Artificiële intelligentie met medische beeldvorming
voor de diagnose van primaire hersentumoren

Artificial Intelligence in Medical Imaging
for the Diagnosis of Primary Brain Tumours

Stijn Bonte

Promotoren: prof. dr. R. Van Holen, prof. dr. I. Goethals
Proefschrift ingediend tot het behalen van de graad van
Doctor in de ingenieurswetenschappen: biomedische ingenieurstechnieken

UNIVERSITEIT
GENT

Vakgroep Elektronica en Informatiesystemen
Voorzitter: prof. dr. ir. K. De Bosschere
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2018 - 2019

ISBN 978-94-6355-168-7
NUR 954, 984
Wettelijk depot: D/2018/10.500/86

Medical Image and Signal Processing (MEDISIP)
Department of Electronics and Information Systems
Faculty of Engineering and Architecture
Ghent University

Department of Radiology and Nuclear Medicine
Faculty of Medicine and Health Sciences
Ghent University

Corneel Heymanslaan 10
Entrance 36, floor 5
9000 Ghent
Belgium

**Promotors**
prof. dr. Roel Van Holen
prof. dr. Ingeborg Goethals

**Examination board**
prof. dr. ir. Paul Kiekens, Ghent University, *chairman*
prof. dr. Karel Deblaere, Ghent University, *secretary*
prof. dr. ir. Wesley De Neve, Ghent University
dr. Mathieu Hatt, French Institute of Health and Medical Research
prof. dr. Michel Koole, KU Leuven

# Acknowledgements

*"Good judgement is the result of experience,*
*experience the result of bad judgement."*
*Mark Twain*

Doing a PhD is - rather than a quest for good scientific output - a learning experience. It not only teaches you how to perform a thorough scientific study, but also to set up your own project, how to deal with the inevitable problems, how to present your work to experts and peers, and how to cooperate with colleagues from different fields of expertise. This book, which is considered the end product of my PhD, can therefore only contain a fraction of the results I've obtained during these four years. I want to thank many people who have guided and supported me through this process.

First of all, a most sincere thanks goes out to my promotors, prof. dr. Roel Van Holen and prof. dr. Ingeborg Goethals, without whose help my PhD would not have been successful. Roel, thank you for the weekly meetings and the ever-constructive feedback. You were always able to ask the right questions in order to improve my work. With your help, we have given shape to this project and this dissertation. Ingeborg, thank you for believing in this topic and in me. I could always come to you with medical questions, you have provided me with a good insight in the biological background and clinical management of brain tumour patients. You have also given me the opportunity to present and discuss my work to the medical experts in the Ghent University Hospital on

multiple occasions. I would also like to thank you for the nice time we had during our conference trips to Mannheim, Vienna and Stockholm.

Next, I would like to thank the co-authors and jury members for reading my manuscripts and providing feedback that has significantly improved the final version of my papers and this thesis. Furthermore, my gratitude goes out to dr. Giorgio Hallaert, dr. Caroline Van den Broecke, dr. Marjan Acou, prof. dr. Tom Boterberg and the other members of the multidisciplinary neuro-oncological staff of the Ghent University Hospital for helping me collecting the data and for the very interesting discussions.

For the last four years, MEDISIP has been my second home. Therefore, I would like to thank prof. dr. ir. Stefaan Vandenberghe, prof. dr. ir. Christian Vanhove, prof. dr. ir. Pieter van Mierlo, dr. Benedicte Descamps and dr. ir. Vincent Keereman for shaping the group into a multidisciplinary team with a broad range of different research projects. Also, all of my many colleagues deserve to be mentioned here, it's you who make the living heart and the great atmosphere of our group. Mariele and Marek, thank you for the bike rides in the Vlaamse Ardennen, the nice evenings and the silly conversations. Paulo, Prakash, Charlotte and Gwennaëlle, thank you for the great times we've spent in the office. Milan, thank you for the advice on machine learning. Jens, Kim, Tim, Emma, thank you for the research coffee breaks (one of our better ideas). I wish you all the best of luck with your PhD projects! Thibault, thank you for introducing me to artificial intelligence. Willeke, thank you for all the advice and for allowing me to use your (and Thibault's) PhD book source code. Nathalie, Carmen, Karen, Ester, Margo and Radek, thank you for being so inspiring.

Saskia and Inge, you make our PhD lives so much easier. Your good care allows us to focus on our research activities. I also want to thank the administrative staff from the nuclear medicine department for their help. Yves and Johan, your Statler-and-Waldorf-wise discussions made the long hours of data collection more pleasant. Pieter Devolder, many thanks for all your efforts in collecting the MGMT-data. Michaël and Sam, your help significantly reduced the time I had to spend in collecting the PET database.

I also want to thank prof. dr. ir. Patrick Segers for his sincere inter-

est and for being so involved with all the IBiTech PhD students. Also a big thank you to all the bioMMeda colleagues for the often hilarious lunch time discussions.

Muziek is al vele jaren mijn grootste uitlaatklep. Daarom wil ik in het bijzonder Elly en Geert bedanken voor de mooie kansen met Jong Symfonisch Gent. Ik hou veel goede vrienden over aan dit orkest. Ook Kevin, het bestuur en de muzikanten van Continuo wil ik bedanken voor de vele mooie muzikale en niet-muzikale momenten. Reinout en Alexander wil ik hier graag vermelden voor de PCM-sessies, dringend tijd om nog eens een nieuwe barbecue te organiseren!

Ten slotte wil ik hier uitdrukkelijk mijn ouders bedanken. Jullie onvoorwaardelijke steun bij alles wat ik onderneem (zelfs het zotte idee om drie benefietconcerten te organiseren) betekent bijzonder veel voor mij. Het is al te makkelijk dit als een evidentie te beschouwen, maar zonder jullie hulp was ik onmogelijk zo ver gekomen. Ook Sofie en David, Iluna en Andreas, mijn grootouders, schoonouders en familie wil ik bedanken voor de liefde, de leuke reisjes en gezellige momenten. En dan blijft de belangrijkste persoon in mijn leven nog over: Valérie – liefje. Jij bent mijn steun en toeverlaat, en ik ben er zeker van dat we nog hele mooie tijden tegemoet gaan.

Bedankt!

Stijn
Gent, 21 november 2018

# Summary

The goal of this PhD dissertation is to develop a computer-aided diagnosis system for primary brain tumours based on medical imaging using techniques from artificial intelligence. We will focus on two tasks: the automated delineation of the tumour on medical images, and tumour classification, with special emphasis on distinguishing between low-grade and high-grade gliomas.

Primary brain tumours are a complex class of neoplasms originating in the brain. With an estimated 10.8 people per 100 000 diagnosed with a form of primary brain tumour per year, they are relatively rare. However, they contribute significantly to the number of cancer-related deaths, mainly because the brain itself is an extremely complex and vital organ. Determining the optimal treatment strategy strongly depends on the accurate diagnosis of the tumour. In the 2016 classification scheme, the World Health Organisation defines over 150 different types of primary brain tumours, based on both histological and genetic findings. Therefore, tumour samples need to be analysed, requiring an invasive and potentially risky surgical procedure. However, in some cases such a radical intervention is not possible, e.g. due to medical co-morbidities, when the tumour is difficult to reach or located in eloquent regions, or when the patient refuses surgery. In contrast, medical imaging forms a non-invasive and repeatable tool towards brain tumour diagnosis.

State-of-the-art imaging procedures are able to map a wealth of both anatomical tumour structures and biological processes. They are however not able to visualise information on the cellular or genetic level,

necessary for an accurate diagnosis. Nevertheless, processes on the microscopic level might be translated into signals on the macroscopic level which can be picked up by dedicated computer algorithms. This is the hypothesis of *radiomics*. In this work, we apply the radiomics workflow to magnetic resonance imaging (MRI) and positron emission tomography (PET) scans of primary brain tumour patients, collected at the Ghent University Hospital and from online repositories.

In a first study, we investigate the problem of automatically distinguishing lower-grade from high-grade gliomas, as this has important consequences for both prognosis and therapy-planning. We use an online dataset consisting of 75 lower-grade and 210 high-grade glioma patients, for which structural MRI scans and a manual tumour delineation are provided. Per patient, 2097 quantitative features are calculated, capturing the appearance of the tumour on the images. These features are then fed to a classification model based on machine learning. We compare different models and obtain an optimal accuracy of 88% correctly classified patients using random forests.

Since manual tumour delineation is a time- and labour-intensive task prone to inter- and intra-performer variability, we next investigate different approaches towards automated tumour segmentation. A first technique models the healthy brain tissues based on prior anatomical knowledge. Locations with a different intensity profile compared to the expected value are considered as abnormal tissue. This algorithm is tested on a dataset consisting of 274 patients, obtaining a median Dice score of 73.3%. The main advantage of this approach is that it is very flexible regarding the number and type of input images, since it is not trained on an annotated dataset. However, the algorithm requires an excellent estimation of the healthy tissues, a task that might not be straightforward when the deviation from the normal anatomy is large. Moreover, the outlier detection method only distinguishes between normal and abnormal appearing tissue, and is therefore not suitable for discriminating between different tumour tissues such as necrosis, contrast-enhancing tumour or oedema.

Therefore, we implement a second tumour segmentation algorithm, where a machine learning model learns to recognise the appearance of different tumour tissues based on an annotated training set of 30 pa-

tients. For every voxel (= volume pixel, smallest element of a 3D picture) in the image, we calculate 52 features capturing the local texture, as well as several measures of abnormality. Based on these features, the voxel is classified into one of five healthy and four tumour tissues. Mainly for high-grade gliomas, the algorithm obtains good results, with Dice scores of 74.8%, 75.0% and 80.1% for segmenting the contrast-enhancing tumour, tumour core and total abnormal region, respectively.

This segmentation method is then applied to MRI scans of 352 patients, collected in eight different centres. Since distinguishing low-grade from high-grade tumours is not sufficient to determine the optimal therapy, the patients are divided into six different tumour classes. We again extract quantitative features in order to classify them in a *multiclass* fashion, in contrast to the *binary* problem that was discussed before. In a first approach, we model two classification algorithms: one for tumour grade (grade I–IV) and one for tumour type (meningioma, astrocytoma, oligodendroglioma, glioblastoma). The first model achieves an overall accuracy of 60.3%, while the model predicting grade achieves 65.6%. However, both models show poor performance for one out of four classes, hampering the applicability in clinical practice.

In general, physicians will already have a good idea about the diagnosis based on the medical images and clinical status. They might however doubt between a few specific possible tumour types. Therefore, in a second approach, we again split up the multiclass problem in a series of fourteen binary classifiers, comparing different tumour groups. Every binary problem can be solved with high accuracy, ranging from 75% to 95%. Afterwards, the binary classifiers are combined in four decision schemes or using machine learning. This yields a best overall accuracy of 52.8%.

In the last part of this dissertation, we aim to improve the performance of the binary grade classifier by complementing the MRI scans with $^{18}$F-FET PET scans. This amino acid radiotracer shows an excellent tumour-to-background contrast. Moreover, the dynamic uptake profile of this radiotracer is a well-known biomarker of malignancy in gliomas. This however requires a long scanning protocol of about 40–60 minutes. In order to increase the patient's comfort, we only use static $^{18}$F-FET images obtained during a 10-minutes protocol. The tumour

masks that are previously segmented on MRI are transferred to the PET scans, and we again extract quantitative features capturing the intensity distribution, shape, texture and environment on both MRI and PET. Validated on 30 patients, we predict a correct tumour grade in 29 patients (accuracy of 96.7%) using only one MRI and four PET features. Further validation on an independent dataset is however necessary to confirm this result.

In conclusion, we have investigated several techniques from artificial intelligence to aid in the diagnosis of primary brain tumours. An automated tumour segmentation algorithm is presented, able to delineate several tumour tissues on MRI scans. Applying this algorithm to clinical scans, we show several approaches towards a computer-aided diagnosis. In particular, fourteen binary classifiers achieving a high accuracy can aid the physician in decision-making. When combining MRI with amino acid PET, the discrimination between low-grade and high-grade gliomas is further improved.

# Samenvatting

Het doel van deze doctoraatsthesis is een computer-geassisteerd diagnosesysteem voor primaire hersentumoren te ontwikkelen gebaseerd op medische beeldvorming en aan de hand van technieken uit artificiële intelligentie. We zullen ons toeleggen op twee taken: het automatisch aflijnen van tumoren op medische beelden, en tumorclassificatie, met bijzondere nadruk op het onderscheid tussen laag- en hooggradige gliomen.

Primaire hersentumoren vormen een complexe klasse van neoplasia die ontstaan in de hersenen. Gezien de incidentie geschat wordt op 10.8 personen per 100 000 per jaar, zijn ze vrij zeldzaam. Ze dragen nochtans sterk bij tot aan kanker gerelateerde sterfte, voornamelijk omdat de hersenen zelf een gecompliceerd en vitaal orgaan zijn. Het bepalen van de optimale behandelingsstrategie hangt sterk af van de accurate diagnose van de tumor. De Wereldgezondheidsorganisatie definieert meer dan 150 types primaire hersentumoren in het classificatiesysteem uit 2016. Dit is gebaseerd op zowel histologische als genetische bevindingen. Om een diagnose te stellen dient tumorweefsel geanalyseerd te worden, wat een invasieve en potentieel riskante chirurgische ingreep vereist. Een dergelijke ingrijpende operatie is echter niet altijd mogelijk, denken we bijvoorbeeld aan andere aandoeningen die een ingreep verhinderen, wanneer de tumor moeilijk te bereiken is, wanneer de tumor zich in eloquente hersengebieden bevindt, of als de patiënt een operatie weigert. Medische beeldvorming daarentegen vormt een niet-invasieve methode om een diagnose te stellen, die bovendien meermaals kan herhaald worden.

Beeldvormingsprocedures zijn tegenwoordig in staat om een grote

verscheidenheid aan zowel anatomische structuren als biologische processen in kaart te brengen. Ze zijn nochtans niet in staat om informatie op cel- of genetisch niveau te visualiseren, wat nodig is voor een accurate diagnose. Processen op de microscopische schaal kunnen zich echter wel vertalen in signalen op macroscopisch niveau die opgepikt kunnen worden door gespecialiseerde computeralgoritmen. Dit is de hypothese van *radiomica*. In deze dissertatie zullen we de workflow van radiomica toepassen op beelden van magnetische resonantie (MRI) en positronenemissietomografie (PET) van patiënten met primaire hersentumoren, verzameld in het Universitair Ziekenhuis Gent en in online databases.

In een eerste studie onderzoeken we het probleem van het automatisch onderscheiden van lagergradige en hooggradige gliomen, aangezien dit belangrijke consequenties heeft voor de prognose en therapie. We maken gebruik van een online dataset bestaande uit 75 lagergradige en 210 hooggradige glioompatiënten, waarvoor anatomische MRI scans en een manuele tumorsegmentatie voorhanden zijn. Per patiënt worden 2097 kwantitatieve parameters of *features* berekend, die het voorkomen van de tumor op een beeld in kaart brengen. Deze features worden vervolgens gebruikt in een classificatiemodel gebaseerd op machinaal leren. We vergelijken verschillende modellen en behalen een optimale accuraatheid van 88% correct voorspelde patiënten gebaseerd op *random forests*.

Aangezien manuele tumordelineatie een tijds- en arbeidsintensieve opdracht is, die bovendien gevoelig is aan variabiliteit tussen verschillende uitvoerders, onderzoeken we vervolgens strategieën om deze taak automatisch uit te voeren. Een eerste techniek modelleert de gezonde hersenstructuren op basis van eerdere anatomische kennis. Zones die een afwijkend intensiteitsprofiel vertonen in vergelijking met de verwachte waarde, worden beschouwd als abnormaal weefsel. Dit algoritme wordt getest op een dataset van 274 patiënten, waarbij we een mediane Dice score van 73.3% bereiken. Het grootste voordeel van deze aanpak is dat het zeer flexibel is qua aantal en type ingevoerde beelden, aangezien het niet getraind is op een geannoteerde dataset. Het algoritme dient echter te beschikken over een uitstekende schatting van de gezonde structuren, wat niet evident is wanneer er een sterke afwijking van de normale anatomie is. Bovendien maakt de methode enkel onderscheid tussen schijnbaar normale en abnormale weefsels, wat het niet geschikt maakt

om tumorgerelateerde structuren als necrose, contrastcapterend weefsel of oedeem te identificeren.

Om die redenen implementeren we een tweede tumorsegmentatie-algoritme, waarin een machinaal leren model het uitzicht van verschillende tumorweefsels leert te herkennen, gebaseerd op een geannoteerde trainingset van 30 patiënten. Voor elke voxel (=volume pixel, kleinste element in een 3D beeld) in het beeld worden 52 features berekend die naast de lokale textuur ook meerdere abnormaliteitswaarden vastleggen. De voxel wordt geclassificeerd in één van de vijf gezonde weefsels of vier tumorklassen op basis van deze features. Dit algoritme behaalt goede resultaten, voornamelijk voor hooggradige glioma, met Dice scores van 74.8%, 75.0% en 80.1% voor het aflijnen van respectievelijk het contrastcapterend weefsel, de tumorkern en de volledige abnormale zone.

Vervolgens wordt deze segmentatiemethode toegepast op MRI beelden van 352 patiënten, verzameld in acht verschillende centra. Aangezien het onderscheid tussen laag- en hooggradige tumoren niet volstaat om de optimale therapie te plannen, worden de patiënten onderverdeeld in zes tumorklassen. Opnieuw worden de kwantitatieve features berekend om op basis hiervan de patiënten te classificeren in een *multiclass* systeem, in tegenstelling tot de *binaire* problemen die voorheen besproken werden. In een eerste studie modelleren we twee classificatiemodellen: een voor tumorgraad (I–IV), een ander voor tumortype (meningioom, astrocytoom, oligodendroglioom, glioblastoom). Het eerste model behaalt een accuraatheid van 60.3%, het model voor type een accuraatheid van 65.6%. Beide modellen vertonen echter een zwak resultaat voor één van de vier klassen, wat de klinische toepasbaarheid beperkt.

Artsen hebben vaak een gegrond vermoeden over de diagnose gebaseerd op de medische beelden en de klinische status van de patiënt. Het kan echter voorvallen dat ze nog twijfelen tussen een aantal mogelijke tumortypes. Daarom splitsen we in een tweede studie het multiclass probleem opnieuw op in een reeks van veertien binaire classificatiemodellen die elk verschillende tumorgroepen vergelijken. Elk binair probleem kan opgelost worden met een hoge precisie, gaande van 75% tot 95%. Nadien worden de binaire modellen ook gecombineerd aan de hand van beslissingsschema's of met machinaal leren. Dit levert een beste accuraatheid van 52.8% op.

In het laatste deel van deze thesis proberen we het resultaat van het binaire tumorgraadmodel te verbeteren door, naast MRI beelden, $^{18}$F-FET PET beelden te incorporeren. Deze aminozuur radiotracer vertoont een uitstekend tumor-achtergrondcontrast. Bovendien is het dynamische opnameprofiel van deze speurstof een gekende biomerker van maligniteit in gliomen. Hiervoor is echter een lang scanprotocol nodig van ongeveer 40–60 minuten. Om het comfort van de patiënt te optimaliseren gebruiken we hier enkel statische $^{18}$F-FET beelden bekomen gedurende een protocol van 10 minuten. De tumormaskers die eerder gesegmenteerd waren op de MRI-beelden worden toegepast op de PET-scans, en opnieuw extraheren we kwantitatieve features die de intensiteitsverdeling, vorm, textuur en omgeving van de tumor bevatten op zowel MRI als PET. Deze methode wordt gevalideerd op 30 patiënten, waarvan 29 correct voorspeld worden (accuraatheid van 96.7%), uitsluitend op basis van 1 MRI en 4 PET parameters. Dit resultaat dient echter gevalideerd te worden op een onafhankelijke dataset.

Samenvattend hebben we verschillende technieken uit de artificiele intelligentie onderzocht om te assisteren in de diagnose van primaire hersentumoren. We stellen een automatisch tumorsegmentatie-algoritme voor dat in staat is om verschillende tumorweefsels af te lijnen op MRI. Vervolgens tonen we verschillende strategieën voor een computer-geassisteerde diagnose waarbij dit algoritme wordt toegepast op klinische beelden. In het bijzonder kunnen veertien binaire modellen met een hoge precisie de artsen helpen bij de therapeutische besluitvorming. Wanneer MRI gecombineerd wordt met aminozuur beeldvorming op basis van PET kan het onderscheid tussen laag- en hooggradige gliomen verder verbeterd worden.

# Contents

# 1 Problem and goal

## 1.1 Context

On Saturday June 30, 2018, a remarkable competition took place in the Chinese capital Beijing. A team of 15 elite physicians was asked to diagnose patients with brain tumours based only on medical images. Every radiologist, specialised in neuroimaging, received a set of state-of-the-art imaging protocols from 15 patients, and they had to write down the diagnosis they deemed most probable within 30 minutes. Their opponent: a computer with a dedicated artificial intelligence (AI) system called BioMind [1, 2]. Similar to the physicians, it had learned to detect abnormalities, extract features and classify them accordingly based on seeing thousands of images before. The computer scored all 225 images in 15 minutes, thereby achieving an accuracy of 87% correctly diagnosed cases. The radiologists, although performing better than average achieving a 63% accuracy, were outperformed by far.

Primary brain tumours are a complex class of neoplasms. Even though they are relatively rare, they are often difficult to treat as the brain is a extremely complex and important organ itself. Therefore, many types of primary brain tumours are not curable today.

Recent advancements in genetic and molecular research have led to new insights in tumour biology. Based on these parameters, patients can receive a better, more detailed diagnosis, which in turn leads to a personalised treatment. This procedure, called precision medicine,

provides the patient with significantly better survival perspectives.

However, in order to obtain an accurate diagnosis, tumour tissue needs to be removed and analysed. This requires an invasive and potentially risky surgical procedure. For many patients, removing the tumour as much as possible while preserving vital structures is a first and crucial step in the treatment, providing sufficient tissue for diagnosis. However, when clinical symptoms and prior imaging do not yield enough evidence to justify such a radical intervention, physicians might opt for a biopsy. In this case, small tumour samples will be removed using a needle. But as tumours can be heterogeneous, small zones of malignancy can be missed, or there might be not enough tissue available for a detailed diagnosis. Furthermore, in some cases a surgical procedure can be impossible, for example due to medical co-morbidities, when the tumour is difficult to reach or located close to vital structures, or when the patient refuses surgery.

In these cases, medical imaging forms a non-invasive and repeatable tool towards tumour diagnosis. State-of-the-art imaging is able to map both the anatomy and several biological processes *in vivo* of the tumour and the surrounding tissues, thereby delivering a wealth of information which can be interpreted by a trained radiologist. Still, no medical scanner is able to display the tumour on the cell level, let alone on the genetic and molecular level.

Nevertheless, recent progress in medical image analysis has shown that processes on a microscopic level might be translated on the macroscopic level. Although not always perceivable to the naked eye, specific patterns in medical images can be picked up by computer algorithms that lead to an improved tumour classification. This is the hypothesis of *radiomics*.

For these and other applications, dedicated computer programs are increasingly being developed and used in the clinical practice. As not only the amount of clinical data, but also their complexity, is rapidly increasing, software aiding clinicians in detecting the right signals will play an ever-increasing role in medical diagnostics. For example, the Watson Oncology system, designed by IBM in collaboration with clinicians from the Memorial Sloan Kettering Cancer Center, scans the medical record of the patient and recommends specific cancer therapy options.

This is based on different treatments that were prescribed by physicians before [3]. Similarly, the London-based company DeepMind worked in close collaboration with the UK National Health Service on a mobile app called Streams. This app monitors the health status of patients with acute kidney injury, and sends out warning signals directly to the treating physicians when urgent assistance is necessary.

By using these tools, physicians do not longer need to spend time on routine tasks. Instead they can turn their focus to more complex and important jobs. As professor Paul M. Parizel, former president of the European Society of Radiology and one of the jury members during the Chinese competition, said:

> *Personally, I believe that AI will become integrated into existing medical work flow environments, more or less like a GPS navigation system guiding the driver of a car. AI software will give proposals and help the doctor to make an accurate diagnosis, thus providing a roadmap towards correct patient management and follow-up. But it will be the doctor who ultimately decides, as there are a number of factors that a machine cannot possibly take into consideration, such as a patient's state of health and family situation.* [2]

## 1.2 Outline

In this PhD dissertation, we will investigate different techniques from AI that can help physicians to obtain a better diagnosis of primary brain tumours based on medical images. We will thereby focus on the most common types of tumours. Moreover, the goal is to enhance the interpretability of the models, by providing probabilities for different possible outcomes and minimising the number of features a decision is based on.

This work is located at the crossroad of five research domains: computer-aided diagnosis, neuro-oncology, medical imaging, radiomics and machine learning. These topics will be introduced in **Chapter 2**.

In **Chapter 3**, we will investigate the important task of discriminating low-grade from high-grade gliomas. Tumour grade does not only serve as a predictor of prognosis, it also plays a major role in determining the optimal therapy. Different features, being quantitative parameters describing the tumour appearance on the scan, will be presented. Next, these features will be incorporated into several classification models, and we will compare their performance. This is done on a public dataset where a delineation of the tumour is provided.

As tumour delineation is a time-consuming task, we will elaborate on automatic solutions in **Chapter 4**. In a first method, the algorithm detects healthy tissue, and regions with abnormal intensities will therefore be regarded as tumour. A second algorithm uses a training set to learn the appearance of different tumour tissues.

The latter method will be applied to clinical scans in **Chapters 5, 6 and 7**. Since brain tumours can be divided into many different categories, we will focus here on multiclass classification problems. In particular, two models will be developed in **Chapter 5**: one for tumour grade and one for tumour type.

In many cases, radiologists have a good idea on the diagnosis, but might doubt between a small number of specific tumour types. To aid in these situations, we develop 14 binary models in **Chapter 6**. These models will each time give probabilities for two possible tumour classes. Afterwards, we try to combine these probabilities again in a multiclass decision scheme.

In the previous chapters, the input of our models are features calculated on anatomical imaging techniques (MRI). In **Chapter 7**, we complement this with a functional imaging modality, mapping the uptake of amino acids in the tumour ($^{18}$F-FET PET).

Finally, **Chapter 8** concludes this dissertation and offers some future perspectives.

# 2 Introduction

In this chapter, several topics that will be adressed throughout the course of this thesis will be introduced. The purpose of this chapter is not to give an in-depth study of all subjects, but rather to provide the reader with enough background knowledge to understand the topics that will be studied later on. We start with the concept of computer-aided diagnosis, combining techniques from artificial intelligence with medical imaging. This is followed by an introduction to primary brain tumours, the main topic of this thesis. Next, a short summary of medical imaging modalities will be given, followed by an introduction to radiomics. This term is a composition of radiology (science of medical imaging) and omics (neologism often used in biology describing a collective characterisation and quantification of a sample). It therefore refers to the quantification of the appearance of a structure (often a tumour) in a medical image. Since artificial intelligence is only possible when a computer is able to learn from past experience, this chapter is concluded with a section on machine learning, explaining the techniques that will be used for data analysis in this dissertation.

## 2.1 Computer-aided diagnosis

In this thesis, we investigate techniques for AI used in medicine, and particularly in computer-aided diagnosis (CAD) of primary brain tumours. Therefore, some definitions are in place:

**Artificial intelligence** "The ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience" - Encyclopaedia Britannica [4]

**Computer-aided diagnosis** "Class of computer systems that aim to assist in the detection and/or diagnosis of diseases through a "second opinion". The goal of CAD systems is to improve the accuracy of radiologists with a reduction of time in the interpretation of images. CAD systems are classified into two groups: Computer-Aided Detection (CADe) systems and Computer-Aided Diagnosis (CADx) systems. CADe are systems geared for the location of lesions in medical images. Moreover, CADx systems perform the characterization of the lesions, for example, the distinction between benign and malignant tumours" - Firmino et al. (2016) [5]

This definition of CAD clearly stresses the importance of the "second opinion". The system is used as a tool by the clinician to aid in decision making, in contrast to "automated computer diagnosis", where the computer is regarded as an independent reader of the data. The increasing importance of both CAD and AI in medicine can be clearly seen in figure 2.1, where the number of publications in recent years is shown. Since the late 1980s, the field has known a spectacular increase in the number of published papers, showing the rapid development and increasing interest in the field. For example, since 2015 more than 500 articles per year are accepted regarding computer-aided diagnosis.

### 2.1.1 History

Probably the very first paper discussing computer-aided diagnosis was published in 1958 by Lipkin et al. [6]. In this article, the use of punch-cards is suggested to store and retrieve medical information to aid in the differential diagnosis of hematological diseases. In these early days, CAD systems mainly consisted of guiding the clinician through a se-

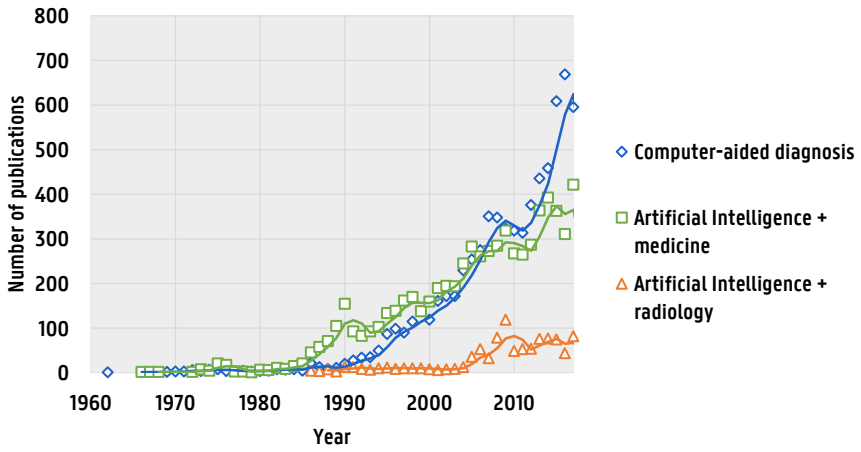



**Figure 2.1:** Articles published on Web of Science with topics related to computer-aided diagnosis or artificial intelligence in medicine and radiology.

ries of steps where the user manually inputs the patient's symptoms or imaging findings [7, 8, 9]. These patterns are evaluated by a statistical algorithm that suggests a diagnosis. Ever since, the interest in CAD and consequently the performance has increased drastically. A main breakthrough was made in the 1980s with the advent of digital imaging systems [10] and the Picture Archiving and Communication System (PACS). This allowed computers to directly process images, rather than the manual input of a human interpreter. Initially, CAD was mainly focussed on the domains where the largest clinical impact was possible, such as the detection of abnormalities on mammography [11, 12, 13] or lung nodules on chest radiography [14, 15, 16]. Already from the 1990s on, commercial systems for CAD-purposes became available for very specific tasks. Their success caused the investments in these systems to rapidly increase, which in turn resulted in a major improvement in performance [10].

In the last decades, the use of medical imaging has sky-rocketed [17, 18]. Moreover, since many of these procedures have become more complex, with an increasing number of 3D images, better image resolutions leading to larger images, and more scans per protocol, the workload

for radiologists has significantly increased [19]. This leads to a need for automated techniques able to assist the radiologist or even take over some routinely performed jobs. In recent years, the availability of large amounts of medical images that can be used as training data and the ever-increasing computing performance has led to the development of dedicated and robust CAD systems that are increasingly being used in clinical practice. Some applications of CAD will be discussed in the next section.

### 2.1.2 Applications

Research on CAD is currently being conducted on a broad range of applications in medicine. Covering them all is nearly impossible and is not the purpose of this thesis. Therefore, a selection of CAD applications is discussed here, which either have a very high medical importance, or have led to successful results in the shape of a commercially available product.

#### Screening mammographies

The largest application of CAD software is in screening mammography [20]. Already in 1998, the US Food and Drug Administration (FDA) approved CAD software for this purpose. Ten years later, 74% of all mammograms in the USA were inspected using CAD [21, 22], by 2016 this number exceeded 90% [23]. A wide variety of techniques is available, both for detection and diagnosis of suspicious lesions (see e.g. [24, 25] for an overview). However, the use of CAD in mammography is also controversial, since it might lead to a higher number of false-positives and therefore the need for a second reading of the images [26, 22]. Further technological progress will probably aid in improving these issues [27, 28].

#### Chest radiography

A chest radiograph is the most commonly performed radiologic procedure worldwide [18]. While providing a lot of information about the

health of the patient, due to the 2D nature and the inherent superposition of different tissues, it is extremely challenging to interpret. CAD methods are applied to enhance the image quality and therefore nodule detection, to detect temporal changes from subsequent radiographs using image subtraction techniques or segment the lungs, rib cages or small nodules [29]. These systems have shown to improve the detection rate by radiologists, but do not yet qualify for a stand-alone standard of diagnosis [30, 31]. Another important application of CAD on chest radiography is the diagnosis of tuberculosis. This disease ranks among the top 10 causes of death worldwide, with over 10 million people being affected every year, mainly in low- and middle-income countries [32]. However, good diagnostic accuracies obtained using CAD on low-cost chest radiography systems might improve this situation [33, 34].

### Bone age

A common way to assess physical development in children is the bone age method, usually evaluated by performing a radiography of the left hand and wrist. It can be used to diagnose growth and endocrine disorders, delayed or advanced stages of puberty and predict the final height of patients presenting with short stature [35]. The images are usually compared to a standard atlas [36], from which the bone age can be estimated [37, 38]. Other methods are possible as well, scoring the maturity level [39, 40]. These manual methods are time-consuming, require a large degree of expertise and might be subjective. A successful example of CAD software for bone age estimation is BoneXpert [41], which is CE-labelled. Several studies show that this automated method performs the bone age estimation with a similar accuracy as manual readers, but with reduced variability and a shorter processing time [42, 43].

### Brain lesions

In the follow-up of Multiple Sclerosis (MS) patients, frequently repeated magnetic resonance imaging (MRI) scans are taken to assess white matter lesions in the brain. Changes in lesion volume, lesion extent and brain volume (atrophy) are meaningful outcomes for disease prognosis. These parameters can be assessed manually, but this requires a time-

consuming segmentation step, suffering from intra- and inter-observer variability. The Belgian company Icometrix [44] provides automated software for lesion segmentation [45] and brain atrophy assessment [46]. Moreover, the software *icobrain* provides tools for the follow-up of dementia [47] and the assessment of traumatic brain injury [48]. *Icobrain* is both CE- and FDA-approved, and as a result over 100 hospitals worldwide are using the program.

## 2.2 Primary brain tumours

This dissertation will be dedicated to a specific application of CAD, namely in neuro-oncology, which is the study of neoplasms (tumours) in the brain and spinal cord. More specifically, the main focus will be on the segmentation and diagnosis of primary brain tumours (PBTs) based on medical images. In contrast to secondary brain tumours or metastases, which have spread from primary tumours elsewhere in the body, PBTs originate in the brain itself. In this section, some background information on this type of tumours is given. Since brain tumours are historically named after the cells or structures from which they arise, we start with a brief overview on neuroanatomy.

### 2.2.1 Neuroanatomy

A schematic overview of the human central nervous system (CNS), consisting of the spinal cord and the brain, is given in figure 2.2. The main components of the CNS are neurons or nerve cells. They are responsible for transmitting and processing information. Neurons are composed of a cell body or soma, dendrites and an axon or nerve fiber. Communication between neurons occurs by transferring chemical compounds called neurotransmitters, usually from the axon of one neuron to the dendrites of the next. The somas process these signals and transmit an electrical signal along the axons. Macroscopically, the brain consists of two tissue types called grey and white matter. Neuronal cell bodies are mainly found in the grey matter, while white matter consists of axons wrapped in the insulator myelin. The brain and spinal cord are surrounded by

cerebrospinal fluid (CSF), which is produced in the cavities or ventricles in the brain. The brain and spinal cord are enveloped by the protective meninges.

