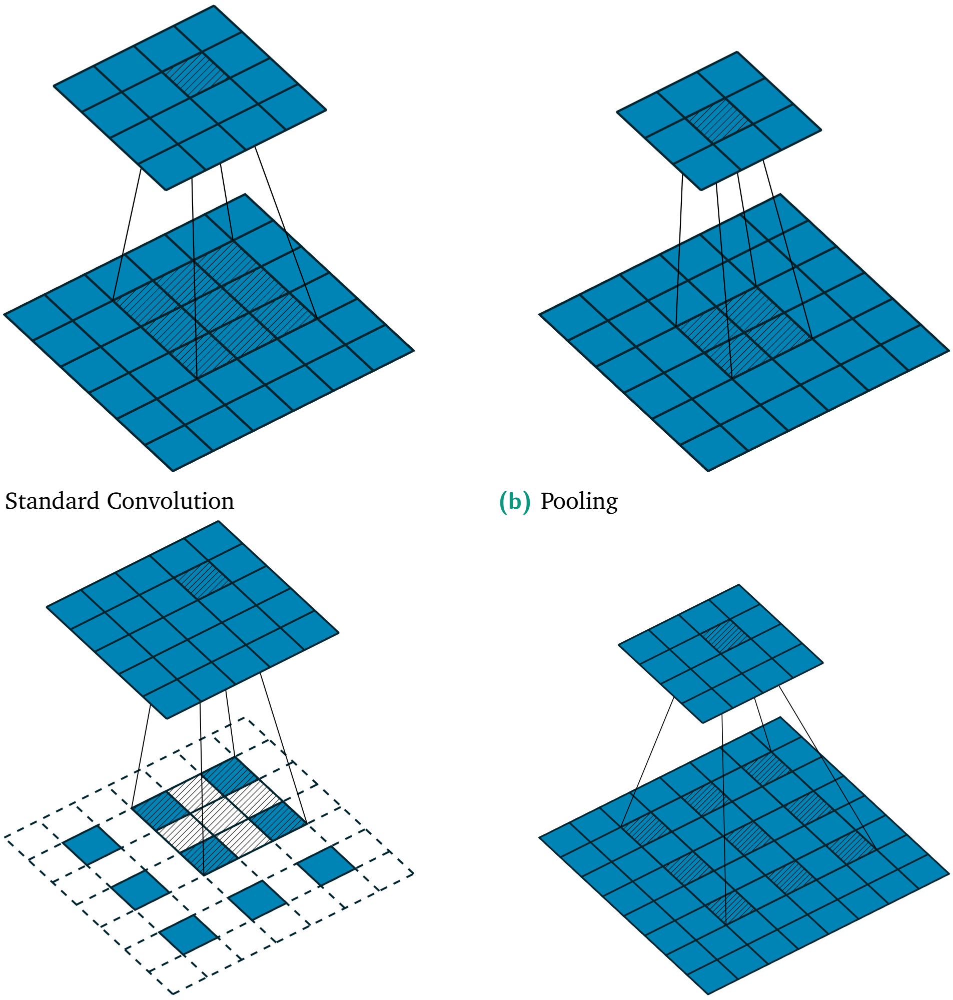
$$S(i,j,k) = (I * K)(i,j) = \sum_{l,m,n} I(l,m,n)K(i-l, j-m, k-n) \tag{2.85}$$

Often cross-correlation is used, which is a convolution with a flipped kernel. This corresponds to replacing the minus signs with pluses in the sum. This does not change the operation, because the index of the weights is arbitrary. Different types of convolutional layers are shown in Fig. 2.7.

If a kernel size larger than one is used, the output feature map will be smaller than the input. For segmentation, the output has the same spatial dimensions as the input image, so the feature maps are padded before the convolution is applied, to preserve the size.

An important hyperparameter is the *stride*. A stride of 1 means that every point in the input is considered. With a stride of 2, only every second input will be considered for that unit.

A *pooling* function replaces the output of the net at a certain location with a summary statistic of the nearby outputs. *Max pooling* is often used, which uses the maximum value. The advantage of pooling is the invariance to small translations in the input. It is possible to space them more than one pixel apart. This will reduce the input size to the next layer. The reduced feature map size also reduces the amount of



(a) Standard Convolution

(b) Pooling

(c) Transposed Convolution

(d) Dilated Convolution

Fig. 2.7: Different types of convolutions. The input feature map is shown on the bottom and the output on the top. Here, only one channel is depicted. The same operation is applied to all channels and summed up (except for depthwise separable convolutions (Guo et al., 2019)). This operation is the repeated for all output channels.
All convolutions shown here have a kernel size of 3 (except for the pooling in (b) with a size of 2). In the standard convolution (a), the kernel is multiplied with the patch of the input feature map. The stride is the step-size in which the kernel is stepped over the feature map.
For the pooling, shown in (b), the maximum or average of the feature map region is taken. This is often used to reduce the spatial dimensions by using a stride of more than one.
To increase the spatial dimensions, the transposed convolution, seen in (c), can be used. This is needed for example in segmentation to increase the size of the feature map to the size of the input image.
The input-feature map is padded before applying the convolution. To increase the field of view per neuron, dilated convolutions are used. As seen in (d), the kernel is padded with zeros, which increases its size and therefore field of view (Yu and Koltun, 2016). Image adapted from Dumoulin and Visin, 2018.

computation that is needed. The feature maps usually get smaller, but deeper (with more channels) further into the networks.

This offset can also be varied in size so that the classification layer always receives the same number of features, no matter the input size. For example, if average or maximum pooling over the whole image can be used as the last layer and the output is then passed onto a dense layer which has a fixed input size.

2.3.6 Regularization

Regularization can also be performed, which poses additional constraints on the parameters, for example, favoring lower weights. Regularization adds prior knowledge and tries to reduce the generalization error, but does not try to reduce the training error. It can help to prevent overfitting. It can be done by adding another term to the cost function. Other methods, such as augmentation, indirectly constrain the parameters, but still use prior knowledge.

Data Augmentation The size of the dataset can be artificially increased by generating synthetic data. This is done by applying transformations to the data. This makes the network invariant to those transforms, and can thus be seen as a type of regularization. For computer vision, geometric transforms (rotating, scaling, shearing) of the input image usually do not change the ground truth. For segmentation, the same transform must also be applied to the ground truth.

Early Stopping At first, the training set loss and the validation set loss will decrease. After some *epochs* (number of iterations over the dataset), the training set loss will continue to decrease, but the validation set loss will start to increase again due to overfitting to the training set.

This can be easily prevented by always storing the model parameters with the lowest validation error. After the model has not improved for a number of iterations, the training is stopped and the model with the lowest loss of the validation set is returned.

Multi-Task Learning It is assumed that the lower parts of the model can be shared between multiple tasks. This will improve the generalization of the model. The advantage is that more examples can be used. Usually, the input and the first few layers are shared, and the models differ later on. For this to work, some factors must be shared, explaining the variation in the data. An example of this is to combine segmentation with predicting treatment response. It can be assumed that most of the important features are the same, so the same encoding path could be used for both tasks. For the response prediction, the features would be fed into a classifier (for example, a series of dense layers), and for the segmentation, they would be upsampled to get a segmentation map.

Dropout The main idea of dropout is to randomly drop units and their connections from the network during training (Srivastava et al., 2014). Dropout makes the network more robust, because not just noise is added, but individual features are erased; this forces the model to rely on multiple features to make its choice. This can also be seen as training an ensemble of thinned networks, which are then averaged during inference when all units are used.

2.4 Radiomics

The goal of radiomics is to extract (semi) quantitative metrics from medical images, which are called radiomic features, with the goal of obtaining predictive or prognostic models. In oncology, tumors are usually analyzed. A good overview is provided in Afshar et al., 2019 and Mayerhoefer, Materka, et al., 2020.

The extracted features can be used, for example, to predict survival, treatment response, or tumor phenotype. Other data sources, such as clinical characteristics, can also be combined with imaging data.

2.4.1 Hand-Crafted Radiomics

The conventional approach is to use a hand-crafted radiomics (HCR) pipeline, which operates in four steps:

1. Image acquisition and reconstruction
2. Image segmentation
3. Feature extraction and quantification
4. Statistical analysis

There can be hundreds of features, so machine learning is usually used in the last step. An approach that is becoming more common is a deep learning-based radiomics pipeline (see Section 2.4.2). A neural network can learn to do the whole process in one step. Hybrid approaches are also possible.

Image acquisition and processing

The metrics and radiomic features derived from the images are sensitive to image acquisition settings, reconstruction algorithms, and image processing. Significant influences can be resolution, scan duration, field strength, and other acquisition parameters (Schurink et al., 2022; Dreher et al., 2020). So a uniform acquisition protocol should be used.

Preprocessing is needed to get the data in a uniform format. Normalization can be used to remove differences between different scanners (Scalco et al., 2020; Um et al., 2019). It may also be necessary to re-sample the data to get the same resolution and scale in all images, although this only partly removes the dependency on the voxel size at acquisition (Shafiq-ul-Hassan et al., 2017).

Image segmentation

Segmentation is very important for radiomics, because it is used to define the Region of interest (ROI), where the features are extracted. The standard approach is manual segmentation, but that is time-consuming and can introduce bias, so different approaches have been developed for (semi-) automatic tumor segmentation (van Heeswijk et al., 2016; Soomro et al., 2019; X. Zhao et al., 2020). The most successful algorithms rely on deep learning (a lot of them are based on the U-Net (Ronneberger et al., 2015; Isensee et al., 2021; Huang et al., 2019)).

Feature Extraction

There are many features; a standardized list of features was compiled by the Image Biomarker Standardization Initiative (IBSI) (Zwanenburg et al., 2020). Some other guidelines, for example, for preprocessing, are also included in the guidelines.

The features can be histogram-, texture-, model-, transform- and shape-based. The underlying image can be 2D or 3D and is grayscale. Only the ROI is analyzed and not the whole image. Before the features are calculated, the image is discretized.

Statistical Analysis

Feature reduction is a critical step in radiomics. Although many quantitative features can be extracted, most of them are highly correlated, irrelevant to the task at hand, and/or contribute to the overfitting of the model. There are supervised and unsupervised techniques for feature reduction.

Supervised methods look at predefined classes and select the features that best distinguish these classes. Examples are filtering and wrapper methods. Unsupervised methods aim to eliminate redundant features. Examples of unsupervised methods are Principal Component Analysis (PCA) (Z. Li et al., 2021), independent component analysis, and zero variance. The main characteristics that are important for the selection of features are the reproducibility, informativeness, relevance, and redundancy.

An important aspect of radiomics is the stability of the extracted features, which quantifies the degree of dependency between the features and the data acquisition and preprocessing steps. There are two methods to test the stability.

Test-Retest can be used when the same imaging exam is performed more than once for the same patient and the images are collected separately. Ideally, the features should be invariant. Another criterion is the inter-observer reliability. For this, different people delineate the ROI on the same image, which should ideally not change the features.

Model Construction and Classification

Statistical analysis and machine learning are commonly used for model construction and classification. It can be a regression problem (for example survival in months), a classification problem (such as tumor subtype), or a classification problem (grouping similar patients together). A prior assumption on the meaning of features (and their importance) can also be included.

For classification and regression models, standard ML approaches such as random forest (Horvat, Veeraraghavan, et al., 2018), support vector machines (Z. Liu et al., 2017) and neural networks can be used.

2.4.2 Deep Learning-Based Radiomics

The difference from the hand-crafted approach is that in Deep Learning-Based Radiomics (DLR), features are automatically extracted according to the predefined task. Common approaches are convolutional neural networks (see Section 2.3.5) and autoencoders (see Section 3.2.3). The extracted features can then be further processed by the neural network for analysis and decision-making, or they may exit the network and go through a different analyzer, such as an SVM or a Decision Tree (Afshar et al., 2019).

One great advantage is that the features are automatically extracted and that no prior knowledge is needed. The model can be trained end-to-end and can be improved by training with more examples. The segmentation process can also be eliminated, which reduces computational time or cost by not having to employ experts to manually annotate the images. This can also reduce the errors that can be introduced by the segmentation process. Another option is to include the segmented image as input or as an additional task (Jin et al., 2021).

The main disadvantages are that much more data is needed to train DL models compared to hand-crafted radiomics. For small datasets, there is no clear advantage in using deep learning (Schelb et al., 2019). Another issue is the lack of robustness to some transformations (for example, changes in the noise structure). Performing a sensitivity analysis is a critical step in explaining the connection between the design choices and the results achieved.

2.4.3 Challenges

There are multiple challenges in using radiomics in practice. One of the main problems is the stability of the features (Scalco et al., 2020). There are many factors that influence the radiomic features (van Timmeren et al., 2020). Controlling all of them is impossible, but they have to be standardized as much as possible. This also makes it difficult to collect large datasets, which are especially needed for deep learning. Better harmonization and normalization techniques can also help with this problem (Isaksson et al., 2020; Saint Martin et al., 2020).

This problem is not limited to image processing. Even when following the IBSI guidelines, not all feature classes are reliable between software platforms, because there are still undefined calculation settings (Fornacon-Wood et al., 2020). Guidelines for a robust radiomics application are provided by Lambin et al. (Lambin et al., 2017).

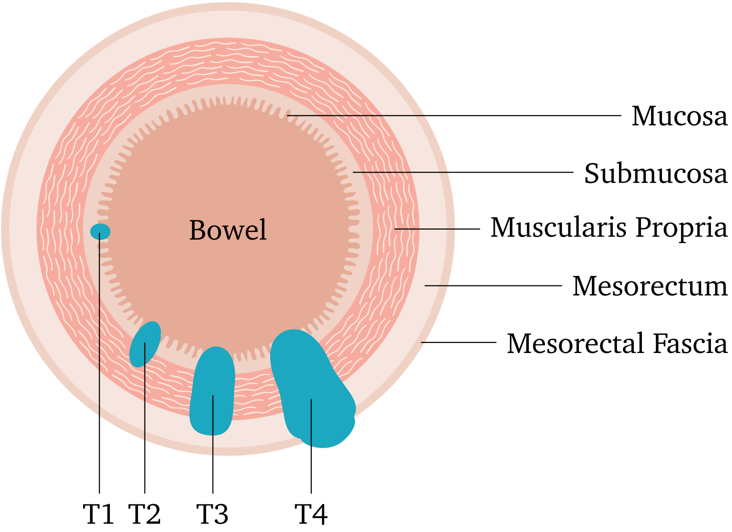


Fig. 2.8: Shown here are the main anatomical structures around the rectum and in the rectal wall. Tumors (in blue) are shown at different stages in their relation to the anatomical structures. Adapted from Cancer Research UK, 2014. 🅭🅯🄎

2.5 Medical Basics

This section provides a brief overview of the medical basics necessary to understand the staging and treatment outcome in rectal cancer.

2.5.1 Anatomy

The rectum refers to the lowest 12 cm to 15 cm of the large intestine. The anatomy can be seen in Fig. 2.8. The cavity of the rectum is surrounded by the mucosa, which forms a protective membrane to protect the body and prevent dehydration. It sits on top of the submucosa, a layer of connective tissue, which contains blood and lymphatic vessels and nerves, which are connected to the surrounding tissue.

Around this is the muscularis propria, a muscular layer that is needed to move fecal matter through the colon by contraction. The colon has a large muscular layer because the fecal matter is harder to move along than the partially liquid semi-digested food in other parts of the intestinal tract.

This muscular layer is then surrounded by the *mesorectum*. It consists of fatty tissue and contains lymphatic vessels and lymph nodes. A very important structure for rectal cancer is the mesorectal fascia (MRF). It surrounds most of the mesorectum.

2.5.2 Staging

Staging is an important step in the diagnosis of cancer. A widespread system is the *TNM* staging system. The mandatory components are the stage of the primary tumor *T*, the spread to local lymph nodes *N* and if there are metastases *M*. The histologic tumor grade score along with the metastatic (whole-body-level cancer spread) staging is used to evaluate each specific cancer patient, develop their individual treatment strategy, and predict their prognosis.

A prefix can also be added. An overview is provided by Horvat, Carlos Tavares Rocha, et al., 2019 for rectal cancer.

Prefixes

c — The prefix c means that the stage is determined from evidence acquired prior to surgery. The c-prefix is implicit in the absence of the p-prefix.

p — The prefix p means that the stage was determined by histopathological examination of a surgical specimen.

Mandatory Parameters

T - primary tumor — Size or direct extent of the primary tumor (see also Fig. 2.8), for rectal cancer, the stages are:

T0 No evidence of a tumor

T1 The tumor involves only the first or second layer of the rectal wall, and no lymph nodes are involved.

T2 The tumor penetrates into the mesorectum, but no lymph nodes are involved.

T3 The cancer has grown into the outermost layers of the colon or rectum but has not gone through them.

T3a Tumor extends <1 mm beyond the muscularis propria

T3b Tumor extends 1–5 mm beyond the muscularis propria

T3c Tumor extends 5–15 mm beyond the muscularis propria

T3d Tumor extends 15 mm beyond the muscularis propria

T4 There is convincing evidence of cancer in other parts of the body, outside the rectal area.

T4a The cancer has grown through the wall of the colon or rectum (including the visceral peritoneum) but has not reached nearby organs

T4b The cancer has grown through the wall of the colon or rectum and is attached to or has grown into other nearby tissues or organs

N - lymph nodes — Number of local lymph node metastases

N0 No lymph node affected

N1 One to three local lymph node metastasizes

N2 Four or more local lymph nodes are metastasized

M - distant metastasis — Describes distant metastases

M0 No distant metastasis

M1 Distant metastasis

2.5.3 Treatment of Rectal Cancer

Treatment depends on many factors. The *TNM* stage is very important, but the location of the tumor and other factors must also be considered. The location is quantified using the height of the tumor (U. I. Attenberger et al., 2020), which measures the distance from the anal verge to the lower border of the tumor.

MRI is essential for staging and determining further treatment, because it offers information about the tumor location, distance to the MRF and infiltration of local structures (*T*-stage) and lymph nodes (*N*-stage) (U. Attenberger and B. Wichtmann, 2015).

There are different ways of treating rectal cancer. Guidelines are given in (Schmoll et al., 2012). For local rectal cancer, the standard of care is a total mesorectal excision (TME) (Hofheinz et al., 2023), which means that the rectum is resected, including the mesorectal fat and lymph nodes. For small tumors (T1 N0), a local excision may be enough.

For a TME, the rectum is removed up to the circumferential resection margin (CRM), which is formed by the MRF. A distance of a tumor of less 1 mm to the CRM means that there may be tumor cells beyond the surgical margin, which is a strong predictor of local recurrence and low survival (Taylor et al., 2011).

Before the TME, preoperative treatment is used to improve operability and reduce the chance of a local relapse. This can be chemotherapy, radiotherapy, or a combination of both, which is called radio-chemo-therapy (CRT). Most of the time, this leads to a shrinkage of the tumor.

For locally advanced rectal cancer (LARC), which is defined as a primary tumor with T3 or T4 and/or metastasis in lymph nodes (N+) (Oronsky et al., 2020), the treatment recommendation is *neoadjuvant* CRT (Ulrike I. Attenberger et al., 2020). Neoadjuvant means that the CRT is in addition to and before primary therapy, which is the TME.

After the CRT, the patients are restaged. This means that another MRI is performed to determine the tumor stage again. The accuracy decreases greatly compared to the initial staging, because it is difficult to distinguish tumor cells from fibrotic tissue (Ulrike I. Attenberger et al., 2020).

The *tumor regression* in response to neoadjuvant treatment can be assessed using the histopathological grading developed by Dworak et al., 1997 (also called Dworak score). The tumor regression grade (TRG) is defined with the following grades:

Grade 0 No response

Grade 1 Minimal response, dominant tumor with obvious fibrosis (forming of scar tissue), vasculopathy (destruction of blood vessels)

Grade 2 Moderate response, easy to find fibrotic changes with few tumor cells or groups

Grade 3 Near complete response, very few tumor cells in fibrotic tissue, which are hard to find microscopically

Grade 4 Complete response, no tumor cells (total regression)

After restaging, about 50 % to 60 % of patients are downstaged. Approximately 20 % show pathologic Complete Response (pCR) according to Benson et al., 2015. pCR is the absence of all signs of cancer in tissue samples removed during TME.

For most patients, there is further adjuvant therapy after surgery, usually CRT. This reduces the chance of a recurrence of the tumor. A patient is considered cured if he has five years of complete remission (no signs of cancer).

Methods

3

This chapter starts with a description of the dataset in Section 3.1, which is used throughout this work. In the following Section 3.2, the neural network architectures that were used in this thesis are described. A new normalization method was developed and is compared with other, statistical methods, in Section 3.4. Parts of this chapter have been published in the article “Comparison of Image Normalization Methods for Multi-Site Deep Learning” in *applied sciences* in a special issue titled ”Deep Learning Application in Medical Image Analysis“ (Albert et al., 2023)[1]. The last Section 3.5 describes the classification of locally advanced rectal cancer (LARC) using a deep neural network. Parts of this chapter have been published in *diagnostics* (B. D. Wichtmann, Albert, et al., 2022)[2].

3.1 Dataset

Most of the dataset used in this work comes from staging MRIs performed for a clinical phase two rectal cancer study. Additional in-house data from clinical routine without study conditions was used.

3.1.1 Study

The dataset used was acquired during the clinical CAO/ARO/AIO-12 study (Fokas, Allgäuer, et al., 2019; Fokas, Schlenska-Lange, et al., 2022). It enrolled 311 patients in 18 centers in Germany. The objective was to study the scheduling of radio-chemotherapy (CRT) and chemotherapy in neoadjuvant therapy for rectal cancer. The patients were divided into two subgroups. Group A received chemotherapy followed by CRT and then surgery. Group B first received CRT and then chemotherapy and surgery. The criteria for the inclusion of patients in the study were the following:

- histologically confirmed rectal adenocarcinoma (up to 12 cm above the anal verge)

[1]
[2]

- TNM stage of T3, T4 or lymph node involvement, for T3 either < 6 cm from the anal verge or spread into the mesorectal fat of at least 5 mm (T3c or T3d)
- At least 18 years old
- No distant metastasis
- Adequate organ function

The patients were treated with neoadjuvant therapy according to their group and after finishing the neoadjuvant therapy and being restaged, they received a total mesorectal excision (TME). All patients were operated without considering the clinical response, except patients who refused surgery (due to clinical Complete Response (cCR) or other reasons).

To assess the long-term outcome, patients were regularly followed up for 5 years after surgery to evaluate the long-term effects (Fokas, Schlenska-Lange, et al., 2022). Survival and recurrence rates were similar in both groups, but group B had lower toxicity during treatment.

All clinical data from the study was made available pseudonymously. MRI data was provided by eight centers, which treated 181 patients in the study. There was no central repository for imaging data, only for clinical data, so all centers had to be contacted individually.

All centers in the study, with more than ten patients were contacted, but not all provided data. The inclusion and exclusion flow chart for patients can be seen in Fig. 3.1.

An additional cohort of 61 patients was used, which was collected in-house. However, the cohort was slightly different. There were some patients with stage T2 and the patients only received CRT and no additional chemotherapy before or afterward.

3.1.2 Patient Data

An overview of the patient cohort can be seen in Tab. 3.1. There was some variation between the centers. For example, only 37 % of the patients taken from the study were women, but at Center 12, 54 % of the patients were women.