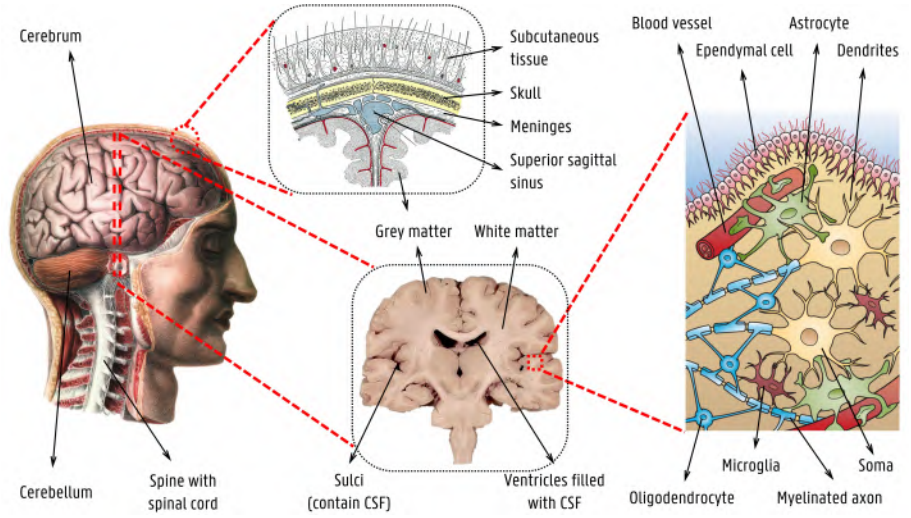



**Figure 2.2:** Schematic overview of the structure of the central nerve system, adapted from [49, 50, 51, 52].

Next to neurons, the CNS consists of glial cells or glia. The most abundant type is astrocytes. They have a star-like appearance, hence the name. Astrocytes have a number of functions, as they are involved in maintaining the blood-brain barrier (BBB) and the neuronal signalling. Oligodendrocytes are responsible for wrapping myelin around the axons. Ependymal cells are involved in secreting and circulating CSF. Finally, microglia are responsible for repairing brain damage and removing cell debris [53].

### 2.2.2 The 2016 WHO classification

The World Health Organisation (WHO) defines more than 150 different types of primary brain tumours. We limit our discussion here to some of the most common forms, as can be seen in figure 2.3.

In the classification scheme of 2016, the WHO prescribes an integrated approach for the diagnosis of primary brain tumours. The histological findings, based on light microscopic features, eosin-stained sec-

**Diffuse astrocytic and oligodendroglial tumours**
Diffuse astrocytoma, WHO grade II
Anaplastic astrocytoma, WHO grade III

Oligodendroglioma, WHO grade II
Anaplastic oligodendroglioma, WHO grade III

Glioblastoma, WHO grade IV

**Other astrocytic tumours**
Pilocytic astrocytoma, WHO grade I

**Ependymal tumours**
Subependymoma, WHO grade I
Ependymoma, WHO grade II or III
Anaplastic ependymoma, WHO grade III

**Neuronal and mixed neuronal–glial tumours**
Ganglioglioma, WHO grade I
Anaplastic ganglioglioma, WHO grade III
Rosette-forming glioneuronal tumour, WHO grade I
Central or extraventricular neurocytoma, WHO grade I

**Embryonal tumours**
Medulloblastoma, WHO grade IV
CNS neuroblastoma, WHO grade IV

**Meningiomas**
Meningioma, WHO grade I
Atypical meningioma, WHO grade II
Anaplastic (malignant) meningioma, WHO grade III

**Lymphomas**
Diffuse large B-cell lymphoma of the CNS

**Germ cell tumours**
Germinoma, WHO grade II

**Tumours of the sellar region**
Craniopharyngioma, WHO grade I

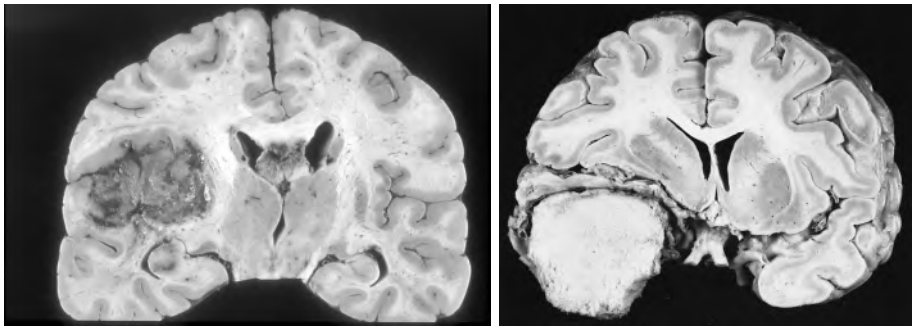
**Figure 2.3:** Selection of the WHO classification of tumours of the central nervous system. Adapted from [54].

tions and immunohistochemical (IHC) expression of proteins, are now complemented with molecular and genetic parameters.

## Histopathological features

Histologically, brain tumours are divided into a set of different categories, among which gliomas and meningiomas are the most important ones. Gliomas arise from glial cells, and were until 2016 also called after the cell type they share histological features with. Astrocytoma, oligendroglioma and ependymoma are the most common gliomas. They are in general heterogeneous and invade healthy brain tissue. Meningiomas arise from the meninges and rarely invade the brain, but rather compress surrounding tissue. They are also more homogeneous compared to gliomas. Since they are slowly growing, meningiomas may become very large before presenting with clinical symptoms. Figure 2.4 shows autopsy slices of a malignant glioma and a meningioma, where the difference in brain invasion and heterogeneity is clear.

Primary brain tumours are further classified into WHO grades, in order of increasing malignancy, according to histological characteristics such as cellularity, mitotic activity, pleomorphism, endothelial prolif-



**(a)** Glioblastoma [55] **(b)** Meningioma [56]

**Figure 2.4:** Coronal slices of gross pathology of patients with different primary brain tumours.

eration (neoangiogenesis) and necrosis [57, 58]. WHO grade I and II tumours are called low-grade glioma (LGG), the more malignant grade III and IV tumours are called high-grade glioma (HGG). LGGs have a low proliferative activity, but whereas WHO grade I tumours might be cured with surgical resection alone, WHO grade II tumours are generally more infiltrative and therefore prone to recurrence. They can also evolve into more malignant types. WHO grade III tumours show evidence of malignancy, such as nuclear atypia and anaplasia. The worst prognosis is linked to patients with WHO grade IV tumours. These are prone to necrosis (premature cell death), show a large degree of heterogeneity, are rapidly evolving and have a fatal outcome. The most malignant form of PBT is a WHO grade IV astrocytoma, also called glioblastoma (GB) or glioblastoma multiforme (GBM). This type of tumour can either evolve from a lower-grade astrocytoma (secondary glioblastoma, about 10%), or *de novo*, meaning that it immediately forms as a primary glioblastoma from healthy tissue (about 90%) [54].

### Genetic features

Histopathological diagnosis of primary brain tumours suffers from inter-observer variation, as some degree of the reporting is based on experience and subjective findings [59, 60, 61]. Moreover, several studies show that gene-expression analysis is able to correlate better with survival than

the histological diagnosis [62, 63].

Therefore, in the 2016 classification scheme, the WHO has added genetic features next to the histopathological findings [54]. Since a complete discussion of the nomenclature would lead us to far, we will focus here on the important example of the diffusely infiltrating gliomas (WHO grade II and III astrocytic tumours, grade II and III oligodendrogliomas and glioblastomas), as illustrated in figure 2.5.

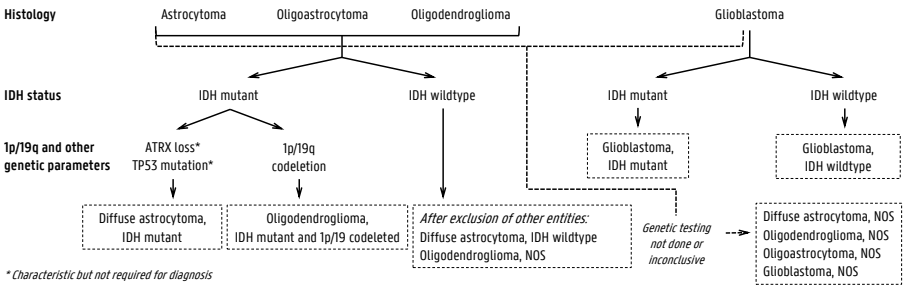



**Figure 2.5:** Classification of the diffuse gliomas based on histological and genetic features. Adapted from [54].

Two genetic parameters play a crucial role in the classification of diffuse gliomas. The first one is the enzyme isocitrate dehydrogenase (IDH), coded in the IDH1 and IDH2 genes, playing an important role in the Krebs cycle of glucose metabolism [64]. The great majority (60–90% [65, 66, 67]) of grade II and III gliomas are IDH-mutant. IDH-wildtype astrocytomas are rare, and are considered as being *glioblastoma-like* [68]. They can be detected with a negative sequencing for the *IDH1* codon 132 and *IDH2* codon 172 gene mutations. The prognosis of IDH-mutant glioma patients is favorable over IDH-wildtype cases, and IDH-status seems a more important predictor than WHO grade [69, 70]. A negative IHC for the mutant R132H IDH1 protein is also an indicator of IDH-wildtype astrocytoma, but is not sufficient for diagnosis without sequencing, as about 10% of IDH-mutated tumours are missed when relying on a negative IHC staining [71]. Astrocytoma or oligodendroglioma cases without the availability for IDH-sequencing should be classified as not otherwise specified (NOS). This is not a separate category, but indicates that the required tests for a specific diagnosis cannot be performed. Glioblastomas are also divided into IDH-wildtype (about 90%),

mostly corresponding to primary glioblastomas, and IDH-mutant (about 10%), frequently secondary glioblastomas with a history of lower-grade glioma. When full IDH evaluation is not possible, the diagnosis would be glioblastoma, NOS.

The second important genetic parameter is the combined whole-arm losses of the chromosome arms 1p and 19q (1p/19q codeletion). This can be detected using fluorescence *in situ* hybridisation (FISH). When a IDH gene family mutation and 1p/19q codeletion is present, the tumour is classified as oligodendroglioma. This type is linked to a favourable outcome, in particular since these tumours have a better response to chemotherapy [72, 73, 74].

When glioma patients are classified according to molecular subtypes, they show a distinct overall survival pattern, as illustrated in figure 2.6. This justifies the 2016 WHO classification scheme. IDH-wildtype (primary) glioblastomas show the worst prognosis, whereas oligodendrogliomas (LGG with IDH mutation and 1p/19q codeletion) present the most optimistic survival changes. It is remarkable that lower-grade gliomas with a wildtype IDH-status show a similar clinical course as glioblastomas.

### 2.2.3 Epidemiology

Primary brain tumours are a relatively rare disease. A recent systematic review and meta-analysis showed that worldwide 10.82 (95% confidence interval (CI): 8.63 - 13.56) people per 100 000 are diagnosed with a form of primary brain tumour per year [75]. However, they contribute significantly to cancer mortality, especially in children and young adults, where they are the leading cause of cancer deaths [76]. In children (age 0-14), the most common type of CNS tumour is pilocytic astrocytoma (WHO grade I, 18%), followed by malignant glioma (14.3%) and embryonal tumours (mainly medulloblastoma, 13.8%). In adults, meningiomas account for about 36.6% of all CNS tumours, followed by tumours of the pituitary (15.9%) and glioblastomas (14.9%) [77].

Ionising radiation is the only established risk factor linked to an increased incidence of both meningeal and glial tumours [78]. Concern about the use of mobile phones (emitting non-ionising radiation) seems

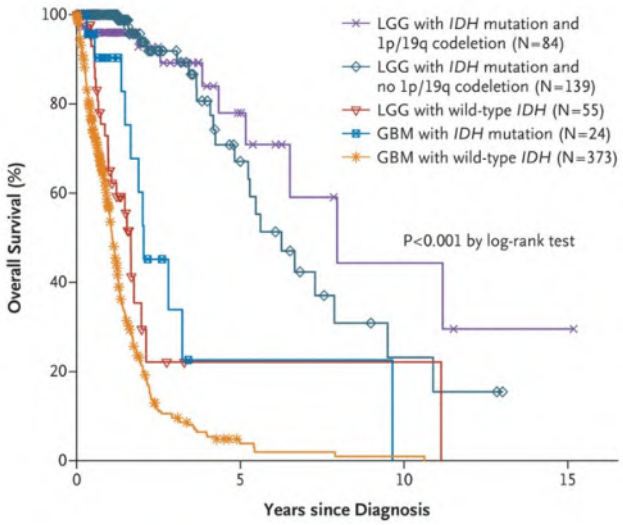



**Figure 2.6:** Kaplan-Meier plot showing overall survival when gliomas are classified according to molecular subtype. Reproduced with permission from [67], Copyright Massachusetts Medical Society.

unfounded, as a large study with over 5000 patients was not able to show a causal relation between mobile phone use and PBTs [79]. Other factors, such as genetic susceptibility, hereditary syndromes, allergies or immune-related conditions have not been established [76].

### 2.2.4 Symptoms

The symptoms of PBTs are strongly dependent on the type, location and size of the tumour. In general, there are four kinds of symptoms. The first one is epileptic seizures [80], which is more frequently the case for low-grade and slowly growing tumours. Secondly, intracranial pressure might lead to headache, nausea, vomiting, drowsiness or visual abnormalities. This is mainly the case for rapidly growing tumours showing a disruption in the blood-brain barrier, leading to a leakage of oedema, or when the CSF circulation is obstructed. A third type of symptoms are focal neurological deficits such as hemiparesis or aphasia. These symptoms can often indicate the tumour location. Lastly, mental-status abnormalities are common in frontal brain tumours or diffuse brain in-

filtration. However, many PBTs, e.g. meningiomas, are slowly growing and show no clinical effects. They are often incidentally discovered, for example after a brain scan due to an accident or stroke.

### 2.2.5 Treatment

There are several therapy options for primary brain tumours. Planning the optimal treatment protocol strongly depends on the tumour type and the presurgical evaluation. Survival is however more dependent on pretreatment variables than treatment itself [80]. The main factors for a good prognosis are, next to the type of tumour, young age and a good performance status.

#### Watch-and-wait

It might sound counter-intuitive, but the first option for some brain tumour patients is not to treat the tumour. When there are clear indications that the lesion is benign (e.g. meningioma WHO grade I) or when there are little to no symptoms, clinicians might opt to postpone an invasive procedure. The patient will however be closely monitored with frequent brain scans to gauge the tumour growth dynamics. A recent study showed that there is no difference in overall survival (OS) when surgery was delayed until signs of growth on follow-up scan compared to early resection in low-grade glioma [81].

#### Surgery

For most PBT patients however, surgical tumour resection is the first form of treatment. This not only enables the characterisation of the tumour using histopathological and molecular analysis, it frequently allows immediate relief of symptoms [80]. Most meningiomas can be cured with surgery alone, particulary patients with WHO grade I tumours in favorable locations [82].

The aim of surgery for all PBT patients is the complete tumour removal, also called gross total resection. Since this is not always feasible, maximum safe resection (MSR) is the goal [83]. An extent of resection (EOR) of at least 78% should be achieved to have a survival benefit

in GBM, and this trend continues with more complete surgical resections [84]. Still, the prevention of neurologic deficits is more important than EOR. Since glioma are infiltrative, tumour cells can be found beyond the lesion, causing these tumours not being curable with resection. Microsurgical techniques and improved imaging modalities, presurgical evaluation and intraoperative techniques for tumour delineation, as well as awake brain tumour surgery have contributed to low failure rates and excellent long-term functional outcomes [85].

If brain imaging is suggestive of a low-grade glioma, but tumour resection is not possible (e.g. due to a deep-seated lesion, located within the eloquent cortex or when other medical co-morbidities obstruct craniotomy), a biopsy might be considered. In this case, small tumour fractions will be extracted using a needle under local anaesthesia [86]. The role of biopsy is however debatable, since analysing small fractions of a heterogeneous tumour might lead to misclassification due to sampling bias [87, 88, 89]. Moreover, biopsy is related to a decreased OS compared to wait-and-scan and resection, as a recent study showed [81]. The authors suggest two possible reasons for this controversial finding. Physicians might be more biased towards a biopsy when suspecting a worse prognosis. Another option is a negative effect of the biopsy itself, such as an acute inflammatory response, or an increased glioma aggressiveness induced by the surgery. In this study, it is recommended to avoid biopsy when possible.

## Radiotherapy

Radiation therapy (RT) consists of a set of different techniques using ionising radiation for the treatment of cancer. These can be divided into external beam radiotherapy (EBRT) or internal radiotherapy. EBRT uses externally applied beams of high-energy particles such as photons, electrons or protons to irradiate the patient. Careful planning is necessary to optimise the radiation dose delivered to the lesion while protecting healthy tissue, especially sensitive regions such as the brainstem or optic nerves. This can be done using intensity-modulated radiation therapy (IMRT), where the particle beams are administered from different angles around the patient, while modifying the shape and intensity

of the beams, as is illustrated in figure 2.7. In fractional radiotherapy, a low administered dose to the tumour is repeated several times (e.g. 30 fractions of 2 Gy), whereas in stereotactical neurosurgery, a very high dose (e.g. 18 Gy) is focused on the lesion once.

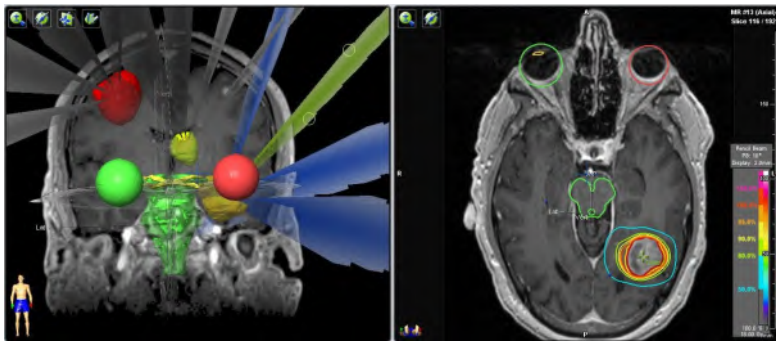


**Figure 2.7:** Example of radiation planning to optimise the delivered dose while protecting sensitive areas such as the eyes and brainstem, in this case for stereotactic radiosurgery [90].

Internal radiotherapy or brachytherapy uses small sealed radiation sources that contain unstable nuclei emitting high-energy photons. The radioactive container is placed next to the tumour, such that the target is radiated locally [91]. Brachytherapy is mainly considered in children or in adults with deeply localised tumours. Studies comparing brachytherapy with conventional techniques using randomised trials are yet to be conducted [83].

In gliomas, radiotherapy is used after surgery to improve local control, preserve function and increase survival at a reasonable risk benefit ratio. The dose and timing will depend on a number of factors, such as type and grade, as well as age, performance score and extent of resection. In meningioma, radiotherapy achieves similar results as far as disease control and survival rates are concerned to gross total resection. Mainly in WHO grade II or III (atypical or anaplastic meningiomas), radiotherapy will play a more important role [82].

### Systemic treatment

Therapy using pharmacological drugs can roughly be divided into three categories. The first type are drugs that do not offer therapeutic benefit, but are able to reduce or relieve the symptoms. These include corticosteroids to decrease cerebral oedema, antiepileptic drugs, anticoagulants (blood thinners) or antidepressants.

In the second category are cytotoxic chemotherapeutics, aiming to destroy fastly reproducing cancer cells. In Europe, temozolomide is the agent of choice for glioma treatment and is part of the standard of care for most patients. It is an oral DNA alkylating agent with limited side effects and good blood-brain barrier penetration potential. The European Association for Neuro-Oncology (EANO) recommends the use of temozolomide (combined with radiotherapy) in IDH-mutant glioblastoma and IDH-wildtype gliomas with $O^6$-methylguanine DNA methyltransferase (MGMT) promotor methylation [83]. A second type of chemotherapeutics are nitrosoureas, most commonly PCV. This is a combination of three agents: procarbazine, lomustine and vincristine. These are mostly administered to patients with anaplastic oligodendrogliomas. A last type of chemotherapy are monoclonal antibodies such as bevacizumab with anti-angiogenic capabilities. This agent is approved for recurrent glioblastoma, albeit outside the European Union.

Other pharmacotherapeutic approaches are targeted and immunological therapies, such as immune checkpoint inhibitors and vaccines. Their efficacy is however to be established in clinical trials.

The standard care of glioma patients is very briefly summarised in table 2.1. For a more detailed overview, we refer to the EANO guidelines on the diagnosis and treatment of astrocytic and oligodendroglial gliomas [83] and meningiomas [82].

## 2.3 Medical imaging

Medical imaging plays an increasing role in the diagnosis and follow-up of brain tumour patients. In this section, a short description of the two main imaging modalities in neuro-oncology will be given, being magnetic

**Table 2.1:** Standard of care in gliomas, adapted from [61, 83].

| Tumour type | Standard of care |
|---|---|
| Astrocytoma grade II | Resection followed by RT (45-55 Gy) and maintenance PCV chemotherapy |
| Astrocytoma grade III | Resection followed by RT (60 Gy) and maintenance temozolomide |
| Oligodendroglioma | Resection followed by RT plus PCV chemotherapy, regardless of tumour grade |
| Glioblastoma | Resection followed by radiotherapy (60 Gy) and 6 cycles of concomitant and maintenance temozolomide chemotherapy |

resonance imaging (MRI) and positron emission tomography (PET). To introduce the different modalities, we start by giving a brief overview of the history of medical imaging, inspired by [92].

## 2.3.1 Brief history

In figure 2.8 some important events that have marked the history of medical imaging are illustrated.

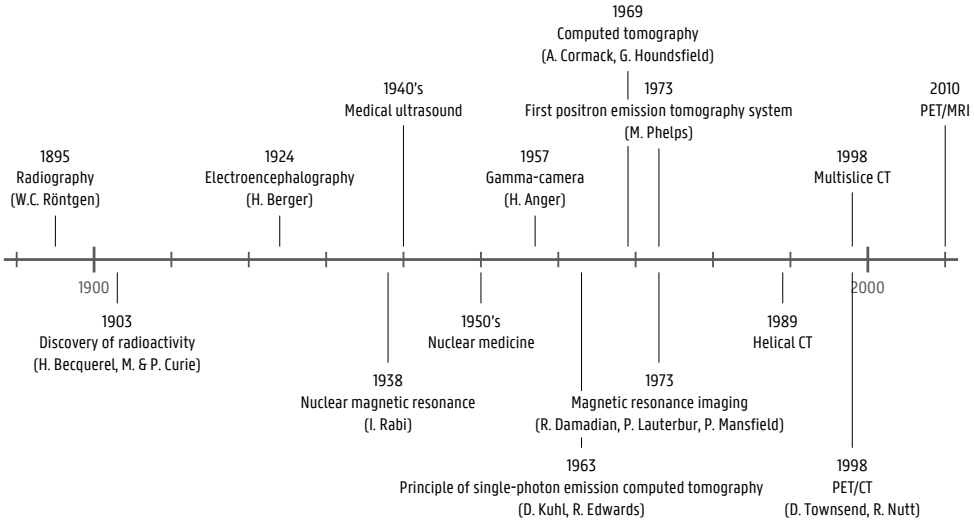



**Figure 2.8:** Important historical events in medical imaging.

## Radiography and Computed Tomography

The very first picture of the interior of the human body, without needing to cut through skin, was made by the German professor Wilhelm Röntgen. He was experimenting with a Crookes tube, a system generating electron beams, when he noticed fluorescence on a nearby barium platinocyanide screen. The invisible radiation coming out of the tube was of an unknown nature, and was therefore called X-ray. The projection of an object could however be recorded through a fluorescent material on a photosensitive film. The first image Röntgen made of his wife's hand (figure 2.9) amazed the scientific community, and the medical potential became soon clear. Since materials with a different density have a different X-ray absorption coefficient, a radiograph shows the mean density along the path between the source and every point of the detector. Radiography, which is the use of X-rays to view inside a body, was soon commercialised. A remarkable story from the early days of radiography is that Marie Curie, who had received the Nobel Prize in Physics in 1903 together with Henri Becquerel and her husband Pierre Curie for the discovery of radioactivity, invented and operated a mobile X-ray vehicle during World War I able to image bullets, shrapnel and fractures inside the body of soldiers.

**Figure 2.9:** First medical X-ray by Wilhelm Röntgen of his wife Anna Bertha Ludwig's hand (1895) [93].

In 1917, the Austrian mathematician Johann Radon proved that a function can be reconstructed from its projections along different angles [94]. This laid the foundation for Allan Cormack to develop the mathematics behind Computed Tomography (CT) in the early 1960's. In short, by rotating the source-detector geometry around the patient, a three-dimensional representation of the interior of the body can be formed. On October 1, 1971, the first patient, a woman with a right frontal lobe brain tumour, was scanned on a system designed by Sir Godfrey Hounsfield. The scanner took about 5 minutes for every slice, with 2.5 hours needed to process the data, resulting in a $80 \times 80$ pixels image, illustrated in figure 2.10. Cormack and Houndsfield received the 1979 Nobel Prize in Physiology and Medicine.

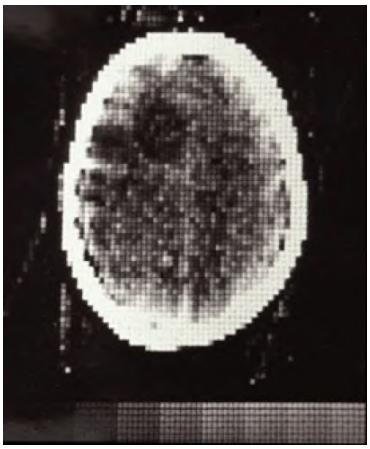


**Figure 2.10:** First clinical CT scan, Atkinson Morley's Hospital, October 1971, adapted from [95].

Ever since, CT technology has dramatically improved, starting with the first full body scanner coming soon after [96]. More recently, helical CT, multislice CT and dual-energy systems have aided in reducing scanning time and patient dose while optimising image quality.

## Nuclear medicine

Ever since radioactivity was discovered at the end of the nineteenth century, scientist have tried to use the energy emitted by unstable nuclei for medical treatment purposes, both inside and outside the body [97].

Whereas in the beginning nuclear medicine experiments were on a small scale, the Atomic Age following the Second World War marked the introduction of its applications to a broader audience. In 1946, a patient was cured of a thyroid cancer with an "atomic cocktail" [98], gaining immense publicity.

Energy from isotopes can not only be used for treatment purposes, since radiation (photons, electrons, positrons or alpha particles) emitted by unstable nuclei can also be detected, and therefore lead to an improved diagnosis. A radioactive atom linked to the right molecule will be taken up by a specific organ, which can be scanned. In the beginning, this was done by positioning a Geiger counter near the organ of interest. The advent of the Anger scintillation camera in 1957 ensured that images of the distribution of the radionuclide within the body could be made. This technique is up to now still used and is known as scintigraphy. Tomographic nuclear imaging, being the reconstruction of the distribution in a cross-sectional plane of the body, was made possible with single-photon emission computed tomography (SPECT) and positron emission tomography (PET) systems, based on the work of David E. Kuhl and Roy Edwards from the 1950's [99]. The concept of annihilation coincidence detection applied in positron emission tomography (PET) was developed in 1973 by Phelps and Hoffman [100], who also built the first PET scanner with a spatial resolution of about 1.2 cm full width at half maximum (FWHM) [101]. After developing 2-deoxy-2-($^{18}$F)fluoro-D-glucose ($^{18}$F-FDG), the same group performed the first human PET scan to visualise cerebral glucose consumption [102], illustrated in figure 2.11.

The next big step in nuclear medical imaging was the hybridisation of PET or single-photon emission computed tomography (SPECT) with Computed Tomography (CT). Since both nuclear imaging modalities suffer from a limited resolution in the range 4–6 mm for PET and 8–12 mm for SPECT, the exact localisation of the radioactivity is hard to trace. When combined with a structural imaging modality in a PET/CT [103] or SPECT/CT system, both functional and anatomical information can be obtained simultaneously. This has caused oncology to become the main application field of nuclear imaging. The last decade has also seen rise to hybrid PET/MRI scanners.

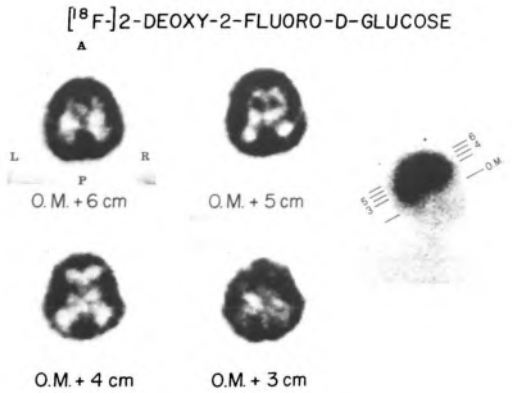



**Figure 2.11:** First human $^{18}$F-FDG PET scan (1979), reproduced with permission from [102].

## Magnetic resonance imaging

Magnetic resonance imaging (MRI) is the most modern imaging modality we will discuss here, with the first clinical scan performed in 1980, illustrated in figure 2.12. However, the theory behind this medical device dates back to 1938, when Rabi discovered nuclear magnetic resonance (NMR) [104]. The real invention of MRI is to be attributed to Paul Lauterbur and Peter Mansfield who suggested to encode spatial information in the NMR signals by varying a large magnetic field with smaller gradients. Some years prior to that, Raymond Damadian already reported that tumours and healthy tissue could be distinguished based on their NMR signal. He also designed a hypothetical MRI machine.

The technological advances in MRI followed rapidly in the 1980's and 1990's. These include improvements to better map the anatomy of the patient, such as the use of contrast agents or fluid-attenuation inversion recovery (FLAIR), both very important techniques in neuro-oncology. But next to anatomical imaging, advanced techniques made it possible to make an image of certain biological processes inside the body, such as the diffusion of water using diffusion-weighted imaging (DWI) or the perfusion of tissues by blood using perfusion-weighted imaging (PWI). Similar techniques allowed to visualise white matter tracts inside the brain using diffusion tensor imaging (DTI) and tractography. Lastly, tracking changes in the blood oxygen-level when the patient is

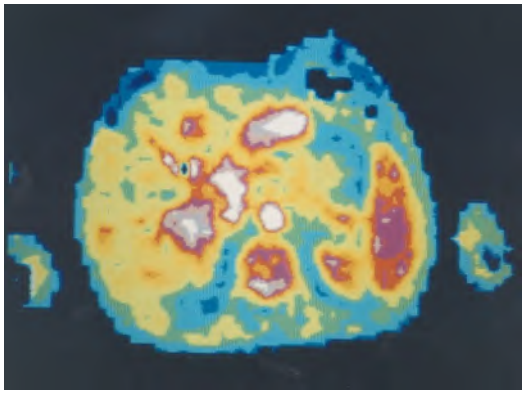


**Figure 2.12:** The very first useful clinical MRI scan showing an abdominal transverse image of malignant deposits in the liver. Copyright IOP publishing, adapted from [105].

performing a certain task, allows to identify specific regions in the brain responsible for this task, a technique called functional magnetic resonance imaging (fMRI).

## Other medical imaging techniques

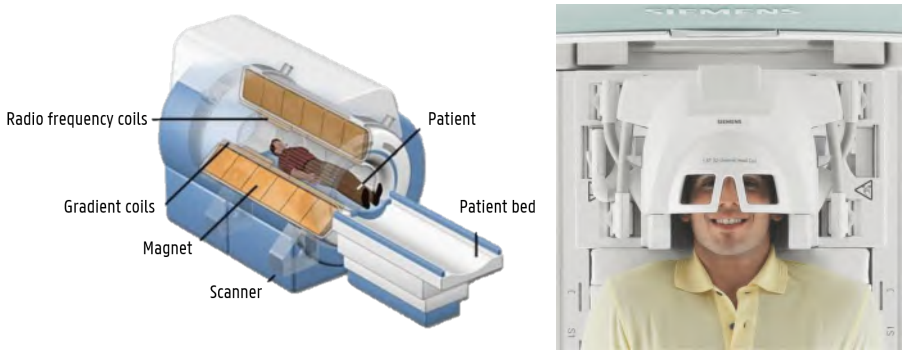
The aforementioned imaging modalities require large and expensive installations. Two other medical imaging devices used in neuro-oncology are much smaller, and can therefore be used in a mobile set-up. The first one is electroencephalography (EEG), a technique that was applied for the first time in humans in 1924 by Hans Berger. Several electrodes are positioned on the scalp of the patient, measuring the small electric signals related to neuronal activity. In this way, abnormal spikes such as epileptic seizures can be recorded, and using advanced techniques, these can be allocated to a specific seizure onset zone inside the brain. The second modality is medical ultrasound. This method is related to the radar, which had known large technical improvements during World War II. In contrast to radar, that makes use of radio-frequent electromagnetic waves, ultrasound uses high-frequency sound waves. In neuro-oncology, this method is both used therapeutically [106] and intra-operatively [107].

### 2.3.2 Magnetic Resonance Imaging

The previous section discussed the history of medical imaging. Two of the most important modalities in neuro-oncology, being MRI and PET will now be elaborated on. We start with the principle of MRI followed by important neuro-oncological applications.

#### Principle

The main component of an MRI scanner is a very large superconducting magnet, cooled with liquid helium and producing a homogeneous magnetic field. Most commonly found are 1.5 T (Tesla) systems, which is about 60 000 times stronger than the earth magnetic field, although 3 T systems are increasingly being used. Next to this static magnet, the scanner has coils in three orthogonal directions that can be switched off an on, producing small variations, called gradients, upon the main magnetic field (see figure 2.13a). Furthermore, there is a set of coils working as an antenna able to emit or receive radio-frequent (RF) waves. For most applications there exist designated receiver coils which can be brought very close to the body part being examined, such as the head coil illustrated in figure 2.13b.




**(a)** MRI scanner design, adapted from [108]

**(b)** Receiver head coil, copyright Siemens Healthineers [109]

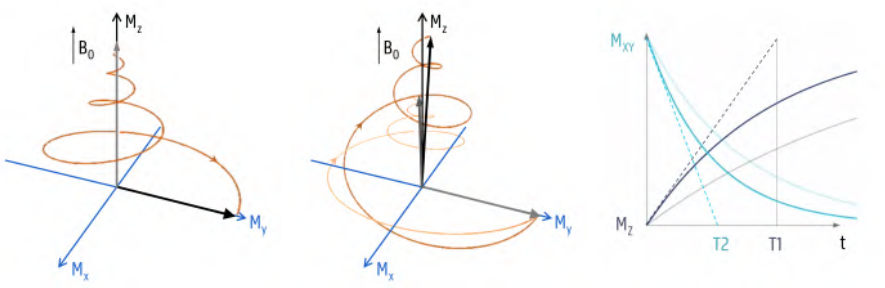
**Figure 2.13:** Magnetic Resonance Imaging: scanner design and head coil.

The principle of NMR is based on the quantum mechanical property spin of subatomic particles. This can be regarded as a rotating motion of the charged particle, creating a tiny magnetic moment. When subatomic particles are combined into nuclei, they can have a zero or non-zero net magnetic moment. Hydrogen nuclei, consisting of one proton, have a non-zero magnetisation. Since soft tissue contains a lot of water and therefore many hydrogen atoms, MRI is ideally suited to visualise it.

Hydrogen nuclei have two independent spin states, called spin-up and spin-down. In the absence of a magnetic field (thermal equilibrium), both states have the same energy. However, inside the strong magnetic field $B_0$ of the MRI scanner, an interaction will occur between the external field and the nuclear magnetic dipole moment. As a result, a small energy difference between the spin states will arise, and there will be a bias towards the lower energy state. A group of spins, also called isochromat, will therefore create a net magnetisation $M$ aligned with the external field.

The RF coil will now emit a pulse carefully tuned to the frequency with which the magnetisation precesses around the magnetic field, also called the Larmor frequency. The isochromat will absorb the energy from this pulse, which is the nuclear magnetic resonance (NMR) effect. As a result of this excitation phase, spin flip will occur, as the magnetisation starts precessing in a wider arc around the static magnetic field. The flip angle will depend on the duration of the excitation pulse, but this is often chosen to be 90°, illustrated in figure 2.14a.