Only 17 % of all patients (including the in-house data) show a pathologic Complete Response (pCR) after neoadjuvant therapy. In the study, it was 17 % of the patients

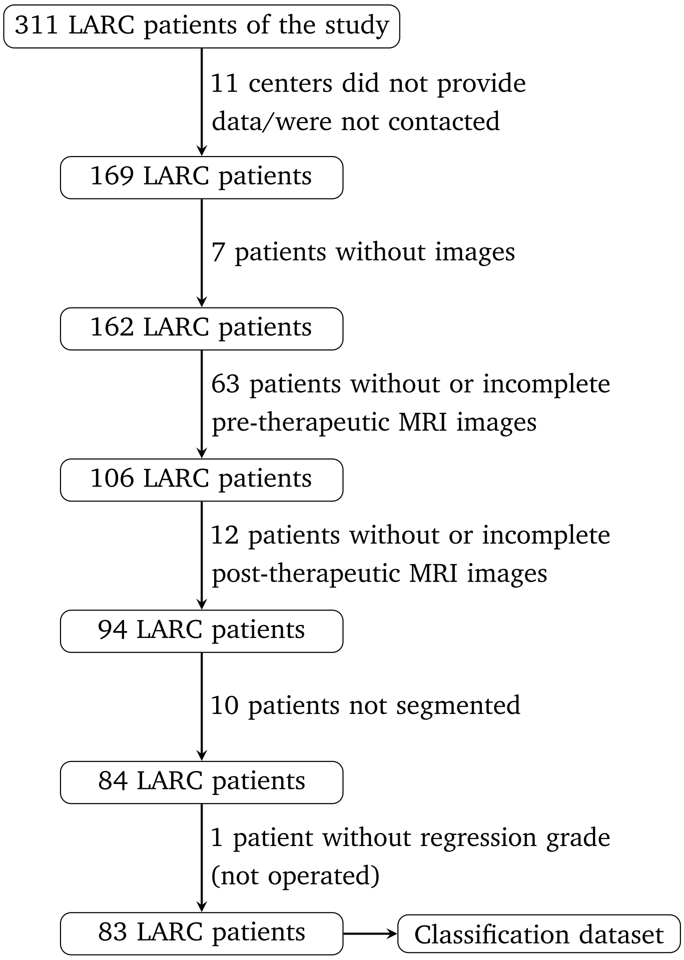

311 LARC patients of the study
11 centers did not provide data/were not contacted
169 LARC patients
7 patients without images
162 LARC patients
63 patients without or incomplete pre-therapeutic MRI images
106 LARC patients
12 patients without or incomplete post-therapeutic MRI images
94 LARC patients
10 patients not segmented
84 LARC patients
1 patient without regression grade (not operated)
83 LARC patients
Classification dataset

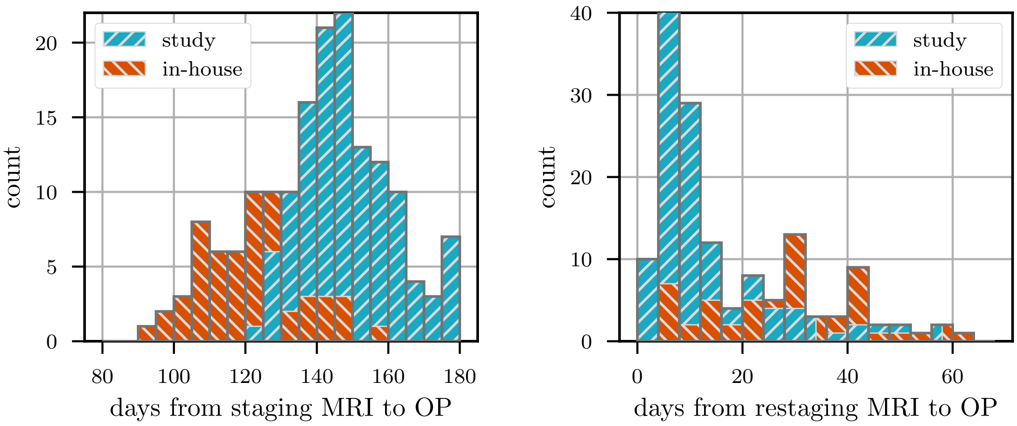
Fig. 3.1: The inclusion and exclusion of patients from the CAO-ARO-AIO-12 study. For some patients, there were no images, due to them dropping out of the study or the images not being found at the center.

After collecting all images, they were visually inspected and removed if there were severe artifacts that obscured the primary tumor or the wrong modality (for example, T2 weighted with fat suppression). The most common reason for exclusion was missing diffusion-weighted images.

Patients were segmented if there was a complete pre- and post-therapeutic MRI examination (some patients without a complete post-therapeutic MRI examination were also missing, so there are 104 segmented pre-therapeutic MRI images). The 10 patients not segmented were all from Center 8, which only provided data after the segmentation was already done. One patient did not receive the surgery, so there was no pathological regression grade.

Tab. 3.1: Overview of the patient cohort, The significance column shows the significance of the difference between the study and the in-house data. Continuous data was compared using the t-Test and ordered categorical data using the Mann-Whitney-U-test. For continuous data, the mean and standard deviation are shown.

Charactersitic	Center 1	Center 5	Center 8	Center 11	Center 12	Center 13	Center 16	In-house	significance
Number Patients	42	5	14	35	35	20	18	61	
Acquisition (years)	2015 – 2018	2016 – 2017	2015 – 2017	2015 – 2017	2015 – 2017	2015 – 2018	2015 – 2017	2009 – 2013	0.00
Age (years)	60(10)	61(7)	52(13)	60(10)	62(9)	63(10)	63(13)	64(11)	0.01
Sex									0.11
Male	30	4	8	25	16	13	11	44	
Female	12	1	6	10	19	7	7	17	
Pre-nCRT T-stage (MRI)									0.00
T1	0	0	0	0	1	0	0	0	
T2	4	0	0	0	3	4	2	13	
T3	31	3	7	5	28	9	13	46	
T4	5	1	0	0	1	0	1	2	
not specified	2	1	7	30	2	7	2	0	
Pre-nCRT N-stage (MRI)									0.00
N-	2	0	0	3	0	4	7	33	
N+	40	5	14	32	35	16	11	28	
Tumor location									0.0
lower third	19	3	3	8	14	7	4	18	
middle third	17	2	3	13	21	7	1	28	
upper third	0	0	1	0	0	2	0	15	
not specified	6	0	7	14	0	4	13	0	
Post-nCRT T-stage (MRI)									0.00
T0	2	1	1	1	11	1	0	0	
T1	2	0	0	0	1	0	0	5	
T2	14	0	0	0	3	0	0	32	
T3	19	3	4	1	20	1	2	24	
T4	5	0	5	0	0	1	1	0	
not specified	0	1	4	33	0	17	15	0	
Post-nCRT N-stage (MRI)									0.23
N-	19	3	1	3	12	1	1	52	
N+	23	1	9	1	23	3	2	7	
not specified	0	1	4	31	0	16	15	2	
Regression Grade									0.14
No response	1	0	0	0	0	0	0	0	
Minimal response	1	0	5	6	4	3	0	10	
Moderate response	21	3	5	8	15	10	9	26	
Near complete respons	10	1	1	7	9	3	6	19	
Complete response	8	1	2	14	5	3	1	6	
not specified	1	0	1	0	2	1	2	0	



(a) The time between the staging MRI before CRT and the surgery

(b) The time between the restaging MRI after CRT and the surgery

Fig. 3.2: Comparison of the time distance between both MRIs and the date of the surgery. In (a), the time between the staging MRI and the surgery was on average larger for the study patients. But for the second MRI in (b), the time between the restaging MRI and the surgery was short for the study patients and distributed for the in-house patients.

in group A and 25 % of the patients in group B. For the in-house patient cohort, only 10 % had a pCR.

When comparing the study data with the in-house data, there were some differences. The data was acquired on average 5.7 years earlier, and patients were on average four years older. There were also significant differences in tumor characteristics before neoadjuvant therapy. There were more tumors located in the upper third of the abdomen and more patients with stage cT2 in the in-house data compared to the study data. However, there were no statistically significant differences in response to treatment.

Another big difference between the study and in-house data was the timing of the treatment. As seen in Fig. 3.2, there was a large difference in the number of days between the first MRI and the surgery.

Imaging

MRI was performed on all patients before therapy and before surgery for both cohorts of patients. Some patients also received MRI during therapy and after surgery.

The MR image acquisition protocol of the study was as follows:

- Localizer images to plan further images
- T2w-Half-Fourier Acquisition Single-shot Turbo spin Echo (HASTE) images in the coronal plane to plan high resolution images (see Section 2.2.2)
- T2w-Turbo Spin Echo (TSE) images in high resolution (max. 3 mm slice thickness and 0.8 mm in-plane resolution) in the sagittal and axial plane (see Section 2.2.2)
- EPI 3D images
- Diffusion images with b=50/400/800 s mm^{-2} (see Section 2.2.3)

Not all of these images were used in this thesis. The HASTE images had lower signal-to-noise ratio (SNR) than the TSE images and the resolution was also reduced. Thus, those images were only used when no TSE image was available. Only the axial TSE images were used, because they show the tumor best, and the diffusion-weighted (DW) images were also acquired axial to the tumor.

For the diffusion images, the b800 and apparent diffusion coefficient (ADC) images were used. Images with a high b-value can improve the performance in the tumor restaging (Horvat, Carlos Tavares Rocha, et al., 2019). Low ADC values reflect areas with high cellular density, which is typical for tumor tissue (Koh and Collins, 2007). The ADC images are also interesting, because it is a quantitative modality.

Many MRI scanners were used to acquire the images. A summary of the scanners used can be seen in Tab. 3.2. Most of the MRIs (97 % of all MRI measurement) were performed on Siemens scanners. A total of 33 scanners and 16 different scanner models were used. All scanners (except one used for three examinations) had a field strength of 1.5 T or 3 T with more patients imaged at 3 T. Patients were also often imaged on different scanners at different visits.

For the characterization of the images, the SNR was measured. This was done using the method presented in Chen et al., 2015, which uses the correlation between the noise level and the eigenvalues of the covariance matrix.

3.1.3 Data Preprocessing

While the required preprocessing depends on the task the images are needed for, some steps were needed for all images and were performed when compiling the dataset.

Tab. 3.2: Scanners used in the study and the in-house data. Shown here is the number of unique patients per scanner and per scanner for the pre- and post-therapeutic MRI examinations. The numbers do not match, because not all patients got their examinations on the same scanner and some patients were missing some MRI examinations or had additional MRI examinations during the neoadjuvant therapy.

Location	Manufacturer	Model	Field Strength (T)	# pre-nCRT patients	# post-nCRT patients	# patients total
Center 1	Siemens	Avanto	1.5	16	15	24
	Siemens	Prisma	3.0	6	21	22
	Siemens	Symph. Tim	1.5	11	1	11
	Siemens	Aera	1.5	1	3	4
	Siemens	Spectra	3.0	2	0	2
	Philips	Ingenia	1.5	1	0	1
	Siemens	Verio	3.0	1	0	1
	Siemens	Espree	1.5	0	1	1
Center 5	Siemens	Symph. Tim	1.5	1	2	4
	Siemens	Skyra	3.0	1	1	3
	Siemens	Avanto	1.5	1	1	1
Center 8	Siemens	Avanto	1.5	5	3	10
	Siemens	Aera	1.5	4	4	6
	Siemens	Vida	3.0	3	5	6
	Siemens	Skyra	3.0	1	0	1
Center 11	Siemens	Skyra	3.0	24	22	33
	Siemens	Avanto	1.5	6	13	24
	Philips	Ingenia	1.5	3	0	3
	Siemens	Aera	1.5	1	0	1
	GE	Signa HDxt	1.5	1	0	1
Center 12	Siemens	Skyra	3.0	5	20	30
	Siemens	Prisma	3.0	11	15	24
Center 13	Siemens	Skyra	3.0	6	13	18
	Siemens	Trio Tim	3.0	9	6	14
	Siemens	Avanto	1.5	1	0	3
	Philips	Intera	1.5	2	0	2
	Siemens	Aera	1.5	1	0	1
	Siemens	Harmony	1.0	1	0	1
Center 16	Siemens	Avanto	1.5	4	5	9
	Siemens	Verio	3.0	4	4	9
	Philips	Achieva	1.5	6	0	8
	Philips	Pan. HFO	1.0	2	0	2
In-house	Siemens	Trio Tim	3.0	61	59	61

Data Selection and Quality Control

A challenge is data extraction. The images adhere to the Digital Imaging and Communications in Medicine (DICOM) standard, but not all tags were always set correctly and can vary from clinic to clinic and even session to session within a clinic. After some manual corrections, the images were exported as a standardized dataset of NIfTI files, which are more suitable for further processing than the DICOM images.

Images were visually inspected. Images with a lot of artifacts or images where another image with better quality was available from the same time point were removed. Some images also used fat suppression, which was specified to not be used in the protocol.

Manual Segmentation

Segmentation is an import step of a hand-crafted radiomics pipeline. Segmentation is also used as a task to evaluate the normalization performance. To train a neural network for segmentation, ground truth is needed.

The images before and after therapy in both cohorts of patients were then segmented by a radiologist with 6 years of experience using ITK-SNAP (Yushkevich et al., 2006). Segmentation was performed on the T2-weighted images, and diffusion-weighted images were used for orientation. For the in-house data, another medical doctor segmented some images, and the first checked and corrected the segmentations if necessary. Only the primary tumor was segmented.

N4-Correction

To remove bias due to inhomogeneities in the B_1 field (see Section 2.1.7 on page 17), N4 correction was performed (Tustison et al., 2010). It estimates a multiplicative bias field, which was removed by multiplying the image intensities by the inverse of the bias field.

The algorithm iteratively estimates the bias field using B-splines. This was done in multiple resolution levels. Each successive level has twice the mesh resolution for the B-splines as the previous one.

For this, the images were first downsampled by a factor of four in the x–y plane. The algorithm was then applied with 100 iterations at each of the four resolution levels. A fixed value threshold was used to generate a mask and only calculate the bias field for the foreground.

Registration

The DW images were registered to the T2-weighted image using the rigid registration algorithm included in the Advanced Normalization Tools (ANTs) (Avants et al., 2011). This algorithm was chosen because it performed the best after visual inspection of randomly chosen images.

In addition to patient movement, DW images might also be spatially deformed. This is due to the high gradient fields being switched on and off rapidly, which can result in eddy currents being induced in surfaces, which in turn results in deformations (see Section 2.2.3 and page 30). Therefore, an elastic registration would have been preferable. However, this was not possible because of low SNR in the DW images.

To avoid deterioration of the image quality, the images were not resampled. Instead, the origin and direction were changed, so the images share the same coordinate system. This avoids having to interpolate the images.

3.2 Network Architectures

The network architectures, which were used throughout this work, are presented here. Some, such as U-Net and ResNet are concrete models, while architectures such as the *Autoencoder* or Generative Adversarial Networks (GAN) are more abstract concepts, which consist of concrete models. The U-Net and ResNet were used for segmentation and classification. For normalization, an autoencoder with adversarial losses was used.

3.2.1 ResNet

The main idea is the introduction of residual connections. The layers are formulated as residual functions with respect to the input of the layers. One or more convolutional layers, including their activations, are seen as a function $F(x)$. The output of this function is then added to the original input and the result is $F(x) + x$. A residual block can be seen in Fig. 3.3.

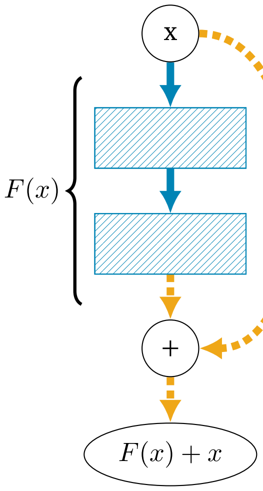


Fig. 3.3: In this exemplary residual block, two convolutional layers including activations are applied to the input x. These two convolutions are seen as the residual function $F(x)$. The output of the block is the residual function plus the input $F(x) + x$.

The *ResNet* (He et al., 2015) is still an often used network for classification, and residual connections are used in many other networks. It consists of four sections with a varying number of residual blocks. Each section is followed by a pooling layer, which halves the spatial resolution and the number of features is doubled in each section. In the original publication, multiple networks with a different number of layers are defined.

The 50-layer ResNet, which was used in this thesis, has a 1×1, 3×3 and 1×1 convolution in each convolutional block and there are 3, 4, 6 and 3 convolutional blocks in each section. This is followed by average-pooling and a dense layer. The output of that layer depends on the task. For classification, there is an output node for each class and softmax is used. For Regression tasks, there is just one output node without activation.

An advantage of this architecture is that this reduces problems with vanishing gradients, and it becomes possible to train deeper networks because the gradients are not reduced along the skip connection. With the ResNet architecture, networks with more than 1000 layers can be trained successfully.

3.2.2 U-Net

Another network architecture is the U-Net (Ronneberger et al., 2015), where the input and output have the same dimensions. Typical applications are image-to-image applications such as segmentation or image generation. This makes the network particularly interesting for medical applications.

The main idea is that the network has an encoding and a decoding path. Skip connections are used to add the information from the encoding path to the decoding path at the level with the same spatial resolution. In this way, high-level features are extracted in the encoding path. At the same time, the skip connections help to recover fine-grained details as the resolution is increased in the decoding path. The architecture of the U-Net can be seen in Fig. 3.4.

It is still widely used with some modifications (Isensee et al., 2021; Wong et al., 2023). For segmentation, the Rectified linear unit (ReLU) activation function is replaced by the ELU function

$$\mathrm{ELU}(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha \left(e^{x} - 1\right), & \text{otherwise.} \end{cases} \tag{3.1}$$

where α is a hyperparameter and was chosen to be 1. For $\alpha = 1$, the first derivative is continuous, making ELU more robust to noise compared to ReLU, whose gradients fluctuate between 0 and 1 for values around zero.

In contrast to the ReLU function, negative input also leads to activation. This solves multiple problems. An issue with ReLUs is that the mean activation is not zero, because the output is never negative, leading to a bias shift for the following units.

Another issue with ReLUs is that the deactivated units (with $x < 0$) do not provide additional information because their gradient is zero. For ELUs, there is a saturating contribution for negative values, which propagates the information to the next layer.

Another modification is the use of batch normalization 2.3.1 on page 36). Residual connections are also used. There are also many other proposed modifications, such as dilated convolutions or attention, but it is not clear if they provide a benefit (Isensee et al., 2021).

3.2.3 Autoencoder

Autoencoders are networks that learn to encode information efficiently in latent space and to transform this information back into the input space. Their structure is similar to that of a U-Net with an encoder and decoder, but without the skip connection, as shown in Fig. 3.5. Autoencoders are usually trained unsupervised.

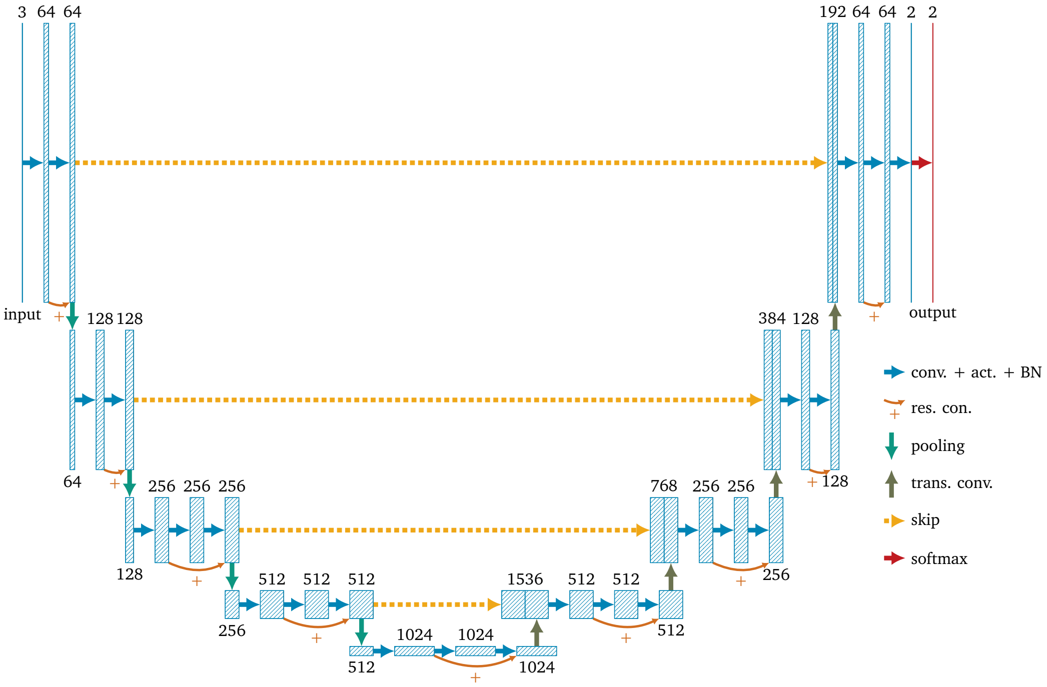


Fig. 3.4: Schematic overview of the U-Net architecture used for segmentation. The blue blocks show the feature maps, with the number of features displayed at the top. The other two dimensions depend on the input dimensions. The features are extracted along the encoding path and pass through the bottleneck. After the bottleneck, they are expanded again in the decoding path to recover the spatial resolution.

There are four blocks in the encoding path. Each block starts with a 3×3 convolution, and the number of features is doubled (except for the first block, which increases the depth from 3 to 64). Each of the two to three convolutions in the block is followed by exponential linear unit (ELU) activation, batch normalization, and spatial dropout. A residual connection adds the result of the first convolution to the result of the last in the block. Then, the maximum pooling is used to halve the resolution in each spatial dimension. The result of the pooling operation is then used as input for the next block.

The fifth block is called the bottleneck, with the lowest spatial resolution, but with the highest number of features, containing the most abstract information. It is built the same as the previous blocks, without the max pooling at the end. Instead, a transposed convolution (see Fig. 2.7c) is applied. This increases the spatial resolution by a factor of two in each dimension.

In the decoding path, there are four blocks again, with a similar structure as in the encoding path. The difference is that the max pooling operations are replaced by transposed convolutions, which increase spatial resolution. The upsampled feature maps from the previous block are concatenated with the result of the corresponding block in the encoding path. The following convolution reduces the number of features rather than increasing them.

To get the output, a convolution is applied to the last feature map, to reduce the number of features down to the number of classes (here tumor and background). A softmax function (see (2.83)) is then used to convert the output into numbers between zero and one, which can be interpreted as probabilities.

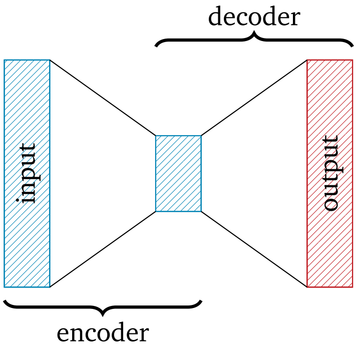


Fig. 3.5: The structure of an autoencoder with the encoder and decoder. The dimensionality of the input is reduced by the encoder to get the information in latent space (small block at the center). The latent space is then expanded by the decoder to get the same dimensions as the output.

If the input image (or other input data) is used as the desired output, the autoencoder learns an efficient representation of the image. This is useful for dimensionality reduction, data compression, anomaly detection, or can be used for pre-training. The input can also be modified, for example, by adding noise or downsampling to train the autoencoder for denoising or upsampling.

If the data in the latent space is varied, autoencoders can also be used to generate images similar to the training data distribution. This can be used for sample generation.

3.2.4 Generative Adversarial Networks

A GAN consists of two networks. There is a generator G that produces some output. As a loss, a second network is used, which is trained to detect if the output was produced by the generator or is from the source distribution $z \sim p_{\text{source}}(z)$. This network is called the discriminator D. It is used as an adversarial loss.

In the original publication (Goodfellow, Pouget-Abadie, et al., 2014), the image was generated from noise. In this work, another image is used as input to translate the image from one style to another. As generator, a U-Net is used here.

The discriminator is trained to differentiate between the generated images and images from the target distribution $x \sim p_{\text{data}}(x)$. Therefore, the loss function $\mathcal{L}_D$ for the discriminator is defined as

$$\mathcal{L}_D(D,G) = -\mathbb{E}_{x\sim p_{\text{data}}(x)} \log\left(D(x)\right) - \mathbb{E}_{z\sim p_{\text{source}}(z)} \log\left(1 - D(G(z))\right). \tag{3.2}$$

After training the discriminator on one batch of data, the same data is used to train the generator. Its loss function $\mathcal{L}_G$ is

$$\mathcal{L}_G(D,G) = \mathbb{E}_{z\sim p_{\text{source}}(z)} \log\left(1 - D(G(z))\right). \tag{3.3}$$

Thus, the discriminator tries to maximize $\log\left(1 - D(G(z))\right)$ and thus to detect which images were generated and which images stem from the original distribution. Meanwhile, the generator tries to minimize the same expression and thus fool the discriminator. This makes training difficult, because as the generator improves, the discriminator has to improve as well, and vice versa.