When the excitation pulse is switched off, the magnetisation is allowed to return to its equilibrium state, a process called relaxation, illustrated in figure 2.14b. While relaxing, the nuclei emit energy in the shape of RF photons, which can be picked up by the receiver coils. The signal in these coils is called free induction decay (FID), a sine-wave at the Larmor frequency but damped with a time constant T2*. Relaxation is not an immediate process, as the nuclei will gradually return to equilibrium, depending on the local environment of the particle. The time at which 63% ($= 1 - 1/e$) of the net longitudinal magnetisation (parallel to the static magnetic field) is recovered is called T1, spin-lattice or longitudinal relaxation time. While the longitudinal magnetisation is increasing during relaxation, the transverse magnetisation (perpendicular




**(a)** excitation **(b)** relaxation **(c)** T1 and T2 time

**Figure 2.14:** Illustration of the RF excitation and relaxation processes, the related longitudinal ($M_Z$) and transverse ($M_{XY}$) magnetisation and the definition of T1 and T2 relaxation times. Adapted from [110].

to the static magnetic field) is decreasing. During the excitation phase, some nuclei will be phase-locked, meaning that their perpendicular component is rotating with the same phase, creating a net magnetisation in the transverse plane. This phase-locking is however quickly lost during relaxation. The time to fall to 37% ($= 1/e$) of the maximum transverse magnetisation is called T2, spin-spin or transverse relaxation time. The previously mentioned T2* relaxation time is related to T2, but whereas T2 is the "true" transversal relaxation time, the shorter T2* is the "effective" or observed transversal relaxation time:

$$\frac{1}{\mathrm{T2^*}} = \frac{1}{\mathrm{T2}} + \frac{1}{\mathrm{T}_i} \,,$$

where $\mathrm{T}_i$ is a term related to inhomogeneities in the main magnetic field, caused by local distortions, small magnet imperfections, chemical shift or magnetic susceptibility.

The relaxation times T1 and T2 depend on the local environment of the nucleus. Typical values at 1.5 T are given in table 2.2.

As a result, different tissues will show a different value on an MRI scan. The precise contrast is determined by two parameters: the time between the excitation pulse and the acquisition of the signal in the receiver coils, called echo time (TE), and the time between two excitation pulses, the repetition time (TR). The latter determines the remaining

**Table 2.2:** Approximate T1 and T2 relaxation times at 1.5 T.

| | **T1 (ms)** | **T2 (ms)** |
|---|---|---|
| grey matter | 920 | 100 |
| white matter | 780 | 90 |
| CSF | 4200 | 2100 |
| fat | 240 | 70 |

magnetisation before applying a new excitation pulse. Short TR and TE will result in T1-weighted images, long TR and short TE in proton density (PD) weighted images, and long TR and TE in T2-weighted images.

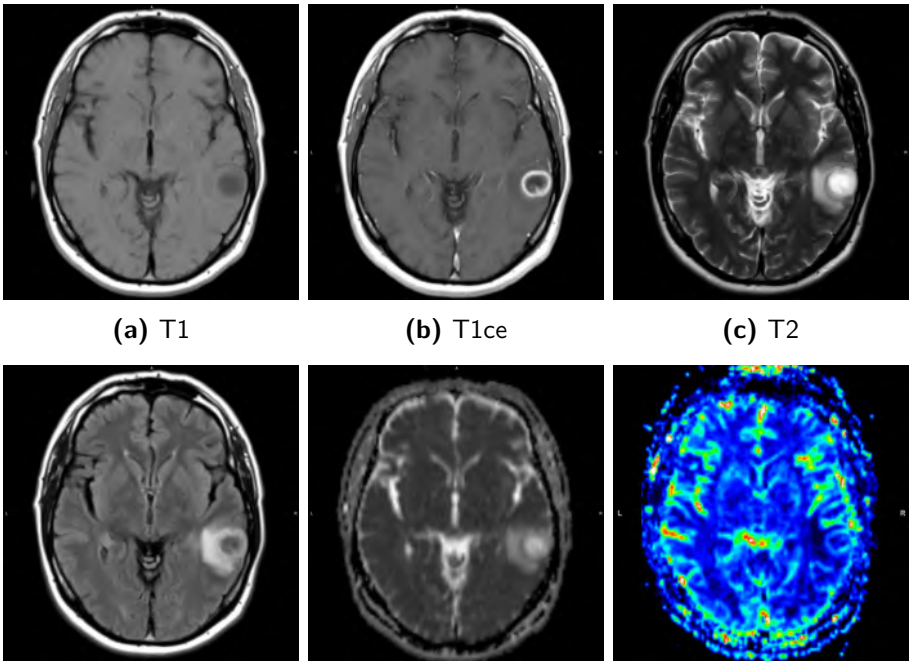
Localisation of the received signal is possible by using small gradients in the magnetic field, induced by the gradient coils. By applying small variations in the $x$, $y$ and $z$ direction within the static magnetic field, the exact position of a voxel can be encoded in the k-space or Fourier domain of the acquired signal. The 3D image can now be obtained by reconstruction of this signal.

In the last decades, hundreds of different MRI protocols have been proposed with different scanning parameters, each time yielding a unique image contrast or highlighting a different biological tissue. Examples are fluid suppression, fat suppression, highlighting blood vessels in magnetic resonance angiography (MRA) or blood perfusion in arterial spin labelling (ASL). Quantitatively comparing images obtained with different acquisition parameters is very difficult, and structural imaging with MRI therefore mainly remains a qualitative rather than a quantitative tool. In the next section, we will discuss the most important sequences in neuro-oncology.

## Applications in brain tumour imaging

MRI is the principle method of choice when diagnosing the presence of a brain tumour. The EANO recommends four different structural images, illustrated in figures 2.15a–2.15d.

In this example several scans obtained during the same imaging protocol are shown of a patient with a glioblastoma, IDH-wildtype. In figure 2.15a a T1-weighted scan is shown. This imaging type shows



**(a)** T1 **(b)** T1ce **(c)** T2

**(d)** FLAIR **(e)** ADC **(f)** rCBF

**Figure 2.15:** MRI scans of a patient with a glioblastoma.

in general a very high resolution and good contrast between the different anatomical structures. The tumour is slightly hypo-intense. In figure 2.15b a contrast-enhanced T1 (T1ce) scan is shown. The same scanning parameters as a T1-scan are used, but the image is obtained after administration of a gadolinium-based contrast agent. This shows a bright signal in the blood vessels, as well as in regions with a disrupted BBB. Typical for GBM is the ring-like contrast enhancement with a necrotic core. The T2-weighted scan of figure 2.15c shows a high intensity in regions containing a lot of water, such as the ventricles and sulci containing CSF. The necrotic tumour-core is highlighted as well. Compared to the T1-weighted scans, T2 shows better the peritumoural oedema, caused by fluid-leakage and tumour invasion into healthy tissue. The fluid-attenuation inversion recovery (FLAIR) scan of figure 2.15d is a T2-weighted scan as well, but the signal of CSF is removed, thereby providing excellent contrast between healthy and pathological

fluids. This sequence is also used for detection of white-matter abnormalities such as MS-plaques.

Next to the four anatomical MRI techniques, figure 2.15 also shows two more advanced scans providing biological information. Figure 2.15e maps the apparent diffusion coefficient (ADC) obtained with diffusion-weighted imaging (DWI). We observe a high degree of water diffusion in the necrotic and oedematous regions, whereas in the heterogeneous contrast-enhancing tissue, diffusion is obstructed. Diffusion MRI is frequently used for brain tumour diagnosis [111] as well as for the early assessment of treatment response [112, 113]. Lastly, a regional cerebral blood flow (rCBF) map is given in figure 2.15f. In this scan, we see that the contrast-enhancing region shows an increased blood flow, evidence for angiogenesis typical for a malignant tumour. Usually, rCBF-maps are obtained using dynamic susceptibility contrast (DSC) imaging which requires exogenous contrast agents, although several studies show the use of ASL, an imaging technique without the need for contrast agents, with similar results [114, 115, 116].

### 2.3.3 Positron Emission Tomography

Next to MRI, positron emission tomography (PET) is used in neuro-oncology to detect metastases, to define metabolic hotspots for biopsy, for post-surgical evaluation and radiotherapy planning. PET is a functional nuclear imaging technique. We will first briefly introduce the principle, illustrated in figure 2.16.

#### Principle

In contrast to an MRI-scanner, a PET-scanner is relatively simple. The most important part is a cylinder with a diameter of approximately 80 cm and a width of about 20–25 cm, consisting of detectors able to detect high-energy photons. A patient coming for a PET-scan is first injected with a radioactive tracer. This tracer consists of a specific biological molecule (e.g. glucose), to which a radioactive atom, most often $^{11}$C, $^{18}$F or $^{68}$Ga, is attached. Depending on the nature of the molecule, the radiotracer will be taken up in specific tissue.

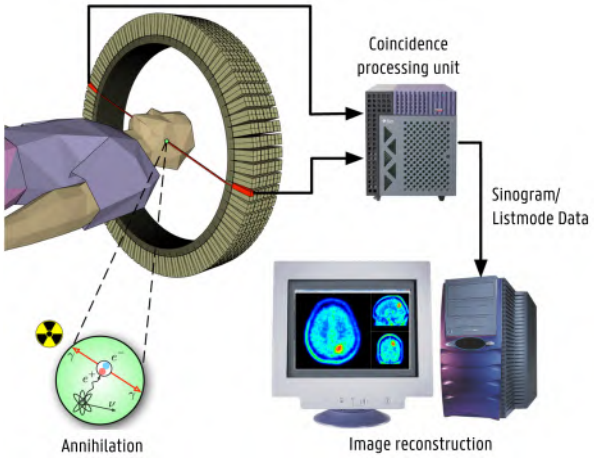



**Figure 2.16:** The principle and workflow of positron emission tomography. Adapted from [117].

When the unstable nucleus of the radioactive molecule decays, it emits a positron. This particle travels typically 1–2 mm through the surrounding tissue, until it annihilates with an electron. This annihilation process creates two photons with an energy of 511 keV that are emitted in opposing directions. Modern PET scanners work in listmode: they save a list of all the detected photons. When two events occur in coincidence (e.g. less than a nanosecond apart), they are assumed to come from the same decay. The position of this decay can now be traced back to the line of response (LOR) between the two detector positions. In this way, the tracer distribution can be reconstructed.

Image reconstruction can happen in two ways. A first option is a static image acquisition, creating a map of all the decays that occurred during a fixed period of time. For example, the radiotracer is first administered intravenously to the patient, who is brought to the PET-scanner 45 minutes later. He is then scanned for 15 minutes to obtain an image of the average distribution of the molecule. The second option is called dynamic scanning. In this case, the tracer is administered to the patient while already positioned in the scanner. The image acquisition starts from the moment of injection and continues for 40–60 minutes. Afterwards, the scan is reconstructed in multiple short time frames. In this

way, the tracer kinetics can be analysed with time-activity curves.

Most modern PET-scanners are hybrid PET/CT scanners. The CT-component serves two purposes. Firstly, the excellent resolution offers anatomical details which help to localise the PET signal. Furthermore, the CT-image can be used during the PET reconstruction algorithm, as it gives an estimate of the attenuation of the emitted photons. More recently, hybrid PET/MRI scanners are being installed. As MRI gives little to no information on bone structures, different techniques have to be used for attenuation correction [118, 119].

## The radiotracer principle

The radiotracer principle was invented in 1913 by George de Hevesy, for which he received a Nobel Prize in 1943 [120]. The use of radioactive tracers must fulfil two properties [121]. First of all, the unaltered molecule and the radioactive version should be undistinguishable for the body, such that they follow the same physiological processes. The concentration of radioactive molecules should also be low enough (typically nanomolar amounts) not to alter the normal biology. Secondly, it should be feasible to detect the radiation emitted by the radiotracer and thereby report on the properties of the system. Since PET has a very high sensitivity, it is ideally suited to image metabolic processes *in vivo*.

## Radiotracers in neuro-oncology

Many different radiotracers can be used to image brain tumours. In the discussion below, we limit ourselves to the three main molecules, and briefly mention some of the other options.

2-DEOXY-2-($^{18}$F)FLUORO-D-GLUCOSE ($^{18}$F-FDG) is a glucose analog where one hydroxyl group is replaced with a radioactive $^{18}$F atom. Since malignant tumours show an increased glucose metabolism, $^{18}$F-FDG is the most commonly used radiotracer in oncology. In the healthy brain, $^{18}$F-FDG shows a high uptake in grey matter, showing the metabolic demands of neurons and glial cells. It therefore does not show specific uptake in brain tumour tissue, and interpreting an $^{18}$F-FDG PET scan

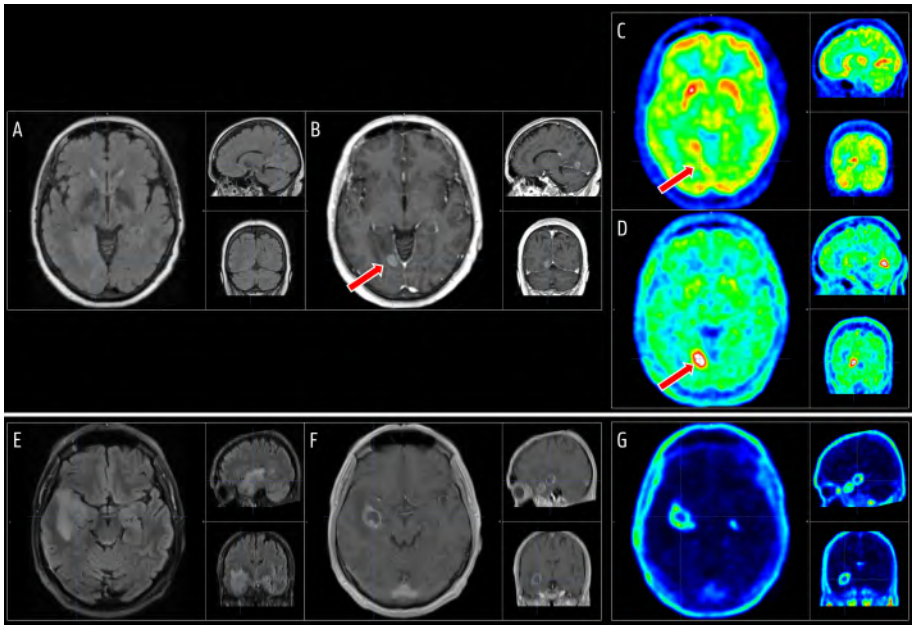



**Figure 2.17:** Representative images of PET scans with different radiotracers. Parts A-B-C-D belong to one patient, parts E-F-G to another. Both patients are diagnosed with a glioblastoma, IDH-wildtype. A&E: FLAIR MRI; B&F: T1ce MRI; C: $^{18}$F-FDG PET, acquired approximately 3 weeks before MRI; D: $^{18}$F-FET PET, acquired approximately 3 weeks after MRI; G: $^{18}$F-cho PET, acquired 1 week after MRI.

without knowledge of the tumour location, e.g. using an MRI scan, is very difficult (see figure 2.17c).

*O*-(2-[$^{18}$F]FLUOROETHYL)-L-TYROSINE ($^{18}$F-FET) is a radiolabeled amino acid tracer and the most commonly used PET tracer in brain tumours. Malignant tumours use amino acids for energy, protein synthesis and cell division [122]. A clear advantage of $^{18}$F-FET is that healthy brain tissue shows a low uptake, whereas the vast majority of high-grade glioma show an increased $^{18}$F-FET uptake (see figure 2.17d). The sensitivity to detect these tumours is therefore very high. The EANO recommends using $^{18}$F-FET over $^{18}$F-FDG when assessing patients with a newly diagnosed brain tumour [123]. Moreover, the kinetic behaviour of this tracer offers additional information, and it can therefore be used

to differentiate between recurrent or progressive glioma from treatment-related nonneoplastic changes with higher accuracy than conventional MRI [124, 125, 126].

$^{18}$F-FLUOROMETHYL-CHOLINE ($^{18}$F-CHO) is a radiolabeled version of choline, an important molecule involved in cell membrane synthesis. In cancer, there is an increase in the cellular transport and phosphorylation of choline, resulting in an increased uptake of $^{18}$F-cho. This tracer therefore offers an excellent discrimination between tumour and healthy tissue (see figure 2.17g). However, there are also some drawbacks, such as the very fast kinetic behaviour, and the fact that the uptake of $^{18}$F-cho is influenced by BBB damage and inflammation [127].

OTHERS Apart from the previously mentioned radiotracers, others are being used in neuro-oncology as well. A detailed discussion of these tracers is beyond the scope of this thesis, but we briefly mention some of them here. $^{11}$C-methionine is an amino-acid tracer, yielding similar images as $^{18}$F-FET [128]. However, due to the longer half-life (109 minutes for $^{18}$F versus 20 minutes for $^{11}$C), $^{18}$F-FET is often preferred. $^{18}$F-fluoro-L-phenylalanine ($^{18}$F-DOPA) is also an amino-acid tracer, but in contrast to $^{18}$F-FET, it shows a high uptake in the healthy striatum [129]. Hypoxia, a shortage of oxygen in tissue, is a typical feature of aggressive tumours. There exist several PET hypoxia radiotracers, such as $^{18}$F-fluoromisonidazole ($^{18}$F-MISO) or $^{18}$F-fluoroazomycin arabinoside ($^{18}$F-FAZA). $^{18}$F-fluorothymidine ($^{18}$F-FLT) assesses DNA-synthesis and therefore tumour proliferation. Lastly, the 18-kDa mitochondrial translocator protein (TSPO) is overexpressed in glioma. Specific TSPO-radiotracers are therefore being proposed, such as $^{18}$F-GE-180 [130], but the discrimination between tumour and inflammatory processes using this radiotracer is to be examined.

## 2.4 Radiomics

Technological improvements in the last decades have optimised the quality of medical images for visual inspection. However, as stated by Gillies

et al., "images are more than pictures, they are data" [131]. Consequently, there might be a lot of information in scans that is difficult to assess with the bare eye. Examples are detecting changes in tumour volume and structure on subsequent scans of the same patient, or analysing a large database of scans of different patients, and looking for correlations between imaging findings and tumour phenotype. The recently developed technique *radiomics* can offer help for these applications. Possible definitions are given below.

**Radiomics**

- high-throughput extraction of quantitative imaging features with the intent of creating mineable databases from radiological images [132]
- the automated quantification of the radiographic phenotype [133].

In other words: by transferring qualitative 2D or 3D images into large vectors of quantitative features, the scans can be analysed by dedicated computer algorithms, which makes it possible to assess large databases. By doing this, hidden correlations between abstract features and tumour biology have been uncovered, paving the way for precision medicine. This means that by examining large groups, patients can be divided into clusters with similar quantitative image feature profiles, often called the *radiomics signature*. Ideally, these clusters correlate well with one or more biological outcome parameters, such as tumour type, treatment response, or survival. Recently, the associations between image-based phenotype data with genomic patterns are being investigated as well. This will give insight in how biological processes are reflected in the image, a technique called radiogenomics [134, 135].

All malignant tumours exhibit intra-tumour heterogeneity, caused by variations in cellularity, angiogenesis, extracellular matrix and necrosis. High intra-tumour heterogeneity plays an important role in tumour behaviour, treatment response and drug resistance [136]. Because of limited spatial resolution, none of the available imaging modalities is close to displaying microscopic substructures in tumour tissue. However, the underlying hypothesis of radiomics is that microscopical or even genetic

heterogeneity is translated into macroscopic heterogeneity assessed on medical images [137]. Although not necessarily visually observable, this might be discovered by calculating complex texture features. In this way, medical imaging makes it possible to quantify the heterogeneity of the entire tumour as a whole. In contrast, histological and genetic analysis is usually performed on very small tumour samples, such that small zones of distinctly different tissue might be missed.

As imaging is non-invasive, it can be often repeated for treatment monitoring. On every follow-up scan, changes in the tumour phenotype can be quantified. In this way, treatment response can be assessed early, and with minimal cost and burden to the patient. In conclusion, radiomics can offer a non-invasive and presurgical tool towards an automated diagnosis, prediction of therapy outcome and early treatment response monitoring.

### 2.4.1 Principle

The workflow of radiomics consists of four steps, illustrated in figure 2.18. These steps will now be further discussed.

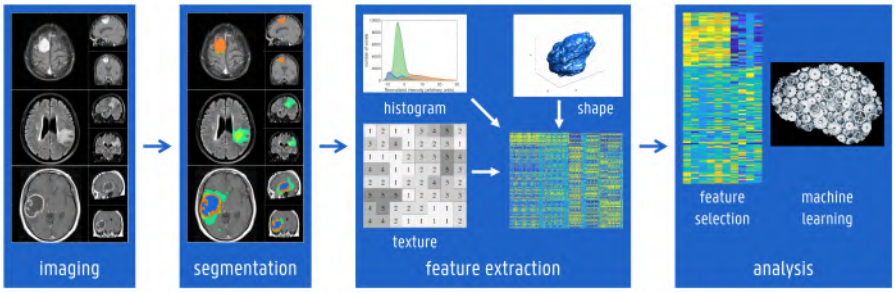



**Figure 2.18:** The workflow of radiomics.

#### Imaging

The first step of radiomics is the acquisition of high-quality medical images. Depending on the application, different modalities such as CT, MRI and PET can be taken into account. Standard clinical imaging protocols are preferred, as this allows to obtain large datasets. Conse-

quently, many hospitals already own a wealth of data that can be used for radiomics analyses.

### Segmentation

In the second step, the macroscopic tumour volume is defined on the images. This can be done manually by an experienced radiologist, or using a (semi-)automatic segmentation algorithm. Also different tumour masks per patient might be used, e.g. to analyse different tissues within the same tumour. This will be the topic of chapter 4.

### Feature extraction

In the third step, the previously defined tumour masks are used to extract many (typically more than 200) quantitative features. These features involve descriptors of the intensity distribution (by analysing the histogram), of the heterogeneity and spatial relationships between intensity levels (texture analysis) or of the tumour volume and shape. This list can even further be complemented by including spatial frequency patterns discovered with wavelet analysis or by looking for relations between the tumour and the surrounding tissues. Feature extraction algorithms will be discussed in chapter 3.

### Analysis

The last step of the radiomics workflow is the analysis of the extracted features. This mostly starts with removing the noisy or redundant features. Usually, the goal is to select a minimal dataset of highly predictive features. This will of course depend on the specific task one seeks to perform. Examples are: classifying tumours into different classes, predicting therapy respons and survival, or early treatment response monitoring. Several machine learning algorithms can be used for feature analysis, as will be discussed in section 2.5.

## 2.4.2 Challenges

There are several issues that should be taken into account when performing a radiomics study. We list some challenges involved with every

step below.

### Imaging

Many image properties influence the extracted features. It is therefore very important to obtain standardised imaging protocols when comparing different scans. However, different vendors have optimised their own scanning parameters, which results in different image qualities. For all imaging modalities, properties such as slice thickness, resolution, contrast and noise can vary significantly. For CT and PET, the intensity is expressed in normalised units, being Houndsfield unit (HU) and standardised uptake value (SUV) respectively. Even then, the observed value might depend on the reconstruction algorithm [132]. For MRI, the image intensities strongly depend on a complex interplay of tissue properties and acquisition parameters. Conventional MRI image signal intensities are therefore very hard to interpret in a quantitative manner.

When performing a radiomics study, standardised acquisition protocols (reconstruction, resolution, acquisition parameters) should be preferred. When data from different sources are used, it is good practice to normalise the images both spatially (common voxel size) and intensity-wise. In this thesis, we will apply the white-stripe normalisation method for MRI, where intensities are normalised to the normal-appearing white matter (NAWM) [138]. Furthermore, one can expect that large numbers of images may be able to overcome some of the heterogeneities inherently present in clinical imaging.

### Segmentation

Since the extracted features are based on the tumour masks, the segmentation step is crucial in the radiomics process. Depending on the specific task, both manual or (semi-)automatic approaches are possible. However, the goal is to find accurate and reproducible tumour boundaries in a time-efficient way. Manual segmentation performed by experienced readers suffers from high inter- and intra-reader variability and is very time- and labour-intensive. For large databases, this approach might therefore be infeasible.

(Semi-)automatic segmentation algorithms should be carefully tested for their accuracy, reproducibility and consistency. To this end, they are often compared to manual performance, which is variable, as mentioned before. Obtaining a gold-standard segmentation mask can therefore only be obtained with "virtual" tumours, where scans of healthy subjects are manipulated to mimic tumour structures. These techniques however often underestimate realistic tumour complexity, leading to over-optimistic segmentation performances [139]. Assessing accuracy is therefore difficult, and reproducibility and consistency might be more important properties of a good segmentation algorithm. To accomplish this, user-interference should be minimised.

### Feature extraction

The third step consists of extracting a large number of features. Most researchers limit themselves to features based on histogram, texture or shape analysis. This list could in theory be expanded unlimitedly. One should however keep in mind that models based on features with an intuitive meaning are easier to explain, and therefore more accepted in a clinical setting. When different segmentation masks are available, robustness of the features should also be tested for. For the sake of reproducibility, the exact definition of all features should be given, since multiple names are being given to the same feature, and conversely, features with the same name are sometimes calculated in a different way.

### Feature analysis

In many cases, the number of extracted features is higher than the amount of samples in a particular study. This increases the probability of overfitting the data, meaning that models are less able to distinguish the important relations in the data from the noise. Moreover, many features will inherently contain the same data and are consequently redundant. Dimensionality reduction and selection of task-specific features are therefore important requirements before modelling the relations with the biological parameters. Machine learning models should always be thoroughly validated, as will be discussed later on.

### 2.4.3 Applications

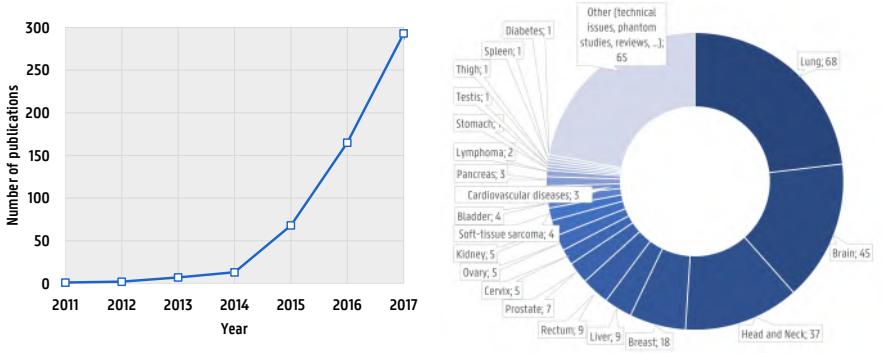
The main application of radiomics lies in oncology, although other areas such as pneumonitis [140] or neurological disorders [141, 142, 143] are possible as well. In 2014, Hugo Aerts and colleagues published the first study using the radiomics workflow in Nature Communications [144]. The results were groundbreaking, in the sense that they built a model based on the analysis of CT scans from non-small cell lung cancer (NSCLC) patients, and applied this model successfully to head and neck (H&N) tumour patients. In total, 1019 subjects were included, and for every patient 440 features were calculated. In a first step, unsupervised clustering (see section 2.5.3) was applied on the raw features of the training set, consisting of 422 patients with NSCLC. This clustering was associated with primary stage, overall stage and histology.

In the second part, test/retest and multiple delineation was used to identify robust features in a subset of 31 and 21 patients, respectively. Next, the previously mentioned training set was used to select four features as radiomics signature. These include the best features from histogram, texture, shape and wavelet analysis. The weights of the signature were optimised in a Cox proportional hazards model for survival. Based on this signature, two distinct survival groups were identified in the independent NSCLC group (225 patients) and, more surprisingly, in the 231 H&N patients. Furthermore, significant associations between signature features and gene-expression patterns were found in a group of 89 H&N patients.

Ever since, the number of radiomics-based publications has skyrocketed, as illustrated in figure 2.19a. The technique is being used in a broad range of pathologies, see figure 2.19b.

### 2.4.4 Radiomics in primary brain tumours

Techniques from radiomics are being used for a number of different tasks in neuro-oncology. The discussion below gives examples of different applications, based on the review paper from Zhou et al. [145]. A more detailed review on the current state-of-the-art in literature will be given in the following chapters.




**(a)** Number of publications on Web of Science with topic "radiomics"

**(b)** Distribution of topics on radiomics from articles published in 2017

**Figure 2.19:** Literature study on Web of Science, articles on radiomics.

## Glioma subtype classification

A first application of radiomics in neuro-oncology is the automated classification of brain tumours according to their grade and/or type. This is mostly done in a binary fashion, where a computer model tries to distinguish between two classes, e.g. low-grade versus high-grade glioma, or metastases versus glioblastoma. However, PBTs can be subdivided into more than hundred different classes, meaning that automated classification is inherently a multiclass problem. This can be solved in several ways: either by combining binary classifiers, or by building a multiclass predictor able to give the probabilities for all classes that are taken into account. Since enough samples per class are needed, most studies are limited to a small number (typically less than 10) of classes. In chapter 3, the automated discrimination between lower-grade glioma and glioblastoma, a binary problem, will be discussed. In chapters 5 and 6, the multiclass problem will be tackled.

## Therapy prediction and survival estimation

Medical scans might contain information that can predict therapy response and survival. An example is a study from 2017 by Zhou et al. [146], where imaging findings could distinguish between glioblastoma pa-

tients with short-term and long-term survival chances. Similarly, several studies were performed to predict the treatment response to antiangiogenic therapy with bevacizumab in recurrent glioblastoma [147, 148].

### Prediction of molecular markers

As the WHO 2016 classification scheme of CNS tumours is both based on histological and molecular parameters, non-invasive determination of these genetic and molecular features could give an early diagnosis of the tumour type. Examples are determining the 1p/19q codeletion status [149, 150, 151], IDH-status [152, 151, 153] or MGMT promotor methylation in glioblastoma [154, 155].

### Discriminating radiation necrosis from tumour recurrence

A major difficulty in neuro-oncology is the differentiation between radiation-induced necrosis from true recurrence of the tumour. Traditional visual inspection of the contrast enhancement on MRI is in most cases not sufficient to discriminate between these phenomena, but quantitative radiomics features have shown to be helpful [156, 157, 158].

## 2.5 Machine learning

The last step of radiomics consists of modelling the extracted features in a model able to detect relations in the data or predict biological parameters for unseen patients, a task where machine learning (ML) is perfectly suited for. ML is a set of statistical techniques designed to learn from data without the underlying relations being explicitly programmed, and to use the newly gained knowledge to make predictions about unseen data.

There are several types of machine learning. In supervised learning, for all samples in the dataset both the input features and the output labels are known. The goal is then to find relations between the data and the labels. In contrast, in unsupervised learning the output labels are unknown or hidden to the computer. In this case, the task is to find

structure or clusters in the data. A variation on these methods is semi-supervised learning, where only a (small) part of the samples is labelled. Finally, reinforcement learning is based on trial-and-error. In this way, a computer can for example be taught how to win a game without being instructed on the best strategies.

Furthermore, all ML problems can be divided into two categories: regression or classification. In regression problems, the output label is a continuous variable (e.g. height in centimetres, percentage decrease of tumour volume, ...), while in classification tasks, the output consists of a limited and discrete number of classes (e.g. benign versus malignant, mutated or wildtype, tumour grade I-IV, ...).

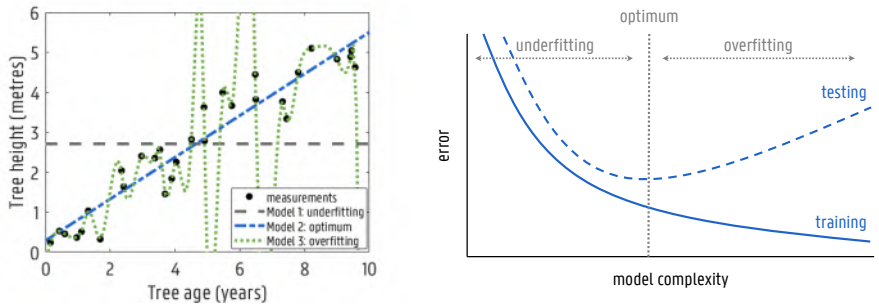
In the next section, we will briefly explain the important concepts of generalisation. This will be followed by a discussion on some of the techniques used in this thesis.

### 2.5.1 Principle of generalisation

Consider the following example: assume we planted 30 trees in a period of ten years. Now, we measure the tree height, plotted in figure 2.20a. This information can be used to predict how tall a new tree will become. The easiest model we can think of, is assuming that all trees have the same height, e.g. the average of the measurements (model 1 in figure 2.20a). This model clearly does not explain all the observed variation in the data, and we say that the *bias* of the model is too high, also called *underfitting*. In a second attempt, we assume a linear dependency between the height and age of the tree (model 2 in figure 2.20a), a so-called linear regression problem:

$$\text{height} = \alpha + \beta \times \text{age} \, .$$

The parameters $\alpha$ and $\beta$ can be estimated from the data. The most popular way to do this is by minimising the summed square of the residuals (i.e. the difference between the measurements and the predicted




**(a)** Example dataset with three different models

**(b)** Error obtained on the training and testing set as function the model complexity

**Figure 2.20:** Machine learning: example and concept of generalisation.

values):

$$\hat{\alpha} = \overline{\text{height}} - \hat{\beta} \times \overline{\text{age}}$$
$$\hat{\beta} = \frac{\sum_{i=1}^{30}(\text{age}_i - \overline{\text{age}})(\text{height}_i - \overline{\text{height}})}{\sum_{i=1}^{30}(\text{age}_i - \overline{\text{age}})^2}$$

with $\hat{\alpha}$ and $\hat{\beta}$ the estimated parameters of the model, and $\overline{\text{age}}$ and $\overline{\text{height}}$ the average age and height. We clearly see that linear regression achieves a good prediction of the measurements, and we say that this model generalises well for this problem. Therefore, *generalisation* can be described as the ability to correctly predict the output values of new, unseen data.

However, we could think of more complex than linear functions to model the influence of tree age. Moreover, other parameters could determine the height as well, e.g. type of soil, amount of water, amount of light, ... . These parameters can be taken into account in the model to achieve even better prediction accuracies. When adding enough features to the model, it will be possible to explain all variation in the training set and in this way reduce the residual error to zero (model 3 in figure 2.20a). But assume that in our limited dataset trees that were planted on a Tuesday incidentally happen to be a bit taller than trees planted on a different day. Incorporating the day of the week into the model will therefore result in a reduced training error. However, we do not expect

that this trend will be true in general. The higher level of complexity in the model will therefore lead to an increased error when predicting an unseen sample. We say that the *variance* is too high, also called *overfitting*.

For every ML-based problem, it is important to estimate the optimal model complexity (bias-variance tradeoff), illustrated in figure 2.20b. Since we can only use the training set to optimise the model, it is difficult to gauge the generalisation capacity to unseen data. A technique to solve this issue, is called cross-validation. For example, in five-fold cross-validation the data is distributed into five equal parts. Four parts are used for training the model, which is validated on the remaining dataset. This experiment is repeated five times as each part is left out once. The average error over the validation sets is then obtained and used to optimise the model parameters. After training the optimal model, its performance should be validated on an independent test set that was not involved in the training process.

In the remainder of this chapter, we will introduce some machine learning techniques that will be used in the following studies. We will focus on classification problems, starting with some supervised learning methods.

### 2.5.2 Supervised learning

As mentioned before, in supervised learning we have obtained a training set with known output labels. Many different algorithms exist to model the relations between input features and output labels, but we limit ourselves here to three techniques: support vector machine (SVM), random forests (RF) and artificial neural network (ANN). We will also briefly discuss a special class of ANNs, namely convolutional neural networks (CNNs), a type of *deep learning*.

#### Support vector machine

Suppose we have two classes that are linearly separable. This means that we can find a hyperplane such that all points of the first class lie on one side of the hyperplane, and all points of the second class

on the other side. The goal of a support vector machine (SVM) is to find the hyperplane that maximises the distance (or *margin*) between said hyperplane and the points that lie closest to it, called the support vectors. This is illustrated in figure 2.21. When evaluating an unseen sample, we only need to calculate on which side of the hyperplane it resides.

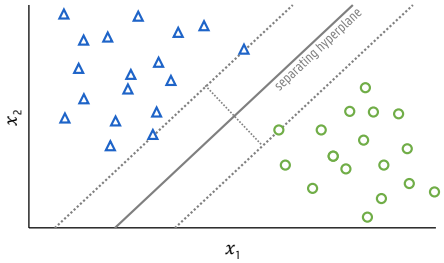



**Figure 2.21:** Example of a support vector machine.

When the data are not separable, it is impossible to find a *hard* margin as before, and the goal is therefore to find a *soft* margin between the two classes. In this case, SVM will look for a hyperplane that separates many, but not all data points. Moreover, some classes are not linearly separable, but can be distinguished when using another separation criterion. This is the motivation behind using the *kernel* method. In this case, the data points are mapped into a feature space using a nonlinear function, followed by the previously explained linear algorithm. Popular kernels are multinomial functions, or the Gaussian kernel (radial basis function).

SVMs are designed as binary classifiers, discriminating between two classes. They can however be used for multi-class problems as well, by combining several one-vs-all models. The computational cost of an SVM rises quadratically with the number of training samples. Therefore, it might not be the ideal technique for very large datasets.

### Random forests

Random forests (RF) is a bagging (bootstrap aggregating) technique. This means that an ensemble of weak learners, in this case decision

trees, is trained, where every learner sees only a subset of the available training data. In this way, noisy and unbiased models are averaged to decrease the variance of the final model. The method is invented by Breiman [159] in 2001, and is illustrated in figure 2.22.

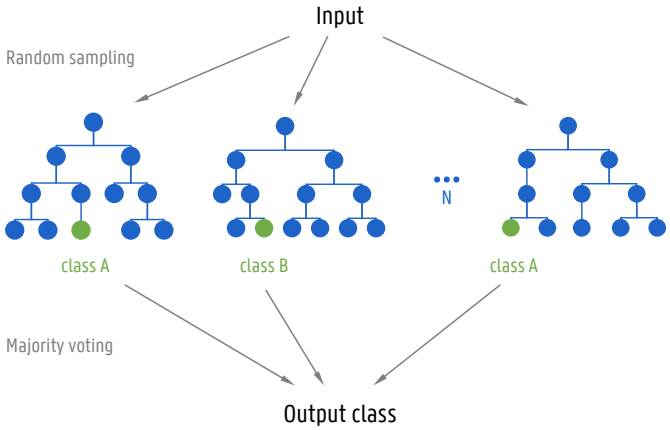



**Figure 2.22:** Principle of random forests.

First, a random subset of samples is selected to train a decision tree. This is a structure with a series of *nodes*, starting with the *root node* which receives the chosen subset of training samples. In a node, the data are split up in two parts according to a certain test, e.g. if $x_2 < 5$ follow the left *branch*, else follow the right branch. Which question is asked at a specific node is determined by the *information gain*, meaning that all possible decisions are tried, and the decision yielding the best discrimination between different classes is chosen. To increase the randomness, only a random subset of features is considered at every node. In this way, the entire tree is built until no further splits are possible. This last node becomes a *leaf* with a corresponding output class. Several stopping criteria can be applied as well to abort growing of the tree. Examples are a maximum number of layers, or similarly a minimal number of samples per leaf. A fully-grown tree can also be *pruned*, meaning that irrelevant branches can be cut off. This process is repeated to build a high number $N$ (typically more than one hundred up to several thousands) of classification trees, forming the forest.

When predicting an unseen sample, every tree in the forest is evalu-

ated from root to leaf, yielding $N$ corresponding classes. Majority voting is then applied to decide on the final output class. Moreover, the output of the $N$ trees can also be used to estimate the probabilities of the output classes.

There are a number of advantages related to random forests. First of all, in contrast to SVM, which is inherently designed for binary problems, random forests handle multi-class datasets very well. Furthermore, because many decision trees that are trained on subsets of the data are combined, random forests do not have the tendency to overfit, even when using many features. Random forests can also be visually inspected to learn on which properties a certain decision is based. This is of course harder when building large and complex forests. Lastly, random forests can be used to estimate the predictor importance. This will be explained in section 3.4.1.

### Artficial neural networks

At the core of many applications in AI is an artificial neural network (ANN), illustrated in figure 2.23. The principle is based on the human neural network, where neurons are communicating with each other by transmitting signals over the axons and synapses. Similarly, an ANN consists of a set of connected nodes called neurons that are organised in layers. Starting from the input layer, every neuron receives signals from neurons in the previous layer. These signals are weighted differently and added up, before being processed in a non-linear way using an activation function. The resulting value is then forwarded to the next neuron. During the training phase, the weights are optimised using back-propagation. Different layers typically perform different tasks. In the last layer, the output is given, such as the corresponding class label in classification tasks, or an estimate of the outcome in regression problems.

Since the use of ANNs is limited in this thesis, we will not go further into detail in the underlying mechanisms. However, a very interesting application of ANNs is discussed in the next section.

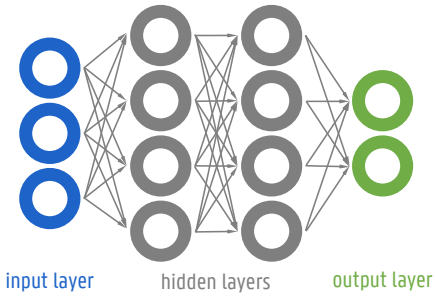



**Figure 2.23:** Principle of artificial neural networks. In this example, the fully-connected network consists of three input nodes, two hidden layers each comprising four nodes, and a output layer with two nodes.

## Deep learning

When the number of hidden layers in a neural network becomes very large, we speak of a deep neural network or deep learning. This has many applications, such as natural language processing, but it is most widely used in computer vision, where the machine is trying to understand the content of digital images. A convolutional neural network (CNN) is highly suitable for this task. The input of this network is not comprised of individual features, but of a 2D or 3D image. This image is passed through a series of layers, mostly starting with a convolutional layer (hence the name). Here, a set of filters or convolutions are applied to the image. Next, pooling layers reduce the dimensions of the image and fully connected layers apply non-linear operations on the input from previous layers, like normal ANNs. In this way, different layers will correspond to different levels of abstraction of the images: starting with the original image, the first layers will for example focus on edges, that are combined in following layers into different shapes (e.g. noses, ears or tails) resulting in object recognition (e.g. cats or dogs) in the last layers. An important difference with the traditional ANN is that features do not need to be hand-engineered, as the network learns the optimal features for a specific task.

An example of a CNN is given in figure 2.24. This is the famous AlexNet architecture by Alex Krizhevsky et al. [160]. This network won the ImageNet Large Scale Visual Recognition Challenge in 2012 with a

margin exceeding 10% over its competitors. The goal of this challenge is to classify a large number of natural images. Currently, the datasets consists of over 14 million pictures, labelled into more than 21000 classes. In 2015, teams from both Microsoft [161] and Google [162] outperformed the human error rate of 5.1% using a CNN. Therefore, the applications of CNNs in medical imaging has gained a lot of attention, both for classification and segmentation tasks. However, because deep neural networks are in general very complex with many weights to optimise, a very large dataset (typically several tens of thousands) of labelled images needs to be available during the training phase. This also requires specific computational hardware such as powerful graphics processing units (GPUs).

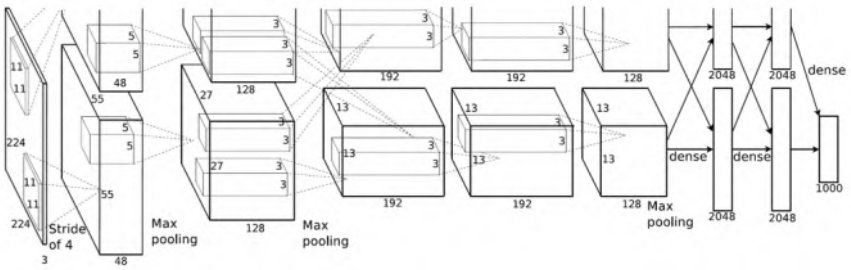



**Figure 2.24:** Illustration of the famous AlexNet CNN architecture [160].

## 2.5.3 Unsupervised learning

Labelling a large dataset is often a very expensive task. When sample labels are not available, we can still look for certain structures in the data using unsupervised methods. These are mainly clustering techniques, trying to separate the data into different classes. The main difficulties are that the number of classes is in general not known, and that a sample can belong to more than one class. Two clustering techniques will be discussed in the last section of this introduction, starting from the following example. Consider the dataset of figure 2.25, where two features ($x_1$ and $x_2$) are measured for 1080 samples. By observing the data, or from background knowledge about the experiment, we assume

that there are three distinct clusters.

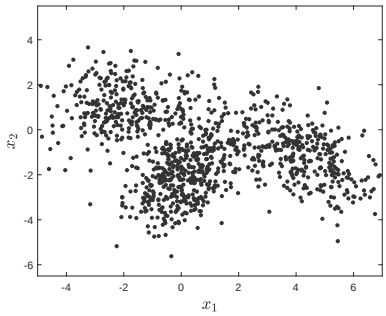


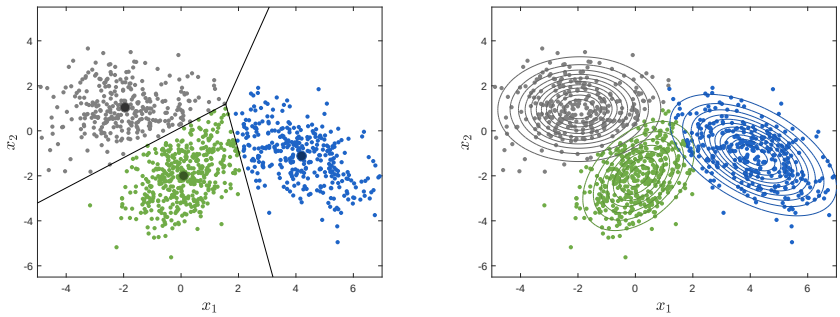

**Figure 2.25:** Example dataset for unsupervised learning. Two values $x_1$ and $x_2$ are measured for 1080 samples. We assume that the data can be divided into three clusters.

### $k$-means clustering

The $k$-means algorithm was proposed by MacQueen et al. [163] and gives an easy solution to dividing data into clusters with similar properties. It is an iterative algorithm, starting with an initial guess of cluster centres (e.g. random seeds). Next, all points are assigned to the cluster with the nearest centre, after which the centre is updated to the centroid of the newly chosen cluster points. This process is repeated until convergence. The result on the example dataset is displayed in figure 2.26a.

### Gaussian mixture model

When fitting a Gaussian mixture model (GMM) to a dataset, we assume that every sample belongs to one of $k$ clusters and every cluster is defined by a multivariate normal distribution with a specific mean and covariance matrix. The sample label, as well as the cluster parameters are unknown or *latent*. Methods such as the *Expectation Maximisation* algorithm can be used to estimate the cluster parameters. Here, we initialise the clusters by assigning a class to every point, either randomly or using an easy algorithm such as $k$-means. Next, we calculate the cluster mean and covariance matrix and determine for every point the probability to belong to all classes given the current estimates of the




**(a)** $k$-means **(b)** Gaussian mixture model

**Figure 2.26:** Results obtained on example dataset with two different unsupervised learning techniques.

class parameters (the $E$-step). Finally, we assign the class with maximal probability to the data (the $M$-step). This process is repeated until convergence. The resulting clustering on the example data is given in figure 2.26b.

The unsupervised learning techniques are here illustrated in two dimensions, but this can of course be generalised to higher dimensions. An issue with both $k$-means clustering and GMM is that they do not always generate a unique solution, as they depend on the (random) initialisation of the process.

# 3
# Radiomics using manual tumour delineation

In this chapter, we will discuss the important binary problem of distinguishing lower-grade gliomas (astrocytoma and oligodendroglioma, WHO grade II or III) from glioblastomas. Tumour grade has both prognostic and therapeutic consequences. We will investigate a non-invasive method using the radiomics workflow on an online dataset where a manual tumour delineation is provided. This will give insights in feature selection, dimensionality reduction and classification. These techniques will then be used in the more complex problems that will be tackled in the following chapters.

Parts of this work have been presented during the 2016 Medical Imaging Summer School [164] and the 2016 EANO meeting [165].

## 3.1 The importance of primary brain tumour grading

As mentioned in the introduction, primary brain tumour grading has important therapeutic consequences. For glioblastomas, surgical resection followed by combined chemotherapy according to the so called "Stupp protocol" (6 weeks of radiotherapy for 60 Gy in total + temozolomide both during and post-radiotherapy) [166] is the current standard-of-care, with the only exception of patients older than 70 with a negative MGMT promotor methylation status, where surgery plus radiotherapy alone is

recommended by the EANO [83]. These patients might also benefit from a 3-week shortened course of combined chemoradiation, although this debate is still ongoing [167].

For lower-grade gliomas (WHO grade II and III), different treatment options according to the histopathologic and molecular profile of the tumours should be considered, as illustrated in figure 3.1. For suspected low-grade (WHO grade II) tumours with limited clinical complaints for the patient, a watch-and-wait strategy can be taken into account. In this case, the surgical resection can be delayed until signs of tumour growth or enhancement on follow-up imaging, without limiting the overall survival (OS) [81]. However, one should take into account that diffuse low-grade glioma are slowly growing and invasive. Therefore, early and maximal surgical resection using advanced techniques (functional mapping-guided resection) is currently the gold-standard to preserve or even improve the quality of life [168, 169].

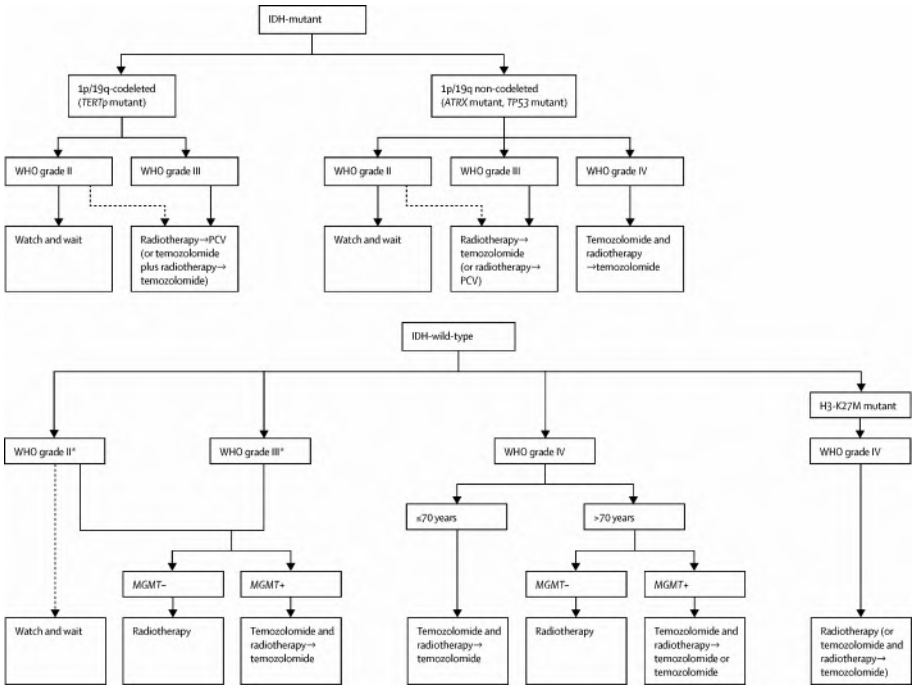



**Figure 3.1:** Therapeutic approaches for diffuse glioma according to the EANO guidelines. Reprinted from [83], Copyright (2017), with permission from Elsevier.

Visual assessment of tumour grade on MRI scans is a very challenging task. High-grade tumours are mostly associated with contrast enhancement on T1ce images, but this is not a perfect predictor, since approximately one third of all nonenhancing gliomas are malignant [170, 171]. Conversely, low-grade gliomas show contrast enhancement in about one out of six cases [172]. This number is expected to be even higher in low-grade oligodendrogliomas [173, 174].

Therefore, in clinical practice the diagnosis is always based on analysis of tumour fragments obtained during biopsy or resection. However, this might not be possible for patients refusing a surgical procedure, when other medical comorbidities obstruct anaesthesia or when the tumour can not be reached without harming the normal functioning of the patient. Moreover, a biopsy suffers from sampling bias, where analysing small fragments of a heterogeneous tumour might lead to an underestimation of the tumour aggressiveness. Compared to a wait-and-scan procedure, biopsy is also related to a reduced OS [81].

For these reasons, non-invasive PBT grading based on medical imaging has gained a lot of attention. Some recent studies on automated brain tumour grading are listed in table 3.1.

**Table 3.1:** Overview of recent studies on automated brain tumour grading.

| author | task | images | method | result |
|---|---|---|---|---|
| Zacharaki (2009) [175] | distinguishing grade II (n=22), grade III (n=18) and grade IV (n=34) gliomas | T1, T1ce, T2, FLAIR, DSC rCBV | linear discriminant analysis, k-nearest neighbour, SVM | sensitivities: 90.9% (grade II), 33.3% (grade III), 41.2% (grade IV) |
| Zöllner (2012) [176] | low-grade (n=38) vs high-grade (n=63) gliomas | DSC rCBV | SVM | 87% accuracy |
| Skogen (2016) [177] | distinguishing grade II (n=27), grade III (n=34) and grade IV (n=34) gliomas | T1ce | ROC analysis on individual features | AUC=0.91 (LGG vs HGG); AUC=0.84 (II vs III), AUC=0.73 (III vs IV) |
| Hsieh (2017) [178] | lower-grade gliomas (n=73) vs glioblastomas (n=34) | T1ce | logistic regression | 88% accuracy |
| Zhou (2017) [151] | grade II (n=35) vs grade III (n=49) gliomas | T1, T1ce, T2, FLAIR | logistic regression, random forests | AUC=0.86 |

Zacharaki et al. [175] designed a multiclass machine learning model to discriminate between five tumour classes, among which 22 grade II, 18 grade III and 34 grade IV gliomas. They extracted 161 quantitative

features from three manually drawn region-of-interests (ROIs) on five different MRI sequences. Using binary SVMs, they achieved accuracies of 75.0% (grade II versus grade III), 96.4% (grade II versus grade IV) and 90.4% (grade III versus grade IV). Afterwards, the binary models were combined into a multiclass scheme with excellent sensitivity for low-grade gliomas (90.9%), but lower sensitivities for high-grade gliomas (33.3% and 41.2% for grade III and grade IV gliomas, respectively).

Zöllner et al. [176] used a binary SVM classifier to distinguish between low-grade and high-grade gliomas. As input to the model, the histogram of the relative cerebral blood volume (rCBV) map was used. This yielded an accuracy of 87%.

Skogen et al. [177] extracted texture parameters on different spatial scales from T1ce images and tested these using receiver operating characteristic (ROC) curves. Their approach reached an area under the curve (AUC) of 0.910 to discriminate low-grade from high-grade gliomas, and lower values to identify individual grades.

Hsieh et al. [178] selected two-dimensional slices of T1ce images. Twenty texture features per scan were fed to a binary logistic regression classifier, achieving an accuracy of 88% in distinguishing between lower-grade gliomas and glioblastomas. Interestingly, this system was used in a follow-up study [179] where three radiologists were asked to grade the same images. Without help from the computer, they independently achieved accuracies of 72%, 73% and 74%. Subsequently, the computer prediction was revealed to them and the radiologists were allowed to revise their decision. This resulted in an improved accuracy with 4% to 9%. Nonetheless, the CAD system outperformed the manual diagnosis.

Zhou et al. [151] investigated radiomics features and machine learning algorithms for a number of tasks, among which grading of lower-grade gliomas. They achieved an AUC score of 0.86.

In the remainder of this chapter, a similar approach as the previously mentioned studies will be followed. Based on the central problem of distinguishing lower-grade gliomas from glioblastomas, we will explain in more detail the different steps of the radiomics workflow. The main focus will be on different feature extraction, dimensionality reduction and classification methods. This will be done on the publicly available BraTS 2017 dataset.

## 3.2 The multimodel BraTS challenge

### 3.2.1 Purpose

The brain tumour segmentation (BraTS) challenge [139, 180] is an annual competition organised since 2012. The goal is to bring together and harmonise the research being conducted on automated brain tumour segmentation on medical images. This field has known a large variability in technical approaches, each time optimised on a specific patient population with its particular imaging characteristics, making it very hard to compare different methods. Therefore, the organisation of the BraTS competition has released several large datasets of preoperative glioma MRI images. Chapter 4 will focus on the actual segmentation problem, for now we will only focus on the data.

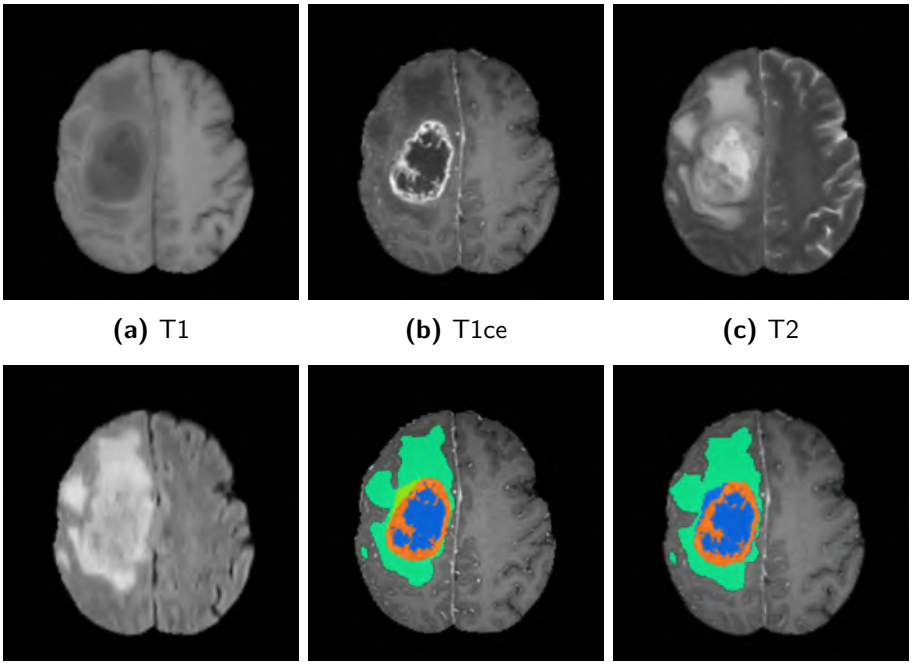
### 3.2.2 Data

The most recent version of the BraTS dataset, edition 2017, consists of MRI scans of 75 lower-grade glioma and 210 glioblastoma patients. For every patient, four sequences are provided: T1, T1ce, T2 and FLAIR. The data are obtained in 19 different clinical centres, on different imaging systems and with variable imaging parameters. All scans are co-registered to the T1ce scan and resampled to a uniform $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$ voxel size with linear interpolation. They are also skull-stripped, meaning that only the brain area is visible on the images. Moreover, for every patient gold standard segmentation labels are provided, as illustrated in figure 3.2.

In the 2012 edition of the competition, contenders only needed to delineate the tumour core and surrounding oedema. From 2013 on, more detailed gold standard labels were given in the training data. To obtain these, four radiographers manually outlined different structures on the scans and the corresponding labels are:

**1** necrotic or fluid-filled core

**2** oedema



**(a)** T1 **(b)** T1ce **(c)** T2

**(d)** FLAIR **(e)** gold standard (2013) **(f)** gold standard (2017)

**Figure 3.2:** Example of a high-grade glioma in the BraTS dataset. For every patient, four MRI sequences are given, as well as gold standard segmentation labels. Colour code: blue = necrosis (2013) or necrosis + non-enhancing tumour (2017); yellow = non-enhancing tumour (2013); orange = enhancing tumour; green = oedema.

**3** non-enhancing or solid core

**4** contrast-enhancing core

The final labels are obtained using a hierarchical majority voting scheme, assigning a voxel to the highest class to which at least half of the raters agree on [139]. Since 2017, the tumour core is only divided into two parts: contrast-enhancing tumour (label 4), and necrotic and non-enhancing tumour core (label 1). The label 2 for peritumoural oedema remained unchanged.

As mentioned in the introduction, conventional MRI is not a quantitative, but a qualitative imaging modality. Therefore, some prepro-

cessing techniques need to be applied before we can extract quantitative features.

### 3.2.3 Preprocessing

In this thesis, the Multi-image Analysis GUI (MANGO, `ric.uthscsa.edu/mango`) is used for visualisation of medical images. This software is also used to convert clinical scans from the standard Digital Imaging and Communications in Medicine (DICOM) format to Neuroimaging Informatics Technology Initiative (NIfTI) 1-1 format [181], thereby removing sensitive patient-related information in order to anonymise the images.

#### Bias field correction

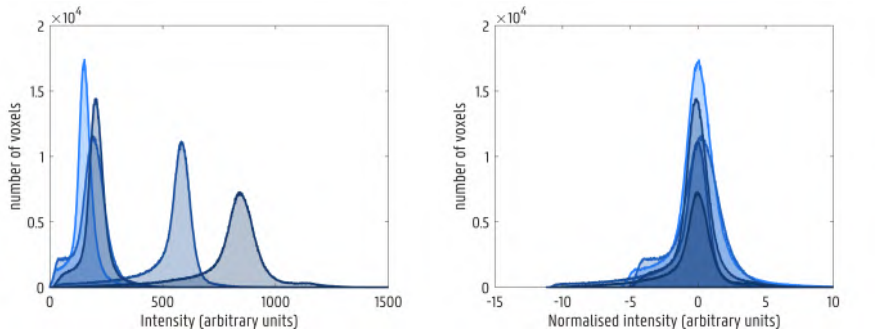
Small inhomogeneities in the magnetic field of the MRI scanner can cause a slight error in the images, called the bias field. This takes the shape of a smooth, low-frequency signal corrupting the image intensities. These artefacts, although not usually a problem for visual inspection, can impede automated processing of the images. Fortunately, many methods exist to correct for this phenomenon. In this thesis, all images are bias field corrected using SPM12 (version 6906, Wellcome Trust Centre for Neuroimaging, London) [182], running on MATLAB R2017b (The MathWorks, Inc., Natick, MA).

#### Intensity normalisation

Image intensities on MRI are difficult to interpret, and highly variable between patients and even between scans, as is clear from 3.3a. Intensity as such can therefore not be used as a quantitative predictor, unless it is carefully modified. In the following studies, the MRI intensities are normalised using the *white-stripe normalisation technique*, developed by Shinohara et al. [138]. For this method, we first use the segmentation module of SPM12 to estimate tissue probability maps (TPMs) for the grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF), based on the T1-weighted image. Next, the largest peak in the WM histogram is assumed to belong to the normal-appearing white matter

(NAWM). The normalised intensity is then given by $(I - \mu)/\sigma$, where $\mu$ and $\sigma$ are the mean and standard deviation of the intensity of the NAWM, respectively.

As a result of this operation, the healthy white matter will have zero mean and unit variance for all images and for all patients. The intensities of other tissues can then be related to the intensity of the NAWM and become interpretable in this way. This is illustrated in figure 3.3b.




**(a)** Before normalisation **(b)** After normalisation

**Figure 3.3:** Example of the white-stripe normalisation technique on the histogram of five randomly selected T1ce scans from the high-grade BraTS dataset.

## Intensity discretisation

For the calculation of some texture features, the intensities of an image need to be discretised to a limited number of values, typically a power of 2. An additional advantage of intensity discretisation is that the noise will be suppressed. Different techniques are possible, which cause a slight variability in the resulting features. In this thesis, intensity discretisation is performed using the following formula:

$$I_{\text{discr}} = \left\lceil (N_g - 1) \frac{I - \min(I)}{\max(I) - \min(I) + 1} \right\rceil \quad ,$$

where $N_g$ is the number of grey-levels (usually 64), and $I$ and $I_{\text{discr}}$ are the original and discretised images, respectively.

## 3.3 Feature extraction

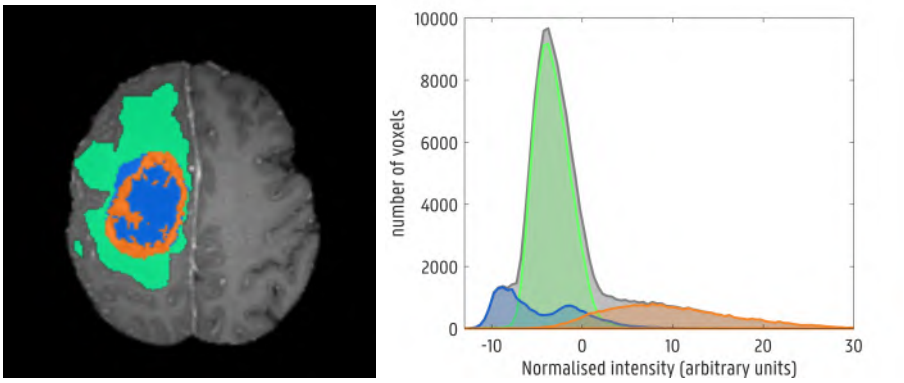
The appearance of a tumour on an image can be quantified using a myriad of different features, each one highlighting a different aspect of the tumour. Here, we will only focus on four feature sets: histogram, shape, texture, and localisation and environment features. A list of all features and their definition is given in appendix A.

### 3.3.1 Histogram features

The histogram describes the intensity distribution within a ROI. Its shape therefore contains a lot of information on the heterogeneity: a homogeneous region (many voxels with similar grey-level) will have a narrow shape, whereas heterogeneous tissue (many different grey-levels) will have a broader appearance. Typical histogram features (sometimes called first-order statistics (FOS)) are minimal, maximal, mean and median value, range, standard deviation and variance. More advanced features include energy (sum of the squared intensities), uniformity (sum of the squared probabilities in the normalised histogram), root mean square (square root of the energy divided by the number of voxels), skewness (measure of asymmetry) and kurtosis (measure for the presence of heavy tails in the histogram).

### 3.3.2 Shape and size features

A second set of features is based on the size and shape of the tumour. The easiest parameters are volume (number of voxels times voxel volume) and maximal 3D diameter (the largest distance between any two points belonging to the tumour). Other features are based on the tumour area, which is calculated by triangulating the volume surface and summing the triangle areas. Surface to volume ratio and several variants (such as compactness $= V/(\sqrt{\pi}A^{2/3})$) can be used as predictors of tumour infiltration, as smaller area to volume ratios will correspond to more spherical tumours (e.g. low-grade meningioma). An example of a 3D-rendered high-grade tumour is given in figure 3.5.




**(a)** T1ce with tumour mask **(b)** Corresponding histogram in different tissues

**Figure 3.4:** Example of a T1ce scan and corresponding histogram of a high-grade glioma from the BraTS dataset. The histogram is calculated on the entire 3D tumour masks, not on this example slice. Colour code: blue = necrosis + non-enhancing tumour; orange = enhancing tumour; green = oedema; grey (on the histogram) = entire tumour.

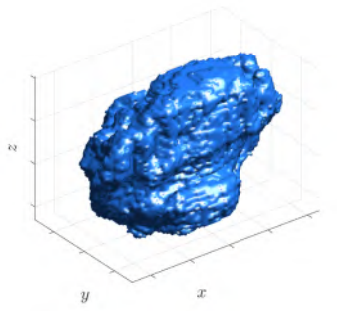



**Figure 3.5:** 3D-rendering of the tumour of figure 3.4a, obtained by triangulating the surface.

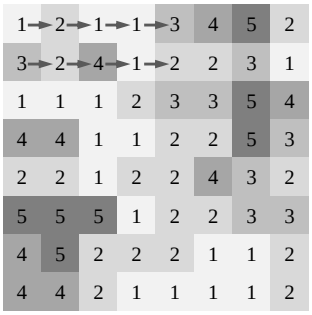
### 3.3.3 Texture features

Although the histogram contains information on the distribution of grey-levels and therefore provides insights in tumour heterogeneity, it does not take into account the spatial distribution of these grey-levels. A

region with gradually increasing intensities could for example result in a similar histogram as a very heterogeneous region. As different spatial distributions of grey-levels can provide valuable biological information, it is important to quantify the heterogeneity with a technique called *texture analysis*. This usually starts with the calculation of different texture matrices, which will be explained with the following four types.

## Grey-level co-occurrence matrix

The use of the grey-level co-occurrence matrix (GLCM) was published for the first time by Robert Haralick in 1973 [183]. It describes the occurrence of pairs of specific voxel intensities: the element $(i, j)$ of the GLCM equals the number of times intensity $j$ occurs at a distance $d$ in direction $\theta$ from intensity $i$. The shape of the GLCM is therefore $N_g \times N_g$, with $N_g$ the number of grey-levels present in the image, for every $d$ and $\theta$. This is illustrated in figure 3.6 for an $8 \times 8$ matrix with five grey-levels. For two-dimensional images, the GLCM is usually calculated for four angles ($\theta = 0°$, $45°$, $90°$, $135°$), while in 3D thirteen spatial directions are customary.




**(a)** Example matrix

| | | *j* | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| | 1 | 8 | 8 | 1 | 0 | 0 |
| | 2 | 4 | 7 | 3 | 2 | 1 |
| *i* | 3 | 1 | 2 | 2 | 1 | 1 |
| | 4 | 2 | 1 | 1 | 2 | 2 |
| | 5 | 1 | 2 | 1 | 1 | 2 |

**(b)** Corresponding GLCM

**Figure 3.6:** Example matrix and corresponding grey-level co-occurrence matrix (GLCM) for distance $d = 1$ and $\theta = 0°$: every element $(i, j)$ of the GLCM depicts the number of times intensity $j$ occurs to the right of intensity $i$.

Starting from the GLCM, many different texture features can be calculated. Some easy examples are *contrast*, which is defined by $\sum_i \sum_j |i-j|^2 \mathrm{GLCM}(i, j)$. This feature will be large when the GLCM contains large

values at positions far from the diagonal, or in other words, when voxels with a large intensity difference often occur close to each other. Conversely, the feature *homogeneity 1*, defined as $\sum_i \sum_j \text{GLCM}(i,j)/(1 + |i-j|)$ will be large when the GLCM is mainly focussed around the diagonal, or indeed when the region mostly contains similar grey-levels close to each other.

Some features are based on variants of the GLCM, such as the sum or difference matrices: $p_{x\pm y}(k) = \sum_i \sum_j P(i,j)$ $, |i \pm j| = k$, and are therefore less intuitively clear. For a fixed distance $d$, every parameter is calculated for every possible direction $\theta$, and the mean and standard deviation are stored as features.

## Grey-level run-length matrix

The grey-level run-length matrix (GLRLM) [184] quantifies runs of grey-levels, being consecutive voxels with the same intensity along a straight line, in an image. In other words, every element $(i,j)$ of the GLRLM contains the number of times $j$ voxels with equal intensity $i$ are found along a straight line defined by the angle $\theta$. An example is given in figure 3.7. The size of the GLRLM is $N_g \times N_r$ with $N_g$ the number of grey-levels and $N_r$ the length of the longest run in direction $\theta$.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 1 | 1 | 3 | 4 | 5 | 2 |
| 3 | 2 | 4 | 1 | 2 | 2 | 3 | 1 |
| 1 | 1 | 1 | 2 | 3 | 3 | 5 | 4 |
| 4 | 4 | 1 | 1 | 2 | 2 | 5 | 3 |
| 2 | 2 | 1 | 2 | 2 | 4 | 3 | 2 |
| 5 | 5 | 5 | 1 | 2 | 2 | 3 | 3 |
| 4 | 5 | 2 | 2 | 2 | 1 | 1 | 2 |
| 4 | 4 | 2 | 1 | 1 | 1 | 1 | 2 |

**(a)** Example matrix

| | | *j* | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| *i* | 1 | 5 | 3 | 1 | 1 |
| | 2 | 8 | 5 | 1 | 0 |
| | 3 | 5 | 2 | 0 | 0 |
| | 4 | 5 | 2 | 0 | 0 |
| | 5 | 4 | 0 | 1 | 0 |

**(b)** Corresponding GLRLM

**Figure 3.7:** Example matrix and corresponding grey-level run-length matrix (GLRLM) for $\theta = 0°$: every element $(i,j)$ of the GLRLM depicts the number of runs in horizontal direction with length $j$ and intensity $i$.

Texture features based on the GLRLM are in general easily inter-

pretable. Typical parameters emphasise the run-length of specific grey-levels. For example, the feature *short run low grey-level emphasis*:

$$\text{SRLGLE} = \frac{\sum_i \sum_j \left( \frac{\text{GLRLM}(i,j)}{i^2 j^2} \right)}{\sum_i \sum_j \text{GLRLM}(i,j)} \ ,$$

will be large when the image mainly contains short runs (small $j$) of low grey-values (small $i$). Again, every feature is calculated for GLRLMs defined in 4 (in 2D) or 13 (in 3D) different directions, and the mean and standard variance are stored.