3.3 Evaluation Metrics

To evaluate the performance of the networks, evaluation metrics are used. They are presented in the following section. The evaluation metrics used depend on the task that is being evaluated. They also depend on the desired outcome for this task.

Segmentation

In segmentation, each voxel in the image is assigned to a class. Thus, there is the predicted segmentation map and the ground truth segmentation map, which have to be compared. For tumor segmentation, it is a binary classification task for each voxel. But popular metrics for classification, such as accuracy, are not well suited for most classification problems.

There is often a *class-imbalance*, with a lot more background voxels than foreground voxels. This means that a network that predicts everything to be background would have very high accuracy. To consider this, the Dice similarity coefficient (DSC), (also called F_1 score or Sørensen-Dice coefficient) is often used (Sudre et al., 2017). It was originally proposed in ecology in 1945 to quantify the similarity between two species by comparing sets of samples X and Y (Dice, 1945). It is defined as

$$DSC = 1 - \frac{2|X \cap Y|}{|X| + |Y|} = 1 - \frac{2TP}{2TP + FP + FN} \tag{3.4}$$

with TP being the number of voxels correctly classified, FP the number of voxels incorrectly classified as tumor and FN is the number of tumor voxels incorrectly classified as background (see also Tab. 3.3.). Correctly classified background voxels

Tab. 3.3: Confusion matrix for binary classification. The total number of samples is $P + N$, with $P = TP + FN$ and $N = FP + TN$.

	Predicted Positive (PP)	Predicted Negative (PN)
Actual Positive (P)	True Positive (TP)	False Negative (FN)
Actual Negative (N)	False Positive (FP)	True Negative (TN)

TN are not considered for the loss. If there are multiple foreground labels, there is a separate DSC for each label and all other labels are considered as background.

Classification

For classification, the goal is to evaluate how accurately the network finds the correct class. For this, there are many evaluation metrics. Here, we focus on binary classification. There, a threshold is chosen to decide if the output is positive or negative (often 0.5 is used). The result is a confusion matrix, as seen in Tab. 3.3.

From this matrix, many metrics can be derived. Accuracy is defined by accuracy $= \frac{TP+TN}{P+N}$, but accuracy is often not a good measure, because it suffers from sample imbalance and can depend greatly on the decision threshold.

Instead, *sensitivity* (also called recall, or true positive rate) and *specificity* are often used. They are defined as

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{3.5}$$

$$\text{specificity} = \frac{TN}{TN + FP}. \tag{3.6}$$

So, sensitivity measures how many of the samples predicted as positive are actually positive. Specificity measures the opposite, that is, how many of the samples predicted as negative are actually negative.

There is usually a trade-off between the two. For example, for diagnosis, high sensitivity means that almost all patients with the disease are predicted as such, but this usually also increases the number of false positives, which lowers the specificity. This trade-off can, for example, be adjusted by changing the decision threshold. To avoid having to choose a specific decision threshold, the area under the receiver operating characteristic curve (AUC) is used as a metric.

The receiver operator curve is calculated by varying the threshold from the minimum to the maximum value and graphing the sensitivity over the false positive rate ($1 -$ specificity). The curve starts at the origin with everything classified as negative and ends at (1,1), when everything is classified as positive. The area under the curve is used as a metric.

For classification, the AUC varies between 0 and 1, but a classifier that only outputs random values will already achieve a value of 0.5. A classifier where there is no overlap of the predicted value ranges for the actual positive and negative sample will have an AUC of 1. Therefore, most classifiers have an AUC between 0.5 and 1.

Regression

For regression, the output is a continuous value. One of the most-used metrics in this case is the root-mean-square error (RMSE). For N samples out of the predicted values X and ground truth values Y, it is defined as

$$\mathrm{RMSE}(X,Y) = \sqrt{\frac{\sum_i^N (x_i - y_i)^2}{N}}. \tag{3.7}$$

A lower RMSE is better. As a loss, the *mean-square-error* is also used, which is the square of the RMSE.

3.4 Normalization

Tumor segmentation in rectal cancer is intrinsically challenging (Trebeschi et al., 2017), even for state-of-the-art deep neural networks, due to hard-to-delineate tumors. Similar problems arise for the prediction of treatment outcome (B. D. Wichtmann et al., 2022).

One problem is that datasets are relatively small and often include only one or two centers (Wong et al., 2023). To apply such models in a broader clinical setting, they must show good generalization, which means that the network should be trained on a set of images from different centers.

For this, it is important to implement a standardized and harmonized protocol. Nevertheless, certain differences arising from, e.g., vendor-specific differences between scanners and sequence implementations exit (Mayerhoefer, Szomolanyi, et al., 2009). To mitigate such effects to a certain extent, image normalization methods could be used for preprocessing prior to training of a machine learning algorithm.

For most MRIs, the acquisition parameters are known and saved in the DICOM image metadata. A deep learning algorithm was proposed that takes advantage of this information toward homogenization of the image data. This method was evaluated against four classical methods from the literature. The techniques were applied to the multicenter dataset.

Furthermore, the influence of different normalization technologies on the performance of deep learning networks was evaluated for tumor segmentation and prediction of the pathological response, sex, and age.

For segmentation, the preprocessed data was used with different normalization methods applied to it. There were 104 segmented patients from the study and 57 from the in-house data. The same data was also used for classification. The registered and B_1 bias corrected images were used.

3.4.1 Normalization Methods

The different normalization methods can be divided into the classical methods derived from statistics of the images and the deep-learning-based methods.

Classical Methods

The percentile (Perc) method is very simple. The 5th and 95th were used as the minimum and maximum values for the input of the network. We set any values outside this range to the corresponding minimum or maximum to eliminate outliers.

The second method is histogram matching (HM), originally developed for brain images (Nyul et al., 2000). The idea is to extract landmarks from each image and then average them over all images. As landmarks, the 1st, 10th, 20th, . . . , 90th and 99th percentiles of the voxel intensities were chosen.

The average of the landmarks is used to define a standard histogram. Then, intensities are interpolated to follow this standard histogram. The 1st percentile was used as the minimum value for the input of the neural network and the 99th percentile as the maximum value. All values outside this range were clipped to that value.

The original paper suggests using Otsu thresholding to separate the brain from the background. Instead, landmarks were extracted from the center volume (measuring $180 \times 180 \times 100\,\mathrm{mm}^3$) because this region does not contain background voxels and there is less variation due to differences in patient anatomy.

A combination of the (percentile and histogram matching (Perc-HM)) methods was also tested. First, the images were normalized using the percentile method, and then the landmarks for the histogram matching were extracted from those images and histogram matching was performed.

As the fourth normalization method, a fixed mean and standard deviation (M-Std) was used. Here, the mean was chosen to be zero and the standard deviation to be one.

The simplest method uses a fixed window (Win). The window was chosen from zero to 3000 for the T2w and ADC images and from zero to 1000 for the b800 DW images. These values were chosen because almost all images lie in that range. The minimum value is subtracted from the images; then they are divided by the maximum value, and the rescaled to the input range of the network.

Deep Learning Method

For the deep learning-based normalization, different auto-encoders were used. Multiple discriminators were added that were trained to predict the acquisition parameters of the DICOM headers, the location, and scanner. The generator architecture is shown in Fig. 3.6.

It has a traditional CNN architecture, but the edge information of the input image is passed to the fully-upsampled output block to improve image quality. A Gaussian filter was applied before edge detection to propagate larger features but not noise.

Three different discriminators were implemented: first, for acquisition parameters, the output of the discriminators applied to the generated image should match the value in the acquisition protocol. If no value was provided, the median value of all images was used.

Second, for other variables, such as the scanner model or location, it was attempted to remove the information using the discriminator as an adversarial loss on the latent features or the generated image. Therefore, the desired result is the same probability for each class in the classification tasks.

Lastly, a real/fake discriminator that tries to detect which images are the original input images and which were generated by the generator was added. The generator tries to fool this discriminator. This was the same as in a standard GAN and prevents the generator from reducing the quality of the images to remove information.

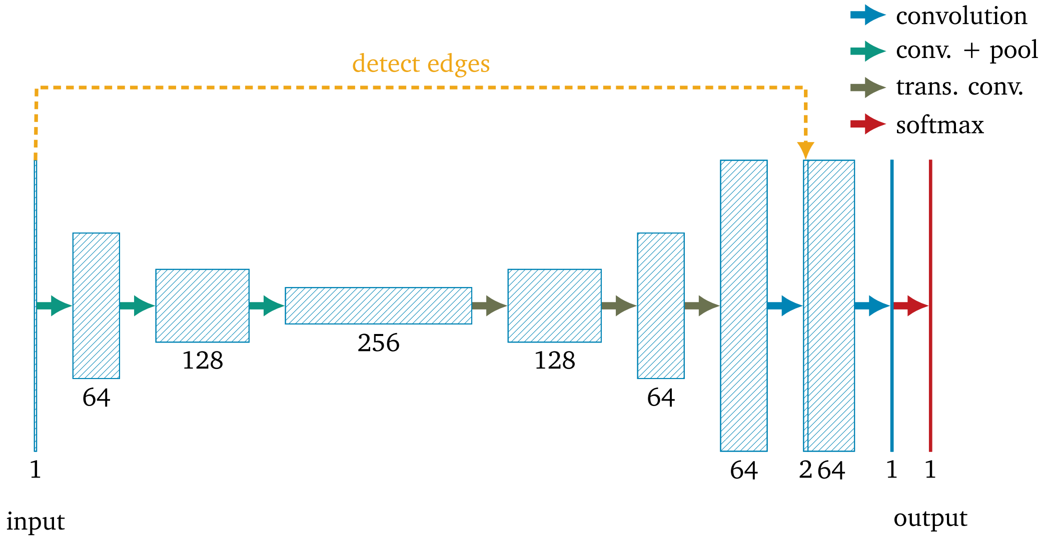


Fig. 3.6: The architecture of the auto-encoder. Each blue box represents a convolutional block consisting of a batch normalization, convolutional layer and activation. The numbers show the number of filters per convolution. On the contracting path, a stride of two halves the feature-map size in each dimension. On the expanding path, a transposed convolution doubles the size again. Once full resolution is reached again, the edge information is concatenated with the feature map, and another convolutional block is applied.

All discriminators are applied to the whole image, but their receptive field does not cover the whole image. All three discriminators consist of three convolutional blocks. Each block has a convolutional layer with a kernel size of 3×3 and a stride of 2. It thus performs a dilated convolution (see Fig. 2.7d). It is followed by a spatial dropout layer. This is a version of dropout, which drops entire feature maps, because dropping a single feature would have little impact for convolutional neural networks. LeakyReLUs are used as the activation function.

This means that the region of the input that produces one feature measures 13×13 pixels. The patch size in training was 128×128 pixels. The discriminator thus does not cover the whole patch, but the style due to different acquisition parameters should be very local.

To achieve the final result with the correct output dimensions, a convolutional layer with a kernel size of 1×1 is employed. This approach allows for multiple predictions for different overlapping regions in the image. For classification tasks, a softmax layer is added at the end. For the real/fake classification, no softmax layer was used. The prediction is then averaged and compared against the ground truth when training the discriminators.

Tab. 3.4: Hyperparameters for the different GANs used for normalization. The first is the default one (GAN-Def). For GAN-Seg, segmentation was added as an additional task to preserve important details. GAN-Img uses all discriminators on the images and not on the latent space. GAN-Win and GAN-No-ed were trained on images with window normalization with and without propagated edge information.

Network	Segmentation loss	Train only on image	Initial Normalization	Skip Edges
GAN-Def	No	No	Perc	Yes
GAN-Seg	Yes	No	Perc	Yes
GAN-Img	No	Yes	Perc	Yes
GAN-Win	No	No	Win	Yes
GAN-No-ed	No	No	Win	No

In each training step, the real/fake discriminator was trained first. Then, the image discriminator was trained on the original input and the generator's output images. The latent space discriminator was trained on the latent information of the original images. The generator already needs images in a certain range as input, so one of the classical normalization methods was used to normalize the images before training the auto-encoder.

An auto-encoder was trained individually for each set of training images and each modality. For training, all the images available for that modality were used, not only the segmented images. Different hyperparameters were used, which are listed in Tab. 3.4.

3.4.2 Experiments

For treatment response, the tumor regression grade (TRG) was predicted. This was done as a classification task with five classes for the different TRGs. Sex and age were given in the patient data and also in the DICOM header.

One network was trained for segmentation and another for classification and regression. A modified 2D U-Net (see Section 3.2.2) with batch normalization and residual connections was used for segmentation. As an architecture for classification and regression, we chose a ResNet50 (see Section 3.2.1), which we used with random initialized weights. Only the last layer of the ResNet was changed to have the desired number of output neurons.

All networks were trained for 100 epochs with a 5-fold cross-validation. The three modalities (T2w, ADC, and b800) were combined into a 3-channel image and 32 random patches per image were extracted in each epoch. For segmentation, at least 40 % of patches had their center inside the tumor volume. The patches were augmented by rotating them in the plane and uniform spatial scaling.

The networks were trained in three configurations:

All In this configuration, the networks were trained on images from all centers and evaluated using cross-validation.

Except-One In this experiment, the networks were trained on all centers except one. The performance of the training center was evaluated using cross-validation, and the networks of each fold were evaluated on the remaining center.

Single-Center In the last configuration, the networks were trained on one center only. Similarly to the Except-One experiment, the performance on that center was evaluated using cross-validation, and all networks were applied to the other centers and evaluated.

The first experiment was performed once, the other two three times, with images from center 1, center 11 and center 13 being excluded from or being used exclusively as training data. The centers were chosen because they had the highest number of usable images.

The three modalities (T2w, ADC, b800) were normalized individually. The Perc, M-Std and Win methods do not need to be trained, so the whole dataset was only normalized once. For HM, Perc-HM and the deep learning method, the methods were trained for each experiment on the patients included in the training and validation set.

After training the networks, they were evaluated using the network from the epoch with the best performance in the validation set. To reduce noise, a moving average with a decay rate of 0.3 was used to smooth the validation loss.

As an evaluation metric, the DSC of the tumor class was used to evaluate the segmentation performance and the AUC for the evaluation of the prediction of sex and TRG. In the case of several classes, the AUC was calculated by the average of the one-versus-others AUCs. The RMSE was used as the metric for age prediction.

The confidence intervals of the AUC were calculated by bootstrapping using 1000 samples. Each sample was generated by drawing from the scores predicted by the network and had the same number of samples, but duplicates were allowed. When comparing two metrics, the Student's t test was used to determine the significance of the mean differences. A p-value of less than 0.05 was considered significant.

3.5 Classification

Accurate prediction of the tumor response to neoadjuvant treatment would be very beneficial for patients, because patients with cCR could be put on a watch-and-wait regimen instead of having a surgery.

The problem is that it is currently not possible to detect complete response with imaging or endoscopic procedures. Therefore, such approaches are not recommended at the moment but is under active investigation (Oronsky et al., 2020). Finding a method to accurately determine cCR could greatly improve the outcome for these patients, by avoiding invasive surgery and adopting a watch-and-wait approach instead. Prediction in an early stage of treatment could also be used to adapt treatment and increase survival.

This poses a task well-suited for deep learning. The pCR (see Section 2.5.3 on page 49) is used as the ground truth, because it is currently the best indicator of the treatment response, but only available after surgery.

For the deep learning-based approach, a Siamese U-Net was used (see Fig. 3.7). The network was developed by Jin et al. (Jin et al., 2021). It is a multitask network, which uses the U-Nets to segment pre- and post-therapy images in addition to classification. This should make training easier, especially for small datasets.

The classification is done by extracting features at multiple levels from both U-Nets. A depthwise convolution is applied to the features before concatenating them and using a dense layer for the final classification.

The code for the network, preprocessing, and training was provided in an online repository[3]. There were some discrepancies between the published code and the article, so the authors were contacted for further clarification. The network was used as published in the repository; only the number of input channels was changed, because Jin et al. used T1 weighted images as an additional modality.

3.5.1 Data Processing

The registered and N4 corrected dataset was used. Diffusion images were denoised using Marchenko-Pastur PCA (Veraart et al., 2016) from the MRtrix3 software package (Tournier et al., 2019). Additionally, to registering the diffusion to the T2w images, which was already done for the dataset, the images from both time points

[3] `https://github.com/Heng14/3D_RP-Net`, retrieved on December 22nd, 2022

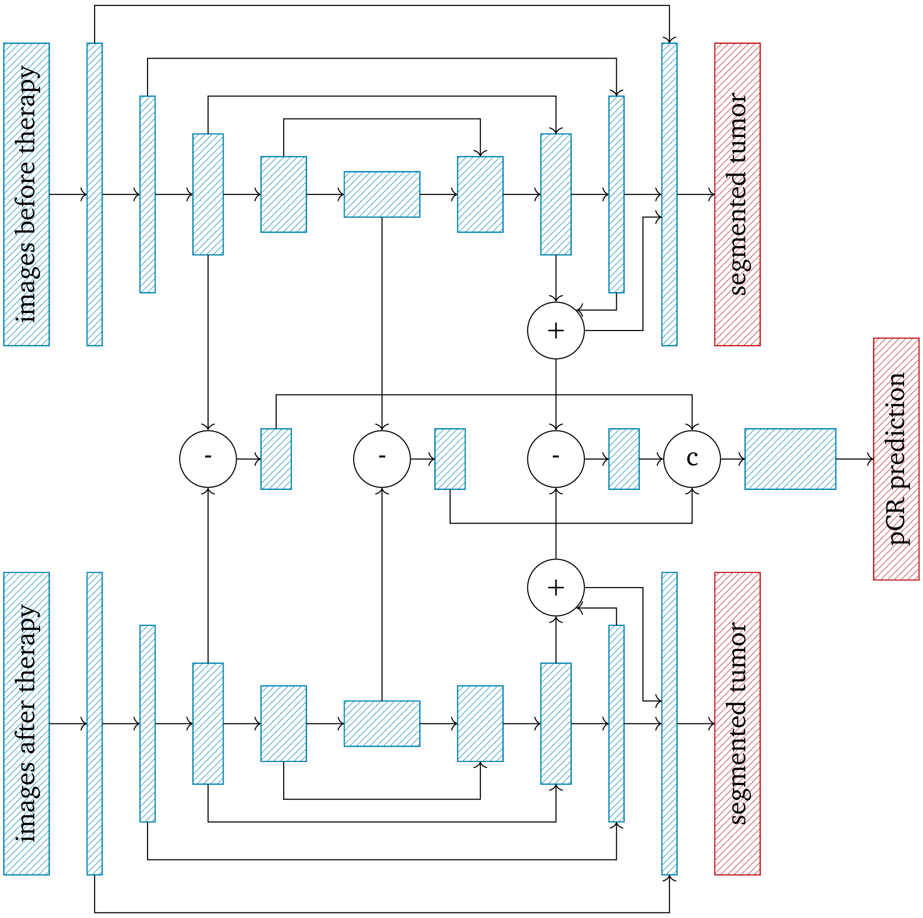


Fig. 3.7: This shows the working principle of the Siamese U-Net used for classification. Two U-Nets were used, which share their weights. The U-Nets have a few modifications. Residual connections are used for each block, and leaky ReLU is used instead of ReLU. After each convolutional layer, instance normalization is used. This is similar to batch normalization, but without averaging over the whole batch and averaging over all the instances in the batch separately. This is done, because the batch size is usually very small for 3D networks. Also, the pooling layers, which reduce the resolution at the end of each block, have been reduced by convolutional layers. The U-Nets are applied to the registered images before and after therapy. Then, the features are extracted at three points in the network. For this, the result of the third encoding block, the bottleneck block, and the added result of the second and third decoding blocks is used. The features of both networks are then subtracted from each other, and a $1 \times 1 \times 1$ convolution is applied to get a feature map with 32 channels. Global average pooling is then used to get a feature vector with 32 entries. All three feature vectors are then concatenated, and a dense layer is used to predict the pCR.

were registered to each other using rigid registration. This was done using ANTs with standard parameters. Afterward, the registration was checked visually.

Further processing was performed using the code provided by Jin et al., 2021. They used center-cropping to remove irrelevant parts of the image and normalized the images by setting the minimum value to zero and the maximum value to one.

For classification, patients were divided into two groups, patients with pCR and patients who did not have pCR. pCR was defined as TRG 4 (complete response / no tumor cells) and non-pCR as TRG of 0-3. This simplifies the prediction of the treatment response to a binary classification problem, making prediction and evaluation easier.

The TRG was available for most patients, except for some who did not undergo surgery. In the study data from the centers from which images were available, seven patients did not have a TRG. Two had surgery, but no grade. One with pathological tumor stage T0 and one with T3. Of the five patients who were not operated, three had a pre-OP tumor stage of T0, one had no tumor stage, and one a stage of T3. Only one of those patients had usable MRI images from both time points.

For training, two different datasets were used. One included diffusion-weighted images, and the other one did not. In this way, the impact of diffusion images on the diagnosis could be investigated.

3.5.2 Training

To test the generalization, the data from Mannheim (Center 13 of the study and the in-house data) was excluded from this dataset. The remaining dataset was divided into three parts. 68 % were used for training and 12 % for validation. 20 % were used for testing. A five-fold cross-validation was performed, so the training was performed five times with five different, non-overlapping test sets.

In this way, there was one set of data from an unseen center of the study, which should be more similar to the data from the rest of the study, and one set of data with similar images from the clinical routine. Thus, three different types of generalization were tested. In the cross-validation test set, there were previously unseen patients from seen centers and scanners. The study dataset from Mannheim contains data which should follow the same protocol (for the treatment and imaging), but from an unseen center and unseen scanners. The routine clinical dataset was similar, but follows a slightly different treatment protocol.

The models were trained for 100 epochs, with an initial learning rate of 0.01. The learning rate was halved if the loss did not improve for more than 30 epochs. To monitor the loss, the validation set was used. This was also used to evaluate the network after each epoch. For the final network, the weights with the lowest validation loss were used. This was evaluated using a moving average to remove random fluctuations.

3.5.3 Analysis

Each of the resulting networks was evaluated on its test set and on the unseen data from Mannheim. In this way, it can be tested if the network generalizes to data from a different center from the same distribution and to data from a slightly different distribution.

Performance was compared using the AUC. Confidence intervals were calculated by bootstrapping using 1000 samples. Samples without patients with pCR were skipped. For comparison between AUCs, the fast implementation of De Long's method (Sun and Xu, 2014) was used.

3.5.4 Baseline using clinical data

As a baseline, two simple machine learning models were trained on the clinical data. A random forest and a logistic model were used, as implemented in Scikit-learn (Pedregosa et al., 2011). As features, the age, sex, pre- and post-therapy T and N stage and the minimum distance to the circumferential resection margin (CRM) were used.

A random forest is an ensemble of tree predictors, which are each trained using a random subset of the features and the dataset (Breiman, 2001). This makes them more robust to noise compared to single decision trees and, according to Breiman, 2001, they do not overfit.

For the random forest used to predict the treatment outcome, the number of trees was set to 100. The criterion for the best split at a node in the decision tree was the Gini impurity. The size of the trees was not limited. Thus, the training data is divided using the features until all leaves contain only one class. The importance of the individual features was calculated using the Tree SHAP algorithm (Lundberg et al., 2020).

A logistic regression is a linear classifier. A sigmoid shaped function is applied to a linear combination of feature values to get an output value between 0 and 1. For the logistic model used here, a L2 penalty term was used with a regularization strength of 1.

For preprocessing, the data was converted into numerical values. The categorical data was binary (such as sex) or ordered (such as the *TNM stage*), so each category was assigned a number. Then each column was normalized by subtracting the mean and dividing by the standard deviation.

The data from the training centers was split into 80 % used as the train dataset and 20 % used as the test dataset. This was done at random 2000 times to generate sufficient error statistics. Folds where there was not at least one patient with pCR in all sets were skipped, as this makes it impossible to calculate AUC values. Each model was also evaluated on the test patients from the study and the in-house test patients.

This was done for all features present in the study and in-house dataset. Two sets of features were used. In the first set, all features were used, and in the second one, only features known before the beginning of therapy were used.

4 Results

This chapter is divided into three sections. In the first Section 4.1, the characteristics of the dataset are presented, measuring the data quality and heterogeneity. The second Section 4.2 reports the results of the normalization. Parts of this chapter have been published in Albert et al., 2023[1]. Finally, the third section presents the results of the network trained for classification of the treatment response. Parts of this chapter have been published in B. D. Wichtmann, Albert, et al., 2022[2].

The following notation is used: If the standard deviation is given, it is written as mean (standard deviation). When the standard error of the mean is provided, it is written as mean ± error. For confidence intervals, the values are written in brackets.

4.1 Characterization of the dataset

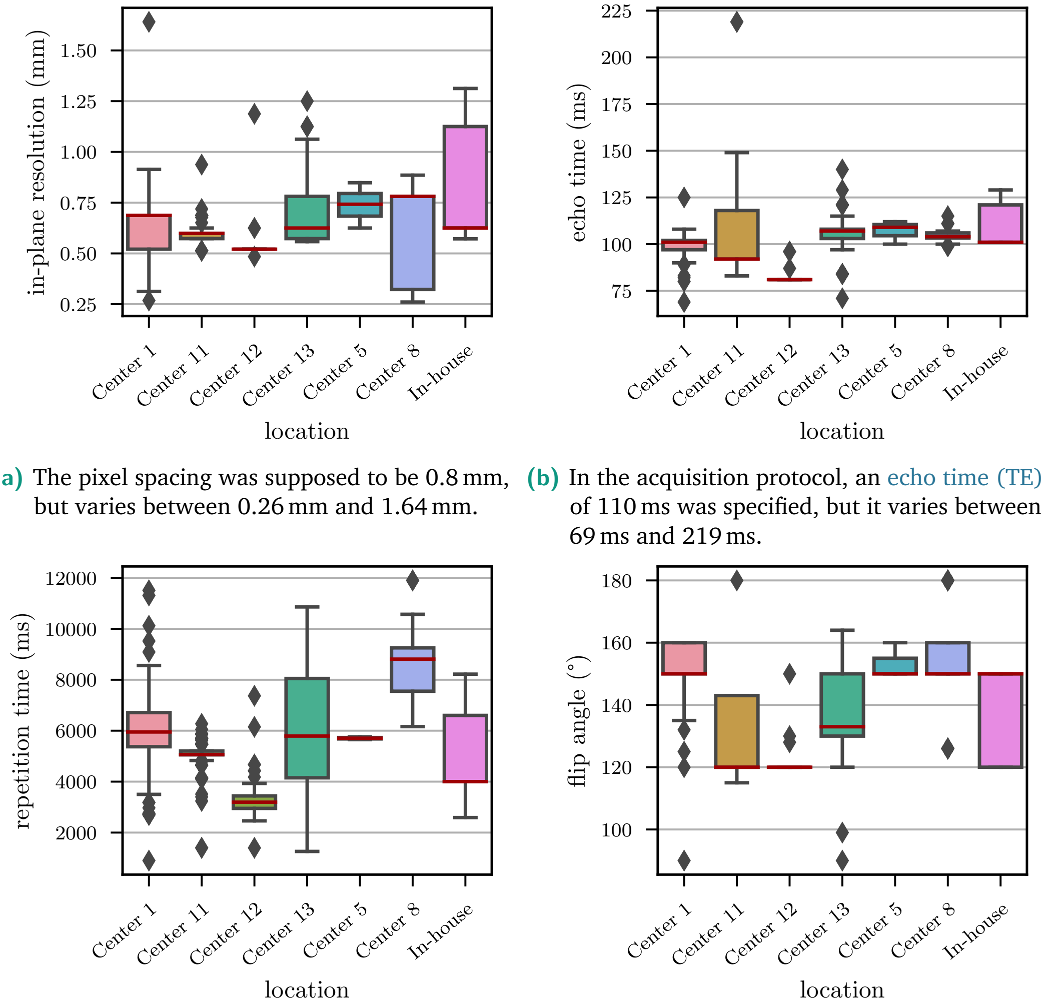
For machine learning, the dataset has a large influence on the results. Neural networks struggle to generalize from one data distribution to another. The heterogeneity of the imaging data is compared using different characteristics in this section.

4.1.1 Images

Although the same imaging protocol was used, the imaging parameters still varied substantially between the different scans. Selected imaging parameters can be seen in Fig. 4.1. It is visible that there was high variation between the centers, but also within one center.

For the DW images, no parameters were specified besides the b-values, and therefore the variation was larger. Due to the lower signal-to-noise ratio, DW images were measured with lower resolution, with an in-plane resolution of 1.7(0.3) mm and

[1] © 2023 by the authors. Licensee MDPI, Basel, Switzerland. 🅭🅯
[2] © 2022 by the authors. Licensee MDPI, Basel, Switzerland. 🅭🅯



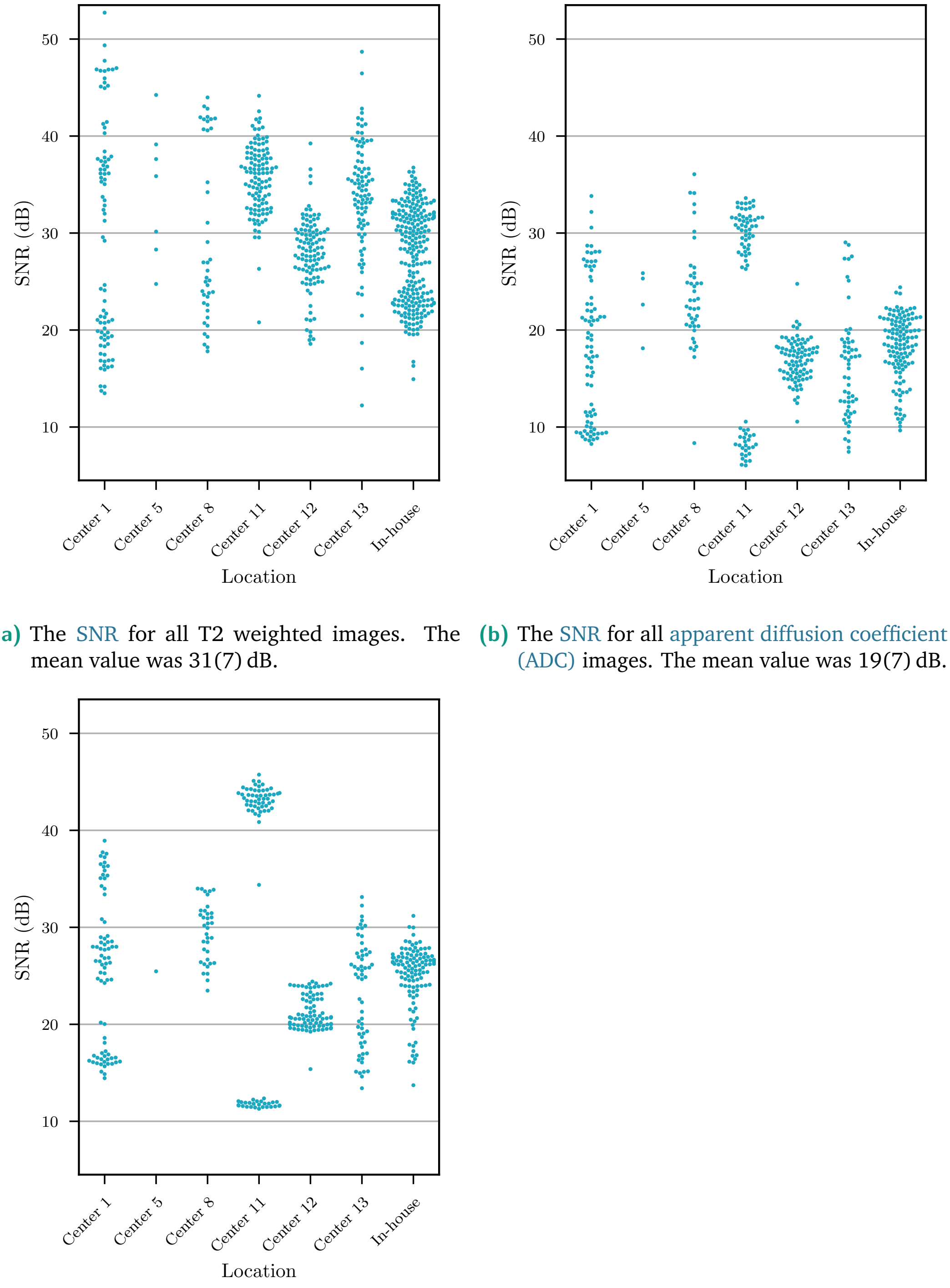
(a) The pixel spacing was supposed to be 0.8 mm, but varies between 0.26 mm and 1.64 mm.

(b) In the acquisition protocol, an echo time (TE) of 110 ms was specified, but it varies between 69 ms and 219 ms.

(c) According to the acquisition protocol, the repetition time (TR) was supposed to be 5600 ms, but varies between 900 ms and 11 900 ms.

(d) The flip angle was not specified in the acquisition protocol, it varies between 90° and 180°.

Fig. 4.1: The distribution of selected data acquisition parameters for the T2-weighted axial images. The boxes show the 25th and the 75th percentile. The median is shown as a red line. The slice thickness varied less between the data of the different centers, and most images had a slice thickness between 3 mm and 4 mm. A maximum slice thickness of 3 mm was specified in the imaging protocol. The parameters were taken from the DICOM headers.

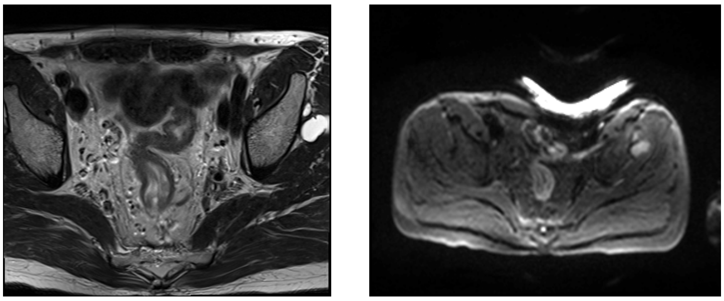


(a) The SNR for all T2 weighted images. The mean value was 31(7) dB.

(b) The SNR for all apparent diffusion coefficient (ADC) images. The mean value was 19(7) dB.

(c) The SNR for all diffusion-weighted (DW) b800 images. The mean value was 26(8) dB.

Fig. 4.2: Signal-to-noise ratios (SNRs) for the different modalities used in this work. On the x-axis, the points are grouped by location. The SNR is shown on the y-axis on the logarithmic decibel scale.



(a) T2 weighted image (b) b800 diffusion weighted image

Fig. 4.3: Example for diffusion artifacts. In the T2w image on the left, no artifacts are visible. The b800 image on the right shows very strong distortions. Artifacts such as this were common in the DW images, but most did not affect the area directly around the tumor, which was located at the center of the volume.

a slice thickness of 5.0(0.9) mm. The TE was lower with 70(10) ms than for the T2-weighted images, which had an TE of 101(15) ms.

The SNRs for the different modalities can be seen in Fig. 4.2. The T2-weighted images had the highest SNR, and there were no big differences between the centers. This was different for the DW images. There, distinct clusters can be seen and the variation between centers was greater.

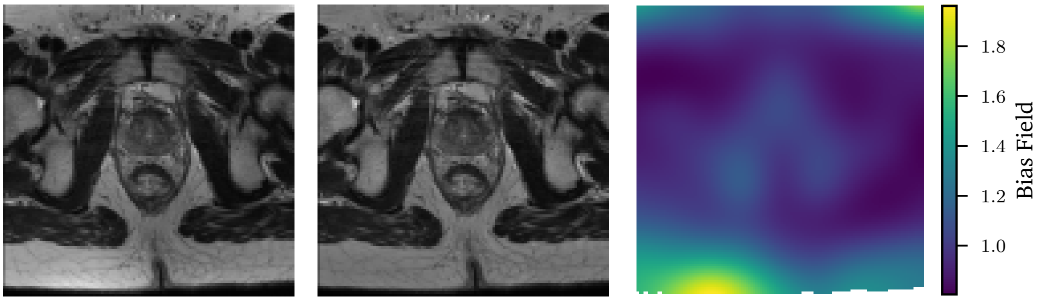
For the T2-weighted images, the image quality was usually good, and most artifacts were motion artifacts due to breathing, which did not affect the region around the tumor. For the DW-images, there were a lot more artifacts in the images (see Fig. 4.3).

The N4-Correction worked well, with small corrections for most images. The resulting images were visually examined and seemed reasonable. An example can be seen in Fig. 4.4.

4.2 Normalization

As seen in the previous section, the dataset is quite heterogeneous, which is a problem for deep-learning, especially for small datasets. The heterogeneity can be reduced by using normalization. This also helps the network by transforming all intensities into a well-defined range and should increase the generalizability of the networks.

The generalization was investigated using three different scenarios. In the first scenario, training and evaluation was performed on data from all centers (*All*). For



(a) The original T2 weighted image **(b)** The image after applying the N4 correction **(c)** The bias field, the units are relative to the intensity

Fig. 4.4: An example for the N4 correction for one image. Visible in the image (a), there is higher signal close to the surface, especially at the bottom left, where the patient is lying on the coil included in the patient bed. These low-frequency components were calculated in the bias field (c) and removed in the corrected image (b).

the second scenario, one center was left out of the training dataset and used to test the generalizability (*Except-One*). In the last scenario, only one center was used for training and the rest for testing (*Single-Center*). For the last two scenarios, the metrics were evaluated separately for the patients from the centers used in training and those just used for testing. The results are presented in the following three subsections.

The images were first normalized using the different methods described in the Methods Section 3.4.1. Fig. 4.5 shows examples of normalized slices compared to unprocessed slices. After normalizing the images, the two networks were trained for the three scenarios.

There are ten different normalization methods, so twenty networks had to be trained when training on all centers and sixty each when leaving one center out or training on just one center. The results obtained when evaluating the data from the training centers are summarized in Fig. 4.6. The generalization performance can be seen by looking at the patients from centers not used in the training set, which can be seen in Fig. 4.7.

4.2.1 All

Looking at all centers, there were no significant differences in segmentation performance when using batch normalization. The best method was percentile and histogram matching (Perc-HM) with a Dice similarity coefficient (DSC) of 0.69 ± 0.01 and the worst was the GAN with segmentation task (GAN-Seg) with a DSC of 0.67 ± 0.01, but the differences were not statistically significant.

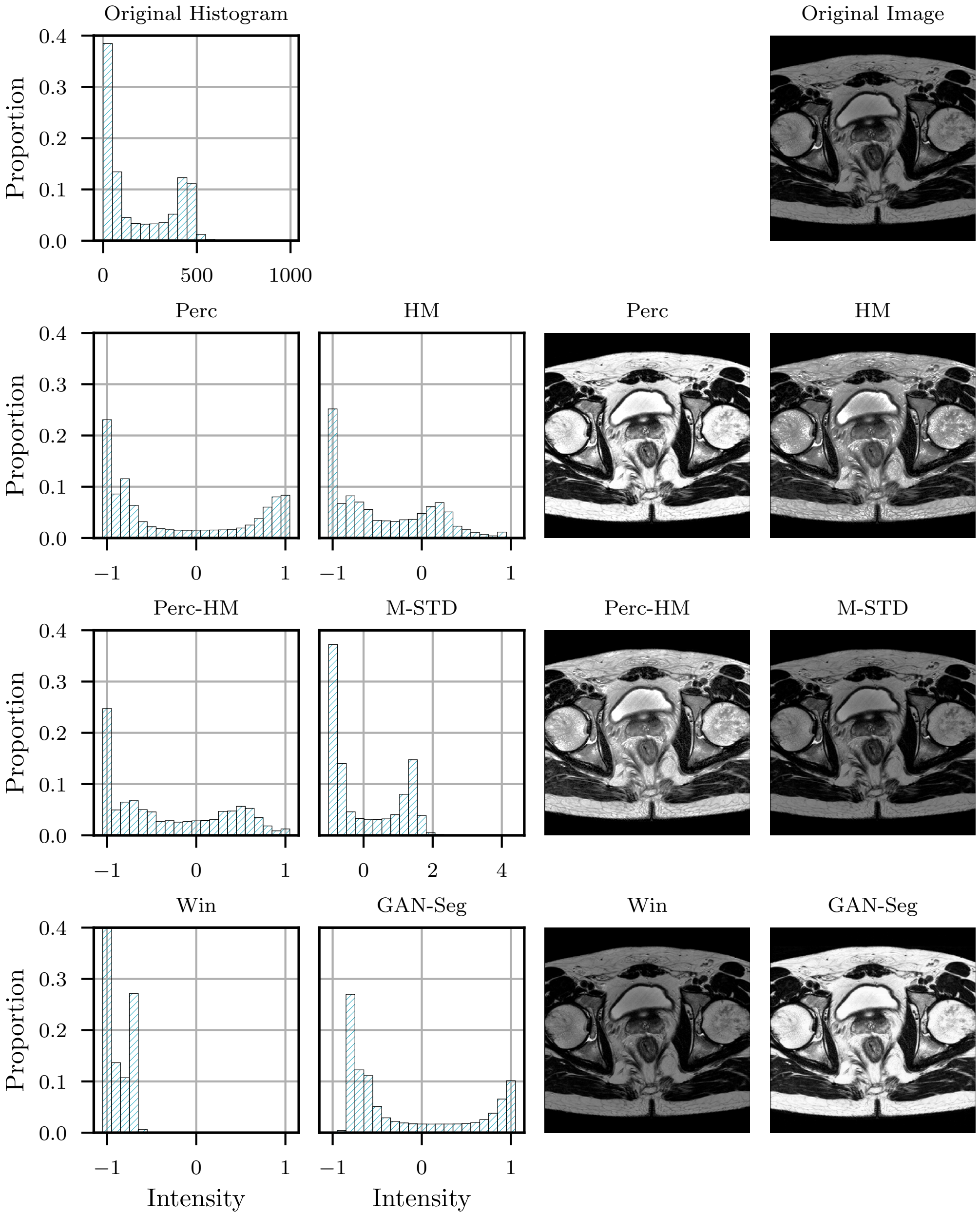


Fig. 4.5: Visualization of six exemplary normalization methods applied to an exemplary image. The upper row shows the original image and a histogram of the intensity, and the other four rows the resulting histograms and images.
The images were normalized to the minimum and maximum values of the resulting slice. Using a fixed window (Win) or subtracting the mean and dividing by the standard deviation (M-Std) only shifts and rescales the values. Thus, the images look the same as the original image. For the fixed window, a maximum value must be selected that is higher than the intensity of most voxels in most of the images; therefore, many images only use a small part of the available range. The other methods result in intensities between -1 and 1 (other values can also be selected). Areas with high intensities are mostly fat, urine, and bone marrow.

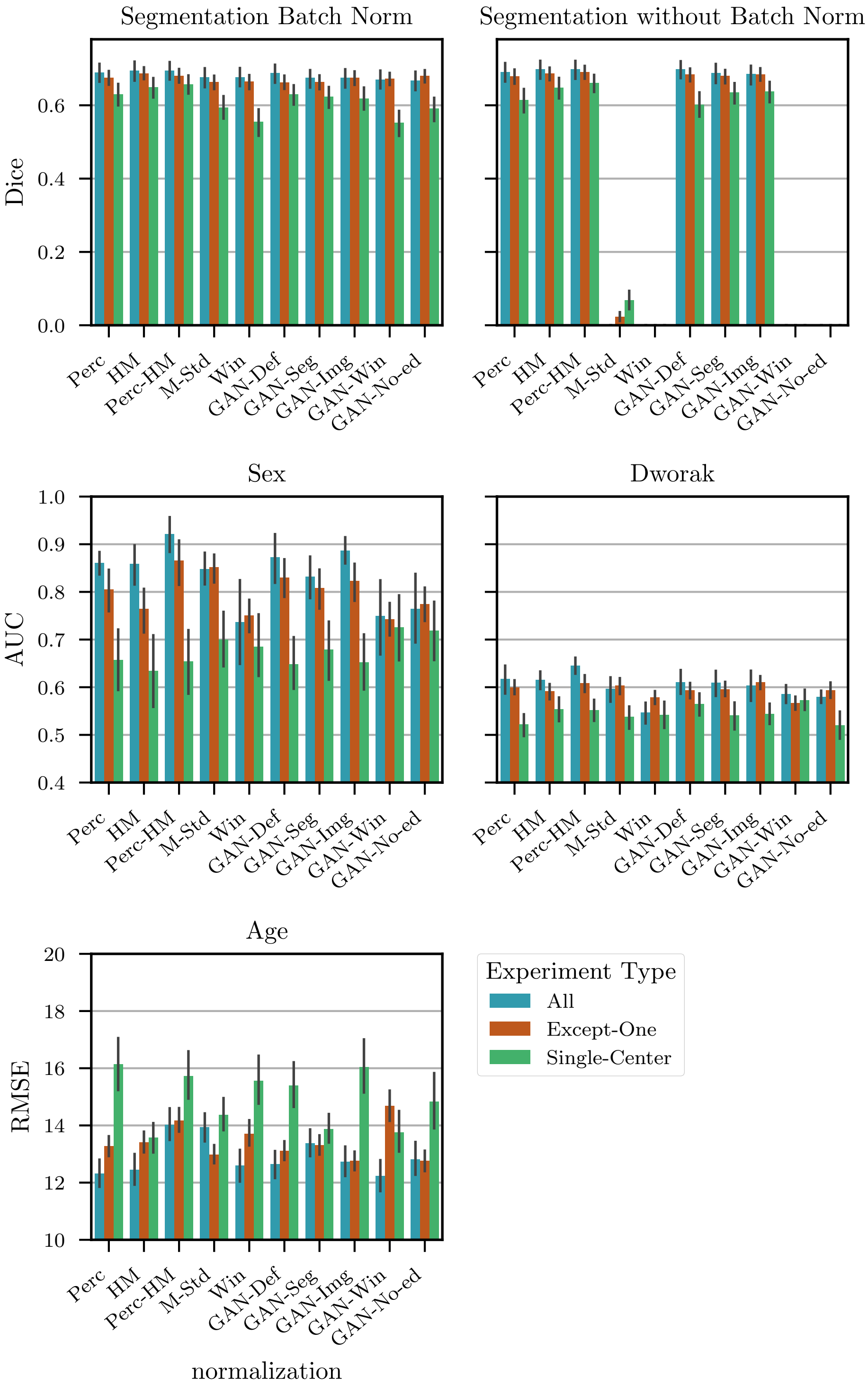


Fig. 4.6: Performance of the different normalization methods for each task and training scenario when being evaluated on the training centers (All, Except-One and Single-Center). The error bars show the 95 % confidence intervals.