## Grey-level size-zone matrix

In contrast to the GLCM and GLRLM, the grey-level size-zone matrix (GLSZM) [185] is not dependent on a specific direction. Its definition is similar to the GLRLM, as the element $(i, j)$ of the GLSZM contains the number of times a zone with size $j$ and intensity $i$ is found in the image, as illustrated in figure 3.8. A zone is a cluster of connected voxels with equal intensity. The size of the GLSZM is $N_g \times N_z$, where $N_z$ is the size of the largest zone.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 1 | 1 | 3 | 4 | 5 | 2 |
| 3 | 2 | 4 | 1 | 2 | 2 | 3 | 1 |
| 1 | 1 | 1 | 2 | 3 | 3 | 5 | 4 |
| 4 | 4 | 1 | 1 | 2 | 2 | 5 | 3 |
| 2 | 2 | 1 | 2 | 2 | 4 | 3 | 2 |
| 5 | 5 | 5 | 1 | 2 | 2 | 3 | 3 |
| 4 | 5 | 2 | 2 | 2 | 1 | 1 | 2 |
| 4 | 4 | 2 | 1 | 1 | 1 | 1 | 2 |

**(a)** Example matrix

| $i$ \ $j$ | 1 | 2 | 3 | 4 | ··· | 6 | ··· | 10 |
|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 0 | 1 | 0 | ··· | 2 | ··· | 0 |
| 2 | 3 | 4 | 0 | 0 | ··· | 0 | ··· | 1 |
| 3 | 4 | 1 | 1 | 0 | ··· | 0 | ··· | 0 |
| 4 | 4 | 1 | 1 | 0 | ··· | 0 | ··· | 0 |
| 5 | 1 | 1 | 0 | 1 | ··· | 0 | ··· | 0 |

**(b)** Corresponding GLCM

**Figure 3.8:** Example matrix and corresponding grey-level size-zone matrix (GLSZM): the element $(i, j)$ contains the number of times a cluster with size $j$ of intensity $i$ occurs in the image.

Features based on the GLSZM have a comparable definition as those based on the GLRLM, as they emphasise the size of zones with specific

grey-levels. For example, the *small zone low grey-level emphasis*:

$$Szonelogl = \frac{\sum_i \sum_j \left( Szpcent^2 \frac{\mathrm{GLSZM}(i,j)}{i^2 j^2} \right)}{\sum_i \sum_j \mathrm{GLSZM}(i,j)} \ ,$$

will be large when many small zones (small $j$) of low intensity (small $i$) are present in the image. *Szpcent* is here a normalisation constant, defined by the total number of zones divided by the number of voxels.

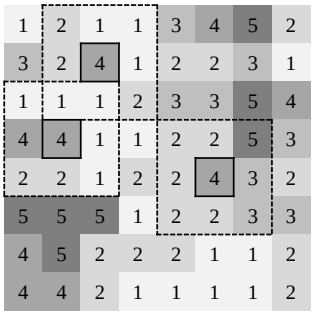
## Neighbourhood grey-tone difference matrix

The neighbourhood grey-tone difference matrix (NGTDM) [186] is a one-dimensional vector containing information on the local differences between intensities, which is constructed in the following way. For every voxel, we first calculate the mean $\bar{A}_i$ of the intensities in a local neighbourhood (excluding the voxel itself). The size of the neighbourhood is determined by a parameter $d$. Next, the element $i$ of the NGTDM equals the sum of the absolute difference between $i$ and $\bar{A}_i$ for all voxels with intensity $i$: $\mathrm{NGTDM}(i) = \sum |i - \bar{A}_i|$. To further clarify this, consider the example matrix of figure 3.9 and assume we want to calculate the NGTDM for $d = 1$, resulting in a $3 \times 3$ voxels environment. This means that we can only use the inner matrix with size $6 \times 6$ for the calculation of the NGTDM, since we cannot define the environment for voxels at the edge. Inside the inner matrix, there are three voxels with $i = 4$. The mean intensity of their neighbourhood (the 8 adjacent voxels) is 1.375, 2.625 and 1.625. Therefore, $\mathrm{NGTDM}(4) = |4 - 1.375| + |4 - 2.625| + |4 - 1.625| = 6.375$.

Typically, only four features are calculated based on the NGTDM: *coarseness*, *contrast*, *complexity* and *strength*. For example, coarseness is defined as:

$$f_{cos} = \left( \varepsilon + \sum_{i=1}^{N_g} p_i \, \mathrm{NGTDM}(i) \right)^{-1} \ ,$$

where $p_i$ is the fraction of voxels with intensity $i$ and $\varepsilon$ a small number (to avoid this feature becoming infinite for a perfectly uniform image). Coarseness can be regarded as homogeneity on a local level: when an image is more coarse, we can expect that the difference between a voxel




**(a)** Example matrix

| $i$ | NGTDM |
|---|---|
| 1 | 9.5 |
| 2 | 7.5 |
| 3 | 2.125 |
| 4 | 6.375 |
| 5 | 9.25 |

**(b)** Corresponding NGTDM

**Figure 3.9:** Example matrix and corresponding neighbourhood grey-tone difference matrix (NGTDM) for distance $d = 1$.

and its direct neighbourhood is small. Therefore, the contributions of the NGTDM will be small, and $f_{cos}$ will be large.

### 3.3.4 Localisation and environment features

We can not only extract information from the isolated tumour, but also from its environment. For example, meningiomas will always stem from the meninges, and will therefore be mainly found near the border of the brain rather than in central areas. Moreover, the contrast between the intensities inside and outside the tumour can have predictive power. Meningiomas will for example be well-defined with a distinct tumour border, whereas glioma have a more infiltrating border, causing the intensities inside and outside the tumour to be similar.

### 3.3.5 Construction of feature matrix

Now that we have explained all the features, we will return to our radiomics problem. For every ROI, we calculate 207 features:

- 14 histogram features
- 8 shape and size features
- 185 texture features:

  - 138 GLCM-based features ($d = 1, 2, 3$, mean and standard deviation over 13 directions)
  - 22 GLRLM-based features (mean and standard deviation over 13 directions)
  - 13 GLSZM-based features
  - 12 NGTDM-based features ($d = 1, 2, 3$)

Moreover, for every patient, we define five tumour ROIs:

- necrosis + non-enhancing tumour (BraTS label 1)
- oedema (BraTS label 2)
- enhancing tumour (BraTS level 4)
- tumour core (BraTS labels 1 and 4)
- total abnormal region (BraTS labels 1, 2 and 4).

Finally, we calculate the features on two images per patient: the T1ce and FLAIR scans, as these provide complementary information. Moreover, these scans are available for most patients in the Ghent University Hospital, which is an important prerequisite since we will apply this methodology to clinical scans collected in our centre in chapters 5 and 6. Therefore, we obtain $207 \times 5 \times 2$ features, to which we add 27 localisation and environment features. In total, we have 2097 quantitative features per patient. However, some features are not calculated for all patients (not all patients have necrotic and/or contrast-enhancing tissue). These parameters can be removed *a priori*. One lower-grade patient has no oedema label in the groundtruth segmentation, and this patient was also removed. In this way, 1908 features are remaining. Since this number still largely exceeds the total number of patients in the dataset (74 lower-grade glioma and 210 glioblastoma patients), we need to reduce the number of parameters for further analysis. This is the topic of the next section.

## 3.4 Dimensionality reduction

Reducing the number of parameters is possible in roughly three ways. A first option are *feature ranking* methods (paragraph 3.4.1). This means that all features are ranked according to a certain criterion, and the top ranked features are then simply selected to build a predictive model. These methods are often preferred for large datasets due to their computational feasibility. However, in feature ranking methods no interactions between different features are taken into account, which might lead to the selection of redundant parameters.

*Feature subset selection* methods are designed to incorporate interactions between parameters and therefore select an ideal subset of features for a specific task. However, as typically many feature combinations need to be analysed, this method comes with a high computational cost for high dimensional problems. Therefore, this is often combined with a feature ranking method to reduce the number of possible features. We will discuss the sequential forward selection method in paragraph 3.4.2.

Lastly, *feature transformation* methods apply a function to a set of features in order to obtain new parameters with higher predictive power. A typical example is principal component analysis (PCA) (paragraph 3.4.3), where the features are transformed into orthogonal components in decreasing order of variance. A disadvantage of this method is that the principal components are linear combinations of the original features and are therefore difficult to interpret.

As different features often have a strongly divergent value range, it is difficult to compare them. Therefore, it is good practice to transform the feature values to their standard score:

$$z = \frac{x - \mu}{\sigma} \ ,$$

where $x$ is the original value, and $\mu$ and $\sigma$ are the mean and standard deviation of the feature, respectively.

### 3.4.1 Feature ranking methods

Many criteria can be used to obtain a ranking of the features. In the next discussion, we will focus on three methods.

#### Two-sample $t$-test

An easy way to check if the mean of two distributions is significantly different, is the two-sample $t$-test. Without loss of generality, we can assume that the variance of a feature is different for lower-grade and high-grade glioma patients, and we can therefore use Welch's $t$-test by setting the `'variance'` setting to `'unequal'` in MATLAB's implementation of the `ttest2` function. This function returns a $p$-value for every feature, which gives the probability that the feature values are observed under the null hypothesis that the mean of the feature distribution is the same in both classes. In other words, a high $p$-value corresponds with a lower discriminative power to separate the two classes, since there is a higher probability that the feature mean is equal. Features are therefore ranked according to increasing $p$-value.

#### Relief

The *relief* algorithm was proposed by Kira and Rendell in 1992 [187]. It is an easy and fast method which calculates weights for every feature in an iterative way. Suppose our dataset is split up in two classes, and initialise the weight vector $\vec{W}^0$ to zero. Then we randomly pick one sample $\vec{x}$ from our dataset, and select two instances, each from one class, closest to it. The instance from the same class is called the *near-hit* sample, the instance from the other class the *near-miss* sample. Next, we update the weight-vector as follows:

$$W_i^{j+1} = W_i^j - (x_i - \textit{near-hit}_i)^2 + (x_i - \textit{near-miss}_i)^2 \ ,$$

where $i$ runs over the entire feature-space. This process is repeated $m$ times, where $m$ is smaller than the number of samples. Ultimately, the most predictive features will have the largest weights, since the distance to *near-miss* samples will be larger according to this feature.

The `relieff` implementation in MATLAB uses the adaptation of the original algorithm by Kononenko et al. [188]. Instead of a quadratic distance function, this algorithm uses the absolute difference as update parameter. The main difference with the original algorithm is however that not one but $k$ *nearest-miss* and *nearest-hit* samples are selected, where $k$ is a tunable parameter. Moreover, the process is repeated over all, instead of $m$, samples in the dataset.

#### OOB-error

As mentioned in the introduction, Random Forests can be used to estimate the predictor importance. To accomplish this, every tree is evaluated by the samples that were not used to train this particular tree, called the out-of-bag (OOB) samples. The corresponding error (OOB-error) is then averaged over all trees in the forest. To test the importance of the $j$-th feature, we permute the values of this feature over all samples and again calculate the error. Features with a large difference between the original and the permuted OOB-error are more important. In MATLAB, this parameter is appropriately called `OOBPermutedPredictorDeltaError`, accessible via the `TreeBagger` function.

### 3.4.2 Sequential forward selection

Sequential forward selection (SFS) is a bottom-up search strategy, meaning that we start from an empty feature vector, and gradually add features that minimise a certain cost function. In classification problems, this is generally the misclassification error using $k$-fold cross-validation. In other words: starting with the first candidate feature, we build a classification model and evaluate it with the cost function. This is repeated for all possible features, and we select the feature that minimises the error. Next, we build a new model with two features: the previously selected one combined with a new candidate, selected from the remaining parameter set. Again, the feature with minimal cost is selected. This process is repeated until a certain criterion is met, such as an increasing cost function, or when a specific number of features is selected.

Using SFS, we make sure that only features with complementary information are selected. However, since in every iteration $k$ models need to be built and validated for every feature in the remaining parameter set, this is a computationally expensive technique. Therefore, we often first limit the candidate feature set using a feature ranking method.

### 3.4.3 Principal component analysis

Suppose we have a $n \times p$ data matrix $\mathbf{X}$ consisting of $n$ samples with $p$ features. The principal component decomposition [189] of $\mathbf{X}$ is then given by:

$$\mathbf{T} = \mathbf{X}\,\mathbf{W} \ ,$$

where the columns of $\mathbf{T}$ are the principal components. The number of components is given by $\min(n, p)$. The first component has the largest possible variance. Each following component has the largest possible variance under the condition that it is orthogonal to all previous components. Therefore, the principal components form an orthogonal basis set. Since the first components are able to explain most of the variability in the data, these can be used as a reduced feature set. For a new patient, we can simply multiply the features with the values of the $i^{th}$ column of $\mathbf{W}$ and sum them to obtain the $i^{th}$ principal component.

By default in MATLAB's `pca` function, the algorithm is implemented using singular value decomposition of $\mathbf{X}$. This means the data matrix is first decomposed as:

$$\mathbf{X} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{W}^T \ ,$$

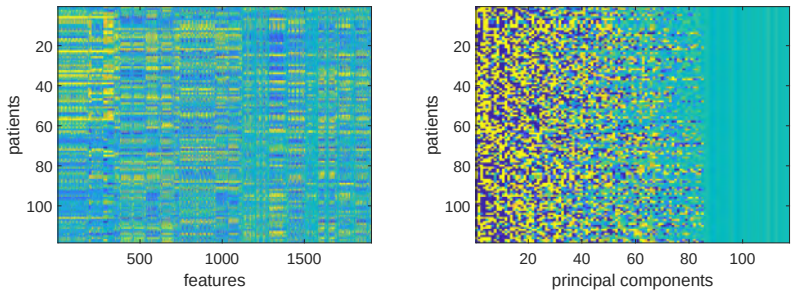
where $\mathbf{\Sigma}$ is a rectangular diagonal matrix, and the columns of $\mathbf{U}$ and $\mathbf{W}$ contain orthogonal unit vectors. It follows that

$$\mathbf{T} = \mathbf{X}\,\mathbf{W} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{W}^T\,\mathbf{W} = \mathbf{U}\,\mathbf{\Sigma} \ .$$

Since the singular value decomposition of a matrix can be calculated very efficiently, this method is often preferred for PCA.

Since PCA calculates a linear combination of features, this can be difficult to interpret. Therefore, the use of PCA is rather limited in

clinical applications. Moreover, the patient classes are not taken into account when calculating the components, and consequently there is no guarantee that they contain the *relevant* variability in the data. This is illustrated in figure 3.10.




**(a)** Before PCA **(b)** After PCA

**Figure 3.10:** Example of a training set with 59 lower-grade glioma (top half) and 59 glioblastoma patients (bottom half). Before PCA, the features are clustered using $k$-means clustering for visual purposes. After PCA, it is clear that the first features contain most of the variance in the data and gradually lose importance, but the distinction between the two classes seems less noticeable.

## 3.5 Binary classification model

We now have all the tools at our disposal for the radiomics study aiming to predict tumour grade based on features extracted from T1ce and FLAIR MRI scans. To this end, we will test the performance of five classification algorithms:

- linear SVM (SVM-1)
- SVM with quadratic kernel (SVM-2)
- SVM with gaussian kernel (SVM-3)
- Random Forests (200 trees per forest)

- simple artificial neural network with one hidden layer with 10 neurons

Every classifier is combined with six previously discussed feature selection methods:

- two-sample $t$-test
- relief (with $k = 5$ or $k = 20$)
- OOB-error
- sequential forward selection
- principal component analysis

This yields a total of 30 different tests. For every test, we first randomly select 55 patients (about 20% of the data), and use the remaining patients to select the features and train the model. Since there are many more glioblastoma patients in the training set compared to lower-grade gliomas, the model might be biased towards this larger class. Therefore, we sub-sample the high-grade class by randomly selecting an equal number of patients for every class to build the model. This model is next applied to the previously selected 55 patients to validate the performance. To avoid any bias that might be present due to the random selection of the patients, we repeat this process 100 times for every test. The average and standard deviation are given in table 3.2.

The best result is obtained when combining the *relieff* feature selection method with Random Forests classification, including 700 features. The accuracy, defined as the number of correct predictions divided by the total number of patients, reaches 88.0% using this model. Overall, many tests achieve a similar performance, with accuracies close to the results reported in literature (table 3.1).

## Influence of classification algorithm

When observing the accuracies in table 3.2, it is clear that the Random Forests classification algorithms achieves the best results. However, it is only slightly better than linear SVM, which outperforms the kernel-based variants. The version with the radial basis function (SVM-3) only

**Table 3.2:** Result of the radiomics study to distinguish lower-grade gliomas from glioblastomas based on T1ce and FLAIR MRI. Five classifications algorithms are combined with six different feature selection algorithms. Maximally allowed number of features: 1000 for the ranking methods (except for ANN: 250), 25 for SFS, all for PCA.

| | **classification algorithm** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **feature** | SVM-1 | | SVM-2 | | SVM-3 | | RF | | ANN | |
| **selection** | *acc* | $n$ | *acc* | $n$ | *acc* | $n$ | *acc* | $n$ | *acc* | $n$ |
| $t$-test | (85.8±4.8)% | 1000 | (83.1±4.6)% | 350 | (80.4±6.0)% | 10 | (87.8±4.3)% | 700 | (82.1±6.9)% | 70 |
| *relieff* $k = 5$ | (86.8±4.7)% | 1000 | **(84.7±4.7)%** | 500 | (82.4±6.6)% | 6 | **(88.0±4.5)%** | 700 | (82.8±5.9)% | 250 |
| *relieff* $k = 20$ | **(87.0±4.7)%** | 1000 | (83.8±4.9)% | 200 | (82.8±6.2)% | 5 | (87.9±4.7)% | 500 | **(82.9±6.5)%** | 200 |
| OOB-error | (83.8±5.0)% | 1000 | (78.9±5.5)% | 400 | (73.6±1.2)% | 20 | (84.0±5.3)% | 900 | (73.8±7.7)% | 200 |
| SFS | (84.2±5.4)% | 25 | (82.6±6.9)% | 3 | **(83.5±5.3)%** | 9 | (85.7±5.4)% | 24 | (81.9±6.7)% | 4 |
| PCA | (86.9±4.5)% | 80 | (81.7±5.1)% | 85 | (73.6±1.0)% | 4 | (85.3±5.2)% | 35 | (80.4±8.3)% | 53 |

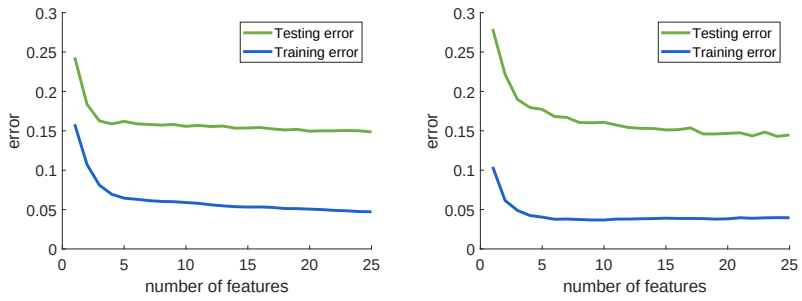
*acc* = accuracy (number of correctly classified samples), expressed as mean ± standard deviation;

$n$ = number of features (or principal components) with best performance

performs well with a low number of features, suggesting that is is more prone to overfitting. For all future studies in this thesis, Random Forests will be used. It not only achieves the best results, it also has the added value of giving probabilities for every decision that was made, a feature that is not standard available when using SVMs.

### Influence of feature selection method

In general, the *relieff* feature ranking method achieves the best results, although many features are necessary for best performance. Ranking the features based on OOB-error can not compete with the other dimensionality reduction techniques, even when combined with Random Forests. Sequential forward selection yields the lowest number of features with similar accuracy as the other selection methods, but is computationally costly. In figure 3.11 the progress of the error when adding features to the classification model is illustrated.




**(a)** SVM-1 **(b)** Random Forests

**Figure 3.11:** Progress of training and testing error during the sequential forward selection (SFS) algorithm. The testing error is still decreasing after 25 features, suggesting that the optimum is not yet found, but adding even more features will make the algorithm computationally too costly.

### Most predictive features

When using the sequential forward selection (SFS) algorithm, we can assess which features are most predictive using a certain classification

algorithm. In the previous experiment, different sets of 25 features were found for 100 repetitions with slightly different training and test sets. We can combine these to find the best features. Therefore, we appoint 25 points to a feature if it select first in the SFS algorithm, 24 points if it is selected second, and so on. Then we add all the points for a specific feature, yielding a ranking of best features for the entire dataset. The top-10 for SVM-1 and Random Forests are listed in table 3.3.

**Table 3.3:** Top-10 best performing features obtained with SFS combined with SVM-1 or Random Forests for the entire dataset.

| | **SVM-1** | **Random Forests** |
|---|---|---|
| 1 | T1ce: tumour core - histogram: mean | T1ce: total - histogram: mean |
| 2 | T1ce: tumour core - histogram: median | T1ce: tumour core - histogram: mean |
| 3 | volume enhancing tumour | volume enhancing tumour |
| 4 | T1ce: total - histogram: mean | T1ce: ratio core / surrounding (5 voxels) |
| 5 | T1ce: tumour core - GLCM: difference entropy ($d = 1$, mean) | T1ce: tumour core - histogram: median |
| 6 | T1ce: tumour core - GLCM: difference entropy ($d = 2$, mean) | T1ce: tumour core - GLCM: entropy ($d = 3$, mean) |
| 7 | T1ce: tumour core - GLCM: entropy ($d = 1$, mean) | FLAIR: necrosis + non-enhancing core - NGTDM: complexity ($d = 2$) |
| 8 | T1ce: tumour core - GLCM: difference entropy ($d = 3$, mean) | T1ce: tumour core - GLCM: difference entropy ($d = 3$, mean) |
| 9 | FLAIR: necrosis + non-enhancing core - NGTDM: complexity ($d = 2$) | T1ce: tumour core - GLCM: inverse difference normalised ($d = 3$, mean) |
| 10 | FLAIR: necrosis + non-enhancing core - NGTDM: complexity ($d = 1$) | T1ce: tumour core - GLCM: entropy ($d = 1$, mean) |

For both classification methods, the volume of the contrast enhancement (a clear radiological parameter) is an important predictor. However, more advanced features including histogram and texture parameters are present as well, mainly based on the T1ce scan. Furthermore, many closely related parameters are selected in this top-10 (e.g. mean and median of the same parameters, different distances for the same texture parameter), showing that they have a similar predictive power. These will however rarely be selected simultaneously for a single training test.

### Example

Table 3.4 gives the confusion matrix obtained for the *relieff* ($k = 5$) feature selection method combined with Random Forests classification. This gives an overview of the predictions for patients of different classes, averaged over 100 repetitions. We observe a high-grade specificity of 90.6% (true positive rate, probability that a high-grade patient is predicted high-grade) and sensitivity of 81.4% (true negative rate, probability that a lower-grade patients is predicted lower-grade). These numbers are quite balanced, since we selected equal amounts of samples from both classes during the training phase. Moreover, the algorithm is more certain when making a correct decision than when picking the wrong choice.

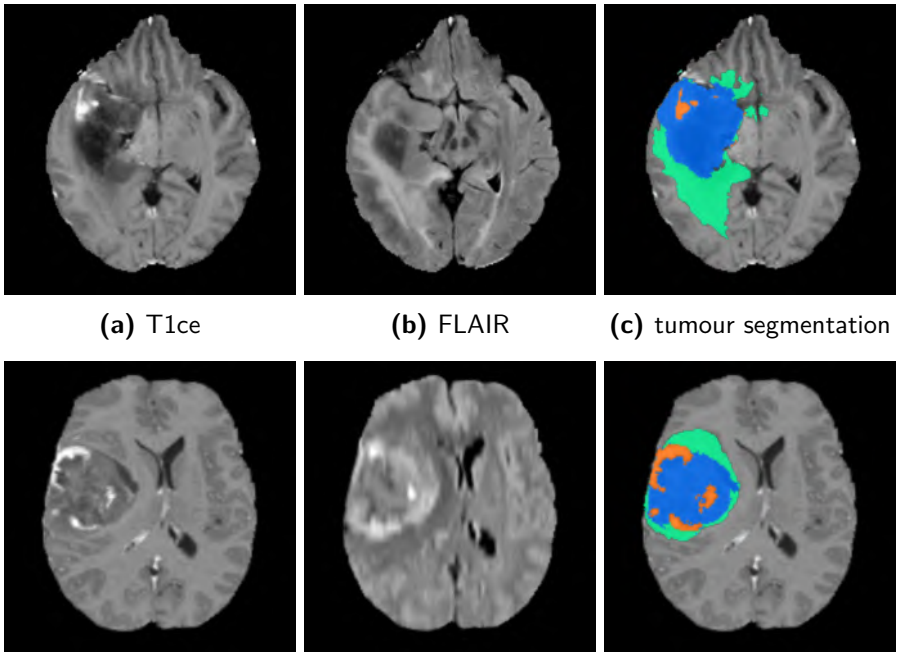
**Table 3.4:** Confusion matrix for Random Forests combined with *relieff* feature ranking ($k = 5$), averaged over 100 iterations. A confusion matrix shows for every class how many samples are correctly and falsely predicted. Also given is the average certainty $\bar{p}$ with which the decision was made. Notice that for the test set, a representative sample of the dataset is taken, and therefore more high-grade than lower-grade patients are present.

| | | predicted | |
|---|---|---|---|
| | | lower-grade | high-grade |
| true | lower-grade | 11.8±1.5 ($\bar{p} = 79.7\%$) | 2.7±1.5 ($\bar{p} = 67.5\%$) |
| | high-grade | 4.0±2.1 ($\bar{p} = 64.5\%$) | 36.6±2.1 ($\bar{p} = 81.6\%$) |

In figure 3.12 two examples are given of wrongly predicted patients: one lower-grade glioma and a glioblastoma. They share some common characteristics, such as a large tumour core and distributed contrast enhancement. Both samples are predicted with a low certainty by the algorithm: 52.2% and 59.5%, respectively. This suggests that a user could interpret these results, and manually make the right choice.

## 3.6 Conclusion

In this chapter some building blocks for a radiomics study are explained in detail, with an emphasis on feature extraction and dimensionality



**(a)** T1ce **(b)** FLAIR **(c)** tumour segmentation

**(d)** T1ce **(e)** FLAIR **(f)** tumour segmentation

**Figure 3.12:** Representative slices of wrongly predicted patients, using *relieff* feature selection ($k = 5$) and Random Forests. Upper row: lower-grade patient predicted as high grade with probability $p = 52.5\%$. Bottom row: high-grade patient predicted as lower-grade with probability $p = 59.5\%$.

reduction. This yields excellent results to automatically distinguish between lower-grade gliomas and glioblastomas, based on only two scans (T1ce and FLAIR MRI). Chapter 7 will further elaborate on this problem and a suggestion will be offered to further improve the performance.

# 4
# Brain tumour segmentation

In the previous chapter, we elaborated on the last two steps of radiomics: feature extraction and analysis. However, before we can calculate features, we first need to define the tumour borders on the image. Therefore, we investigate different approaches to perform this task automatically in this chapter.

This work has been presented during the 18th Symposium of the Belgian Society of Nuclear Medicine (2017) [190] and the European Congress of Radiology 2018 [191].

## 4.1 Introduction

In clinical practice, segmentation is mostly performed manually, as an experienced radiologist delineates the tumour on several slices of a 3D brain scan. Apart from being time and labour intensive, manual delineation is prone to inter- and intra-observer variability. In order to obtain groundtruth labels for the BraTS 2013 database, four manual performers segmented scans from 10 low-grade and 20 high-grade glioma patients [139]. The authors report average inter-rater Dice scores of 67% and 93% for segmenting the low-grade and high-grade tumour core, respectively. Delineating the entire tumour region, including oedema, was considered easier, with inter-rater Dice scores of 84% and 88% for low- and high-grade glioma respectively. The Dice score - also called Dice similarity coefficient (DSC) or Sörensen index - is a measure of overlap between

two sets $X$ and $Y$:

$$\mathrm{D} = \frac{2|X \cap Y|}{|X| + |Y|} \ .$$

If $X$ and $Y$ are perfectly overlapping, their Dice score equals 1.

The tumour masks of the individual expert performers were next fused to obtain the gold standard labels. The individual contributions were again compared to the fused result. yielding significantly higher rater-versus-fused Dice scores. This could be expected, since the masks of individual experts are used to obtain the gold standard. This is illustrated in figure 4.1.

| **Expert annotation** | whole | | core | | active |
|---|---|---|---|---|---|
| Dice (in %) | | *LG / HG* | | *LG / HG* | |
| **Rater vs. Rater** | | | | | |
| mean ± std | 85±8 | *84±2 / 88±2* | 75±24 | *67±28 / 93±3* | 74±13 |
| median±mad | 87±6 | *83±1 / 88±3* | 86±11 | *82±7 / 94±3* | 77±9 |
| **Rater vs. Fused** | | | | | |
| mean±std | 91±6 | *92±3 / 93±1* | 86±19 | *80±27 / 96±2* | 85±10 |
| median±mad | 93±3 | *93±3 / 94±1* | 94±5 | *90±6 / 96±2* | 88±7 |

**Figure 4.1:** Inter-rater and rater-versus-fused Dice scores when creating the BraTS 2013 gold standard labels. Here, 10 low-grade (LG) and 20 high-grade (HG) gliomas were manually segmented by four expert raters. Figure adapted from [139], © 2015 IEEE.

For these reasons, a lot of research has been conducted in recent years to automate brain tumour delineation on medical images. Giving a complete overview on the state-of-the-art in automated brain tumour segmentation has almost become infeasible, and is not the goal of this dissertation, but in the following discussion some important examples are given. Pardillo et al. [192] provided us with a good overview on this topic until 2013, but a more recent review paper is unfortunately lacking. However, since 2013 the field has seen an exponential increase in proposed methods, mainly due to two reasons: the annual brain tumour segmentation (BraTS) competition, and the increased performance of computational models such as deep learning. Still, the automated delineation of tumour tissue remains a challenging task due to the large variety in tumour shape, position, appearance, scanning modalities and scanning parameters.

Automated brain tumour segmentation algorithms can be roughly divided into three broad categories, which are in chronological order

generative, discriminative and deep learning approaches.

### 4.1.1 Generative methods

In generative methods, prior knowledge about the appearance of the brain and brain tumours is used to discriminate between healthy and abnormal tissue. For example, in the approach of Khotanlou et al. [193], it is assumed that the healthy brain shows a large degree of symmetry between both hemispheres. Deviations from this symmetry, detected using histogram analysis, can then be attributed to tumour tissue. Afterwards, the segmentation is refined using a deformable model, which can be regarded as deforming an inflated balloon around the tumour, where the deformations are driven by intensity differences between the tumour and its environment. The authors report a mean Dice score of 92% for delineating both full-enhanced, ring-enhanced and non-enhanced cases.

Another generative approach is published by Prastawa et al. [194]. Starting from the approximate location of healthy GM, WM and CSF, the algorithm distinguishes between healthy and abnormal tissue by detecting intensity outliers. The original implementation is based on T1- and T2-weighted MRI images. The algorithm can however be extended to any number and type of scans, as we will demonstrate in section 4.2.

Generative methods have a number of advantages. Since they do not need a pre-trained model, they are easily generalisable as they learn the optimal parameters based on the given images themselves. This means that they can also be used when a large annotated dataset is lacking. However, it is not always possible to encode all the necessary prior information, such as mass effect or the number of tumour tissues.

### 4.1.2 Discriminative methods

Discriminative methods learn to identify the appearance of different tissues based on an annotated training set. This means that such a large set needs to be available in order to yield robust segmentation results. Moreover, thorough tuning of every scan is necessary to match the characteristics of the images in the training set, making discriminative methods more rigid than generative methods. Still, excellent results are

obtained in literature using these algorithms. They generally consist of the same workflow: a preprocessing step is followed by transforming the brain scans into high-dimensional feature vectors for every voxel. These vectors are now fed to a pre-trained machine learning algorithm, such as Random Forests, which calculates voxel-based probabilities to belong to tumour tissue or the healthy brain, from which the tumour masks are deduced. Zikic et al. [195] use local and distant intensity differences between the four MRI-sequences, as well as Gaussian Mixture Model (GMM) based tissue probabilities as input to classification forests. They achieve average Dice scores of 70% and 71% for segmenting oedema and tumour core in high-grade cases, 44% and 62% for segmenting low-grade cases. Menze et al. [196] use channel-specific tumour and tissue probabilities combined with intensity differences and distance measures between voxels of interest as input of a Random Forests classification model. They achieve mean Dice scores of 71% and 70% for segmenting the tumour core and oedema of high-grade glioma, and 23% and 49% in low-grade cases. Festa et al. [197] extract MRI intensities, neighbourhood information, context information and texture features. Their Random Forests approach achieves average Dice scores of 83% and 70% for segmenting the complete abnormal region and the tumour core in high-grade glioma, and 72% and 47% in low-grade glioma. Reza et al. [198] use intensities and intensity differences combined with texture features as input for the Random Forests classification. These authors report average Dice scores of 92% and 91% for segmenting the entire tumour region and the tumour core for both low-grade and high-grade glioma.

In section 4.3, we will demonstrate a similar approach towards automated brain tumour segmentation. This method is only based on a minimal dataset consisting of a T1ce and FLAIR MRI scan.

### 4.1.3 Deep learning

Deep learning using a convolutional neural network (CNN) can be regarded as a discriminative approach, in the sense that it relies on a training set to be able to discriminate between different tissues. However, in contrast to the previously mentioned methods, the algorithm itself de-

termines the optimal local features. Most state-of-the-arts methods are therefore based on CNNs, such as the BraTS 2016 [199] and 2017 [200] competition winners. They achieve Dice scores approaching 90% for segmenting the tumour core and total abnormal region. These methods surpass the inter-rater variability, and are therefore already more robust than manual segmentation.

Although offering superior performance, CNNs only work for images with the exact same characteristics as the training set. Methods based on the BraTS dataset always make use of four MRI sequences: T1, T1ce, T2 and FLAIR. However, in clinical practice, these four images are not always available for every patient. For instance, a T2-weighted scan might be omitted due to timing constraints. We retrospectively collected preoperative structural MRI-scans of 253 patients with low-grade meningioma and astrocytic or oligodendroglial glioma in the Ghent University Hospital. Only 76.3% of the patients received a T2-weighted scan, whereas in 96.9% of the cases a T1ce scan was available. All four scans were collected for 68.9% of the patients, while in 87.9% T1ce and FLAIR was available. In this thesis, we therefore focus on flexible methods regarding the number and type of input scans (section 4.2), or methods where a minimal dataset is necessary (section 4.3).

## 4.2 Flexible segmentation algorithm using outlier detection

In this section, we demonstrate a segmentation algorithm which fulfils three conditions: it is fully automatic, there is no need for annotated training data, and most importantly, the algorithm is flexible regarding the amount and type of input scans. The method will be illustrated and validated on the BraTS 2015 dataset, consisting of 54 low-grade and 220 high-grade glioma patients. The algorithm consists of four steps: 1. preprocessing, 2. estimating the abnormal regions using outlier detection, 3. isolating the largest abnormal region using morphological operations, 4. obtaining the final tumour mask by clustering the neighbouring voxels with similar intensities. Each step is illustrated in figure 4.2 and will

now be explained in more detail.

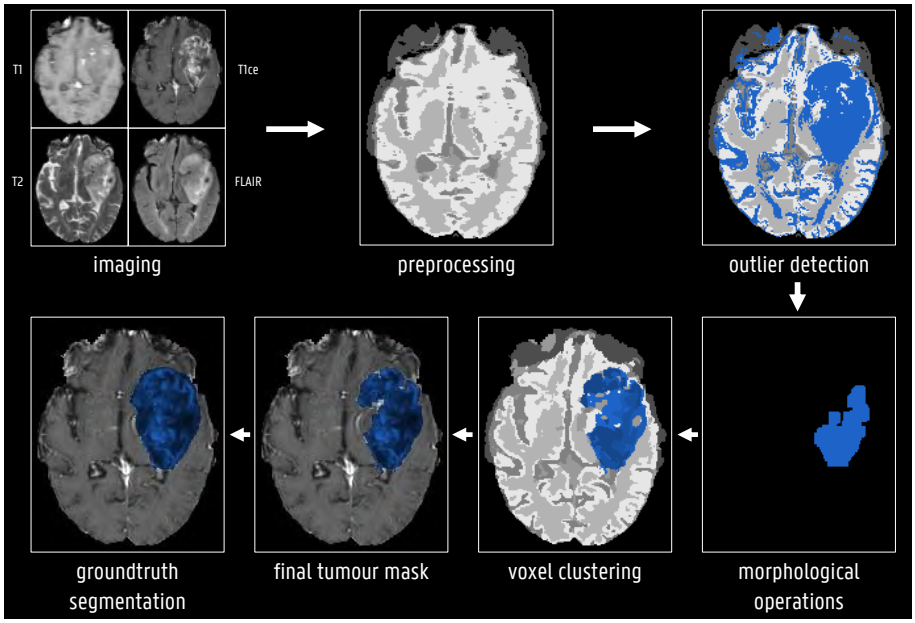



**Figure 4.2:** Illustration of the workflow of the automated segmentation algorithm using outlier detection. In this example, the obtained Dice score was 0.827.