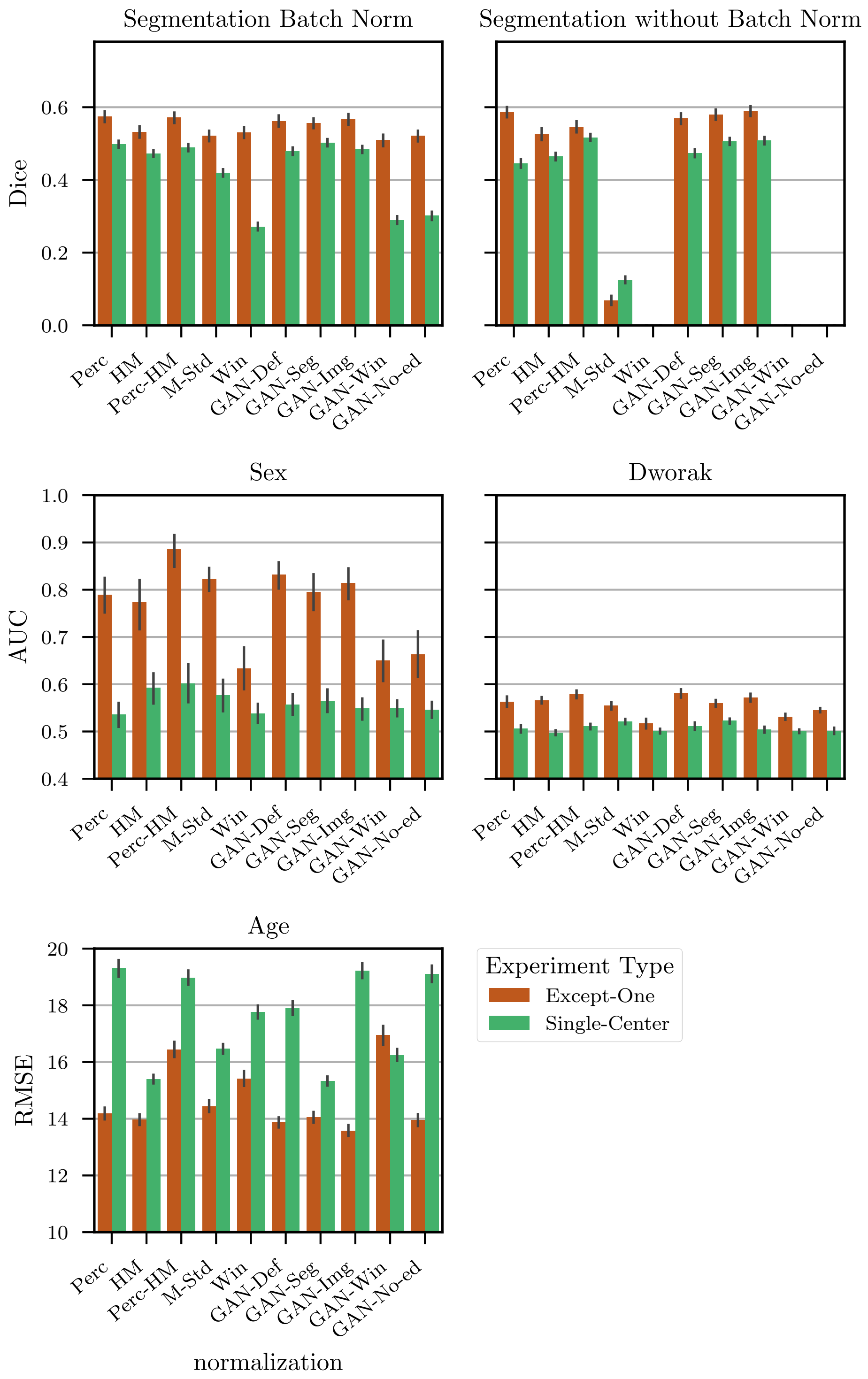


Fig. 4.7: Performance of the different normalization methods for each task and training scenario when being evaluated on the test centers (Except-One and Single-Center). The error bars show the 95 % confidence intervals.

Without batch normalization, segmentation did not work for M-Std, Win, GAN-Win and GAN-No-ed, which all had a DSC of zero. The performance of the other normalization methods did not change significantly. The best method was GAN-Def with a DSC of 0.70 ± 0.01.

For the sex classification, Perc-HM was significantly better than all other methods with an area under the receiver operating characteristic curve (AUC) of 0.94 ± 0.02. In general, the networks achieved good scores for sex classification, with a mean AUC of 0.85 ± 0.07. The worst methods were Win, GAN-Win, and GAN-No-ed with AUCs between 0.75 and 0.79.

Perc-HM was also significantly the best method for the prediction of pathologic Complete Response (pCR) with an AUC of 0.67 ± 0.01. The mean score of all the methods was 0.62 ± 0.03. The worst methods were Win, M-Std, GAN-Win and GAN-No-ed, without significant differences between the methods.

When predicting age, percentile (Perc), Win, GAN-Def, GAN-Img, GAN-Win, and GAN-No-ed were the best methods without significant differences, with Perc being the best with an root-mean-square error (RMSE) of 12.2 ± 0.2 years. M-Std was the worst method with an RMSE of 13.7 ± 0.2 years. The mean RMSE was 12.7 ± 0.5 years.

4.2.2 Except-One

When leaving out one center, all normalization methods achieved a similar DSC between 0.66 and 0.69 for unseen patients from the training centers when using batch normalization. For the test center, Perc, Perc-HM, GAN-Def, GAN-Seg, and GAN-Img were the best methods with no significant differences. The best was Perc with a DSC of 0.58 ± 0.01.

Without batch normalization, M-Std, Win, GAN-Win and GAN-No-ed performed very badly again. M-Std achieved an DSC of 0.02 ± 0.06 in training and 0.07 ± 0.01 on data from the test centers, the rest had a DSC of zero. The other methods did not perform significantly better or worse than those with batch normalization on images from the training and testing centers.

Sex classification worked best if images were normalized using Perc-HM for data from the training and test centers. For images from training centers, the AUC was 0.87 ± 0.04, but only histogram matching (HM), Win, GAN-Win, and GAN-No-ed were significantly worse. For images from the test centers, Perc-HM was significantly the best method, with an AUC of 0.88 ± 0.02.

When classifying pCR for patients from the same centers used in training, there were no significant differences between normalization methods. The mean AUC was 0.59 ± 0.01. When evaluating on data from the test centers, Perc-HM, GAN-Def, and GAN-Img were the best methods. GAN-Def has the highest DSC of 0.581 ± 0.004.

For age prediction, GAN-No-ed performs best with an RMSE of 12.7 ± 0.2 years for patients from the same center, but was not significantly better than GAN-Def, GAN-Img and M-Std. GAN-Img was the best method for data from the test center with an RMSE of 13.6 ± 1.0.

4.2.3 Single Center

When looking at the performance of the segmentation of images from the training center for networks using batch normalization, Perc-HM achieved the highest mean DSC of 0.66 ± 0.01. However, it was not significantly better than all other methods, besides M-Std, Win, GAN-Win and GAN-No-ed (see Fig. 4.7).

For evaluation of images from all other centers, Perc and GAN-Seg were the best methods with a DSC of 0.50 ± 0.01 (for both). But only GAN-Seg was significantly better than Perc-HM (with a p-value of 0.0496 barely significant), which achieved a DSC of 0.49 ± 0.01.

Networks not using batch normalization performed similar to those with batch norm except for the ones using M-Std, Win, GAN-Win or GAN-No-ed as normalization methods which had very low DSCs.

For the sex classification, there were no significant differences for the data from the training center. The mean AUC was 0.68 ± 0.03. For data from the test centers, Perc-HM was the best (AUC of 0.60 ± 0.02), but HM, M-Std, GAN-Def and GAN-Seg were not significantly worse. The networks achieved a mean AUC of 0.56 ± 0.02.

When classifying pCR, the best method was GAN-Win with an AUC of 0.57 ± 0.01, but only Perc, M-Std, and GAN-No-ed were significantly worse. For patients from the other centers, GAN-Seg and M-Std were significantly the best, with an AUC of 0.522 ± 0.003 and 0.520 ± 0.003.

However, for age prediction, the best normalization method was HM with an RMSE of 13.6 years, but not statistically significantly better than GAN-Win and GAN-Seg for images from the training centers. HM and GAN-Seg were the best for testing centers with RMSEs of 15.4 ± 0.8 and 15.3 ± 0.8.

4.3 Classification

For the pCR prediction, two different networks were trained. One just using the T2 weighted images and the other one using the DW b800 and ADC images as well. The networks were trained using five-fold cross validation, so ten different networks had to be trained.

The receiver-operator curves can be seen in Fig. 4.8. Shown are the receiver operator characteristics for the five folds, as well as the curve for all folds, with and without averaging the predictions for each patient. The AUCs can be seen in Fig. 4.9.

Training Centers

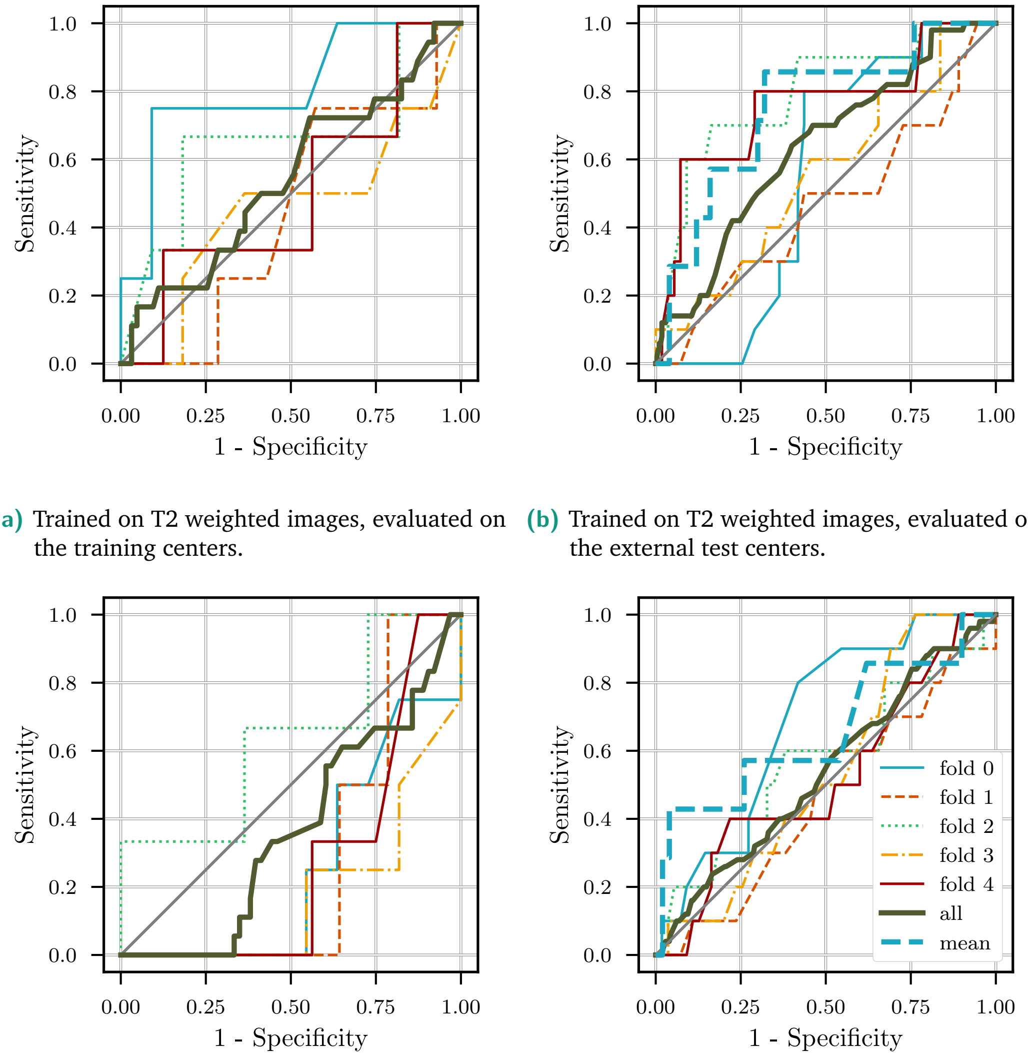
The networks achieved a higher AUC, when using the test set from the training centers of 0.58 (95 % CI: 0.42 to 0.75) using just T2w images, compared to an AUC of 0.39 (95 % CI: 0.26 to 0.53) when all modalities were used. For the external test set, the result using only the T2w images were also better with an AUC of 0.64 (95 % CI: 0.55 to 0.72) compared to 0.54 (95 % CI: 0.46 to 0.63) for all three modalities.

Averaging the predictions from all five folds further increases these scores. The AUC using the T2w images was then 0.75 (95 % CI: 0.52 to 0.92) and 0.66 (95 % CI: 0.41 to 0.89) when also including the DW images.

External Test Data

For the T2w images, the classification score was 0.64 (95 % CI: 0.55 to 0.72) when not averaging the scores and 0.75 (95 % CI: 0.52 to 0.92) when averaging them. If all three modalities were used, the AUC when evaluating on all external data was 0.54 (95 % CI: 0.46 to 0.62) without and 0.66 (95 % CI: 0.41 to 0.89) with averaging predicted scores.

When looking at the external study data, the AUC in the study data was very high for the T2w images with 0.8 (95 % CI: 0.7 to 0.9) without and 1.0 (95 % CI: 1.0 to 1.0) with averaging of the predictions. Using the T2w, b800 and ADC images, the results were 0.49 (95 % CI: 0.36 to 0.63) without and 0.42 (95 % CI: 0.18 to 0.81) with averaged predictions.



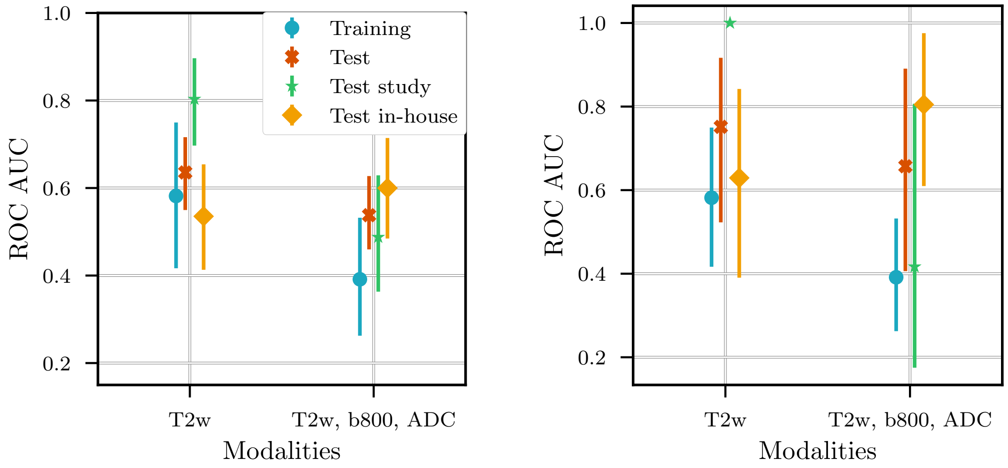
(a) Trained on T2 weighted images, evaluated on the training centers.

(b) Trained on T2 weighted images, evaluated on the external test centers.

(c) Trained on T2 weighted, ADC and DW b800 images, evaluated on the training centers.

(d) Trained on T2 weighted, ADC and DW b800 images, evaluated on the external test centers.

Fig. 4.8: Receiver operator curves for pCR classification using different modalities for the training and evaluated on unseen data from the training and external test centers. The curves of the individual models from each fold are shown as thin lines.
The thick green line shows the curve when using data from all folds without averaging the prediction for one patient, and the thick blue dashed line shows the curve when averaging the predictions from all folds for each patient. Averaged over all folds was done only for data from the external test centers, because each patient was only in the test set once for the training centers.



(a) Without averaging the predictions. **(b)** With averaging the predictions.

Fig. 4.9: AUC scores for the networks with the different input modalities. The AUC was calculated when evaluating on data from the training centers. For the test center, the models were evaluated on the data from the study and the in-house data (combined and separately). The error bars show the confidence intervals.

For the data from clinical practice, it was the other way around. There, the result was better when using all modalities with an AUC of 0.60 (95 % CI: 0.48 to 0.71) without and 0.80 (95 % CI: 0.61 to 0.98) with averaging of the predictions. When using the T2w images, the AUCs were only without 0.54 (95 % CI: 0.41 to 0.65) and with 0.63 (95 % CI: 0.39 to 0.84) averaging.

4.3.1 Baseline using clinical data

The classification results using clinical data can be seen in Fig. 4.10. In training centers, the best model was the random forest using all features with an AUC of 0.73(0.18), but the result using the pre-therapy features were only slightly worse with an AUC of 0.72(0.16).

According to the SHAP analysis, the most important feature was the patient age, for both sets of features. When using all features, the pre-OP T-stage was the second most important one, and the distance to the circumferential resection margin (CRM) was the third most important. For the pre-therapy features, the pre-OP T-stage was removed and the distance to the CRM was the second most important feature.

The results were the worst when using the entire test set, with the data from the study and the in-house data from the clinical routine combined. The AUCs

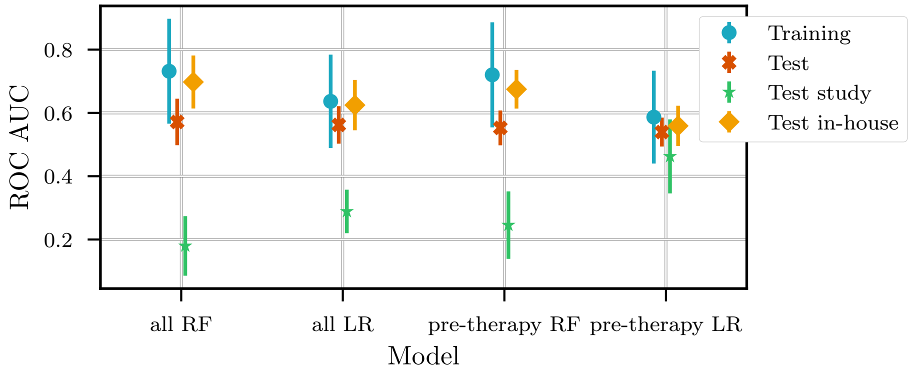


Fig. 4.10: The area under the receiver operator curve when using a random forest and logistic regression on the clinical data. The experiments in the left two columns use all available features to train the random forest (RF) and logistic regression (LR) models. On the right, the results when using only the features available before therapy are shown. The error bars show the standard deviation.

were between 0.54(0.05) (for logistic regression using the features available before therapy) and 0.57(0.08) (for the random forrest using all features). This reduction was even stronger for the study data in the test set, with a maximum AUC of just 0.46(0.12).

When evaluating on the in-house data, the results were better. The best results were achieved by the random forest, using all features with an AUC of 0.69(0.09). The random forest with the pre-therapy features was only slightly worse, with an AUC of 0.68(0.07). The model that performed the worst was the linear regression model using the pre-therapy features, with an AUC of 0.56(0.06).

5 Discussion

The goal of this project was to develop a pipeline for the processing of the images and prediction of the tumor response to neoadjuvant therapy, instead of developing a single network for classification or segmentation.

Using a retrospective dataset from a clinical study and clinical practice, which is discussed in the first section, different normalization methods were tested in Section 5.2. Parts of this chapter have been published in Albert et al., 2023[1].

The large heterogeneity, the small amount of available annotated data, and the large biases in the data are the main problems of machine learning in medicine. Many models look promising for the test data, but perform poorly in the clinic (Varoquaux and Cheplygina, 2022). Thus, the focus was not to develop methods that achieve a slightly better score on a benchmark, but to make them more robust to heterogeneous data.

5.1 Dataset

The heterogeneity of the dataset has a large influence on the deep learning performance. Thus, it is important to quantify the heterogeneity of the dataset, which was quite high for this dataset.

The dataset was close to clinical practice, but not very standardized. There was an acquisition protocol in place, but it was followed only loosely. The imaging parameters varied between centers, but also within one center.

In MRI, small changes in parameters, such as repetition time or echo time, can greatly alter the contrast of the images. Many images also had to be removed, because the wrong sequences were used (mostly fat suppression or with a contrast agent).

The differences were especially large for the apparent diffusion coefficient (ADC) images. Although the ADC is a quantitative value, the intensity values can still differ

[1] © 2023 by the authors. Licensee MDPI, Basel, Switzerland. 🅭🅯

greatly, even when the same protocol was used on the same scanner (Michoux et al., 2021). For different scanners with non-standardized post-processing, the variations were even larger.

The images were also recorded on a variety of scanners. Some patients were imaged on different scanners for pre- and post-therapy images, sometimes even with different field strengths. There were a total of 33 different scanners used for the study data. Only for the in house data, all images were acquired on the same scanner. This further increased the heterogeneity of the data.

The differences can also be seen in the signal-to-noise ratios, which vary widely. As expected, signal-to-noise ratio (SNR) was much higher for the T2w images, because the diffusion weighting greatly decreases the signal, especially for high b-values. Despite the larger voxels in the diffusion-weighted (DW) images (compared to the T2w images), the signal was still much weaker, especially for high b-values.

Differences in preprocessing can also be observed in the SNR. This was especially visible for Center 11. There were two clusters of SNRs for DW images. Some images were noisy, and others contained little noise, but also fewer details, so they were probably smoothed in post-processing.

Nevertheless, no distinct clusters were found. Many techniques rely on the presence of distinct clusters in the data. Substantial variation within individual centers further complicates normalization, as they cannot reliably define distinct domains within the dataset. Techniques, such as the StarGAN could thus not be used for normalization.

A large heterogeneity in the training data is a problem for all machine learning algorithms. It makes generalization harder (Mårtensson et al., 2020) and overfitting more likely. Larger datasets are needed to achieve the same accuracy, if the data is less standardized (George et al., 2020).

This problem is prevalent in the medical field, and the selection of research topics is partly determined by the availability of data (Varoquaux and Cheplygina, 2022). This is because the creation of annotated datasets is costly because of the expertise needed for annotation. There are also a limited number of patients with a given condition available.

This dataset presents a good test case for the generalizability of neural networks, because it is close to the data found in clinical practice, which will have to be used to create large datasets for the training of neural networks. Normalization is especially important in this case, because the data is quite heterogeneous.

5.2 Normalization

In the normalization experiments, a deep learning-based approach was proposed that incorporated image sequence parameters for image normalization. Furthermore, the influence of the normalization strategies implemented, including deep learning-based approaches to rectal cancer segmentation, classification, and regression from multimodal MRI acquired in a multicenter study was investigated.

For segmentation, the different normalization methods only led to minor differences in performance. For classification and regression, there were larger differences, and the best performing method was a combination of the percentile and histogram matching (Perc-HM) methods.

The intensity of the MRI signal depends mainly on the tissue properties of the imaged voxel. All normalization methods use local information (especially convolutional neural network (CNN) based methods) and/or global information (for most statistical methods) to standardize the images. This was not sufficient because voxels have different tissue properties that were not accounted for by normalization. This limits how well normalization models could correct for anomalies.

Therefore, normalization probably mostly helped the neural network by providing prior information (for example, the mean intensity distribution for histogram matching), which explains why the dependence of the performance on the normalization was less for larger datasets but did have a large impact for smaller datasets. This could also prove useful if the dataset used to train the normalization is much larger (by orders of magnitude) than the annotated dataset.

In addition to different acquisition parameters, there were many other parameters that might have hindered the generalizability of the trained network. There were differences in the patient population and treatment. For example, the time difference between the end of neoadjuvant treatment and the surgery was 30(6) d for Center 1 and 37(6) d for Center 2.

This difference is especially large when comparing the study and in-house data, as seen in Fig 3.2. For the study data, there was a lot less time between the restaging MRI and the surgery. The study data was acquired later, and it is now recommended to wait longer with the restaging MRI, because tumor regression can still occur after the end of therapy (West et al., 2016). Thus, the second MRI was performed closer to the therapy. Differences like this cannot be corrected by normalization, but can be a big challenge, especially for treatment outcome prediction.

5.2.1 Classical Normalization Methods

Segmentation

The best performing tumor segmentation network (Perc-HM) reached a DSC of 0.69 ± 0.01 and was in a similar range (0.68–0.85) as DSC values reported in the literature (J. Wang et al., 2018; Huang et al., 2019; J. Lee et al., 2019; Soomro et al., 2019; Trebeschi et al., 2017). There were no significant differences in segmentation performance between normalization methods when training on all data. For the other scenarios, the percentile (Perc) method, Perc-HM and the GAN with segmentation task (GAN-Seg) performed the best.

Classification

The best performing model, Perc-HM, for the pathologic Complete Response (pCR) classification had a lower AUC of 0.67 ± 0.01 compared to Shin et al., 2022 with an AUC of 0.82 ± 0.05, but our dataset was only a quarter of the size and was more heterogeneous. For the classification of sex and age, we could not find respective studies to compare.

Classifying the sex of the patients resulted in a high AUC of 0.94 ± 0.02 for Perc-HM. In some images, the sexual organs were visible and could be picked up by the network. Thus, it should be an easy task for the neural network to perform compared to pCR prediction.

Compared to the segmentation task, there were fewer examples, since the whole volume was classified, while for the segmentation, each voxel was assigned a label thus contributing to the overall performance, though certainly the voxels were not independent. This could explain why there were fewer differences between the normalization methods for segmentation compared to classification.

Regression

For the age prediction, the results were inconclusive. Using all data, the best models had an error of (12.2 ± 0.2) years, which was comparable to the standard deviation of the age over the entire dataset of 10.7 years. No study was found that was comparable to prediction of age from pelvic MRIs. Age could be predicted quite well (with a mean average error of (2.94 ± 0.03) years) from liver and pancreatic MRIs (Le Goallec et al., 2022), but those organs were not fully visible in the images.

Performance decreased for all tasks as the size of the dataset decreased, as expected. This was seen when leaving out one center, and especially when training only on one center. Here, the largest differences could be observed between the different normalization methods.

For data from unseen centers, there was a large generalization error. One of the reasons for this error was that there were larger differences in the data acquisition parameters between centers than within one, as mentioned in the previous section. Some normalization methods could better reduce these differences than others. Especially methods relying on fixed intensities and noise, such as using a fixed window (Win) or fixed mean and standard deviation (M-Std), had a problem in this case. Perc-HM worked quite well, maybe because the percentile normalization first reduces the number of outliers and the histogram matching then normalizes the intensity for the different types of tissue.

5.2.2 Deep-learning Normalization Methods

For segmentation and Dworak classification, the default GAN (GAN-Def) outperformed all other DL methods when training on all centers. Segmentation performance (0.69 ± 0.01) was comparable to classical methods, and literature (J. Wang et al., 2018; Trebeschi et al., 2017), but the pCR classification (0.64 ± 0.01) was worse than classical methods and literature (Shin et al., 2022).

The DL methods were superior to the classical ones in only a few cases when used in the other two scenarios. For example, GAN-Seg, which used segmentation as an additional task, was among the best methods for the Single-Center scenario and for segmentation in all scenarios. It achieved an AUC of 0.522 ± 0.003 for the pCR classification in the single-center scenario, and scores lower than the respective best classical model and the literature reference (Shin et al., 2022).

One of the issues was that the dataset used to train the GAN normalization was only slightly larger than the dataset used to train segmentation and classification. Thus, the U-Net and ResNet probably learned an encoding similar to that of the auto-encoder. The advantage of deep-learning-based methods is that they can be trained on a larger dataset without the need for manual annotations.

5.2.3 Summary

In summary, for tumor segmentation, pCR and sex classification, Perc-HM was the best method in the All and Except-One scenario. In the single center scenario, Perc-HM also performed well for the sex classification, but for the segmentation and the pCR classification, GAN-Seg was significantly better. For age prediction, the results were inconclusive; no method was superior in all three scenarios.

Histogram matching has also been shown to be useful in other MRI applications, especially in the brain (Shah et al., 2011; Carré et al., 2020; Um et al., 2019), but also in the prostate (Isaksson et al., 2020). But normalization in the abdomen is more challenging than in the brain, since there is more variation between patients, because of larger anatomical differences (such as the amount of visceral fat), and other variations.

For example, the intensity distribution in T2w and diffusion images can greatly change depending on, for example, how full the bladder is. Urine, which has a T2 on the order of seconds (Yoshimura et al., 2022), has a very high intensity in T2-weighted images. Because the bladder allows for mainly unhindered diffusion, it also has a high ADC.

This makes it complicated for any normalization method relying on statistics of the intensity distribution, and could explain why normalization in general has a stronger impact and works better in the brain. The intensity of different tissues in the abdomen is also not consistent, therefore the intensity of one tissue cannot be used well for normalization (Scalco et al., 2020).

In general, there were no huge differences between the normalization methods, except for very simple methods such as using a fixed window or subtracting the mean and dividing by the standard deviation. Therefore, intensity normalization should be performed, but a method that is easy to implement with low computational complexity can be chosen, such as histogram matching or histogram matching combined with the percentile method.

The overall performance was limited by the lack of an extensive hyperparameter search. In this experiment, only two networks were tested for their performance using different normalization methods. Although U-Net and ResNet are widely used, there are many other architectures, for example transformers, that are becoming more and more widespread in computer vision (Dosovitskiy et al., 2021).

For segmentation, most neural networks follow the encoder-decoder structure and should behave similarly, but might achieve better performance with more fine-tuning and more sophisticated architecture.

The maximum Dice similarity coefficient (DSC) of 0.69 ± 0.01 is in the range of the inter-observer DSC from the literature (0.68 in van Heeswijk et al., 2016, 0.71(0.13) in J. Wang et al., 2018 and 0.83(0.13) in Trebeschi et al., 2017). The performance could probably be further increased with more parameter tuning and more training data.

5.3 Classification

In addition to the ResNet used to test the normalization, the response to neoadjuvant treatment was also predicted by the network developed by Jin et al., 2021, which used segmentation as an additional task.

The network showed promising results in the original publication, with an AUC of 0.95 (95 % confidence interval: 0.91–0.98) and 0.92 (0.87–0.96) on two independent test cohorts, but did not translate well to the dataset presented in this work. There, only an AUC of 0.75 ± 0.01 could be achieved. The ResNet used in the normalization test only achieved an AUC of 0.66 ± 0.01 for the best-performing normalization method. Using the siamese U-Nets and the segmentation as additional tasks improved the classification performance considerably.

The dataset was very heterogeneous and also smaller than the one used in the publication, therefore lower performance was expected. Weights were not publicly available, thus the pre-trained model could not be tested. This makes it difficult to pinpoint what exactly the problem was.

With 321 patients in the training cohort, the training set used by Jin et al., 2021 was larger, but not by a factor of four. Therefore, the size of the dataset was probably not the only cause of the decreased performance.

Without access to their dataset, heterogeneity was difficult to assess, but all training data was collected at one hospital with an external validation set from a second hospital. There was also some variation in the acquisition parameters with systematic differences between the internal and external patient cohort, but the parameters varied less.

For example, the pixel spacing was the same for all patients in one cohort and the differences were only small between the cohorts (0.55 mm compared to 0.43 mm in-plane resolution for the T2-weighted images in the internal and external cohorts).

A big problem was the imbalance between patients with pathological complete response and without. Because of this, there were not many positive examples in the training and test cohort.

This also made the evaluation difficult. For T2-weighted images, the model achieved an AUC of 1 for the study data in the external dataset. But in that dataset there were only two patients with pCR, which means that the good result may very well be due to random fluctuations and might not perform as well if the patient cohort changes slightly.

Because of the previously mentioned problems, the model was unable to beat simple machine learning models using clinical data. For the training cohort, the random forest was better than the deep learning models, even when their predictions were averaged. On the training set, the random forest achieved an AUC of 0.73(0.18), which was 0.18 higher than the deep learning model.

For the in-house data, the results were similar when only T2-weighted images were used, with both traditional machine learning models being superior to deep learning models. When all the modalities were used and the predictions were averaged for all folds, the deep learning models were superior.

For the external study set, the deep learning models showed better performance, but as mentioned before, there were only two patients with pCR in that dataset. So, the low performance of the classical machine learning models and high performance of the deep learning models could be due to random characteristics of that dataset.

This was not very surprising, because even with a large dataset, it is hard to beat a model that uses a few well-tested clinical characteristics. For example, the distance to the circumferential resection margin has been shown to be a good predictor of the response and outcome of treatment (S.-H. Lee et al., 2005). Another important characteristic is the involvement of lymph nodes, according to Walker and Quirke, 2002.

These simple models also have other properties that are beneficial for clinical use: they are interpretable, easy to use, and have a strong scientific foundation, which are all desired properties (Shortliffe and Sepúlveda, 2018).

It is also unclear whether there is a benefit in adding the diffusion weighted images. For the external in-house data, the AUC increased when adding the DW images, but for the training dataset and the external dataset with patients included in the study, it decreases when using all three modalities.

DW images have the potential to improve the treatment outcome prediction (U. Attenberger and B. Wichtmann, 2015; U. Attenberger, Pilz, et al., 2014). So, it would have been interesting to see their benefit using a deep learning model.

A problem with the DW images is their heterogeneity, which was higher than for T2w images. This, combined with the low number of patients with pCR made it impossible to say whether DW images provide a benefit for tumor restaging. To obtain statistically significant results, a larger dataset is needed.

5.4 Limitations

As in many machine learning applications in medicine, the performance is limited by the size and heterogeneity of the dataset. This is especially true for the prediction of the tumor response.

One of the issues is that only about 20 % of patients achieve pathological complete response (Benson et al., 2015). Therefore, in the dataset, there were only 40 patients with pathological complete response (18 % of the patients). In addition to reducing the number of positive examples available, this also creates a large class imbalance, which can be problematic for CNNs (Buda et al., 2018).

There is an additional bias, because there were patients who did not undergo surgery after a good response to neoadjuvant therapy and instead opted for watch-and-wait. This further decreases the available data for patients with pCR and can change the distribution of the patients.

When predicting the pCR, there was no sufficient data to achieve a performance, which would be sufficient for the model to be used for clinical decision-making. This task is even impossible in many cases for physicians to achieve using only magnetic resonance images (Jang et al., 2020) or with CT and PET images (Guillem et al., 2013).

The best method achieved an AUC of 0.75 ± 0.01. This is not sufficient to inform treatment decisions. Even in studies with a larger training cohort of 592 patients, only an AUC of 0.82 was achieved (Shin et al., 2022).

Other studies claim better results, with an AUC as high as 0.97 in Z. Liu et al., 2017, but this study lacks an external validation cohort, so it is not clear how well their results would translate to a different cohort of patients. A hand-crafted radiomics model was used. Many radiomic features are known to have low repeatability (Schurink et al., 2022; Michoux et al., 2021; Dreher et al., 2020).

For segmentation, there was less of a gap compared to the literature. There, the small dataset size was less of a problem because there were a lot more data points. There were approximately ten million voxels in the tumor class for the segmented patients. Although these are not independent samples, there was still much more data available than for the classification.

The difference in available data for the different tasks was probably also the reason there were much larger differences in performance using different normalization methods for classification and regression than for segmentation.

The limitations due to the size of the dataset are difficult to overcome because it is difficult to collect more data. Collecting more data faces many regulatory and technical challenges. There is also a limited return on adding more data.

Techniques that greatly improve the size of the dataset, such as augmentation, can substantially improve performance, but the data set size needs to be increased by orders of magnitude (Brigato and Iocchi, 2021). Just slightly increasing the size of the dataset does not have a large impact on performance.

Due to the limited available data in-house and in the small uncontacted centers in the study, it would not have been possible to increase the size of the dataset more than by a few dozens of patients, which would probably not have greatly increased the performance. It was also very time-consuming to organize the data transfer from all centers and bring all data into a common format. Additionally, not all centers contacted were able to find the imaging data from the study.

6 Conclusion and Outlook

The goal of this thesis was to create a radiomics pipeline for the prediction of the treatment outcome for neoadjuvant therapy in locally advanced rectal cancer (LARC), which can work with heterogeneous datasets. All components were successfully implemented and tested.

It was possible to predict LARC tumor regression after neoadjuvant therapy better than chance, but for clinical applications, it is crucial to improve sensitivity for tumor response prediction and specificity. It was possible to achieve a similar result using a very simple predictor trained on a few clinical variables, which have been shown to be reliable in predicting the tumor response.

For normalization, performance was evaluated using six different normalization methods for different deep learning tasks in a multicenter setting with data from six different centers. A novel deep learning based approach was compared against five statistical methods. Different scenarios were tested with training on data from all centers, from all centers except one, and data from a single center. In this way, the influence of the normalization method on generalizability was assessed.

Normalization is vital when the data is inhomogeneous, especially if the dataset is small and the network is applied to data from a site not included in the training set. It was more important in classification and regression than it was in segmentation.

The results showed that percentile normalization followed by histogram matching performed the best for tumor segmentation and prediction of treatment outcome in locally advanced rectal cancer. Setting the mean and standard deviation to a fixed value, which is often done for images, performed significantly worse than most other methods.

The deep learning approach utilized an autoencoder trained with three adversarial networks, to remove location and scanner dependent information and transform the images to the style of the acquisition protocol. This method was only slightly better when training on data from a single center, but did not add any improvements over classical methods when training on all data or when leaving out data from one center.

Normalization improved the performance of the network on unseen data for different tasks. However, there are limits to what can be corrected with normalization, which is why that gap cannot be completely closed. It is essential to standardize data acquisition for routine clinical imaging for the widespread application of deep learning in clinical practice.

Outlook

The two main limitations of performance are the inhomogeneity of the data and the small size of the dataset. There are many approaches to solving those problems, which must overcome many technical, legal, and organizational challenges.

For data availability, there is a lot of hope in federated learning. The basic idea is that instead of sharing the data, the models will be trained at the different institutions and only the weights and not the data is shared.

An example is the joint imaging platform (Scherer et al., 2020). It can be integrated into the IT system of the participating clinics (currently 10 university hospitals) and the data can be sent directly from the Picture Archiving and Communication System (PACS). This would also make data sharing much easier. One of the hospitals sent the data for this thesis on CDs by mail, so there is a lot of room for improvement.

Although this has a lot of potential, many challenges remain. This does not fully address problems with the heterogeneity of the data, there can be a large bias between centers due to differences in the acquisition or patient population (Rieke et al., 2020). It also still has to be annotated, which is often a bottleneck.

Medical image datasets could also become larger because huge datasets are being created, such as the UK Biobank (Littlejohns et al., 2020) or the national cohort in Germany (Peters et al., 2022), which could provide much more available medical image data in the future.

These large datasets could play a similar role as ImageNet (Russakovsky et al., 2015) played for natural images and could be used to generate pre-trained models for various tasks in the future. This would mean that they would just have to be fine-tuned for specific tasks. But the question is whether these datasets are close to the data found in clinical practice.

With larger datasets, the deep learning based normalization could be greatly improved. The big advantage is that it can be trained unsupervised. Therefore, the normalization could be trained on a dataset that is much larger than the annotated dataset. In this thesis, the training dataset for the normalization was only about twice as large as the annotated dataset, but with a larger difference, the deep learning based models would probably outperform the statistical methods.

For LARC, it is questionable whether reliable treatment outcome prediction will be possible using only MRI images, which is also not possible for experienced clinicians to perform (Jang et al., 2020). After neoadjuvant treatment, it is very hard to tell fibrotic tumor tissue from vital tumor tissue, so the diagnostic accuracy is reduced greatly (U. Attenberger and B. Wichtmann, 2015).

However, medical imaging is not the only method to predict tumor response. For example, cell-free DNA can be used to detect residual tumor cells (W. Liu et al., 2022). Combining the genomic approach with radiomics can lead to better predictions than both approaches individually (Y. Wang et al., 2021; Chiloiro et al., 2021). Combining radiomics with blood-based tumor markers also increases prediction performance (Jin et al., 2021).

With newer functional imaging techniques, the tumor response prediction could also be improved. Diffusion imaging can be improved for example by using Kurtosis imaging, but this is challenging because very high b-value images are needed, which suffer from low signal-to-noise ratio (SNR) (Zhang et al., 2020).

Thus, hopefully it will be possible in the future to predict a complete tumor response, spare those patients from surgery, and thus improve their quality of life. In the future, deep learning-based treatment outcome prediction could still be viable, by combining the radiomics data with other omics and clinical data.

List of Publications

Journal Papers

- **Albert, S.**, Wichtmann, B. D., Zhao, W., Maurer, A., Hesser, J., Attenberger, U. I., Lothar R. Schad and Zöllner, F. G. (2023). "Comparison of Image Normalization Methods for Multi-Site Deep Learning". In: *Applied Sciences*, 13(15), 8923, DOI: 10.3390/app13158923.

- Wichtmann, B. D., **Albert, S.**, Zhao, W., Maurer, A., Rödel, C., Hofheinz, R. D., J.Hesser, F. G. Zöllner and Attenberger, U. I. (2022). "Are we there yet? The value of deep learning in a multicenter setting for response prediction of locally advanced rectal cancer to neoadjuvant chemoradiotherapy". In: *Diagnostics*, 12(7), 1601, DOI: 10.3390/diagnostics12071601.

Conference Contributions

- **Albert, S.**, Wichtmann, B. D., Zhao, W., Hesser, J., Attenberger, U. I., Schad, L. R., and Zöllner, F. G. "Comparison of Image Normalization Techniques for Rectal Cancer Segmentation in Multi-Center Data: Initial results" [Accepted as oral presentation], *Joint Annual Meeting ISMRM-ESMRMB & ISMRT 31st Annual Meeting*, London, 2022, `https://archive.ismrm.org/2022/0619.html`.

- Wichtmann, B. D., **Albert, S.**, dos Santos, D. P., Attenberger, U. I., and Baessler, B. "Test-retest repeatability of radiomic features derived from T2w MRI in prostate cancer patients" [Accepted as poster presentation], *Joint Annual Meeting ISMRM-ESMRMB & ISMRT 31st Annual Meeting*, London, 2022, `https://archive.ismrm.org/2022/2789.html`.

Supervised Thesis

- I. Gineitaite, “Rectal cancer treatment outcome prediction using radiomics”, Master’s Thesis, 2022.
- J. Hirsch, “Umwandlung T2-gewichteter MRT-Bilder in diffusionsgewichtete mittels einer CycleGAN”, Bachelor’s Thesis, 2022.

List of Abbreviations

ADC	apparent diffusion coefficient. 28, 30, 56, 79, 87, 88, 91
ANTs	Advanced Normalization Tools. 59, 74
AUC	area under the receiver operating characteristic curve. v, 65, 66, 71, 75, 85–87
cCR	clinical Complete Response. 52, 72
CNN	convolutional neural network. 32, 39, 40, 93
CRM	circumferential resection margin. 49, 75, 89
CRT	radio-chemo-therapy. 2, 50–52, 55
DICOM	Digital Imaging and Communications in Medicine. 58, 66, 70
DLR	Deep Learning-Based Radiomics. 46
DSC	Dice similarity coefficient. v, 37, 64, 65, 71, 81, 85, 86, 97
DW	diffusion-weighted. 25, 30, 56, 59, 68, 77, 79, 80, 87, 88, 92, 98, 99
ELU	exponential linear unit. 61, 62
EPI	echo-planar imaging. 19, 30, 31
FID	free induction decay. 18, 23
FOV	Field of View. 22
FSE	Fast Spin Echo. 25
GAN	Generative Adversarial Networks. 59, 63, 68
GAN-Def	default GAN. 70, 85, 86, 95
GAN-Img	GAN not using latent space. 70, 85, 86
GAN-No-ed	GAN without edge information. 70, 85, 86
GAN-Seg	GAN with segmentation task. 70, 81, 85, 86, 94–96
GAN-Win	GAN train with fixed window images. 70, 85, 86
GPU	graphical processing unit. 32
HASTE	Half-Fourier Acquisition Single-shot Turbo spin Echo. 25, 56
HCR	hand-crafted radiomics. 43

HM	histogram matching. 67, 71, 85, 86
IBSI	Image Biomarker Standardization Initiative. 44, 46
KL	Kullback-Leibler divergence. 37
LARC	locally advanced rectal cancer. v, 2, 3, 50, 51, 101, 103
M-Std	fixed mean and standard deviation. 68, 71, 82, 85, 86, 95
MRF	mesorectal fascia. 47, 49
MRI	Magnetic Resonance Imaging. 1, 3, 7, 9, 19, 49
NMR	Nuclear Magnetic Resonance. 7, 9, 14, 19
NN	Neural Network. 31, 33, 34, 36, 38
PACS	Picture Archiving and Communication System. 102
PCA	Principal Component Analysis. 45
pCR	pathologic Complete Response. 50, 52, 55, 72, 74, 75, 85–87, 94–96, 98, 99
Perc	percentile. 67, 71, 85, 86, 94
Perc-HM	percentile and histogram matching. 68, 71, 81, 85, 86, 93–96
PGSE	pulsed gradient spin echo. 25
ReLU	Rectified linear unit. 39, 61
RF	radio-frequency. 10
RMSE	root-mean-square error. 66, 71, 85, 86
ROI	Region of interest. 44
SGD	Stochastic gradient descent. 34–36
SNR	signal-to-noise ratio. 18, 30, 56, 59, 79, 80, 92, 103
TE	echo time. 25, 78, 80
TME	total mesorectal excision. 49, 50, 52
TR	repetition time. 25, 78
TRG	tumor regression grade. 2, 3, 5, 50, 70, 71, 74
TSE	Turbo Spin Echo. 25, 56
Win	fixed window. 68, 71, 82, 85, 86, 95

List of Figures

List of Tables

Bibliography

Afshar, Parnian, Arash Mohammadi, Konstantinos N. Plataniotis, Anastasia Oikonomou, and Habib Benali (July 2019). “From Hand-Crafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities”. In: *IEEE Signal Processing Magazine* 36.4, pp. 132–160. ISSN: 1053-5888, 1558-0792. DOI: 10.1109/MSP.2019.2900993. arXiv: 1808.07954 (cit. on pp. 1, 43, 46).

Schurink, Niels W., Simon R. van Kranen, Sander Roberti, et al. (Mar. 2022). “Sources of Variation in Multicenter Rectal MRI Data and Their Effect on Radiomics Feature Reproducibility”. In: *European Radiology* 32.3, pp. 1506–1516. ISSN: 0938-7994, 1432-1084. DOI: 10.1007/s00330-021-08251-8 (cit. on pp. 1, 44, 100).

Michoux, Nicolas F., Jakub W. Ceranka, Jef Vandemeulebroucke, et al. (July 2021). “Repeatability and Reproducibility of ADC Measurements: A Prospective Multicenter Whole-Body-MRI Study”. In: *European Radiology* 31.7, pp. 4514–4527. ISSN: 0938-7994, 1432-1084. DOI: 10.1007/s00330-020-07522-0 (cit. on pp. 1, 92, 100).

Dreher, C., T.A. Kuder, F. König, et al. (Oct. 2020). “Radiomics in Diffusion Data: A TestRetest, Inter- and Intra-Reader DWI Phantom Study”. In: *Clinical Radiology* 75.10, 798.e13–798.e22. ISSN: 00099260. DOI: 10.1016/j.crad.2020.06.024 (cit. on pp. 1, 44, 100).

Avanzo, Michele, Lise Wei, Joseph Stancanello, et al. (2020). “Machine and Deep Learning Methods for Radiomics”. In: *Medical Physics* 47.5, e185–e202. ISSN: 2473-4209. DOI: 10.1002/mp.13678 (cit. on p. 1).

Varoquaux, Gaël and Veronika Cheplygina (Apr. 12, 2022). “Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future”. In: *npj Digital Medicine* 5.1, p. 48. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00592-y (cit. on pp. 2, 91, 92).

Clark, Kenneth, Bruce Vendt, Kirk Smith, et al. (Dec. 2013). “The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository”. In: *Journal of Digital Imaging* 26.6, pp. 1045–1057. ISSN: 0897-1889, 1618-727X. DOI: 10.1007/s10278-013-9622-7 (cit. on p. 2).

Deng, Jia, Wei Dong, Richard Socher, et al. (June 2009). “ImageNet: A Large-Scale Hierarchical Image Database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cit. on p. 2).

Fitzmaurice, Christina, Daniel Dicker, Amanda Pain, et al. (July 1, 2015). “The Global Burden of Cancer 2013”. In: *JAMA Oncology* 1.4, p. 505. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2015.0735 (cit. on p. 2).

Rödel, Claus, Ullrich Graeven, Rainer Fietkau, et al. (Aug. 2015). "Oxaliplatin Added to Fluorouracil-Based Preoperative Chemoradiotherapy and Postoperative Chemotherapy of Locally Advanced Rectal Cancer (the German CAO/ARO/AIO-04 Study): Final Results of the Multicentre, Open-Label, Randomised, Phase 3 Trial". In: *The Lancet Oncology* 16.8, pp. 979–989. ISSN: 14702045. DOI: 10.1016/S1470-2045(15)00159-X (cit. on p. 2).