### 4.2.1 Preprocessing

Since the different scans in the BRATS 2015 database are already aligned and interpolated to a uniform $1 \times 1 \times 1$ mm$^3$ voxel size, coregistration of the images is not necessary. For bias field correction and brain tissue segmentation SPM12 (Wellcome Trust Centre for Neuroimaging), running on MATLAB R2017b (The MathWorks Inc., Natick, MA, 2000), is used. This results in bias field corrected images and personalised probability maps for GM, WM, CSF, bone and soft tissue (together: "non-brain") and background or air. By default, SPM12 fits two Gaussian distributions to the CSF-class. We therefore also propose to model the CSF as two different classes, by fitting a Gaussian mixture model (GMM) to the CSF datapoints.

To distinguish between normal and pathological tissue, we also coreg-

ister normal tissue probability maps (available in SPM12) to the T1 scans. This is done in two phases: first, an affine transformation is applied using SPM12. Next, the tissue maps are slightly adapted to the individual anatomy using a non-rigid registration with mutual information, implemented in the Medical Image Registration Toolbox (MIRT [201]) for MATLAB. We used 2 hierarchical levels with an initial mesh window size of 8 voxels and maximum 200 iterations.

### 4.2.2 Outlier Detection

The second and most crucial step is estimating the abnormal regions using outlier detection, based on the method by Prastawa et al. [194]. We consider $N$ image channels (usually $N = 4$ i.e. T1, T2, T1ce, FLAIR; but in general any set of coregistered images can be considered).

#### Detection of normal tissue

First, we construct training samples for the different tissues (GM, WM, CSF, non-brain, outside) by assigning to each voxel the tissue with the highest probability according to the non-rigidly coregistered tissue maps. Next, we determine a subset of normal samples for the different brain tissues (GM, WM, CSF) using the Minimum Covariance Determinant estimator. This algorithm is implemented with a series of C-steps: first, given an $N$-dimensional subset of the samples, calculate the mean $\vec{\mu}$ and covariance $\vec{\Sigma}$; next, calculate for all the voxels belonging to the tissue, the Mahalanobis distance:

$$D(\vec{I}(\vec{x})) = \sqrt{(\vec{I}(\vec{x}) - \vec{\mu})' \, \vec{\Sigma}^{-1} \, (\vec{I}(\vec{x}) - \vec{\mu})} \ ,$$

where $\vec{I}$ is the $N$-dimensional intensity at position $\vec{x}$. Finally, select half of the points, which have the shortest distance, as the new subset. Iterating these steps several times (in our case 10 times), lowers the determinant of covariance. Using different random subsets as starting point, iterating the C-steps and choosing the subset with the minimal determinant of covariance, yields a robust estimate of the healthy positions of GM, WM and CSF. This step is illustrated for the GM and WM classes in figure 4.3.

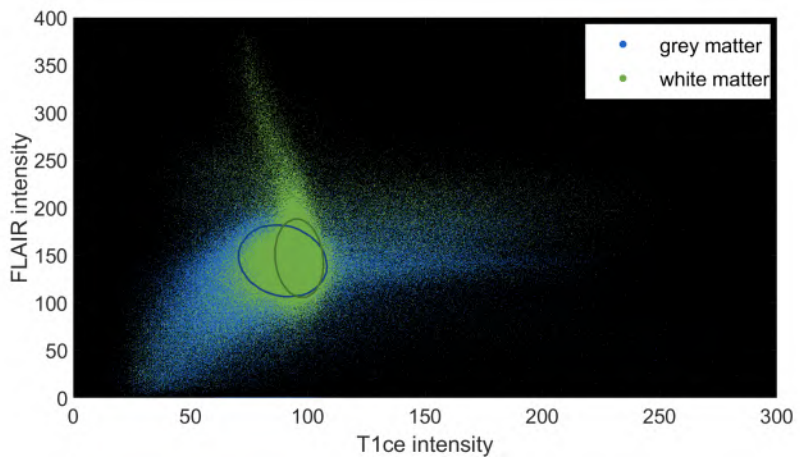



**Figure 4.3:** Illustration of the result of C-steps in two dimensions (intensity on T1ce and FLAIR). All the data points inside the gray and white matter are plotted. The C-steps estimate the ellipses which contain half the data points with minimal variability. These voxels will be used as training points for the healthy tissues.

## Non-parametric model

Next, we use a non-parametric model to determine probability density functions for the classes $\Gamma = \{\text{GM, WM, CSF, non-brain, background, abnormal}\}$, which consists of an Expectation Maximisation loop of three steps. In the first step, voxels belonging to the GM or WM classes that exceed a certain threshold on the Mahalanobis distance to the respective normal tissue subset are assigned to the abnormal class. We chose not to use a fixed threshold, but rather the first local minimum exceeding four standard deviations of the Mahalanobis distance histogram. Secondly, we randomly select a subset of 300 voxels belonging to each class in $\Gamma$ and construct the $300 \times N$-dimensional training sets $\vec{T}_i$ $(i \in \Gamma)$. For every voxel at location $\vec{x}$ with vector of intensities $\vec{I}(\vec{x})$ we now calculate the probability density function for class label $\Gamma_i$:

$$pdf\left(\vec{I}(\vec{x})|\Gamma_i\right) = \underset{j=1\ldots300}{\text{mean}}\left(\frac{1}{(2\pi)^{N/2}}\prod_{n=1}^{N}\frac{1}{\lambda_n}\exp\left(-\frac{(I_n(\vec{x})-T_{i,j,n})^2}{2\lambda_n^2}\right)\right) \quad ,$$

where $\lambda_n$ is chosen as 4% of the intensity range for each channel. The posterior probability for class $\Gamma_i$ is now calculated as:

$$P(\Gamma_i|\vec{I}(\vec{x})) = \frac{pdf\left(\vec{I}(\vec{x})|\Gamma_i\right)\ Pr\left(\Gamma_i, \vec{x}\right)}{\sum_j pdf\left(\vec{I}(\vec{x})|\Gamma_j\right)\ Pr\left(\Gamma_j, \vec{x}\right)}\ ,$$

where the spatial priors $Pr$ are given by the non-rigidly coregistered tissue maps. For the abnormal class, we chose the prior to be the sum of GM and WM, since the tumour usually only occurs in these regions. In the third and last step of each iteration, we assign the class with maximal probability to each voxel. We iterate these steps six times.

### 4.2.3 Morphological Operations

After the abnormality detection, we select the largest abnormal region and assume this is the tumour. Therefore we first construct a mask consisting of all voxels attributed to the abnormal class. We apply an erosion operation of 5 voxels, followed by a dilation with 5 voxels. This removes all loosely-connected parts. Now we consider all the disconnected parts and isolate the cluster with the highest cumulative $P_{\text{tumour}} = P(\Gamma_i = \text{abnormal})$.

### 4.2.4 Voxel Clustering

In the last step, voxels with similar intensities are clustered to yield the final tumour boundaries. This is done in a similar fashion as the construction of fuzzy levels in Hatt et al. [202], combined with the multi-channel approach of Doyle et al. [203]. Therefore we consider the initial segmentation by SPM12 as input. Voxels that are assumed to be tumour after the morphological operations are initially clustered in four classes using $k$-means clustering. These classes can for a high-grade tumour represent enhancing core, non-enhancing core, necrosis and oedema. This results in a total of nine classes: GM, WM, CSF, non-brain, background and four tumour classes. These classes are now iteratively improved assuming a multivariate Gaussian distribution for each class and taking into account the local neighbourhood. In each

iteration four steps are performed. First, for each class $\Gamma_i$, calculate the mean $\vec{\mu}_i$ and covariance-matrix $\vec{\Sigma}_i$. Next, compute the probability density function for each class in each voxel:

$$pdf\left(\vec{I}(\vec{x})|\Gamma_i\right) = (2\pi)^{-N/2}\left\|\vec{\Sigma}_i\right\|^{-1/2}\exp\left(-\frac{1}{2}(\vec{I}(\vec{x})-\vec{\mu}_i)'\,\vec{\Sigma}_i^{-1}\,(\vec{I}(\vec{x})-\vec{\mu}_i)\right) \quad .$$

In the third step, we incorporate the prior information using the neighbourhood for every voxel. For every class $\Gamma_i$ and voxel at position $\vec{x}$, we calculate the linear neighbourhood function $Nb(\Gamma_i, \vec{x})$ which is zero if none of the 26 surrounding voxels were assigned the same class in the previous iteration, and equals 1 if all neighbouring voxels have class $\Gamma_i$. In the last step, the posterior probability is calculated:

$$P(\Gamma_i|\vec{I}(\vec{x})) = \frac{pdf(\vec{I}(\vec{x})|\Gamma_i)\ Nb(\Gamma_i,\vec{x})}{\sum_j pdf(\vec{I}(\vec{x})|\Gamma_j)\ Nb(\Gamma_j,\vec{x})} \quad ,$$

and to each voxel the class with maximal probability is attributed. After 20 iterations the largest connected region with tumour labels is chosen as the final tumour segmentation.

### 4.2.5 Results

We validated this method on the BRATS 2015 dataset for different combinations of input images: all images (T1, T2, T1ce, FLAIR) and combinations of three images (without T2, T1ce or FLAIR). The results of these analyses, expressed in Dice scores, are given in the boxplots of figure 4.4. The corresponding median Dice scores are 73.3% if all images are considered, 56.2% for all images without T2, 64.4% for all images without T1ce, and 65.8% for all images without FLAIR. The algorithm completely failed to detect the tumour (Dice score lower than 1%) in 5 out of 274 cases when all scans are considered, in 19 cases for all scans without T2 or T1ce, and in 26 cases for all images without FLAIR.

### 4.2.6 Examples

Apart from the quantitative validation, we can also assess the performance in a qualitative way. Therefore, in figure 4.5, the result of the

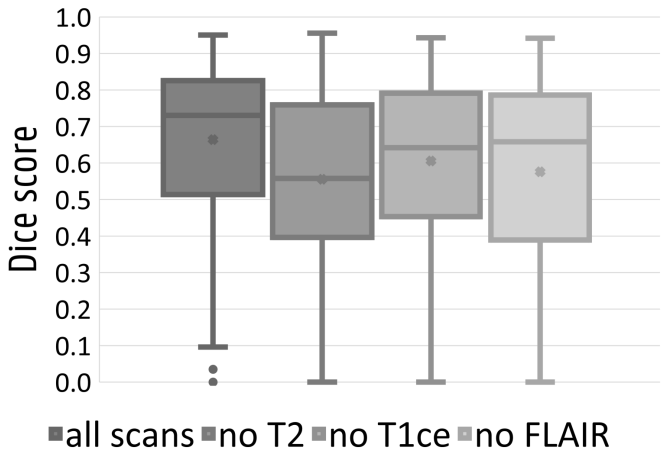


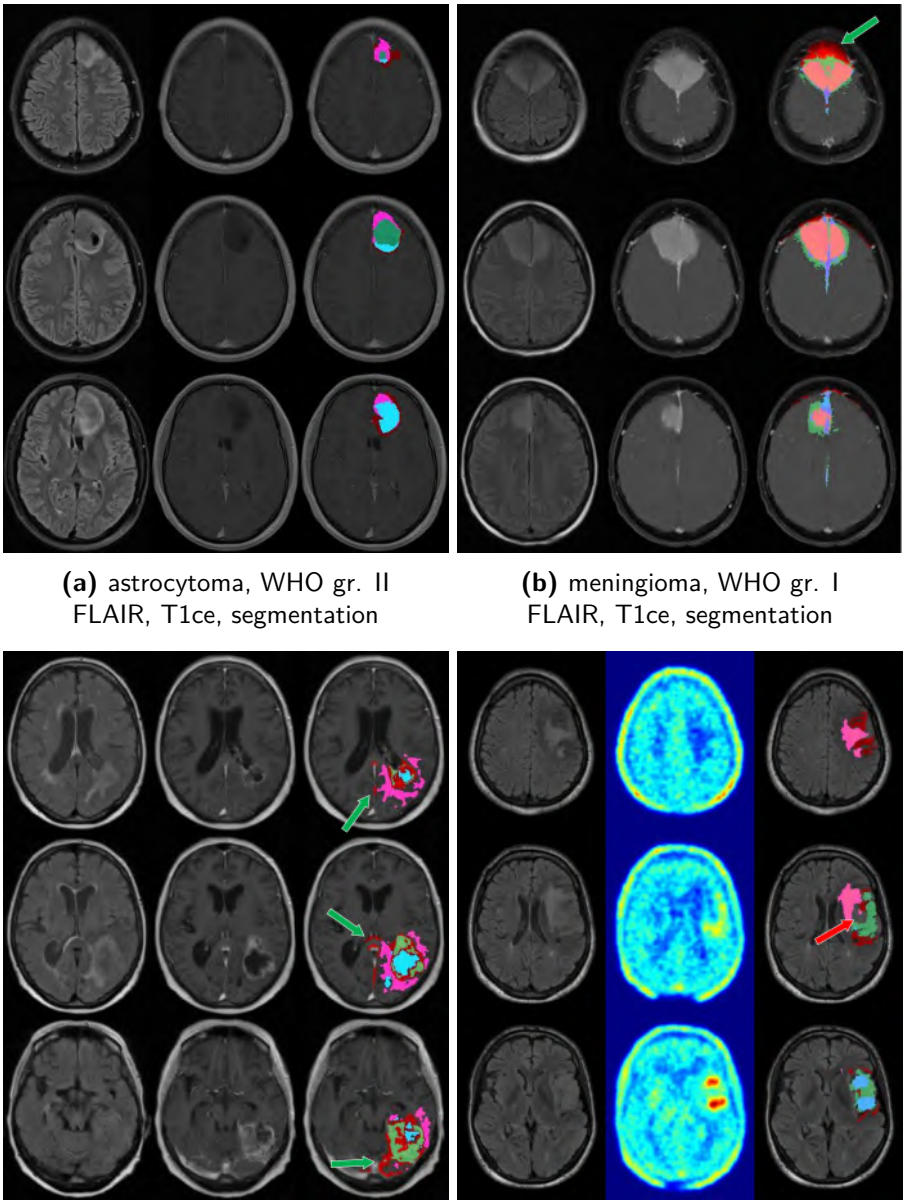

**Figure 4.4:** Results obtained on the BraTS 2015 dataset using the outlier detection based segmentation algorithm.

segmentation algorithm for four different pathologies is illustrated. In this figure, the resulting classes from the voxel clustering part are displayed in different colours, which can however not be directly linked to specific tumour tissues. From these images, it is clear that the tumour is well delineated, although both under-segmentation (i.e. tumour classified as healthy tumour) as over-segmentation (i.e. healthy tissue classified as tumour) are present.

### 4.2.7 Discussion

In this study, we develop a segmentation algorithm that is fully automatic, and provides reproducible results without the need for a large training dataset. More importantly, it is very flexible regarding the number and type of input images, as we have shown with four different set-ups. We achieved median Dice scores around 70%, which is comparable to other generative methods validated on the same dataset [204, 205].

The main drawback of this method is the strong dependency on the initial segmentation into healthy tissues. This especially is problematic when a large tumour causes the ventricles to shift, or in the case of



**(a)** astrocytoma, WHO gr. II
FLAIR, T1ce, segmentation

**(b)** meningioma, WHO gr. I
FLAIR, T1ce, segmentation

**(c)** glioblastoma
FLAIR, T1ce, segmentation

**(d)** oligodendroglioma, WHO gr. III;
FLAIR, $^{18}$F-FET PET, segmentation

**Figure 4.5:** Qualitative results of the segmentation algorithm, performed on scans showing different pathologies. The different colours have no direct meaning, apart from being clusters with a similar intensity profile. Regions with under-segmentation are indicated with a red arrow, regions with over-segmentation with a green arrow.

expanded ventricles due to CSF-obstruction. When the deviation from the normal anatomy is large, the TPMs obtained with SPM12 become unreliable. Conversely, when a large fluid-filled part of the tumour is present next to the ventricles, SPM12 will consider this structure to be part of the ventricles. In these cases, it is very hard for the outlier detection method to distinguish between healthy and abnormal tissue, since normal intensities will be present in abnormal regions and vice versa. However, if the TPMs can be accurately estimated, our method can easily achieve Dice scores of 80% and higher.

In chapter 3, we showed that features calculated on subregions of the tumour can accurately determine the tumour grade. However, our implementation of the segmentation method based on outlier detection yields only a single tumour mask comprising the entire abnormal region. During the last step of the algorithm, tumour voxels with a similar intensity profile are clustered into four classes. But since the method is not trained on an annotated dataset, it is impossible to assign a specific tumorous tissue to every class.

Moreover, compared to pre-trained machine learning algorithms this approach is quite slow. The outlier detection, morphological operations and voxel clustering steps finish in about 15 minutes per patient, but the preprocessing steps, especially the non-rigid coregistration with MIRT, take up to 40 minutes per patient. The entire pipeline was implemented in Matlab R2017b, running on a Intel Xeon CPU E5620 with 4 cores, 2.40GHz and 64.0 GB of installed physical memory. Because of these limitations, a new approach towards brain tumour segmentation is implemented, as will be explained in the following section.

## 4.3 Segmentation based on local texture and abnormality features

In this section, the goal is to implement an automated brain tumour segmentation algorithm able to delineate multiple tumour tissues on a minimal dataset. In our case, this is a T1ce and FLAIR scan, since these MRI sequences provide complementary information and are simultaneously available for many patients in the Ghent University Hospital. This study was published as "Machine learning based brain tumour segmentation on limited data using local texture and abnormality", *Computers in Biology and Medicine* 98 (2018): 39-47 [206].

### 4.3.1 Principle and implementation

The workflow of our segmentation method consist of the following steps: preprocessing of the images, feature calculation, Random Forests classification and voxel clustering, as is shown in figure 4.6. First, we describe our training and validation data. Next, each of these steps in the segmentation process will be explained in more detail.

### 4.3.2 Training and validation data

#### Training and validation set

The BraTS 2013 dataset [139] is used for optimising and training our model. This dataset can freely be downloaded from the Virtual Skeleton Database [207] and consists of 10 lower-grade and 20 high-grade glioma patients. The larger BraTS 2017 dataset is used for validating the method. In this collection, 75 lower-grade and 210 high-grade patients are included. These datasets are already discussed in detail in section 3.2.2.

#### Ghent University Hospital database

In the BraTS databases, only astrocytomas and oligodendrogliomas with WHO grade II, III or IV (glioblastomas) are included. However, we

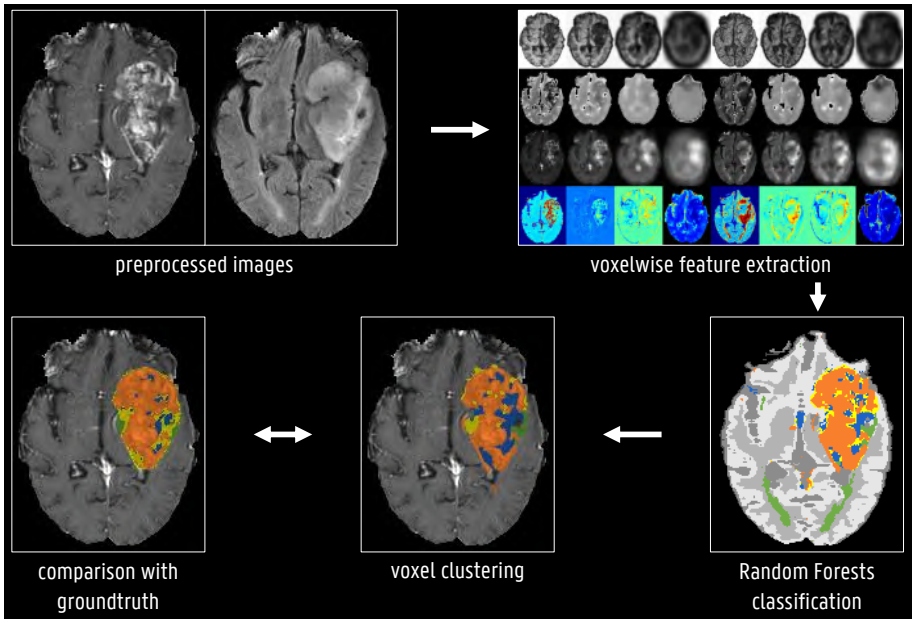



**Figure 4.6:** Illustration of the workflow of our method. A detailed illustration of the voxelwise texture and abnormality features is given in figure 4.7.

also want to perform our segmentation method on other primary brain tumour types, such as meningioma, ependymoma or medulloblastoma. Therefore, we retrospectively collected preoperative brain MRI from 257 patients in our centre. This was done with permission from the local ethics committee, and informed consent was waived (Belgian registration number B670201524727 2015/0521).

### 4.3.3 Preprocessing

The BraTS images are already coregistered and resliced. To mimic these preprocessing steps for the scans acquired in our centre, SPM12 running on MATLAB R2017b is used for co-registering the FLAIR image to the T1ce scan of the same patient. Next, the scans are spatially normalised to MNI-space (Montreal Neurological Institute [208]) and trilinearly interpolated to a $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$ voxel size. Furthermore, for both the BraTS and Ghent University Hospital scans, the segmentation mod-

ule in SPM12 is applied to the T1ce scan to calculate tissue probability maps (TPMs) for five healthy tissues (GM, WM, CSF, skull and soft tissue). This technique is based on a Gaussian mixture model and prior spatial probabilities. These TPMs will later on be used as normality features. During the SPM segmentation, bias field correction is applied to correct for magnetic field inhomogeneities, and the corrected images are also saved and used for all following analyses.

### 4.3.4 Feature extraction

For every patient, 275 feature maps are calculated based on the T1ce and coregistered FLAIR scans, showing the local value for every voxel of a certain textural or (ab)normality property. To capture the local texture, 30 features are calculated on both the T1ce and the FLAIR scan and on four different spatial scales, contributing a total of 240 features. Next, there are 5 normality features capturing the healthy regions of the brain, and 30 abnormality features showing the deviation from normality. An overview of all image features is given in Appendix B. A graphical illustration of some of these is also given in figure 4.7.

#### Texture features

We calculate three types of texture parameters for a total of 30 different features. Every feature contains for a certain voxel information from the $3 \times 3 \times 3$ voxels environment surrounding this voxel. To account for distant interactions, the scans are also downsized with a factor 2, 4 or 8 using MATLAB's `imresize3` function. The same texture features are again calculated on the smaller images, followed by upscaling with cubic interpolation to the original matrix size. For the calculation of the texture parameters, we first discretise all images to 64 grey-levels using

$$I_{\text{discr}} = \left\lceil (N_g - 1) \frac{I - \min(I)}{\max(I) - \min(I) + 1} \right\rceil \ ,$$

where $I$ is the original image, $I_{\text{discr}}$ the discretised image and $N_g$ is the discretisation level, here chosen to be 64. This discretised image is chosen as the first, and most simple, texture map.

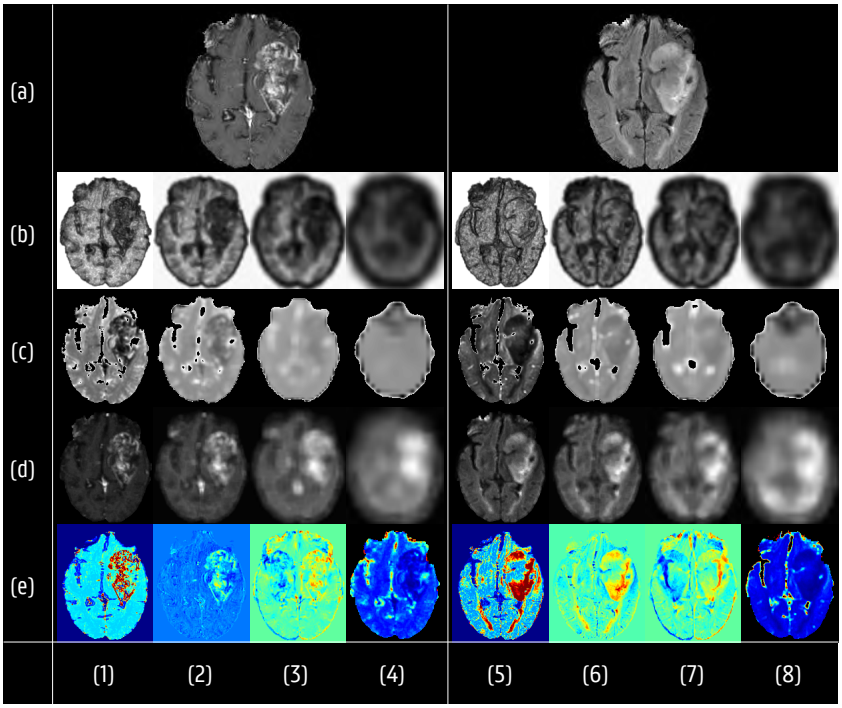



**Figure 4.7:** Illustration of the transformation of the scans into feature maps. (a) original images (left: T1ce, right: FLAIR); (b) homogeneity; (c) low grey-level run emphasis (LGLRE); (d) high grey-level zones emphasis (ZnHiGL); (e) abnormality features. For the texture features: (1,5) downsizing level (DS) 1; (2,6) DS 2; (3,7) DS 4; (4,8) DS 8. For the abnormality features: (1,5) $P_{tumour}$; (2,6) Z-map; (3,7) symmetry; (4,8) abnormal zones with low grey-levels (aZnLoGL).

GLCM FEATURES The grey-level co-occurrence matrix (GLCM) [183] describes the occurrence of pairs of voxel intensities. We only consider a distance of 1 voxel to determine a voxel pair. More distant interactions are accounted for using the downsizing step. In 3D, voxel pairs can be determined in 13 directions. We determine 9 GLCM-based features, which are the averages over these 13 directions: autocorrelation, cluster tendency, correlation, dissimilarity, energy, homogeneity, maximum probability, sum average and variance, according to the definitions in [144].

GLRLM FEATURES The grey-level run-length matrix (GLRLM) quantifies one-dimensional runs, being a set of consecutive, collinear voxels having the same grey level, in the image [184]. Again, these runs can be calculated in 13 different directions, such that the GLRLM-based features are the averages over these 13 directions. We calculate 10 features: short/long run emphasis, grey level non-uniformity, run-length non-uniformity, low/high level run emphasis and short/long run with low/high grey levels emphasis, based on the definitions from [144].

GLSZM FEATURES The grey-level size-zone matrix (GLSZM) quantifies zones or clusters of a certain grey level in the image, and is therefore independent of direction [185]. Again, we calculate 10 features following the definitions in [209]: small/large zone emphasis, grey-level non-uniformity, size-zone non-uniformity, low/high grey-level emphasis and small/large zones with low/high grey-levels emphasis.

### Normality and abnormality features

Next to the texture features, where only local information is included, we also include normality and abnormality features. In this way anatomical information can be incorporated in the model.

TISSUE PROBABILITY MAPS The five TPMs calculated during the segmentation step in SPM12 give the probability for every voxel to belong to GM, WM, CSF, skull or soft tissue using prior anatomical probabilities and assuming a healthy intensity distribution. These TPMs can therefore identify normal appearing regions in the brain.

ABNORMALITY FEATURES An MRI scan presenting a brain tumour will in general show strong deviations from the normal appearing intensities. Therefore, we also calculate abnormality maps starting from the TPMs. These include five probability maps for GM, WM, CSF, non-brain regions and tumour using outlier detection [194], as explained in section 4.2.2. Moreover, we include Z-maps for T1ce and FLAIR: the image $I$ is divided into GM, WM and CSF according to the maximal probability in the TPMs, and the Z-score $Z = (I - \mu)/\sigma$ is calculated, where $\mu$ and $\sigma$ are the mean and average values in the respective tissues.

Similarly, we look for abnormal zones of low or high grey-levels in T1ce and FLAIR. Here, the starting point is the GLSZM-features low/high grey-level emphasis. The parameters $\mu$ and $\sigma$ are chosen as the mean and standard deviation of the largest peak in the feature histograms in GM or WM. In this way we obtain six more maps: abnormal zones of low/high grey-levels compared to normal appearing GM/WM based on FLAIR, and abnormal zones of low grey-levels compared to normal appearing GM/WM based on T1ce. We do not calculate abnormal zones of high grey-levels based on T1ce, since the highest intensities in this image will in general belong to either blood vessels or heterogeneous contrast enhancing tumour tissue. In both cases, we do not expect to see large zones of high intensities on T1ce.

The aforementioned features are all based on either the T1ce or FLAIR scan. We can however also use information from both scans simultaneously to calculate six more abnormality features. These are the probability maps for GM, WM, CSF, non-brain regions and tumour using outlier detection [194], and the multivariate distance in different tissues:

$$d = \sqrt{\left(\frac{I_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{I_2 - \mu_2}{\sigma_2}\right)^2} \ ,$$

where the indices 1 and 2 refer to the T1ce and FLAIR scan respectively.

Symmetry The healthy brain shows a large degree of symmetry, and deviations from this symmetry can therefore be a marker of abnormality, as is used in several brain tumour segmentation approaches [193, 210]. We use a robust intensity-based method to estimate the midsagittal plane [211]. Next, we calculate three symmetry-based features on both T1ce and FLAIR scans: the original intensity difference between the image and the mirrored version, the intensity difference after intensity discretisation, and after intensity discretisation and downsizing with a factor 4.

### 4.3.5 Random forests classification

The goal of the Random Forests classification algorithm is to estimate the probability of a voxel belonging to a tissue type based on the calcu-

lated features in that voxel. We consider 9 different classes, divided into 5 normal and 4 tumour types: GM, WM, CSF, non-brain, background, necrosis, oedema, non-enhancing tumour and enhancing tumour. During the training phase, 1000 voxels per tissue (or if less voxels are present in a certain tissue, we chose the size of the smallest class) are randomly selected per patient in the training set, and the feature values are stored together with the corresponding tissue class. The training matrix is balanced since there is an equal amount of training samples for every class, which implies that there is no bias towards a certain class when predicting an unknown voxel.

Since calculating and storing 275 feature maps is both time and memory consuming, and to reduce overfitting, we try to find an optimal subset of features. For this, we apply sequential forward selection (SFS) using three-fold cross-validation on the training set. This algorithm starts from an empty feature set, and predictors are sequentially added to the model until no further improvement is obtained. As indicator for the model performance, we apply two criteria: total accuracy over all classes, and accuracy of the tumour classes. Finally, we combine these two feature sets in the final model.

Next to feature selection, the number of trees and tree depth influence the model performance. As we use MATLAB's `TreeBagger` implementation of Random Forests, the tree depth can indirectly be controlled using the `MinLeafSize` option, being the minimal number of samples in every leaf. We empirically find that a `MinLeafSize` equal to 20 yields the best results in our model. We use 100 trees per forest, since we found that increasing the forest size above 100 trees does not improve the classification accuracy.

### 4.3.6 Post processing

After estimating the tissue probabilities using the Random Forests model, voxels are assigned to the tissue with highest probability. The final tumour masks are obtained using the morphological operations explained in section 4.2.3 and the voxel clustering explained in section 4.2.4.

### 4.3.7 Results

#### Feature selection

Our final model takes into account 52 features after feature reduction. There are 3 TPM's (for GM, CSF and non-brain), 16 FLAIR texture features, 7 FLAIR abnormality features, 19 T1ce texture features, 5 T1ce abnormality features and 2 combined T1ce/FLAIR abnormality features. The final features are depicted in bold in Appendix B.

#### Random Forests performance

In figure 4.8, the performance of the Random Forests model is graphically illustrated using a confusion matrix, obtained using three-fold cross-validation. It is clear that for healthy tissues, there is a high probability of being correctly classified, with a minimal accuracy of 90.0% for GM. This performance decreases however for tumour tissues, with accuracies of 35.0%, 47.8%, 23.6% and 61.3% for necrosis, oedema, non-enhancing tumour and enhancing tumour tissue, respectively. It is clear that to improve the segmentation accuracy a dedicated post-processing step is required.

#### Segmentation result

We applied the segmentation procedure, including preprocessing, feature extraction, Random Forests classification and post processing, to all scans in the training and test set. The results for different segmentation tasks is given in table 4.1.

**Table 4.1:** Obtained Dice scores on the BraTS 2013 training set and the BraTS 2017 test set for different tumour tissues.

| | | ncr | oed | n-enh | oed+n-enh | ncr+n-enh | enh | core | total |
|---|---|---|---|---|---|---|---|---|---|
| BraTS 2013 low-grade | median | 0.0% | 61.7% | 55.7% | 77.7% | 75.2% | 0.0% | 75.0% | 80.2% |
| | average | 9.8% | 61.3% | 62.5% | 75.4% | 67.0% | 18.9% | 66.6% | 78.4% |
| BraTS 2013 high-grade | median | 71.2% | 73.5% | 28.2% | 75.0% | 59.6% | 83.5% | 83.3% | 85.8% |
| | average | 68.7% | 73.3% | 25.7% | 75.6% | 59.5% | 81.2% | 84.4% | 86.1% |
| BraTS 2017 low-grade | median | 12.2% | 45.7% | | 45.1% | 29.2% | 13.5% | 40.9% | 68.4% |
| | average | 17.3% | 46.2% | | 44.9% | 30.8% | 24.8% | 40.1% | 65.6% |
| BraTS 2017 high-grade | median | 52.9% | 66.1% | | 65.9% | 47.1% | 74.8% | 75.0% | 80.1% |
| | average | 46.8% | 61.0% | | 61.5% | 42.1% | 70.4% | 69.0% | 76.2% |

ncr = necrosis, oed = oedema, n-enh = non-enhancing tumour, enh = enhancing tumour, core = tumour core, total = total abnormal region

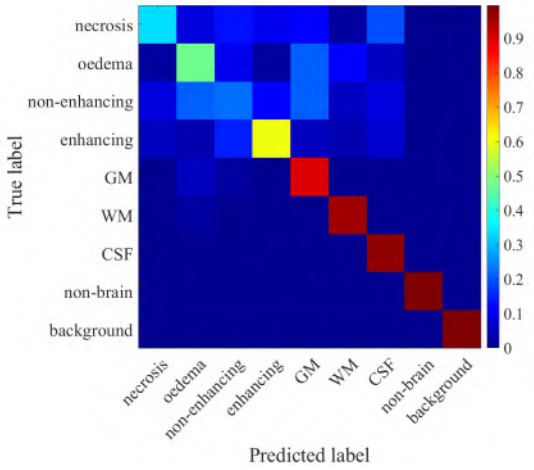


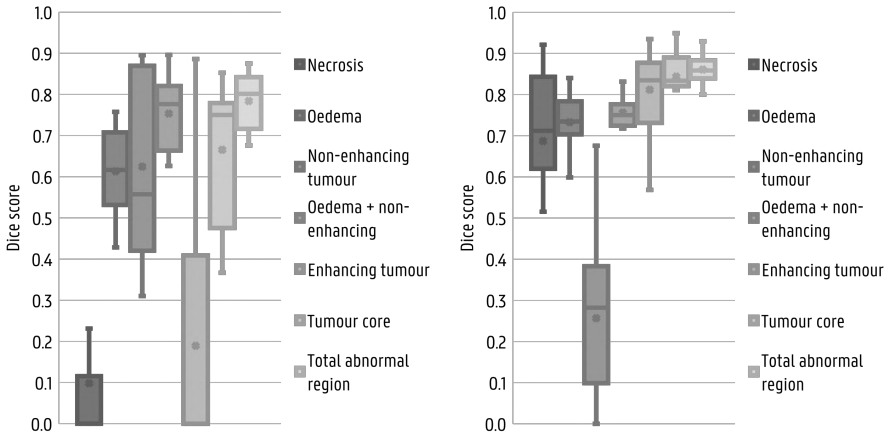

**Figure 4.8:** Illustration of the performance of the Random Forests model, using 3-fold cross-validation on the training set. This confusion matrix shows the probability of correct (diagonal) and incorrect (off-diagonal) predictions.