Benson, Al B., Alan P. Venook, Tanios Bekaii-Saab, et al. (June 2015). "Rectal Cancer, Version 2.2015". In: *Journal of the National Comprehensive Cancer Network* 13.6, pp. 719–728. ISSN: 1540-1405, 1540-1413. DOI: 10.6004/jnccn.2015.0087 (cit. on pp. 2, 50, 99).

Smith, J. Joshua and Julio Garcia-Aguilar (June 1, 2015). "Advances and Challenges in Treatment of Locally Advanced Rectal Cancer". In: *Journal of Clinical Oncology* 33.16, pp. 1797–1808. ISSN: 0732-183X. DOI: 10.1200/JCO.2014.60.1054. pmid: 25918296 (cit. on p. 2).

Horvat, Natally, Camila Carlos Tavares Rocha, Brunna Clemente Oliveira, Iva Petkovska, and Marc J. Gollub (Mar. 2019). "MRI of Rectal Cancer: Tumor Staging, Imaging Techniques, and Management". In: *RadioGraphics* 39.2, pp. 367–387. ISSN: 0271-5333, 1527-1323. DOI: 10.1148/rg.2019180114 (cit. on pp. 3, 29, 48, 56).

Coppola, Francesca, Valentina Giannini, Michela Gabelloni, et al. (Apr. 23, 2021). "Radiomics and Magnetic Resonance Imaging of Rectal Cancer: From Engineering to Clinical Practice". In: *Diagnostics* 11.5, p. 756. ISSN: 2075-4418. DOI: 10.3390/diagnostics11050756 (cit. on p. 3).

Horvat, Natally, Harini Veeraraghavan, Monika Khan, et al. (June 2018). "MR Imaging of Rectal Cancer: Radiomics Analysis to Assess Treatment Response after Neoadjuvant Therapy". In: *Radiology* 287.3, pp. 833–843. ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.2018172300 (cit. on pp. 3, 45).

Ryan, J. E., S. K. Warrier, A. C. Lynch, and A. G. Heriot (2015). "Assessing Pathological Complete Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer: A Systematic Review". In: *Colorectal Disease* 17.10, pp. 849–861. ISSN: 1463-1318. DOI: 10.1111/codi.13081 (cit. on p. 3).

Liu, Zhenyu, Xiao-Yan Zhang, Yan-Jie Shi, et al. (Dec. 1, 2017). "Radiomics Analysis for Evaluation of Pathological Complete Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer". In: *Clinical Cancer Research* 23.23, pp. 7253–7262. ISSN: 1078-0432, 1557-3265. DOI: 10.1158/1078-0432.CCR-17-1038 (cit. on pp. 3, 45, 99).

Jang, Jong Keon, Sang Hyun Choi, Seong Ho Park, et al. (Apr. 2020). "MR Tumor Regression Grade for Pathological Complete Response in Rectal Cancer Post Neoadjuvant Chemoradiotherapy: A Systematic Review and Meta-Analysis for Accuracy". In: *European Radiology* 30.4, pp. 2312–2323. ISSN: 0938-7994, 1432-1084. DOI: 10.1007/s00330-019-06565-2 (cit. on pp. 3, 99, 103).

Jin, Cheng, Heng Yu, Jia Ke, et al. (Dec. 2021). "Predicting Treatment Response from Longitudinal Images Using Multi-Task Deep Learning". In: *Nature Communications* 12.1, p. 1851. ISSN: 2041-1723. DOI: 10.1038/s41467-021-22188-y (cit. on pp. 3, 6, 46, 72, 74, 97, 103).

Wichtmann, Barbara D., Steffen Albert, Wenzhao Zhao, et al. (June 30, 2022). "Are We There Yet? The Value of Deep Learning in a Multicenter Setting for Response Prediction of Locally Advanced Rectal Cancer to Neoadjuvant Chemoradiotherapy". In: *Diagnostics* 12.7, p. 1601. ISSN: 2075-4418. DOI: 10.3390/diagnostics12071601 (cit. on pp. 3, 51, 66, 77).

Guan, Hao and Mingxia Liu (Mar. 2022). "Domain Adaptation for Medical Image Analysis: A Survey". In: *IEEE Transactions on Biomedical Engineering* 69.3, pp. 1173–1185. ISSN: 1558-2531. DOI: 10.1109/TBME.2021.3117407 (cit. on pp. 4, 5).

Mårtensson, Gustav, Daniel Ferreira, Tobias Granberg, et al. (Dec. 2020). "The Reliability of a Deep Learning Model in Clinical Out-of-Distribution MRI Data: A Multicohort Study". In: *Medical Image Analysis* 66, p. 101714. ISSN: 13618415. DOI: 10.1016/j.media.2020.101714 (cit. on pp. 4, 92).

Reinhold, Jacob C., Blake E. Dewey, Aaron Carass, and Jerry L. Prince (Mar. 15, 2019). "Evaluating the Impact of Intensity Normalization on MR Image Synthesis". In: *Medical Imaging 2019: Image Processing*. Image Processing. Ed. by Elsa D. Angelini and Bennett A. Landman. San Diego, United States: SPIE, p. 126. ISBN: 978-1-5106-2545-7. DOI: 10.1117/12.2513089 (cit. on p. 4).

Shah, Mohak, Yiming Xiao, Nagesh Subbanna, et al. (Apr. 2011). "Evaluating Intensity Normalization on MRIs of Human Brain with Multiple Sclerosis". In: *Medical Image Analysis* 15.2, pp. 267–282. ISSN: 13618415. DOI: 10.1016/j.media.2010.12.003 (cit. on pp. 4, 96).

Carré, Alexandre, Guillaume Klausner, Myriam Edjlali, et al. (Dec. 2020). "Standardization of Brain MR Images across Machines and Protocols: Bridging the Gap for MRI-based Radiomics". In: *Scientific Reports* 10.1, p. 12340. ISSN: 2045-2322. DOI: 10.1038/s41598-020-69298-z (cit. on pp. 4, 96).

Shinohara, Russell T., Elizabeth M. Sweeney, Jeff Goldsmith, et al. (Jan. 1, 2014). "Statistical Normalization Techniques for Magnetic Resonance Imaging". In: *NeuroImage: Clinical* 6, pp. 9–19. ISSN: 2213-1582. DOI: 10.1016/j.nicl.2014.08.008 (cit. on p. 4).

Nyul, L. G., J. K. Udupa, and Xuan Zhang (Feb. 2000). "New Variants of a Method of MRI Scale Standardization". In: *IEEE Transactions on Medical Imaging* 19.2, pp. 143–150. ISSN: 1558-254X. DOI: 10.1109/42.836373 (cit. on pp. 4, 67).

Johnson, W. Evan, Cheng Li, and Ariel Rabinovic (Jan. 1, 2007). "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods". In: *Biostatistics* 8.1, pp. 118–127. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj037 (cit. on p. 4).

Eshaghzadeh Torbati, Mahbaneh, Davneet S. Minhas, Ghasan Ahmad, et al. (Dec. 2021). "A Multi-Scanner Neuroimaging Data Harmonization Using RAVEL and ComBat". In: *NeuroImage* 245, p. 118703. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2021.118703 (cit. on p. 4).

Fortin, Jean-Philippe, Drew Parker, Birkan Tunç, et al. (Nov. 2017). "Harmonization of Multi-Site Diffusion Tensor Imaging Data". In: *NeuroImage* 161, pp. 149–170. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2017.08.047 (cit. on p. 4).

Mali, Shruti Atul, Abdalla Ibrahim, Henry C. Woodruff, et al. (Aug. 27, 2021). “Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods”. In: *Journal of Personalized Medicine* 11.9, p. 842. ISSN: 2075-4426. DOI: 10.3390/jpm11090842 (cit. on p. 4).

Dewey, Blake E., Can Zhao, Jacob C. Reinhold, et al. (Dec. 2019). “DeepHarmony: A Deep Learning Approach to Contrast Harmonization across Scanner Changes”. In: *Magnetic Resonance Imaging* 64, pp. 160–170. ISSN: 0730725X. DOI: 10.1016/j.mri.2019.05.041 (cit. on p. 4).

Modanwal, Gourav, Adithya Vellal, and Maciej A. Mazurowski (2021). “Normalization of Breast MRIs Using Cycle-Consistent Generative Adversarial Networks”. In: *Computer Methods and Programs in Biomedicine* 208, p. 106225. ISSN: 01692607. DOI: 10.1016/j.cmpb.2021.106225 (cit. on p. 5).

Li, Yajun, Guoqiang Han, Xiaomei Wu, et al. (Mar. 25, 2020). “Normalization of Multicenter CT Radiomics by a Generative Adversarial Network Method”. In: *Physics in Medicine & Biology*. ISSN: 0031-9155, 1361-6560. DOI: 10.1088/1361-6560/ab8319 (cit. on p. 5).

Bashyam, Vishnu M., Jimit Doshi, Guray Erus, et al. (Sept. 25, 2021). “Deep Generative Medical Image Harmonization for Improving CrossSite Generalization in Deep Learning Predictors”. In: *Journal of Magnetic Resonance Imaging*, jmri.27908. ISSN: 1053-1807, 1522-2586. DOI: 10.1002/jmri.27908 (cit. on p. 5).

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, et al. (2016). “Domain-Adversarial Training of Neural Networks”. In: *The journal of machine learning research* 17.59. Ed. by Gabriela Csurka, pp. 2096–2030. DOI: 10.1007/978-3-319-58347-1_10 (cit. on p. 5).

Cackowski, Stenzel, Emmanuel L. Barbier, Michel Dojat, and Thomas Christen (Mar. 2023). “ImUnity: A Generalizable VAE-GAN Solution for Multicenter MR Image Harmonization”. In: *Medical Image Analysis*, p. 102799. ISSN: 13618415. DOI: 10.1016/j.media.2023.102799 (cit. on p. 5).

Joshi, Niranjan, Sarah Bond, and Michael Brady (Aug. 2010). “The Segmentation of Colorectal MRI Images”. In: *Medical Image Analysis* 14.4, pp. 494–509. ISSN: 13618415. DOI: 10.1016/j.media.2010.03.002 (cit. on p. 5).

Van Heeswijk, Miriam M., Doenja M.J. Lambregts, Joost J.M. van Griethuysen, et al. (Mar. 2016). “Automated and Semiautomated Segmentation of Rectal Tumor Volumes on Diffusion-Weighted MRI: Can It Replace Manual Volumetry?” In: *International Journal of Radiation Oncology*Biology*Physics* 94.4, pp. 824–831. ISSN: 03603016. DOI: 10.1016/j.ijrobp.2015.12.017 (cit. on pp. 5, 25, 44, 97).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Vol. 9351. Cham: Springer International Publishing, pp. 234–241. ISBN: 978-3-319-24573-7. DOI: 10.1007/978-3-319-24574-4_28 (cit. on pp. 5, 44, 60).

Isensee, Fabian, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein (Feb. 2021). “nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation”. In: *Nature Methods* 18.2, pp. 203–211. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-020-01008-z (cit. on pp. 6, 44, 61).

Wang, Jiazhou, Jiayu Lu, Gan Qin, et al. (2018). “Technical Note: A Deep Learning-Based Autosegmentation of Rectal Tumors in MR Images”. In: *Medical Physics* 45.6, pp. 2560–2564. ISSN: 2473-4209. DOI: 10.1002/mp.12918 (cit. on pp. 6, 94, 95, 97).

Huang, Yi-Jie, Qi Dou, Zi-Xian Wang, et al. (Feb. 15, 2019). “3D RoI-aware U-Net for Accurate and Efficient Colorectal Tumor Segmentation”. DOI: 10.1109/TCYB.2020.2980145. arXiv: 1806.10342 [cs] (cit. on pp. 6, 44, 94).

Lee, Joohyung, Ji Eun Oh, Min Ju Kim, Bo Yun Hur, and Dae Kyung Sohn (2019). “Reducing the Model Variance of a Rectal Cancer Segmentation Network”. In: *IEEE Access* 7, pp. 182725–182733. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2960371 (cit. on pp. 6, 94).

Soomro, Mumtaz Hussain, Matteo Coppotelli, Silvia Conforto, et al. (Jan. 31, 2019). “Automated Segmentation of Colorectal Tumor in 3D MRI Using 3D Multiscale Densely Connected Convolutional Neural Network”. In: *Journal of Healthcare Engineering* 2019, pp. 1–11. ISSN: 2040-2295, 2040-2309. DOI: 10.1155/2019/1075434 (cit. on pp. 6, 44, 94).

Gerlach, Walther and Otto Stern (Dec. 1922). “Der experimentelle Nachweis der Richtungsquantelung im Magnetfeld”. In: *Zeitschrift für Physik* 9.1, pp. 349–352. ISSN: 1434-6001, 1434-601X. DOI: 10.1007/BF01326983 (cit. on p. 7).

Rabi, I. I., J. R. Zacharias, S. Millman, and P. Kusch (Feb. 15, 1938). “A New Method of Measuring Nuclear Magnetic Moment”. In: *Physical Review* 53.4, pp. 318–318. ISSN: 0031-899X. DOI: 10.1103/PhysRev.53.318 (cit. on p. 7).

Bloch, F. (Oct. 1, 1946). “Nuclear Induction”. In: *Physical Review* 70.7-8, pp. 460–474. ISSN: 0031-899X. DOI: 10.1103/PhysRev.70.460 (cit. on pp. 7, 14).

Purcell, E. M., H. C. Torrey, and R. V. Pound (Jan. 1, 1946). “Resonance Absorption by Nuclear Magnetic Moments in a Solid”. In: *Physical Review* 69.1-2, pp. 37–38. ISSN: 0031-899X. DOI: 10.1103/PhysRev.69.37 (cit. on p. 7).

Brown, Robert W., Yu-Chung N. Cheng, E. Mark Haacke, Michael R. Thompson, and Ramesh Venkatesan, eds. (Apr. 22, 2014). *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. Chichester, UK: John Wiley & Sons Ltd. ISBN: 978-1-118-63395-3. DOI: 10.1002/9781118633953 (cit. on pp. 7, 8, 15, 18, 24).

Pauli, W (1927). “Zur Quantenmechanik Des Magnetischen Elektrons”. In: *Zeitschrift für Physik* 43, p. 601 (cit. on p. 9).

Bojorquez, Jorge Zavala, Stéphanie Bricq, Clement Acquitter, et al. (Jan. 2017). “What Are Normal Relaxation Times of Tissues at 3 T?” In: *Magnetic Resonance Imaging* 35, pp. 69–80. ISSN: 0730725X. DOI: 10.1016/j.mri.2016.08.021 (cit. on p. 15).

Hayes, Cecil E. and Leon Axel (Sept. 1985). “Noise Performance of Surface Coils for Magnetic Resonance Imaging at 1.5 T: Surface-coil Noise for MRI at 1.5 T”. In: *Medical Physics* 12.5, pp. 604–607. ISSN: 00942405. DOI: 10.1118/1.595682 (cit. on p. 18).

Lauterbur, P. C. (Mar. 1973). “Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance”. In: *Nature* 242.5394, pp. 190–191. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/242190a0 (cit. on p. 19).

– (Jan. 1, 1974). “Magnetic Resonance Zeugmatography”. In: *Pure and Applied Chemistry* 40.1-2, pp. 149–157. ISSN: 1365-3075, 0033-4545. DOI: 10.1351/pac197440010149 (cit. on p. 19).

Mansfield, P. and P. K. Grannell (Nov. 1, 1975). “"Diffraction" and Microscopy in Solids and Liquids by NMR”. In: *Physical Review B* 12.9, pp. 3618–3634. ISSN: 0556-2805. DOI: 10.1103/PhysRevB.12.3618 (cit. on p. 19).

Hennig, J., A. Nauerth, and H. Friedburg (Dec. 1986). “RARE Imaging: A Fast Imaging Method for Clinical MR”. In: *Magnetic Resonance in Medicine* 3.6, pp. 823–833. ISSN: 07403194. DOI: 10.1002/mrm.1910030602 (cit. on p. 25).

Stejskal, E. O. and J. E. Tanner (Jan. 1965). “Spin Diffusion Measurements: Spin Echoes in the Presence of a TimeDependent Field Gradient”. In: *The Journal of Chemical Physics* 42.1, pp. 288–292. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.1695690 (cit. on pp. 25, 28).

Einstein, A. (1905). “Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen”. In: *Annalen der Physik* 322.8, pp. 549–560. ISSN: 00033804, 15213889. DOI: 10.1002/andp.19053220806 (cit. on p. 27).

Han, Chengkun, Long Zhao, Shan Zhong, et al. (Oct. 2015). “A Comparison of High *b* -Value *vs* Standard *b* -Value Diffusion-Weighted Magnetic Resonance Imaging at 3.0 T for Medulloblastomas”. In: *The British Journal of Radiology* 88.1054, p. 20150220. ISSN: 0007-1285, 1748-880X. DOI: 10.1259/bjr.20150220 (cit. on p. 28).

Basser, P.J., J. Mattiello, and D. Lebihan (Mar. 1994). “Estimation of the Effective Self-Diffusion Tensor from the NMR Spin Echo”. In: *Journal of Magnetic Resonance, Series B* 103.3, pp. 247–254. ISSN: 10641866. DOI: 10.1006/jmrb.1994.1037 (cit. on p. 28).

Skudlarski, Pawel, Kanchana Jagannathan, Vince D. Calhoun, et al. (Nov. 2008). “Measuring Brain Connectivity: Diffusion Tensor Imaging Validates Resting State Temporal Correlations”. In: *NeuroImage* 43.3, pp. 554–561. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2008.07.063 (cit. on p. 28).

Minati, Ludovico and Wadysaw P. Wglarz (Sept. 2007). “Physical Foundations, Models, and Methods of Diffusion Magnetic Resonance Imaging of the Brain: A Review”. In: *Concepts in Magnetic Resonance Part A* 30A.5, pp. 278–307. ISSN: 15466086, 15525023. DOI: 10.1002/cmr.a.20094 (cit. on p. 30).

Iima, Mami and Denis Le Bihan (Jan. 2016). “Clinical Intravoxel Incoherent Motion and Diffusion MR Imaging: Past, Present, and Future”. In: *Radiology* 278.1, pp. 13–32. ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.2015150244 (cit. on p. 30).

Jensen, Jens H. and Joseph A. Helpern (May 19, 2010). "MRI Quantification of Non-Gaussian Water Diffusion by Kurtosis Analysis". In: *NMR in Biomedicine* 23.7, pp. 698–710. ISSN: 09523480. DOI: 10.1002/nbm.1518 (cit. on p. 30).

Le Bihan, Denis, Cyril Poupon, Alexis Amadon, and Franck Lethimonnier (Sept. 2006). "Artifacts and Pitfalls in Diffusion MRI". In: *Journal of Magnetic Resonance Imaging* 24.3, pp. 478–488. ISSN: 1053-1807, 1522-2586. DOI: 10.1002/jmri.20683 (cit. on pp. 30, 31).

Rohde, G.K., A.S. Barnett, P.J. Basser, S. Marenco, and C. Pierpaoli (Jan. 2004). "Comprehensive Approach for Correction of Motion and Distortion in Diffusion-Weighted MRI". In: *Magnetic Resonance in Medicine* 51.1, pp. 103–114. ISSN: 0740-3194, 1522-2594. DOI: 10.1002/mrm.10677 (cit. on p. 31).

López-Muñoz, Francisco, Jesús Boya, and Cecilio Alamo (Oct. 2006). "Neuron Theory, the Cornerstone of Neuroscience, on the Centenary of the Nobel Prize Award to Santiago Ramón y Cajal". In: *Brain Research Bulletin* 70.4-6, pp. 391–405. ISSN: 03619230. DOI: 10.1016/j.brainresbull.2006.07.010 (cit. on p. 31).

Hodgkin, A. L. and A. F. Huxley (Aug. 28, 1952). "A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve". In: *The Journal of Physiology* 117.4, pp. 500–544. ISSN: 0022-3751, 1469-7793. DOI: 10.1113/jphysiol.1952.sp004764 (cit. on p. 31).

McCulloch, Warren S. and Walter Pitts (Dec. 1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *The Bulletin of Mathematical Biophysics* 5.4, pp. 115–133. ISSN: 0007-4985, 1522-9602. DOI: 10.1007/BF02478259 (cit. on p. 31).

Rosenblatt, F. (1958). "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." In: *Psychological Review* 65.6, pp. 386–408. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/h0042519 (cit. on p. 32).

Fukushima, Kunihiko (Apr. 1980). "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position". In: *Biological Cybernetics* 36.4, pp. 193–202. ISSN: 0340-1200, 1432-0770. DOI: 10.1007/BF00344251 (cit. on pp. 32, 39).

LeCun, Y., B. Boser, J. S. Denker, et al. (Dec. 1989). "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4, pp. 541–551. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1989.1.4.541 (cit. on pp. 32, 39).

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (Oct. 1986). "Learning Representations by Back-Propagating Errors". In: *Nature* 323.6088, pp. 533–536. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/323533a0 (cit. on pp. 32, 35).

Oh, Kyoung-Su and Keechul Jung (June 2004). "GPU Implementation of Neural Networks". In: *Pattern Recognition* 37.6, pp. 1311–1314. ISSN: 00313203. DOI: 10.1016/j.patcog.2004.01.013 (cit. on p. 32).

Russakovsky, Olga, Jia Deng, Hao Su, et al. (Dec. 2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3, pp. 211–252. ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-015-0816-y (cit. on pp. 32, 102).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (May 24, 2017). “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Communications of the ACM* 60.6, pp. 84–90. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3065386 (cit. on p. 32).

Ciregan, Dan, Ueli Meier, and Jürgen Schmidhuber (June 2012). “Multi-Column Deep Neural Networks for Image Classification”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3642–3649. DOI: 10.1109/CVPR.2012.6248110 (cit. on p. 32).

Deng, Li, Geoffrey Hinton, and Brian Kingsbury (May 2013). “New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. ICASSP 2013 - 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, BC, Canada: IEEE, pp. 8599–8603. ISBN: 978-1-4799-0356-6. DOI: 10.1109/ICASSP.2013.6639344 (cit. on p. 32).

Dahl, George E., Navdeep Jaitly, and Ruslan Salakhutdinov (June 4, 2014). *Multi-Task Neural Networks for QSAR Predictions*. arXiv: 1406.1231 [cs, stat]. URL: http://arxiv.org/abs/1406.1231 (visited on Feb. 14, 2023). preprint (cit. on p. 32).

Jumper, John, Richard Evans, Alexander Pritzel, et al. (Aug. 26, 2021). “Highly Accurate Protein Structure Prediction with AlphaFold”. In: *Nature* 596.7873, pp. 583–589. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-03819-2 (cit. on p. 32).

Zhao, Wayne Xin, Kun Zhou, Junyi Li, et al. (Sept. 11, 2023). *A Survey of Large Language Models*. arXiv: 2303.18223 [cs]. URL: http://arxiv.org/abs/2303.18223 (visited on Sept. 27, 2023). preprint (cit. on p. 32).

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press (cit. on pp. 32, 36, 39).

Robbins, Herbert and Sutton Monro (Sept. 1951). “A Stochastic Approximation Method”. In: *The Annals of Mathematical Statistics* 22.3, pp. 400–407. ISSN: 0003-4851. DOI: 10.1214/aoms/1177729586 (cit. on p. 34).

Cauchy, Augustin (1847). “Méthode Générale Pour La Résolution Des Systemes déquations Simultanées”. In: *Comp. Rend. Sci. Paris* 25.1847, pp. 536–538 (cit. on p. 34).

Curry, Haskell B. (1944). “The Method of Steepest Descent for Non-Linear Minimization Problems”. In: *Quarterly of Applied Mathematics* 2.3, pp. 258–261. ISSN: 0033-569X, 1552-4485. DOI: 10.1090/qam/10667 (cit. on p. 34).

Nesterov, Yurii Evgen’evich (1983). “A Method of Solving a Convex Programming Problem with Convergence Rate O(1/K$\hat{2}$)”. In: *Doklady Akademii Nauk* 269.3, pp. 543–547. ISSN: 0869-5652 (cit. on p. 35).

Kingma, Diederik P. and Jimmy Ba (Jan. 29, 2014). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs]. URL: http://arxiv.org/abs/1412.6980 (visited on Feb. 15, 2023). preprint (cit. on pp. 35, 36).

Schmidt, Robin M., Frank Schneider, and Philipp Hennig (July 7, 2020). “Descending through a Crowded Valley - Benchmarking Deep Learning Optimizers”. arXiv: 2007.01547 [cs, stat] (cit. on p. 35).

Ioffe, Sergey and Christian Szegedy (2015). “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. Version 3. In: DOI: 10.48550/ARXIV.1502.03167 (cit. on p. 36).

Yeung, Michael, Evis Sala, Carola-Bibiane Schönlieb, and Leonardo Rundo (Jan. 2022). “Unified Focal Loss: Generalising Dice and Cross Entropy-Based Losses to Handle Class Imbalanced Medical Image Segmentation”. In: *Computerized Medical Imaging and Graphics* 95, p. 102026. ISSN: 08956111. DOI: 10.1016/j.compmedimag.2021.102026 (cit. on p. 37).

Gilbert, Charles D. and Wu Li (May 2013). “Top-down Influences on Visual Processing”. In: *Nature Reviews Neuroscience* 14.5, pp. 350–363. ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn3476 (cit. on p. 38).

Shelhamer, Evan, Jonathan Long, and Trevor Darrell (May 20, 2016). “Fully Convolutional Networks for Semantic Segmentation”. arXiv: 1605.06211 [cs] (cit. on p. 40).

Guo, Yunhui, Yandong Li, Rogerio Feris, Liqiang Wang, and Tajana Rosing (Feb. 19, 2019). *Depthwise Convolution Is All You Need for Learning Multiple Visual Domains*. arXiv: 1902.00927 [cs]. URL: http://arxiv.org/abs/1902.00927 (visited on Feb. 17, 2023). preprint (cit. on p. 41).

Yu, Fisher and Vladlen Koltun (Apr. 30, 2016). *Multi-Scale Context Aggregation by Dilated Convolutions*. arXiv: 1511.07122 [cs]. URL: http://arxiv.org/abs/1511.07122 (visited on Feb. 17, 2023). preprint (cit. on p. 41).

Dumoulin, Vincent and Francesco Visin (Jan. 11, 2018). *A Guide to Convolution Arithmetic for Deep Learning*. DOI: 10.48550/arXiv.1603.07285. arXiv: 1603.07285 [cs, stat]. URL: http://arxiv.org/abs/1603.07285 (visited on Feb. 17, 2023). preprint (cit. on p. 41).

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *The journal of machine learning research* 15.1, pp. 1929–1958. ISSN: 1532-4435 (cit. on p. 43).

Mayerhoefer, Marius E., Andrzej Materka, Georg Langs, et al. (Apr. 2020). “Introduction to Radiomics”. In: *Journal of Nuclear Medicine* 61.4, pp. 488–495. ISSN: 0161-5505, 2159-662X. DOI: 10.2967/jnumed.118.222893 (cit. on p. 43).

Scalco, Elisa, Antonella Belfatto, Alfonso Mastropietro, et al. (Apr. 2020). “T2wMRI Signal Normalization Affects Radiomics Features Reproducibility”. In: *Medical Physics* 47.4, pp. 1680–1691. ISSN: 0094-2405, 2473-4209. DOI: 10.1002/mp.14038 (cit. on pp. 44, 46, 96).

Um, Hyemin, Florent Tixier, Dalton Bermudez, et al. (Aug. 21, 2019). “Impact of Image Preprocessing on the Scanner Dependence of Multi-Parametric MRI Radiomic Features and Covariate Shift in Multi-Institutional Glioblastoma Datasets”. In: *Physics in Medicine & Biology* 64.16, p. 165011. ISSN: 1361-6560. DOI: 10.1088/1361-6560/ab2f44 (cit. on pp. 44, 96).

Shafiq-ul-Hassan, Muhammad, Geoffrey G. Zhang, Kujtim Latifi, et al. (Mar. 2017). "Intrinsic Dependencies of CT Radiomic Features on Voxel Size and Number of Gray Levels". In: *Medical Physics* 44.3, pp. 1050–1062. ISSN: 00942405. DOI: 10.1002/mp.12123 (cit. on p. 44).

Zhao, Xingyu, Peiyi Xie, Mengmeng Wang, et al. (June 2020). "Deep LearningBased Fully Automated Detection and Segmentation of Lymph Nodes on Multiparametric-Mri for Rectal Cancer: A Multicentre Study". In: *EBioMedicine* 56, p. 102780. ISSN: 23523964. DOI: 10.1016/j.ebiom.2020.102780 (cit. on p. 44).

Zwanenburg, Alex, Stefan Leger, Martin Vallières, and Steffen Löck (Mar. 10, 2020). "Image Biomarker Standardisation Initiative". In: *Radiology*, p. 191145. ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.2020191145. arXiv: 1612.07003 (cit. on p. 44).

Li, Zhihui, Xiaolu Ma, Fu Shen, et al. (Dec. 2021). "Evaluating Treatment Response to Neoadjuvant Chemoradiotherapy in Rectal Cancer Using Various MRI-based Radiomics Models". In: *BMC Medical Imaging* 21.1, p. 30. ISSN: 1471-2342. DOI: 10.1186/s12880-021-00560-0 (cit. on p. 45).

Schelb, Patrick, Simon Kohl, Jan Philipp Radtke, et al. (Dec. 2019). "Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment". In: *Radiology* 293.3, pp. 607–617. ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.2019190938 (cit. on p. 46).

Van Timmeren, Janita E., Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler (Dec. 2020). "Radiomics in Medical ImagingHow-to Guide and Critical Reflection". In: *Insights into Imaging* 11.1, p. 91. ISSN: 1869-4101. DOI: 10.1186/s13244-020-00887-2 (cit. on p. 46).

Isaksson, Lars J., Sara Raimondi, Francesca Botta, et al. (Mar. 2020). "Effects of MRI Image Normalization Techniques in Prostate Cancer Radiomics". In: *Physica Medica* 71, pp. 7–13. ISSN: 11201797. DOI: 10.1016/j.ejmp.2020.02.007 (cit. on pp. 46, 96).

Saint Martin, Marie-Judith, Fanny Orlhac, Pia Akl, et al. (Nov. 12, 2020). "A Radiomics Pipeline Dedicated to Breast MRI: Validation on a Multi-Scanner Phantom Study". In: *Magnetic Resonance Materials in Physics, Biology and Medicine*. ISSN: 0968-5243, 1352-8661. DOI: 10.1007/s10334-020-00892-y (cit. on p. 46).

Fornacon-Wood, Isabella, Hitesh Mistry, Christoph J. Ackermann, et al. (Nov. 2020). "Reliability and Prognostic Value of Radiomic Features Are Highly Dependent on Choice of Feature Extraction Platform". In: *European Radiology* 30.11, pp. 6241–6250. ISSN: 0938-7994, 1432-1084. DOI: 10.1007/s00330-020-06957-9 (cit. on p. 46).

Lambin, Philippe, Ralph T.H. Leijenaar, Timo M. Deist, et al. (Dec. 2017). "Radiomics: The Bridge between Medical Imaging and Personalized Medicine". In: *Nature Reviews Clinical Oncology* 14.12, pp. 749–762. ISSN: 1759-4774, 1759-4782. DOI: 10.1038/nrclinonc.2017.141 (cit. on p. 46).

Cancer Research UK (July 30, 2014). *Diagram Showing T Stages of Bowel Cancer.* URL: https://commons.wikimedia.org/wiki/File:Diagram_showing_T_stages_of_bowel_cancer_CRUK_276.svg (cit. on p. 47).

Attenberger, U. I., J. Winter, F. N. Harder, et al. (Jan. 28, 2020). “Height of Rectal Cancer: A Comparison between Rectoscopic and Different MRI Measurements”. In: *Gastroenterology Research and Practice* 2020, pp. 1–7. ISSN: 1687-6121, 1687-630X. DOI: 10.1155/2020/2130705 (cit. on p. 49).

Attenberger, U.I. and B. Wichtmann (Feb. 2015). “MR-Bildgebung des Rektumkarzinoms: Bedeutung der MRT für Staging, Therapie-Monitoring und potenzielle Therapiestratifizierung”. In: *Der Onkologe* 21.2, pp. 129–135. ISSN: 0947-8965, 1433-0415. DOI: 10.1007/s00761-014-2763-6 (cit. on pp. 49, 99, 103).

Schmoll, H. J., E. Van Cutsem, A. Stein, et al. (Oct. 1, 2012). “ESMO Consensus Guidelines for Management of Patients with Colon and Rectal Cancer. A Personalized Approach to Clinical Decision Making”. In: *Annals of Oncology* 23.10, pp. 2479–2516. ISSN: 0923-7534. DOI: 10.1093/annonc/mds236 (cit. on p. 49).

Hofheinz, Ralf-Dieter, Dirk Arnold, Markus Borner, et al. (2023). *Rektumkarzinom*. DGHO Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie (cit. on p. 49).

Taylor, F G M, P Quirke, R J Heald, et al. (Apr. 26, 2011). “One Millimetre Is the Safe Cut-off for Magnetic Resonance Imaging Prediction of Surgical Margin Status in Rectal Cancer”. In: *British Journal of Surgery* 98.6, pp. 872–879. ISSN: 0007-1323, 1365-2168. DOI: 10.1002/bjs.7458 (cit. on p. 50).

Oronsky, Bryan, Tony Reid, Chris Larson, and Susan J. Knox (Feb. 2020). “Locally Advanced Rectal Cancer: The Past, Present, and Future”. In: *Seminars in Oncology* 47.1, pp. 85–92. ISSN: 00937754. DOI: 10.1053/j.seminoncol.2020.02.001 (cit. on pp. 50, 72).

Attenberger, Ulrike I., Ralf D. Hofheinz, and Barbara D. Wichtmann (2020). “Klinischer Stellenwert der Bildgebung nach neoadjuvanter Therapie”. In: *MRT-basierte Chirurgie des Rektumkarzinoms*. Ed. by Martin E. Kreis and Patrick Asbach. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 95–105. ISBN: 978-3-662-58158-2. DOI: 10.1007/978-3-662-58159-9_9 (cit. on p. 50).

Dworak, O., L. Keilholz, and A. Hoffmann (Mar. 24, 1997). “Pathological Features of Rectal Cancer after Preoperative Radiochemotherapy”. In: *International Journal of Colorectal Disease* 12.1, pp. 19–23. ISSN: 0179-1958, 1432-1262. DOI: 10.1007/s003840050072 (cit. on p. 50).

Albert, Steffen, Barbara D. Wichtmann, Wenzhao Zhao, et al. (Aug. 3, 2023). “Comparison of Image Normalization Methods for Multi-Site Deep Learning”. In: *Applied Sciences* 13.15, p. 8923. ISSN: 2076-3417. DOI: 10.3390/app13158923 (cit. on pp. 51, 77, 91).

Fokas, Emmanouil, Michael Allgäuer, Bülent Polat, et al. (Dec. 1, 2019). “Randomized Phase II Trial of Chemoradiotherapy Plus Induction or Consolidation Chemotherapy as Total Neoadjuvant Therapy for Locally Advanced Rectal Cancer: CAO/ARO/AIO-12”. In: *Journal of Clinical Oncology* 37.34, p. 14. ISSN: 1527-7755. DOI: 10.1200/JCO.19.00308 (cit. on p. 51).

Fokas, Emmanouil, Anke Schlenska-Lange, Bülent Polat, et al. (Jan. 20, 2022). "Chemoradiotherapy Plus Induction or Consolidation Chemotherapy as Total Neoadjuvant Therapy for Patients With Locally Advanced Rectal Cancer: Long-term Results of the CAO/ARO/AIO-12 Randomized Clinical Trial". In: *JAMA Oncology* 8.1, e215445. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2021.5445 (cit. on pp. 51, 52).

Koh, Dow-Mu and David J. Collins (June 2007). "Diffusion-Weighted MRI in the Body: Applications and Challenges in Oncology". In: *American Journal of Roentgenology* 188.6, pp. 1622–1635. ISSN: 0361-803X, 1546-3141. DOI: 10.2214/AJR.06.1403 (cit. on p. 56).

Chen, Guangyong, Fengyuan Zhu, and Pheng Ann Heng (Dec. 2015). "An Efficient Statistical Method for Image Noise Level Estimation". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, pp. 477–485. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.62 (cit. on p. 56).

Yushkevich, Paul A., Joseph Piven, Heather Cody Hazlett, et al. (July 2006). "User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability". In: *NeuroImage* 31.3, pp. 1116–1128. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2006.01.015 (cit. on p. 58).

Tustison, N. J., B. B. Avants, P. A. Cook, et al. (June 2010). "N4ITK: Improved N3 Bias Correction". In: *IEEE Transactions on Medical Imaging* 29.6, pp. 1310–1320. ISSN: 1558-254X. DOI: 10.1109/TMI.2010.2046908 (cit. on p. 58).

Avants, Brian B., Nicholas J. Tustison, Gang Song, et al. (Feb. 2011). "A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration". In: *NeuroImage* 54.3, pp. 2033–2044. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2010.09.025 (cit. on p. 59).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (Dec. 10, 2015). "Deep Residual Learning for Image Recognition". arXiv: 1512.03385 [cs] (cit. on p. 60).

Wong, Chinting, Yu Fu, Mingyang Li, et al. (Jan. 2023). "MRIBased Artificial Intelligence in Rectal Cancer". In: *Journal of Magnetic Resonance Imaging* 57.1, pp. 45–56. ISSN: 1053-1807, 1522-2586. DOI: 10.1002/jmri.28381 (cit. on pp. 61, 66).

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, et al. (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc. (cit. on p. 63).

Sudre, Carole H., Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso (2017). "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations". In: vol. 10553, pp. 240–248. DOI: 10.1007/978-3-319-67558-9_28. arXiv: 1707.03237 [cs] (cit. on p. 64).

Dice, Lee R. (July 1945). "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3, pp. 297–302. ISSN: 00129658. DOI: 10.2307/1932409 (cit. on p. 64).

Trebeschi, Stefano, Joost J. M. van Griethuysen, Doenja M. J. Lambregts, et al. (Dec. 2017). "Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR". In: *Scientific Reports* 7.1, p. 5301. ISSN: 2045-2322. DOI: 10.1038/s41598-017-05728-9 (cit. on pp. 66, 94, 95, 97).

Mayerhoefer, Marius E., Pavol Szomolanyi, Daniel Jirak, Andrzej Materka, and Siegfried Trattnig (Mar. 18, 2009). "Effects of MRI Acquisition Parameter Variations and Protocol Heterogeneity on the Results of Texture Analysis and Pattern Discrimination: An Application-Oriented Study: Effects of MRI Acquisition Parameters on Texture Analysis". In: *Medical Physics* 36.4, pp. 1236–1243. ISSN: 00942405. DOI: 10.1118/1.3081408 (cit. on p. 66).

Veraart, Jelle, Dmitry S. Novikov, Daan Christiaens, et al. (Nov. 2016). "Denoising of Diffusion MRI Using Random Matrix Theory". In: *NeuroImage* 142, pp. 394–406. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2016.08.016 (cit. on p. 72).

Tournier, J-Donald, Robert Smith, David Raffelt, et al. (Nov. 2019). "MRtrix3: A Fast, Flexible and Open Software Framework for Medical Image Processing and Visualisation". In: *NeuroImage* 202, p. 116137. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2019.116137 (cit. on p. 72).

Sun, Xu and Weichao Xu (Nov. 2014). "Fast Implementation of DeLongs Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves". In: *IEEE Signal Processing Letters* 21.11, pp. 1389–1393. ISSN: 1558-2361. DOI: 10.1109/LSP.2014.2337313 (cit. on p. 75).

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, et al. (2011). "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830 (cit. on p. 75).

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324 (cit. on p. 75).

Lundberg, Scott M., Gabriel Erion, Hugh Chen, et al. (Jan. 2020). "From Local Explanations to Global Understanding with Explainable AI for Trees". In: *Nature Machine Intelligence* 2.1 (1), pp. 56–67. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0138-9 (cit. on p. 75).

George, Allan, Ruben Kuzniecky, Henry Rusinek, Heath R. Pardoe, and for the Human Epilepsy Project Investigators (2020). "Standardized Brain MRI Acquisition Protocols Improve Statistical Power in Multicenter Quantitative Morphometry Studies". In: *Journal of Neuroimaging* 30.1, pp. 126–133. ISSN: 1552-6569. DOI: 10.1111/jon.12673 (cit. on p. 92).

West, M.A., B.D. Dimitrov, H.E. Moyses, et al. (Sept. 2016). "Timing of Surgery Following Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer A Comparison of Magnetic Resonance Imaging at Two Time Points and Histopathological Responses". In: *European Journal of Surgical Oncology (EJSO)* 42.9, pp. 1350–1358. ISSN: 07487983. DOI: 10.1016/j.ejso.2016.04.003 (cit. on p. 93).

Shin, Jaeseung, Nieun Seo, Song-Ee Baek, et al. (Feb. 8, 2022). “MRI Radiomics Model Predicts Pathologic Complete Response of Rectal Cancer Following Chemoradiotherapy”. In: *Radiology*, p. 211986. ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.211986 (cit. on pp. 94, 95, 99).

Le Goallec, Alan, Samuel Diai, Sasha Collin, et al. (Apr. 13, 2022). “Using Deep Learning to Predict Abdominal Age from Liver and Pancreas Magnetic Resonance Images”. In: *Nature Communications* 13.1, p. 1979. ISSN: 2041-1723. DOI: 10.1038/s41467-022-29525-9 (cit. on p. 94).

Yoshimura, Sho, Hisashi Tanaka, Shuichi Kawabata, et al. (Sept. 2022). “Effect of Urinary Glucose Concentration and pH on Signal Intensity in Magnetic Resonance Images”. In: *Japanese Journal of Radiology* 40.9, pp. 930–938. ISSN: 1867-1071, 1867-108X. DOI: 10.1007/s11604-022-01273-2 (cit. on p. 96).

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, et al. (June 3, 2021). *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv: 2010.11929 [cs]. URL: http://arxiv.org/abs/2010.11929 (visited on Sept. 7, 2023). preprint (cit. on p. 96).

Lee, Suk-Hawn, Enrique Hernandez De Anda, Charles O. Finne, Robert D. Madoff, and Julio Garcia-Aguilar (Dec. 2005). “The Effect of Circumferential Tumor Location in Clinical Outcomes of Rectal Cancer Patients Treated With Total Mesorectal Excision”. In: *Diseases of the Colon & Rectum* 48.12, pp. 2249–2257. ISSN: 0012-3706. DOI: 10.1007/s10350-005-0186-6 (cit. on p. 98).

Walker, J. and P. Quirke (May 2002). “Prognosis and Response to Therapy in Colorectal Cancer”. In: *European Journal of Cancer* 38.7, pp. 880–886. ISSN: 09598049. DOI: 10.1016/S0959-8049(02)00044-8 (cit. on p. 98).

Shortliffe, Edward H. and Martin J. Sepúlveda (Dec. 4, 2018). “Clinical Decision Support in the Era of Artificial Intelligence”. In: *JAMA* 320.21, pp. 2199–2200. ISSN: 0098-7484. DOI: 10.1001/jama.2018.17163 (cit. on p. 98).

Attenberger, U.I., L.R. Pilz, J.N. Morelli, et al. (July 2014). “Multi-Parametric MRI of Rectal Cancer Do Quantitative Functional MR Measurements Correlate with Radiologic and Pathologic Tumor Stages?” In: *European Journal of Radiology* 83.7, pp. 1036–1043. ISSN: 0720048X. DOI: 10.1016/j.ejrad.2014.03.012 (cit. on p. 99).

Buda, Mateusz, Atsuto Maki, and Maciej A. Mazurowski (Oct. 2018). “A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks”. In: *Neural Networks* 106, pp. 249–259. ISSN: 08936080. DOI: 10.1016/j.neunet.2018.07.011 (cit. on p. 99).

Guillem, José G., Jeannine A. Ruby, Tobias Leibold, et al. (Aug. 2013). “Neither FDG-PET Nor CT Can Distinguish Between a Pathological Complete Response and an Incomplete Response After Neoadjuvant Chemoradiation in Locally Advanced Rectal Cancer: A Prospective Study”. In: *Annals of Surgery* 258.2, pp. 289–295. ISSN: 0003-4932. DOI: 10.1097/SLA.0b013e318277b625 (cit. on p. 99).

Brigato, Lorenzo and Luca Iocchi (Jan. 2021). “A Close Look at Deep Learning with Small Data”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2020 25th International Conference on Pattern Recognition (ICPR), pp. 2490–2497. DOI: 10.1109/ICPR48806.2021.9412492 (cit. on p. 100).

Scherer, Jonas, Marco Nolden, Jens Kleesiek, et al. (Nov. 2020). “Joint Imaging Platform for Federated Clinical Data Analytics”. In: *JCO Clinical Cancer Informatics* 4, pp. 1027–1038. ISSN: 2473-4276. DOI: 10.1200/CCI.20.00045 (cit. on p. 102).

Rieke, Nicola, Jonny Hancox, Wenqi Li, et al. (Sept. 14, 2020). “The Future of Digital Health with Federated Learning”. In: *npj Digital Medicine* 3.1, p. 119. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00323-1 (cit. on p. 102).

Littlejohns, Thomas J., Jo Holliday, Lorna M. Gibson, et al. (May 26, 2020). “The UK Biobank Imaging Enhancement of 100,000 Participants: Rationale, Data Collection, Management and Future Directions”. In: *Nature Communications* 11.1, p. 2624. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15948-9 (cit. on p. 102).

Peters, Annette, German National Cohort (NAKO) Consortium, Annette Peters, et al. (Oct. 2022). “Framework and Baseline Examination of the German National Cohort (NAKO)”. In: *European Journal of Epidemiology* 37.10, pp. 1107–1124. ISSN: 0393-2990, 1573-7284. DOI: 10.1007/s10654-022-00890-5 (cit. on p. 102).

Liu, Wenyang, Yifei Li, Yuan Tang, et al. (Apr. 2022). “Response Prediction and Risk Stratification of Patients with Rectal Cancer after Neoadjuvant Therapy through an Analysis of Circulating Tumour DNA”. In: *eBioMedicine* 78, p. 103945. ISSN: 23523964. DOI: 10.1016/j.ebiom.2022.103945 (cit. on p. 103).

Wang, Yaqi, Lifeng Yang, Hua Bao, et al. (Aug. 31, 2021). “Utility of ctDNA in Predicting Response to Neoadjuvant Chemoradiotherapy and Prognosis Assessment in Locally Advanced Rectal Cancer: A Prospective Cohort Study”. In: *PLOS Medicine* 18.8. Ed. by Sarah-Jane Dawson, e1003741. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1003741 (cit. on p. 103).

Chiloiro, G., Davide Cusumano, Paola Franco, et al. (Nov. 1, 2021). “Does Restaging MRI Radiomics Analysis Improve Pathological Complete Response Prediction in Rectal Cancer Patients? A Prognostic Model Development”. In: *La radiologia medica* 127. DOI: 10.1007/s11547-021-01421-0 (cit. on p. 103).

Zhang, Xiao-Yan, Lin Wang, Hai-Tao Zhu, et al. (July 2020). “Predicting Rectal Cancer Response to Neoadjuvant Chemoradiotherapy Using Deep Learning of Diffusion Kurtosis MRI”. In: *Radiology* 296.1, pp. 56–64. ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.2020190936 (cit. on p. 103).

Declaration

This thesis is the result of my independent investigation under supervision. Where my work is indebted to the work or ideas of others, for example from the literature or the internet, I have acknowledged this within the thesis.

I declare that this study has not already been accepted for any degree, nor is it currently being submitted in candidature for any other degree.

I am aware that a false declaration could have legal implications.

Erklärung

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Mannheim, October 2nd, 2023

Steffen Albert