## Training data

First, we evaluate the final model on the 30 patients in the BraTS 2013 training set. This gives us an upper boundary of the performance of the technique, since a model will in general perform better on the dataset on which it was trained than on independent data. The obtained Dice scores in several tumour compartments are given in figure 4.9.

TUMOUR CORE AND ENTIRE ABNORMAL REGION For low-grade glioma, we obtain median Dice scores of 75.0% and 80.2% for segmenting the tumour core and the entire tumour region, respectively. For high-grade glioma, these scores are 83.3% and 85.8%.

TUMOUR SUB-REGIONS For low-grade glioma, good scores are obtained for segmenting the combination of oedema and non-enhancing tumour tissue, with a median Dice score of 77.7%. For necrosis and enhancing tissue however, the method does not perform well. This can be explained by acknowledging that low-grade glioma in general do not present necrosis nor contrast enhancement. For high-grade glioma, we obtain median



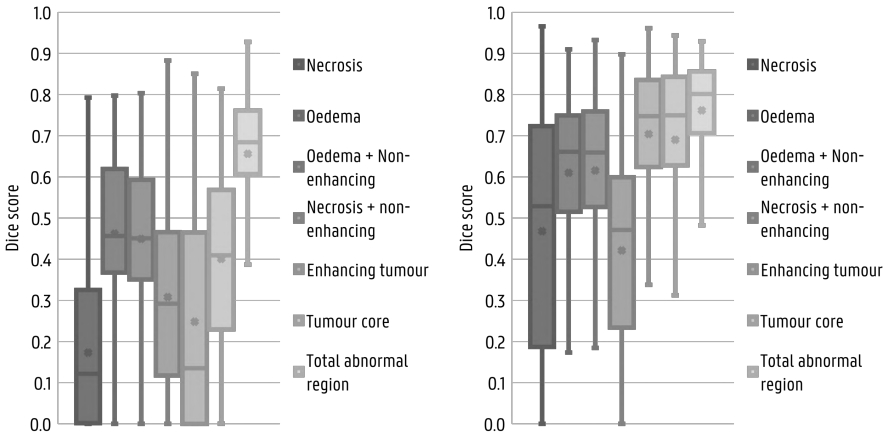

**(a)** Low-grade glioma **(b)** High-grade glioma

**Figure 4.9:** Dice scores obtained on the BraTS 2013 dataset used for training the model.

Dice scores of 73.5%, 83.5% and 75.0% for oedema, enhancing tissue and the combination of oedema and non-enhancing tissue, respectively. The worst results are obtained for segmenting the non-enhancing tumour, with a median Dice score of 28.2%. This can be expected from the Random Forests performance (see figure 4.8), where we also observe poor results for non-enhancing tumour tissue.

## Test data

The manual segmentation masks in the BraTS 2017 dataset do not longer contain separate labels for non-enhancing tumour. These voxels are combined with the necrotic region. Being trained on the BraTS 2013 dataset, our model will however still predict non-enhancing tumour voxels, which is why we choose to calculate Dice scores on the combination oedema + non-enhancing tumour and necrosis + non-enhancing tumour. The obtained Dice scores of the segmentation result on the BraTS 2017 dataset are graphically given in figure 4.10. As can be expected, the method does not perform as well on this dataset compared to the training set.




**(a)** Low-grade glioma **(b)** High-grade glioma

**Figure 4.10:** Dice scores obtained on the BraTS 2017 dataset used for testing the model.

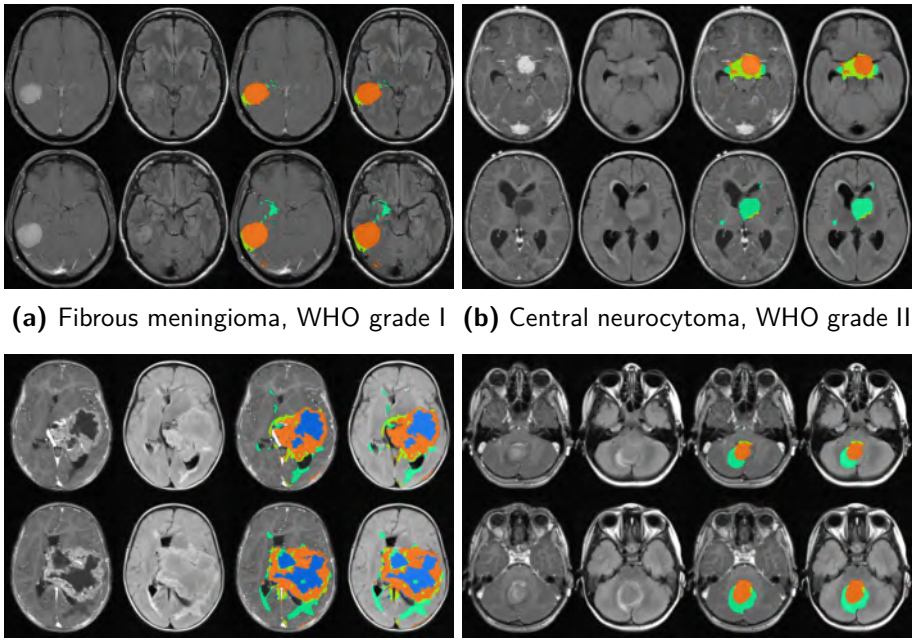
Tumour core and entire abnormal region The median scores scores for low-grade glioma are 40.9% and 68.4% for the tumour core and total tumour region, respectively. These values increase to 75.0% and 80.1% in high-grade cases.

Tumour sub-regions We discover unsatisfying results for segmenting separate tumour tissues in low-grade glioma, with no median Dice scores exceeding 50%. Again, the algorithm performs better for high-grade glioma. We obtain median Dice scores for oedema, enhancing tissue and the combination of oedema and non-enhancing tissue of 66.1%, 74.8% and 65.9%, respectively.

## Clinical scans

To conclude the results section, we illustrate the segmentation performance obtained on scans of other than astrocytic and oligodendroglial tumours collected in our institution. This shows the versatility of our approach, since the model is only trained on these types of primary brain tumours. In figure 4.11 we show results for four non-astrocytic and non-oligodendroglial tumour types. We do not have a manual de-

lineations for these tumours to compare with, so we can only gauge the segmentation result in a qualitative way.



**(a)** Fibrous meningioma, WHO grade I **(b)** Central neurocytoma, WHO grade II

**(c)** Anaplastic ependymoma, WHO grade III **(d)** Medulloblastoma, WHO grade IV

**Figure 4.11:** Illustration of the segmentation performance on four non-astrocytic and non-oligodendroglial primary brain tumours. For every patient a T1ce MRI, FLAIR MRI, segmentation result fused with T1ce and segmentation result fused with FLAIR are shown on two different slices. Colour code: orange = contrast-enhancing tumour, yellow = non-contrast enhancing tumour, green = oedema, blue = necrosis.

In general, we can see that most tumour tissues are well delineated. However, despite the morphological operations, still isolated healthy regions are marked as tumour, such as the enhancing blood vessels posterior in the ependymoma case, or oedema in the right medial temporal lobe in the meningioma case. On the other hand, secondary tumour regions not connected to the main tumour will not get segmented due to these morphological operations, such as the small enhancing lesion in the cerebellum of the neurocytoma case. Moreover, some tumour labels get

mixed up, such as the oedema spot in the center of the necrotic voxels in the ependymoma case. For this patient, the FLAIR-hyperintensities posterior to the lesion, as well as adjacent to the right frontal and right occipital horn of the lateral ventricles, might also be due to infiltrating low grade tumour. Therefore, for further research, we chose to combine the oedema and non-enhancing tumour labels into one class.

### 4.3.8 Discussion

In this study, we have evaluated a new Random Forests based segmentation method for delineating different brain tumour compartments, starting from texture and abnormality features on contrast-enhanced T1-weighted and FLAIR MRI. To the best of our knowledge, no brain tumour segmentation method is available using only these two MRI sequences. The complementary of T1- and T2-weighted scans for brain tumour segmentation has been shown in several studies. Prastawa et al. [194] use outlier detection on T1 and T2 scans. Liu et al. [212] use a fuzzy connectedness framework on T1, T1ce and FLAIR scans. Iftekharuddin et al. [213] segment pediatric tumors on T1ce, T2 and FLAIR using texture and image features in a Self-Organizing Map. A combination of T1 and FLAIR scans has also been used for delineating other brain lesions, such as white matter lesions [45, 214].

We start from an initial feature set of 275 features, but after sequential forward selection, only 52 are incorporated in the final model. We can expect the texture features to mainly help in detecting heterogeneous structures such as enhancing tissue, whereas the abnormality features provide information on homogeneous regions such as low-grade tumours and oedema. We incorporate distant interactions by downsizing the images, calculating texture features and again interpolating these maps. However, no features of subsample level 2 are present in the final model, which suggests they do not provide added value over the local features and downsampled images with factor 4 or 8. Only one FLAIR texture feature is present on the local scale, whereas 14 FLAIR texture features from downsizing scales 4 and 8 are incorporated, suggesting that the local texture on FLAIR contains less information than local texture on T1ce. This might be explained by the lower resolution

of the T2-weighted FLAIR scan. Moreover, neither the original MRI scans nor the discretised versions are included. MRI scans are recorded in arbitrary units, such that image intensities cannot be directly compared. Previous methods often use single scan intensities or intensity differences between different scans as features [195, 196, 197, 198], but this requires a robust intensity normalisation step such as histogram matching or white stripe normalisation [138].

We obtain similar results as the previously mentioned studies, with Dice scores around 75% for segmenting the tumour core and approaching 80% for segmenting the whole tumour region of high-grade tumours. However, we train our method only on two MRI sequences, being T1ce and FLAIR. Moreover, we have only one model for segmenting both low-grade and high-grade tumours, whereas previous studies often train a separate model on low-grade cases. In clinical practice, it will however not always be known a priori what tumour type is analysed. Moreover, we applied our method to other than astrocytic or oligodendroglial brain tumour types. A qualitative analysis shows satisfying results, such that the method can be used for advanced image processing techniques such as radiomics or radiogenomics.

State-of-the-art methods on brain tumour segmentation mostly use deep learning approaches such as CNNs [215, 216, 217]. Many of these studies obtain Dice-scores approaching 90%, using four MRI sequences (T1, T1ce, T2, FLAIR). These methods require a huge annotated data set, as well as powerful hardware for training. CNNs take an image as input and return a label or segmentation mask as output. In between is a complex network of hidden layers. These can be convolutional, pooling, activation or fully connected layers, with a large number of weights that need to be tuned during training. Deep learning for segmentation can roughly be divided into two different approaches. In the first approach, features are extracted from a local patch for every voxel using convolutional layers. These features are then classified with a fully connected neural network to obtain a label for every voxel. The second approach uses fully convolutional networks such as U-Net [218], where the local information is incorporated using up- and downsizing steps. Our method applies similar steps as the first approach, albeit with hand-engineered features: the texture features simulate the convolutional transforma-

tions, we use downsizing of the images instead of pooling layers, and Random Forests is used as a non-linear classifier instead of the fully-connected layers. This enables a less complex training scheme, such that a lower amount of training data is necessary.

Visual inspection of our results often shows oversegmentation, being healthy tissue assigned a tumour label, which causes the Dice score to decrease. A more advanced post-processing step, for instance able to remove enhancing blood vessels, might improve the delineation. Furthermore, we see a lower performance on the test set compared to the training set. This shows overfitting of the Random Forests, although measures have been taken to avoid this, such as feature selection and limiting the tree depth.

Furthermore, our method might be improved by modelling the larger spatial context of a single voxel in a more advanced way. Now we calculate texture features on downsized versions of the image, but in this way some information is lost. Calculating the texture matrices on a larger neighbourhood of every voxel (e.g. $5 \times 5 \times 5$ or $7 \times 7 \times 7$ voxels) might yield better results. However, this was not considered due to the increased computational complexity. Another limitation of our method is the strong dependency on the SPM segmentation when calculating the abnormality features. In large tumours or when a large degree of mass effect is present, this might give rise to unsatisfying results. However, as we see a better performance in high-grade tumours, this effect is limited.

## 4.4 Conclusion

In this chapter, we discussed different approaches towards the automated segmentation of brain tumours on medical images. A first method, based on outlier detection, is flexible regarding the number and type of input images, but yields only a single tumour mask for the entire abnormal region. A second method, where texture and abnormality maps are used as input for a Random Forests classification method, requires a minimal dataset of a T1ce and FLAIR MRI and returns tissue maps for different tumour subcompartments. A third approach using deep learning was also mentioned, but this method was not implemented in the course of

this thesis. The Random Forests based segmentation algorithm will in the following chapters be applied to clinical scans in several radiomics problems.

# 5 The multiclass problem of primary brain tumour diagnosis

In chapter 3 we discussed a binary classification problem for a dataset where manual tumour segmentation labels were available. In clinical practice however, we come across more difficult situations, as the tumour delineation is not routinely performed for all patients. Therefore, we will apply the automated Random Forests based segmentation algorithm developed in chapter 4 to clinical scans. Moreover, classifying patients into lower-grade or high-grade glioma is not sufficient to determine the optimal treatment strategy, as was discussed in the introduction. Primary brain tumours are a heterogeneous group of neoplasms, and in order to obtain a more detailed diagnosis, we will here discuss the multiclass problem.

This study has been presented during the 2017 IEEE Nuclear Science Symposium and Medical Imaging Conference [219].

## 5.1 The importance of the multiclass problem

Few studies in literature perform a multiclass classification of primary brain tumours based on medical imaging. The most important articles are listed in table 5.1.

Herlidou-Meme et al. [220] applied texture analysis on 2D slices from T1 and T2 weighted MRI, aiming to differentiate between several

**Table 5.1:** Overview of the literature on multiclass classification of brain tumours.

| Author | Task | Images | Segmentation | Method | Result |
|---|---|---|---|---|---|
| Herlidou-Meme et al. (2003) [220] | Discriminating texture of healthy brain, solid tumour, kystic/necrotic tumour, whole tumour and oedema in healthy volunteers (n=10) and meningioma (n=15), lymphoma (n=10) and glioma (n=38) patients | T1, T2 | Manual | Correspondence factorial analysis, hierarchical ascending classification | no discrimination between different types of tumours |
| Luts et al. (2007) [224] | Classifying voxels of 4 healthy volunteers and 25 patients, divided into 10 classes | T1, T1ce, T2, PD, MRSI | Manual (selection of voxels) | Support vector machine, linear discriminant analysis | 98.3% overall accuracy |
| Georgiadis et al. (2008) [221] | Distinguishing metastases (n=21), meningiomas (n=19), gliomas (n=27) | T1ce | Manual | Probabilistic neural network | 71.4% accuracy (metastases), 72.2% accuracy (gliomas), 81.3% accuracy (meningiomas) |
| García-Gómez et al. (2009) [225] | Distinguishing between glioblastoma ($n = 84 + 28$), meningioma ($n = 57 + 17$), metastasis ($n = 37 + 32$) and LGG ($n = 33 + 20$) | MRS | Manual (selection of voxels) | Ten classification methods | Accuracies of 78 - 94% for binary problems in independent test set |
| Zacharaki et al. (2009) [175] | Distinguishing metastases (n=24), meningiomas (n=4), grade II gliomas (n=22), grade III gliomas (n=18) and glioblastomas (n=34) | T1, T1ce, T2, FLAIR, DSC rCBV | Manual | Linear discriminant analysis, k-nearest neighbour, Support vector machine | 63.3% global accuracy; Sensitivities: 90.9% (grade II), 33.3% (grade III), 41.2% (grade IV), 91.7% (metastases) |
| Zacharaki et al. (2011) [222] | Distinguishing metastases (n=24), meningiomas (n=4), grade II gliomas (n=22), grade III gliomas (n=17) and glioblastomas (n=34) | T1, T1ce, T2, FLAIR, DSC rCBV | Manual | Three feature selection methods, three search methods, five classification algorithms | 76.3% global accuracy, Sensitivities: 81.8% (grade II), 29.4% (grade III), 82.4% (grade IV), 95.8% (metastases) |
| Skogen et al. (2016) [177] | Distinguishing grade II (n=27), grade III (n=34) and grade IV (n=34) gliomas | T1ce | Manual | ROC analysis on individual features | AUC=0.91 (LGG vs HGG), AUC=0.84 (II vs III), AUC=0.73 (III vs IV) |
| Sachdeva et al. (2016) [223] | Distinguishing LGG (n=118), glioblastoma (n=59), meningioma (n=97), medulloblastoma (n=88), metastases (n=66); In total: 55 patients, different 2D-slices are considered independent | T1ce | Semi-automated | Support vector machine, artificial neural network | 94% global accuracy, Sensitivities: 96.6% (LGG), 86.6% (GBM), 93.3% (medulloblastoma), 97% (metastasis) |

healthy and pathological regions. However, no discrimination between different tumour types was obtained. Skogen et al. [177] were able to make a significant distinction between grade II, grade III and grade IV tumours based on a single intensity-based parameter. They included 95 patients and manually segmented the tumour on a single slice of T1ce scans. Georgiadis et al. [221] proposed a two-level hierarchical decision-tree structure to first distinguish between metastases and primary brain tumours, and in a second step between gliomas and meningiomas. ROIs were manually placed on a post-contrast T1-weighted magnetic resonance imaging scan. Including 67 patients, they achieved accuracies of 71.4%, 72.2% and 81.3% for metastases, gliomas and meningiomas respectively. Zacharaki et al. [175] examined 98 patients divided into five tumour classes (metastases, meningiomas, and gliomas of WHO grade II, III and IV). Six different MRI sequences were used and manual segmentation masks for four tumour tissues were acquired. Binary classification tasks were combined with majority voting for multiclass problems. They obtained excellent results for identifying metastases and low-grade glioma, with accuracies of 91.7% and 90.9% respectively, while distinguishing high-grade glioma was considered more difficult. In a follow-up study [222], the same authors tested ten pairwise problems using different feature selection methods and classification algorithms. The best results were obtained when combining a forward feature selection method with voting feature intervals classification, achieving an overall multiclass accuracy of 76.3%. Sachdeva et al. [223] analysed 428 T1ce MRI slices of 55 patients with astrocytoma, glioblastoma, meningioma, medulloblastoma and metastases. The tumour was delineated semi-automatically. Their multiclass neural network based method achieved an overall accuracy of 94%. This result might however be over-optimistic, since they considered different slices from the same patient as independent samples.

In contrast, Luts et al. [224] tried to classify individual voxels into ten different classes (ranging from healthy tissue to grade IV glioma) using both MRI and magnetic resonance spectroscopy (MRS) data from 25 patients and 4 healthy volunteers. They first constructed 45 pairwise classifiers, and afterwards combined this information for multiclass prediction. Similarly, García-Gómez et al. [225] designed multiclass classi-

fiers to distinguish between meningioma, low-grade glioma, glioblastoma and metastases based on single-voxel MRS acquired in multiple centers. All pairwise discriminations could be made with accuracies of around 90%, except for distinguishing between glioblastoma and metastases.

Most previous studies use either manual or semi-automatic delineations of the tumour region. This might however introduce a certain degree of variability in the classification results, as the BraTS reference paper showed [139]. Furthermore, Parmar et al. [226] proved that quantitative features extracted from (semi-)automatic segmentation have a significantly higher reproducibility and robustness compared to manual delineation.

Therefore, in this chapter we propose a fully automatic pipeline for the quantification of structural MRI scans, with the purpose of a non-invasive and multiclass classification of primary brain tumours. To maximize the clinical applicability, we only use routinely acquired T1ce and fluid-attenuation inversion recovery (FLAIR) MRI scans. Moreover, we collected data from a large number of patients acquired in eight different centres. In this way, the aim is to find features able to overcome the heterogeneity inherently present in the data due to different imaging systems and scanning parameters. In this way, our method can be used not only in a research setting, but also in clinical practice.

## 5.2 Data

To acquire a maximum number of images, patient scans from the Ghent University Hospital are combined with data from online repositories.

### 5.2.1 The Ghent University Hospital data

The local ethics committee of the Ghent University Hospital granted permission for a retrospective study, and informed consent was waived (Belgian registration number: B670201524727, local registration number: 2015/0521). This made it possible to search the PACS and the electronic patient files. To get a notion of the available images in the centre, the PACS was scanned with specific queries such as “MR brain

(stereotaxy-neuronavigation)", "MRS brain (tumour)" or "F-18 FET PET brain oncology". This yielded an initial list of 1331 patients in the period January 2005 – May 2017. For every patient, we browsed the electronic patient file for the presence of the anatomical pathological diagnosis, verified by a senior neuropathologist. Next, patients were included if they had a primary brain tumour status at the time of scanning, and a structural MRI protocol before the initial surgery. In this way, we obtained a list of 347 patients with different primary brain tumours, as is shown in figure 5.1.

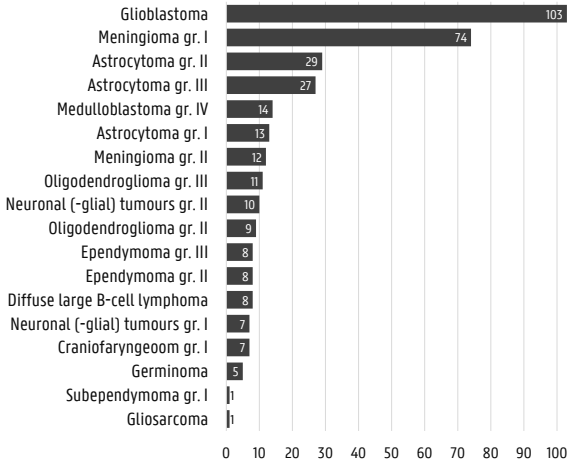



**Figure 5.1:** Verified diagnosis of patients in the Ghent University Hospital primary brain tumour dataset.

Furthermore, we only selected patients with a preoperative T1ce and FLAIR MRI of sufficient quality available, and belonging to a tumour class with at least 25 patients. Since this would mean that only four classes (glioblastoma, meningioma grade I and astrocytoma grade II-III) could be included, we complemented our data with images from online repositories.

### 5.2.2 Additional data: The Cancer Imaging Archive

The Cancer Imaging Archive (`www.cancerimagingarchive.net`) [227] is a continuously growing database of public datasets of cancer-related medical images. We included images from two studies. The

Repository of Molecular Brain Neoplasia Data (REMBRANDT) [228, 229] is a large database consisting of 874 glioma patients aimed to correlate clinical and genomic characterisation data. Presurgical MRI scans are available for 130 patients. As a follow-up study, the VASARI (Visually AcceSAble Rembrandt Images) feature set was developed [230, 231]. This is a set of 24 (qualitative) observations familiar to neuroradiologists to describe the morphology of brain tumours. Examples of such features are "thickness of enhancing margin" or "cortical involvement".

The second database we included is The Cancer Genome Atlas Low Grade Glioma (TCGA-LGG) data collection [232, 233, 67]. This is part of a larger project bringing together genome sequencing data from a large variety of different cancer types. In this database, MRI scans from 199 patients with lower-grade gliomas are included.

In total, we selected 352 patients: 162 patients from the Ghent University Hospital database, 84 cases from the REMBRANDT collection, and 106 patients from the TCGA-LGG database. All patients belong to one of six tumour classes, being meningioma (WHO grade I, $n = 43$), astrocytoma (WHO grade II, $n = 81$; or WHO grade III, $n = 79$), oligodendroglioma (WHO grade II, $n = 29$ or WHO grade III, $n = 29$) or glioblastoma (WHO grade IV, $n = 91$), see also table 5.2.

**Table 5.2:** Distribution of the patients into different classes.

| | | grade | | | | |
|---|---|---|---|---|---|---|
| | | I | II | III | IV | |
| type | M | 43 | | | | 43 |
| | A | | 81 | 79 | | 160 |
| | O | | 29 | 29 | | 58 |
| | G | | | | 91 | 91 |
| | | 43 | 110 | 108 | 91 | |

M = meningioma, A = astrocytoma,
O = oligodendroglioma, G = glioblastoma

The distinction between astrocytoma and oligodendroglioma is made retrospectively based on 1p/19q-codeletion status, according to the WHO 2016 guidelines [54]. However this information was not available for the REMBRANDT collection. For these patients, the histopathological

findings are followed to make the distinction between astrocytoma and oligodendroglioma. In total, patients are acquired in eight different centers, on different imaging systems (different vendors, both 1.5 T and 3 T scanners) and with different scanning parameters. This causes a large degree of variability in image resolution, voxel size, slice spacing and contrast.

## 5.3 Tumour segmentation and feature extraction

Let us look back at the workflow of radiomics, depicted in figure 5.2.

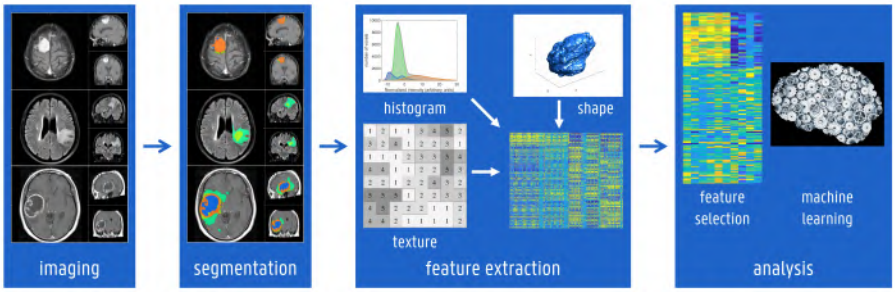



**Figure 5.2:** The workflow of radiomics.

We have already discussed the clinical imaging data, so the next step is segmenting the tumour for all patients. To this end, the scans are first preprocessed in order to match the characteristics of the BraTS dataset. Using SPM12 running on MATLAB 2017b, the FLAIR scans are coregistered to the T1ce images. Next, the scans are spatially normalized to MNI-space, trilinearly interpolated to a $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$ voxel size, and corrected for magnetic field inhomogeneities using bias field correction. Afterwards, the tumour is segmented using the Random Forests based software, validated in chapter 4. All segmentation results are visually inspected and manually adjusted if necessary. For this purpose, a dedicated graphical user-interface (GUI) was built in MATLAB which enables the user to manually draw a ROI inside the tumour on several slices, as shown in figure 5.3. This step replaces the morphological op-

erations, as the tissue probabilities calculated by the Random Forests model are again used as input for the voxel clustering step. Manual adjustment (83 patients, 23.6% of the cases) can therefore aid to avoid healthy tissue such as large blood vessels being considered as tumour.

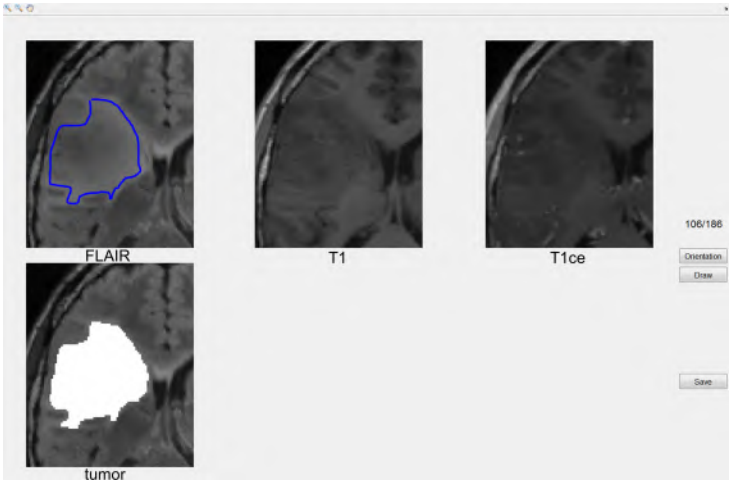



**Figure 5.3:** Screenshot of the MATLAB graphical user-interface for manual tumour contouring. In this example, a zoomed-in slice of T1, T1ce and FLAIR MRI of a glioblastoma patients is shown. The user can draw a contour inside the hyperintense region on FLAIR, which replaces the morphological operations. Afterwards, the tissue probabilities obtained by the Random Forests model are used to cluster the voxels into tumour masks.

Afterwards, we extract features for five tumour masks:

- oedema
- enhancing tumour
- non-necrotic/non-enhancing tissue
- tumour core
- total abnormal region including oedema

on both the T1ce and FLAIR scan. The same 207 features as explained in chapter 3 are obtained: 14 histogram features, 8 shape and size features and 185 second-order (texture) features, consisting of 138 grey-level

co-occurrence matrix parameters, 22 grey-level run-length matrix parameters, 12 neighbourhood grey-tone difference matrix parameters and 13 grey-level size-zone matrix parameters. Before calculating the histogram features, we apply the robust white-stripe normalization method [138], where the intensities are normalised to the normal appearing white matter. For the texture features, the intensities are discretised to 64 grey-levels. Additionally, there are 27 features containing information about the location of the tumour and the difference between the intensities of the tumour and the surrounding tissues. This adds up to a total of 2097 quantitative features per patient. An illustration of the resulting feature matrix, normalised to have zero mean and unit variance for every feature, is given in figure 5.4.

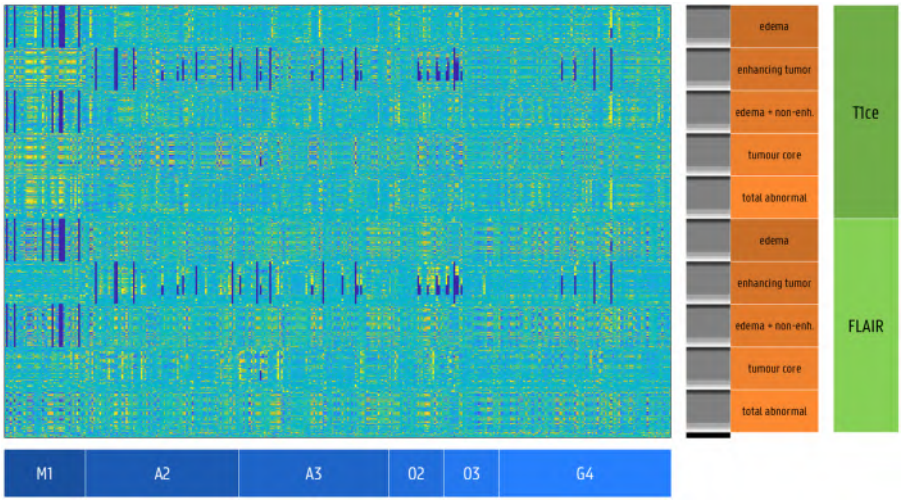



**Figure 5.4:** Structure of the extracted feature matrix. Patient diagnosis code: type + grade; M = meningioma, A = astrocytoma, O = oligodendroglioma, G = glioblastoma.

## 5.4 Multiclass random forests

Now, the extracted features can be analysed using machine learning algorithms. Two multiclass models are trained: one for tumour grade, another for tumour type. For every model, we first hold out a random

subset of 35 patients as test set. On the training set, we first rank the best performing features using the *relief* algorithm. Next, we train a multiclass Random Forests classifier with 200 trees on the highest ranked features. Since the number of training samples per class is highly variable, a cost matrix is applied to avoid bias towards the larger classes. Suppose we train a model to predict grade, then the cost matrix $C$ is calculated as follows:

$$\begin{bmatrix} n_I \\ n_{II} \\ n_{III} \\ n_{IV} \end{bmatrix} = \begin{bmatrix} 39 \\ 99 \\ 97 \\ 87 \end{bmatrix} \Rightarrow C = \begin{bmatrix} 0 & \frac{1+\frac{99}{39}}{2} & \frac{1+\frac{97}{39}}{2} & \frac{1+\frac{87}{39}}{2} \\ \frac{1+\frac{39}{99}}{2} & 0 & \frac{1+\frac{97}{99}}{2} & \frac{1+\frac{87}{99}}{2} \\ \frac{1+\frac{39}{97}}{2} & \frac{1+\frac{99}{97}}{2} & 0 & \frac{1+\frac{87}{97}}{2} \\ \frac{1+\frac{39}{87}}{2} & \frac{1+\frac{99}{87}}{2} & \frac{1+\frac{97}{87}}{2} & 0 \end{bmatrix},$$

where $n_i$ is the number of training samples from grade $i$. In MATLAB's `TreeBagger` function, classes with a high penalty will be oversamples when training the Random Forests, and in this way the bias towards larger classes is reduced. However, when the penalty becomes too large, bias towards the small classes might occur, which is why we added the additional averaging with 1.

The trained model is now evaluated on the 35 independent samples from the test set, and the confusion matrix is stored. We repeat this process 50 times to avoid selection bias. The results for grade prediction are given in table 5.3. We observe a mean accuracy of (60.3±5.7)% with 400 features included in the model.

The results for tumour type prediction are given in table 5.4, where a mean accuracy of (65.6±8.5)% is obtained using the best 800 features.

**Table 5.3:** Performance of multiclass Random Forests to predict tumour grade. A random subset of 35 patients is selected as test set over 100 iterations. Also given is the mean certainty $\bar{p}$ with which every decision is made.

| | | predicted | | | |
|---|---|---|---|---|---|
| | | I | II | III | IV |
| true | I | 3.94±0.77 ($\bar{p}$=62.1%) | 0.08±0.27 ($\bar{p}$=33.0%) | 0.16±0.42 ($\bar{p}$=30.8%) | 0.24±0.43 ($\bar{p}$=35.4%) |
| | II | 0.38±0.60 ($\bar{p}$=41.3%) | 7.30±1.47 ($\bar{p}$=49.9%) | 2.34±1.29 ($\bar{p}$=44.4%) | 0.46±0.54 ($\bar{p}$=44.2%) |
| | III | 0.06±0.24 ($\bar{p}$=30.5%) | 4.82±1.51 ($\bar{p}$=47.7%) | 2.88±1.41 ($\bar{p}$=45.7%) | 2.76±1.04 ($\bar{p}$=49.7%) |
| | IV | 0.64±0.80 ($\bar{p}$=41.8%) | 0.82±0.75 ($\bar{p}$=41.8%) | 1.12±0.92 ($\bar{p}$=42.1%) | 7.00±1.41 ($\bar{p}$=57.3%) |

**Table 5.4:** Performance of multiclass Random Forests to predict tumour type. M = meningioma, A = astrocytoma, O = oligodendroglioma, G = glioblastoma.

| | | predicted | | | |
|---|---|---|---|---|---|
| | | M | A | O | G |
| true | M | 4.10±0.76 ($\bar{p}$=62.7%) | 0.20±0.40 ($\bar{p}$=39.0%) | 0.04±0.20 ($\bar{p}$=34.3%) | 0.24±0.48 ($\bar{p}$=38.1%) |
| | A | 0.32±0.51 ($\bar{p}$=41.9%) | 10.72±1.96 ($\bar{p}$=50.8%) | 1.72±1.53 ($\bar{p}$=38.3%) | 2.70±1.54 ($\bar{p}$=50.3%) |
| | O | 0.0±0.0 ($\bar{p}$=0.0%) | 2.78±1.36 ($\bar{p}$=48.5%) | 1.64±1.31 ($\bar{p}$=47.7%) | 0.96±0.97 ($\bar{p}$=46.3%) |
| | G | 0.78±0.84 ($\bar{p}$=38.7%) | 1.78±1.22 ($\bar{p}$=44.3%) | 0.52±0.86 ($\bar{p}$=35.1%) | 6.35±1.66 ($\bar{p}$=57.9%) |

From these confusion matrices, it is clear that the algorithm is more certain when correctly predicting the meningioma/grade I and glioblastoma/grade IV classes than when predicting one of the lower-grade gliomas. Moreover, we see that most classes are predicted rather well, with the exception of grade III gliomas, which show a very high bias towards grade II and to a lesser extent grade IV, and oligodendrogliomas, which are often being predicted as astrocytomas.

Although confusion matrices offer a very detailed view on the performance of a model, they can be difficult to interpret. This is why in table 5.5 the sensitivities (true positive rate) and specificities (true negative

rate) for every class are given.

**Table 5.5:** Detailed analysis of the two multiclass analyses.

**(a)** grade

| | sensitivity | specificity |
|---|---|---|
| I | 89.1% | 96.5% |
| II | 69.7% | 76.7% |
| III | 27.4% | 85.2% |
| IV | 73.1% | 86.4% |

**(b)** type

| | sensitivity | specificity |
|---|---|---|
| M | 89.5% | 96.4% |
| A | 69.3% | 75.6% |
| O | 30.5% | 92.3% |
| G | 67.9% | 84.7% |

From these tables, we again see that both models perform well for three out of four classes, and show poor performance for the fourth class (grade III gliomas and oligodendrogliomas, respectively). This however makes that this method cannot be directly be used for CAD. Suppose the models predict a grade II astrocytoma, then we have little certainty about this result, since grade III gliomas will more often be predicted grade II than grade III, and oligodendrogliomas have a very high probability of being predicted as astrocytomas.

## 5.5 Discussion and conclusion

In this chapter, the goal was to predict tumour grade and histology of primary brain tumour patients based on quantitative features determined on structural MRI scans. Compared to the previous studies mentioned in the introduction of this chapter, we obtain lower classification performances, even though our dataset is much larger than these studies. Three factors may play in a role in the reduced performance.

First of all, some errors might be due to the high degree of heterogeneity in the data. We obtained scans from three different sources, with a high variability in scanning protocols and imaging parameters. Although all images are normalised to have equal voxel size, differences in scanning resolution and slice thickness will lead to variability in some of the texture features [234, 235].

A second source of variability may lie in the gold standard labels.

As mentioned in the introductory chapter, histopathological diagnosis is prone to inter- and intraobserver variability. Furthermore, we based the difference between the oligodendroglioma and astrocytoma labels on 1p/19q status rather than on histological findings. As this information was not available for the 55 lower-grade glioma patients from the REMBRANDT collection included in our study, several might have got an incorrect tumour label.

Lastly, our study shows that the hardest task is distinguishing between the different lower-grade glioma labels (astrocytoma and oligodendroglioma WHO grade II and III). This can be partly explained by the variability in the pathological diagnosis, but also due to the very nature of this classification. IDH-status is not taken into account in this analysis, and IDH-mutant (or glioblastoma-like) lower-grade gliomas will therefore have got a lower-grade glioma label, although showing a significantly different clinical course. Moreover, IDH-mutant astrocytomas show only a little difference in survival [70] between grade II and grade III, suggesting that they are similar entities. In the Ghent University Hospital, IDH-status is assessed using immunohistochemical (IHC) analysis. However, a negative IDH-status using IHC analysis of a lower-grade tumour does not necessarily mean an IDH-wildtype (or glioblastoma-type) tumour [54]. Sequencing for IDH gene mutations is not routinely performed, so we cannot confidently identify IDH-wildtype glioma.

In conclusion, we have tried to solve the multiclass diagnosis problem by splitting it up into two models: for grade (mean accuracy: 60.3%) and for type (mean accuracy: 65.6%). Since patients are evaluated by the two classification models independently, this can lead to undesired results, such as a patient having a high probability of being both grade I and having a glioblastoma. As this offers only a limited source of information, a more elaborated solution will be given in the next chapter.

# 6 Transforming multiclass to multiple binary problems

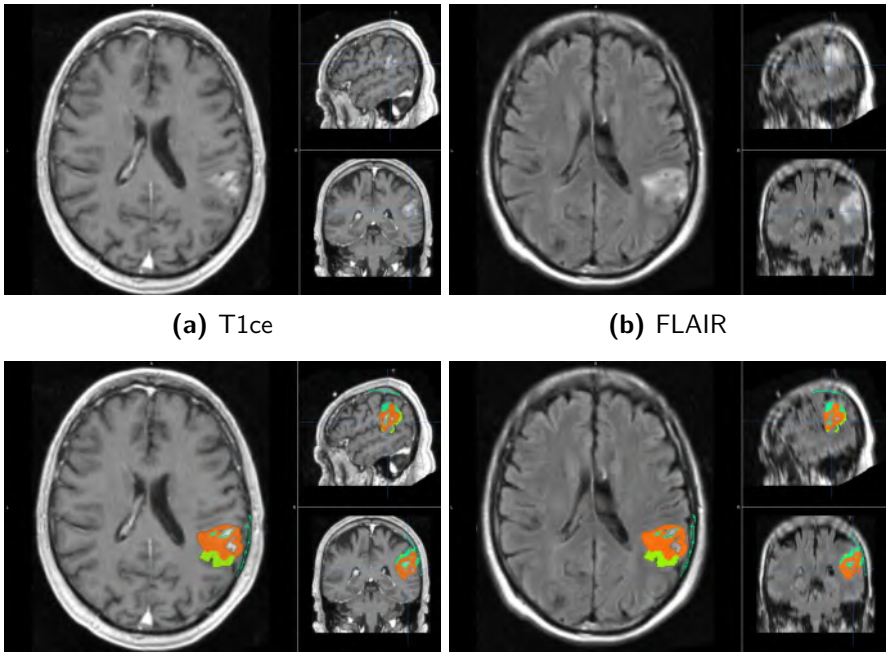
In the previous chapter, we have built multiclass models to predict tumour grade and type of primary brain tumours based on medical images. However, in many situations a clinician will already have an idea on the diagnosis based on clinical parameters or the radiological appearance of the tumour. For example, when the clinical symptoms are highly suggestive of a glioma, clinicians might be mostly interested in determining whether the tumour is low-grade or high-grade, making meningioma probabilities irrelevant. Therefore, in this chapter we try to solve the multiclass diagnosis problem by providing answers to fourteen binary problems, which together can lead to the correct diagnosis.

This work has been presented during the 2018 European Conference on Clinical Neuroimaging [236].

## 6.1 The advantage of binary classification problems

Consider the example scans of figure 6.1. This patient (TCGA-FG-5964 from the TCGA-LGG dataset) shows diffuse contrast enhancement and a clear hyperintense tumour region on FLAIR. Histopathological analysis showed a diffuse oligodendroglioma, WHO grade II at the time of scanning. If we evaluate this patient by the two models from the previous chapter then we obtain the following probabilities for grade: 11.5%

grade I, 42.0% grade II, 24.5% grade III and 22.0% grade IV; and for type: 25.5% meningioma, 34.0% astrocytoma, 15.0% oligodendroglioma an 25.5% glioblastoma. This would suggest a low-grade astrocytoma, although we know from last chapter's analysis that oligodendroglioma cases are often predicted as astrocytoma. Assuming that this patient is indeed a low-grade glioma case, it would therefore be interesting to have a specific model able to distinguish between astrocytomas and oligodendrogliomas.



**(a)** T1ce **(b)** FLAIR

**(c)** T1ce + segmentation **(d)** FLAIR + segmentation

**Figure 6.1:** Example of scans from a diffuse oligodendroglioma, WHO grade II patient (TCGA-FG-5964). The T1ce scan shows diffuse contrast enhancement, with a clear FLAIR hyperintense tumour region. Also given are the automatically generated tumour regions.

In chapter 3, we have shown that a binary classification problem can be solved with high accuracy. Therefore, we split up our multiclass problem into fourteen binary classifiers, with the additional purpose of minimising the number of features per model, according to the workflow

of figure 6.2.

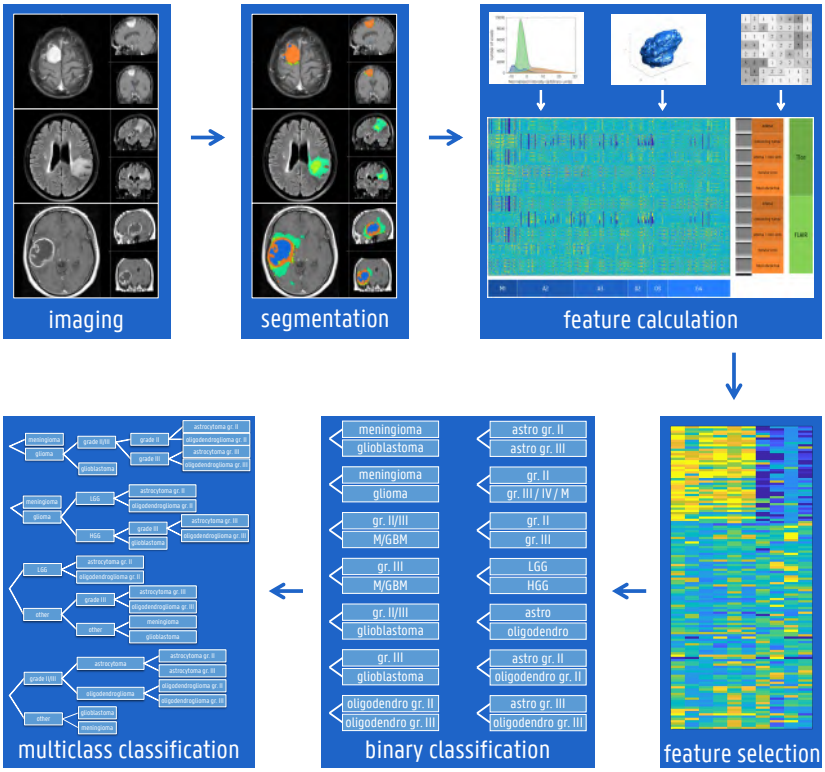



**Figure 6.2:** Workflow of the multiple binary classification. After acquiring the feature matrix and feature selection, fourteen binary models are created, which can afterwards be combined into decision schemes. A more detailed view on the four multiclass decision schemes is given in figure 6.4.

The same data are used as in the previous chapter, resulting in the same feature matrix. Now, patients are first grouped into smaller, not mutually exclusive subsets such as "grade II gliomas", "astrocytomas" or "meningioma and glioblastoma". Every subset consists of tumour types with similar properties, such as an equal grade, histological type or a mutual appearance (for example: both meningiomas and glioblastomas show a large degree of contrast enhancement). Relevant subsets are pairwise compared in binary classification problems. For every model, we test the individual predictor performance using a two-sample $t$-test

with unequal variance. We rank all predictors according to increasing $p$-values and select only the first 200 for every task. Next, these parameters are used to train a Random Forests classification model, implemented in MATLAB's `TreeBagger` function. Only complementary features are used, found by sequential forward selection. This is done using five-fold cross-validation on a balanced dataset, meaning that equal amounts of samples from both classes in each binary problem are randomly selected. This step is repeated 100 times with random selection of the samples. Performance is assessed by the total accuracy of the model and the AUC of the ROC curve, as is given in table 6.1.

**Table 6.1:** Performance of the individual binary problems, ranked according to accuracy. For every problem, the most predictive feature is given as well. Every classification task takes 15 features into account.

| | Name binary problem | Most predictive feature | acc. | AUC |
|---|---|---|---|---|
| 1 | Meningioma vs. glioblastoma | T1ce - ratio core/surrounding (5 voxels) | 95.0% | 0.989 |
| 2 | Meningioma vs. glioma | T1ce - ratio core/surrounding (5 voxels) | 92.6% | 0.976 |
| 3 | Gr. II/III vs. gr. IV/meningioma | T1ce – total abnormal – histogram: mean | 90.7% | 0.952 |
| 4 | Gr. III vs. gr. IV | T1ce - ratio core/surrounding (5 voxels) | 88.8% | 0.944 |
| 5 | Gr. III vs gr. IV/meningioma | T1ce – total abnormal – histogram: mean | 88.5% | 0.937 |
| 6 | Gr. II/III vs. gr. IV | T1ce - ratio core/surrounding (3 voxels) | 88.0% | 0.934 |
| 7 | Oligodendroglioma gr. II vs. oligodendroglioma gr. III | T1ce – oedema – histogram: median | 87.8% | 0.937 |
| 8 | Astrocytoma gr. II vs. astrocytoma gr. III | T1ce – tumour core – histogram: uniformity | 84.3% | 0.919 |
| 9 | Gr. II vs. gr. III/IV/meningioma | T1ce - ratio core/surrounding (5 voxels) | 83.5% | 0.901 |
| 10 | Gr. II vs. gr. III | FLAIR - enhancing tumour – GLCM: difference entropy (d=1), mean | 83.1% | 0.907 |
| 11 | Gr. II vs. gr. III/IV | T1ce - ratio core/surrounding (3 voxels) | 83.0% | 0.906 |
| 12 | Astrocytoma gr. II vs. oligodendroglioma gr. II | FLAIR - total abnormal – GLCM: difference entropy (d=3), std | 80.5% | 0.883 |
| 13 | Astrocytoma vs. oligodendroglioma | y-coordinate centre of mass | 80.3% | 0.892 |
| 14 | Astrocytoma gr. III vs. oligodendroglioma gr. III | T1ce – oedema – histogram: mean | 75.1% | 0.829 |

We identify the optimal model parameters using grid search (see Appendix figure C.1). This shows that optimal results are obtained by incorporating only 15 out of 2925 features for every binary model. An overview of these features is given in Appendix C. The model complex-

ity is indirectly controlled by the `MinLeafSize` parameter, being the minimal amount of samples in every leaf of individual classification trees in the Random Forest. Only a modest regularisation is required to optimise the classification result fixing this parameter to 2, while the number of trees per forest is 200. We see that mainly features calculated on the entire abnormal region, the tumour core and the enhancing tissue are taken into account. This can be explained by acknowledging that the most robust segmentation results are obtained on these tumour masks.

In table 6.1, the performance of the fourteen binary problems is given. Every classification task takes 15 predictors into account, and the best performing feature is also given. It is clear that all classification tasks can be performed with good results, with accuracies exceeding 75% and AUC-scores exceeding 0.82. For distinguishing between lower-grade gliomas and glioblastomas, we obtain an accuracy of 88.0%, which is slightly higher than the 85.7% accuracy obtained using the same method but on different data in chapter 3.

The easiest problems are automatically distinguishing meningioma from glioma, while discriminating astrocytoma from oligodendroglioma is considered harder. The ratio of the T1ce-intensities of the tumour core and the surrounding tissue is the most predictive feature for six binary problems. Moreover, there are five histogram features and two texture features, both GLCM-based on FLAIR images, in the list as most predictive features. Remarkable is that the most important predictor for distinguishing astrocytoma from oligodendroglioma is the $y$-coordinate of the tumour centre of mass, meaning that oligodendroglioma have a slightly higher probability of being located more anterior in the brain compared to astrocytoma ($p = 0.0068$).

Looking back at the example patient from the beginning of this section, we obtain the probabilities for the fourteen binary problems of figure 6.3. Based on these probabilities, we can exclude a meningioma diagnosis with a probability of 91.5%, in favour of a glioma. Taking this into account, there is a high probability (93%) that it concerns a lower-grade glioma instead of a glioblastoma. Finally, the tumour is most probably (87.5%) from the oligodendroglioma type, with again a high probability (86%) of being a diffuse low-grade oligodendroglioma, WHO grade II. In this way, the probabilities from the binary models

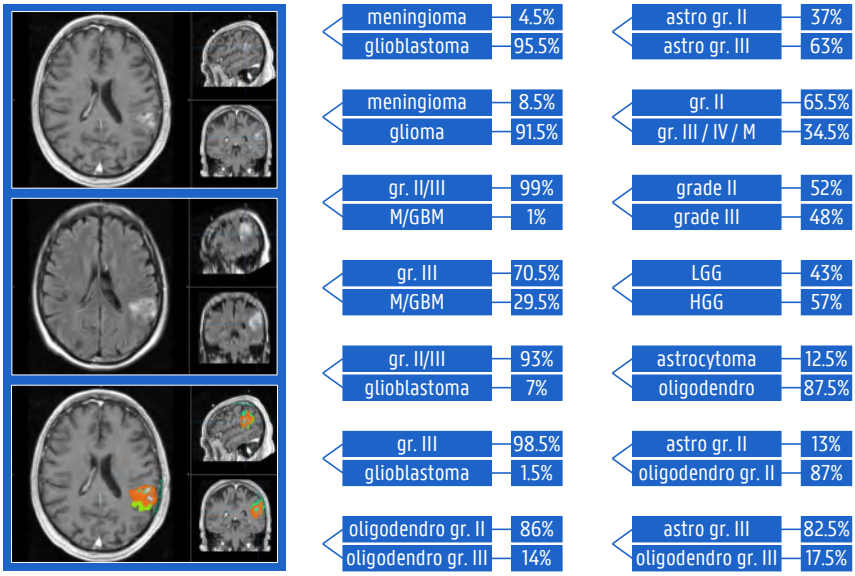



**Figure 6.3:** Probabilities for fourteen binary classification problems obtained for the example patient.

can be used to obtain the right diagnosis. In the next section, we will try to automatically perform this multiclass problem.

## 6.2 Brain tumour classification as a sequence of binary problems

In a second step, we concatenate the binary problems in a hierarchical way to obtain four different decision trees, as displayed in 6.4. The probabilities for every binary step are multiplied to obtain the final probabilities per tumour class. The tumour type with highest probability is then chosen as the final prediction.

This is validated by randomly selecting seven patients from each of the six tumour classes and training balanced Random Forests models using the others patients and the previously found features. This procedure is repeated 100 times to minimise bias from the random selection procedure.

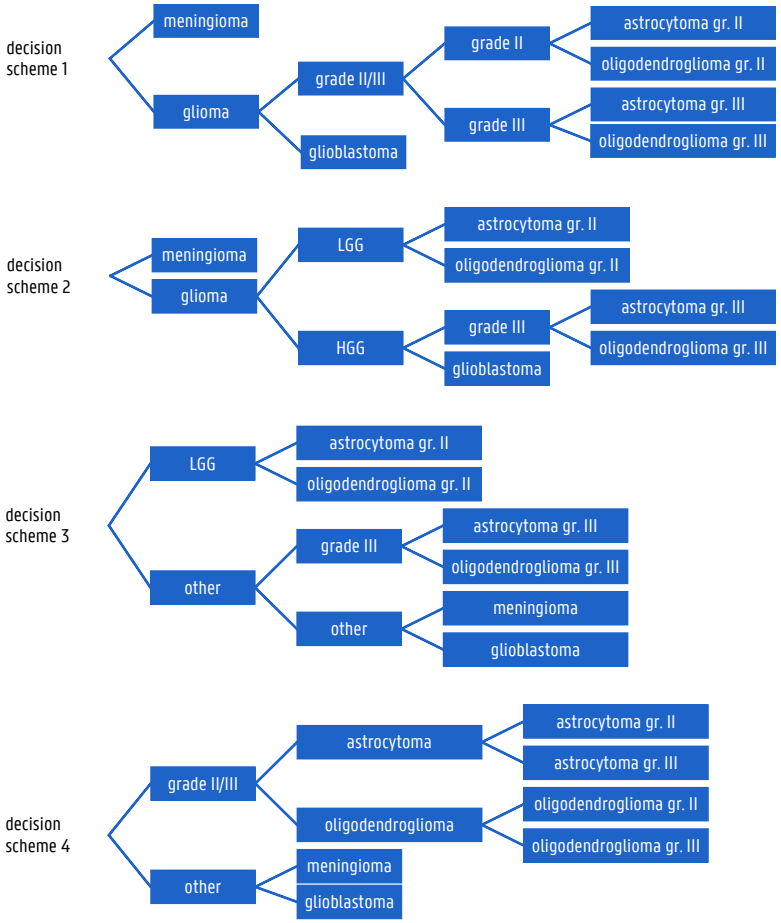



**Figure 6.4:** Concatenation of binary classifiers into four multiclass decision schemes.

In table 6.2 the results are presented for the different decision schemes of figure 6.4. These are obtained by averaging over 100 iterations, to avoid bias by the random selection of training and testing samples. For every method, the results are displayed in confusion matrices, where the real tumour classes are displayed in horizontal and the predicted classes in vertical direction.

Not all decision schemes lead to an equally good result, and some schemes perform better in detecting specific tumour types. Decision scheme 4, where grade II and III tumours are first divided from meningioma and glioblastoma, and next astrocytoma from oligodendroglioma,

**Table 6.2:** Confusion matrices giving the performance of the four concatenated decision schemes of figure 6.4. To validate this method, seven patients from each class are held out as test samples while the method is trained on the remaining patients. The results are displayed as the average over 100 iterations.

**(a)** decision scheme 1: accuracy = 45.2%

| true \ predicted | M1 | A2 | A3 | O2 | O3 | G4 |
|---|---|---|---|---|---|---|
| M1 | 6.32 | 0.01 | 0.01 | 0.01 | 0.00 | 0.65 |
| A2 | 0.47 | 3.05 | 0.82 | 1.27 | 0.32 | 1.07 |
| A3 | 0.73 | 1.58 | 1.29 | 0.46 | 0.44 | 2.50 |
| O2 | 0.43 | 1.18 | 0.93 | 2.82 | 0.83 | 0.81 |
| O3 | 0.63 | 1.18 | 0.83 | 0.57 | 0.69 | 3.10 |
| G4 | 1.68 | 0.19 | 0.19 | 0.04 | 0.07 | 4.83 |

**(b)** decision scheme 2: accuracy = 46.5%

| true \ predicted | M1 | A2 | A3 | O2 | O3 | G4 |
|---|---|---|---|---|---|---|
| M1 | 6.44 | 0.05 | 0.01 | 0.00 | 0.00 | 0.50 |
| A2 | 0.50 | 3.91 | 0.36 | 1.67 | 0.20 | 0.36 |
| A3 | 0.86 | 2.73 | 0.88 | 1.06 | 0.38 | 1.09 |
| O2 | 0.36 | 1.87 | 0.57 | 3.50 | 0.64 | 0.06 |
| O3 | 0.78 | 2.50 | 0.62 | 0.96 | 0.67 | 1.47 |
| G4 | 1.94 | 0.52 | 0.16 | 0.19 | 0.07 | 4.12 |

**(c)** decision scheme 3: accuracy = 47.9%

| true \ predicted | M1 | A2 | A3 | O2 | O3 | G4 |
|---|---|---|---|---|---|---|
| M1 | 6.01 | 0.08 | 0.23 | 0.08 | 0.03 | 0.57 |
| A2 | 0.11 | 4.11 | 0.45 | 1.67 | 0.20 | 0.46 |
| A3 | 0.01 | 2.91 | 0.77 | 1.30 | 0.39 | 1.62 |
| O2 | 0.00 | 1.90 | 0.72 | 3.56 | 0.67 | 0.15 |
| O3 | 0.19 | 2.47 | 0.57 | 1.00 | 0.73 | 2.04 |
| G4 | 0.55 | 0.68 | 0.38 | 0.29 | 0.17 | 4.93 |

**(d)** decision scheme 4: accuracy = 52.8%

| true \ predicted | M1 | A2 | A3 | O2 | O3 | G4 |
|---|---|---|---|---|---|---|
| M1 | 6.12 | 0.10 | 0.01 | 0.03 | 0.01 | 0.73 |
| A2 | 0.13 | 3.47 | 1.00 | 1.06 | 0.48 | 0.86 |
| A3 | 0.02 | 1.39 | 1.76 | 0.83 | 0.48 | 2.52 |
| O2 | 0.01 | 1.18 | 0.76 | 3.64 | 0.84 | 0.57 |
| O3 | 0.21 | 0.85 | 0.69 | 1.13 | 1.59 | 2.53 |
| G4 | 0.57 | 0.30 | 0.23 | 0.08 | 0.24 | 5.58 |

leads to the best results, with an overall accuracy of 52.8%. Using this model, we can automatically detect meningioma with a sensitivity of 87.4% and specificity of 97.3%, and glioblastoma with a sensitivity of 79.7% and specificity of 79.4%. Lower accuracies are obtained for the lower-grade glioma (astrocytoma and oligodendroglioma WHO grade II-III). Astrocytoma and oligodendroglioma WHO grade III are particularly difficult to identify, since these will mostly be predicted as glioblastoma.

Let us return one last time to the oligodendroglioma example of figure 6.1. In figure 6.5 the output of the binary decisions are combined with the decision schemes. Three out of four models correctly classify the tumour, while the fourth model predicts a grade III astrocytoma (42.5%), followed by the correct diagnosis of grade II oligodendroglioma

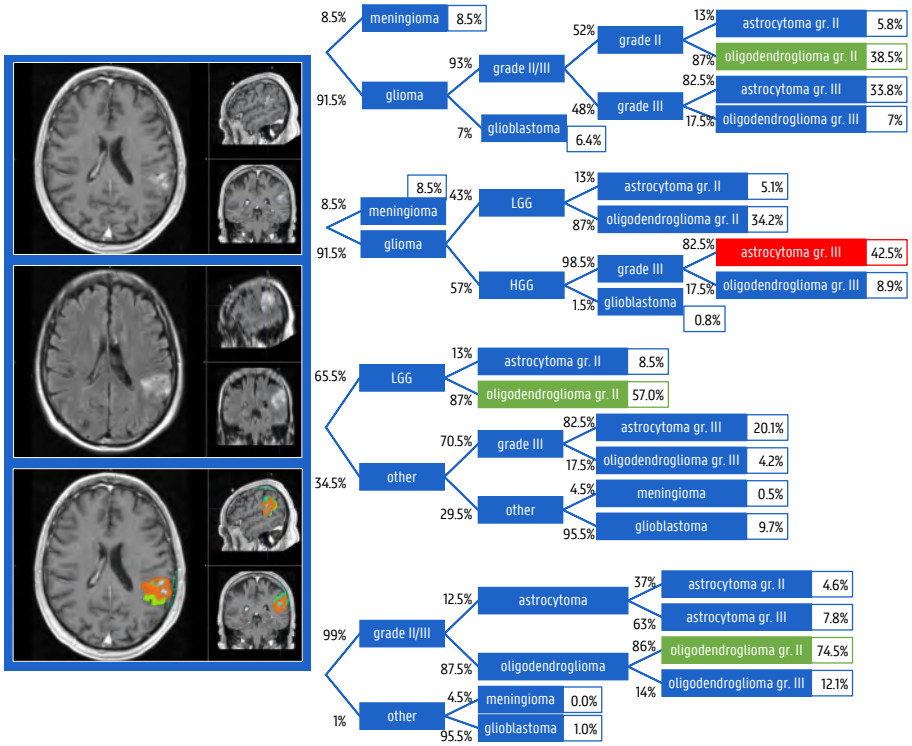



**Figure 6.5:** Example of decision schemes. Three out of four models correctly classify the tumour type in this case.

(34.2%).

## 6.3 Machine learning

Instead of designing decision schemes to solve the multiclass problem, we can also let the computer decide what the best way is to combine the probabilities from the binary models. Therefore, we will test the performance of two machine learning models: multinomial logistic regression and Random Forests.

Multinomial logistic regression can be regarded as a generalisation of linear regression into a multiclass classification method. This means that a linear combination of the scores for the different binary models is calculated which can best predict the tumour class. To train this

model, we evaluate the training data by the fourteen binary classifiers. The output probabilities, combined with the known tumour labels of the training data, is now used to estimate the coefficients of the logistic regression model.

As a second technique, we use Random Forests once more to combine the output scores from the binary problems into a multiclass classifier. Again, the probabilities from the binary classifiers on the training set are used as input for the model, and by randomly selecting equal numbers from each class, we avoid bias towards a certain label.

The results of these techniques are displayed in table 6.3. We obtain accuracies of 49.4% and 52.0% for multinomial logistic regression and Random Forests, respectively. This is comparable to the performance of the concatenated decision schemes. An important remark however is that the scores on which these models are trained could be significantly different from the scores of the test set. After all, the binary classification models are evaluated on the data on which they are trained, possibly giving overly optimistic scores.

**Table 6.3:** Confusion matrices giving the performance of machine learning approaching combining the output of the binary classifiers.

**(a)** multinomial logistic regression: accuracy = 49.4%

| | | predicted | | | | | |
|---|---|---|---|---|---|---|---|
| | | M1 | A2 | A3 | O2 | O3 | G4 |
| true | M1 | 5.52 | 0.21 | 0.20 | 0.00 | 0.03 | 1.04 |
| | A2 | 0.10 | 4.15 | 1.44 | 0.43 | 0.23 | 0.65 |
| | A3 | 0.00 | 2.05 | 2.74 | 0.33 | 0.41 | 1.47 |
| | O2 | 0.01 | 1.82 | 2.36 | 1.99 | 0.60 | 0.22 |
| | O3 | 0.13 | 1.63 | 1.63 | 0.51 | 1.24 | 1.86 |
| | G4 | 0.26 | 0.65 | 0.73 | 0.02 | 0.22 | 5.12 |

**(b)** Random Forests: accuracy = 51.3%

| | | predicted | | | | | |
|---|---|---|---|---|---|---|---|
| | | M1 | A2 | A3 | O2 | O3 | G4 |
| true | M1 | 6.00 | 0.12 | 0.14 | 0.01 | 0.02 | 0.71 |
| | A2 | 0.10 | 3.73 | 1.45 | 0.83 | 0.47 | 0.42 |
| | A3 | 0.01 | 1.74 | 2.72 | 0.60 | 0.74 | 1.19 |
| | O2 | 0.00 | 1.42 | 1.69 | 2.84 | 1.05 | 0.00 |
| | O3 | 0.23 | 1.22 | 1.48 | 0.82 | 1.66 | 1.59 |
| | G4 | 0.51 | 0.38 | 0.76 | 0.06 | 0.38 | 4.91 |

## 6.4 Discussion

In this chapter, we have approached the multiclass primary brain tumour diagnosis problem in an alternative way. Instead of directly modelling the tumour grade and type, we have transformed the automated diag-

nosis into a series of binary problems, which can be solved with great accuracy. Our best method reaches an accuracy of 52.8%, which seems unsatisfying. However, for every patient the probabilities for fourteen separate binary problems are calculated, which can guide manual diagnosis. We achieve accuracies exceeding 85% for all binary problems that do not distinguish between the four lower-grade glioma labels (see table 6.1). The multiclass classification is much harder than binary problems due to propagation of errors. As an additional test, we used MATLAB's `ClassificationLearner`, an easy tool to quickly test a large number of supervised machine learning methods, to evaluate the multiclass performance on our data. This did not yield an improved result, suggesting that the features we calculated do not contain enough information for a better classification.

Our results confirm the finding of the previous chapter, that it is difficult to distinguish between the lower-grade gliomas (astrocytoma and oligodendroglioma WHO grade II and III). The same three possible reasons might explain this finding: data heterogeneity, variability in the gold standard labels, and a large degree of similarity between the different lower-grade gliomas.

An improved segmentation algorithm (e.g. using deep learning approach) might result in a better classification accuracy. However, a deep learning method based on T1ce and FLAIR scans does not yet exist. The segmentation method we applied might yield unsatisfying results for low-grade glioma, since manual interaction was required for some patients to guide the tumour delineation. This was followed by automatic voxel clustering to yield tumour masks with similar properties as the fully automatically obtained segmentations. Still, some enhancing blood vessels might be regarded as enhancing tumour tissue, which can explain the bias towards glioblastoma for the grade III glioma. Moreover, cystic structures will be regarded as necrotic tissue.

For clinical purposes, more tumour types should be incorporated in the model. We collected scans from pilocytic astrocytoma ($n = 13$), ependymoma grade II ($n = 8$) and grade III ($n = 8$), neuronal-glial tumours grade I ($n = 7$) and grade II ($n = 10$), high-grade meningioma ($n = 12$) and medulloblastoma ($n = 14$) in our centre. However, we deemed these numbers too low to be included in this study.

## 6.5 Conclusion

In conclusion, we have further elaborated on a method to non-invasively asses the diagnosis of primary brain tumour patients, based on two routinely acquired medical images. This maximises the clinical applicability. Quantitative features, calculated in different tumour sub-regions are used as input for fourteen binary model achieving a high accuracy. The output probabilities of these binary problems can furthermore guide classification into six different primary brain tumour classes. This can be done using decision schemes or with machine learning models, both yielding a similar total accuracy.

In the last two chapters, we have only included structural MRI scans as information source. In the next chapter, we will try to improve the performance by incorporating a different imaging modality.

# 7

# The added value of $^{18}$F-FET PET for primary brain tumour diagnosis

In the previous chapters, we addressed a non-invasive tool towards automated primary brain tumour diagnosis based on structural magnetic resonance imaging (MRI) scans. This led to the conclusion that low-grade meningioma and aggressive glioblastoma could be identified with high accuracy, but discriminating between the lower-grade tumour labels was considered more difficult. One of the possible explanations for this finding is that these tumours might have a similar appearance on structural MRI, such that quantitative features calculated on these images do not possess enough discriminative power. Therefore, we now incorporate information from a different imaging modality: $^{18}$F-FET positron emission tomography (PET). The goal is to improve the diagnostic accuracy for one binary problem, namely the important distinction between low-grade glioma (LGG, WHO grade II) and high-grade glioma (HGG, WHO grade III-IV).

This work has been presented during the 2018 EANO conference [237] and is being prepared for publication.

## 7.1 Introduction

In chapter 2, we briefly introduced $^{18}$F-FET for PET. It is the most frequently used radiotracer in neuro-oncology due to its excellent tumour-to-background contrast, practical half-life of 110 minutes and efficient chemical synthesis process [238]. A 2D representation of the structure

is displayed in figure 7.1.

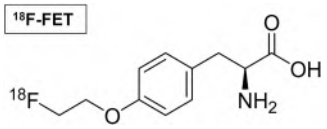



**Figure 7.1:** 2D representation of the chemical structure of *O*-(2-[$^{18}$F]fluoroethyl)-L-tyrosine ($^{18}$F-FET). Reprinted from [239], with permission from Elsevier.

## 7.1.1 The biology of $^{18}$F-FET

*O*-(2-[$^{18}$F]fluoroethyl)-L-tyrosine ($^{18}$F-FET) is an analogue of tyrosine, an important amino acid involved in protein synthesis and signal transduction processes. $^{18}$F-FET is not metabolised and not incorporated into proteins [240]. Its uptake in tumour cells is mediated by the L (leucine preferring) amino acid transport system. The "large neutral amino acid transporter" or "L-system amino acid transporter" LAT1, a membrane transport protein, is responsible for the accumulation of $^{18}$F-FET in tumour cells [241]. Since these amino acid transporters are overexpressed in glioma, $^{18}$F-FET is highly specific in brain tumour imaging. As an additional advantage, disruption of the blood-brain barrier (BBB) appears not to be required for $^{18}$F-FET uptake [240].

Another popular radiolabelled amino acid is L-[methyl-$^{11}$C]methionine ($^{11}$C-MET). This amino acid is incorporated into proteins, and its imaging specificity is therefore caused by a high rate of protein synthesis in fast dividing tumour cells. Both $^{18}$F-FET and $^{11}$C-MET show a similar tumour-to-background contrast, but $^{11}$C-MET has a more rapid uptake [240, 128]. However, due to the short half-life of $^{11}$C (20 minutes), $^{18}$F-FET is often preferred, especially when an on-site cyclotron for radiotracer synthesis is not available.

## 7.1.2 $^{18}$F-FET PET in neuro-oncology applications

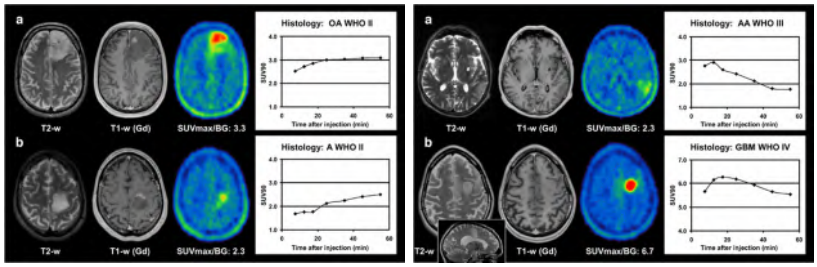
$^{18}$F-FET PET is recommended for a number of different tasks by the European Association for Neuro-Oncology (EANO) [123], which can roughly be divided into the following categories.

## Diagnosis and classification

$^{18}$F-FET PET is in the first place used for the diagnosis of cerebral tumours. Gliomas show mostly an increased uptake compared to healthy tissue [242], although uptake below the background level might also be sign of malignancy [243]. Hot spots on $^{18}$F-FET PET provide an added value over MRI for biopsy guiding and can therefore lead to an improved histological and genetic tumour diagnosis [244, 245]. $^{18}$F-FET is also preferred over $^{18}$F-FDG for biopsy guidance and treatment planning [246].

A large deal of research has been conducted on non-invasive and presurgical grading and classification of primary brain tumours based on $^{18}$F-FET PET. Table 7.1 provides an overview on recent articles addressing grading of untreated patients.

Most studies agree on a number of findings. By analysing static PET scans, they often find significantly higher uptake in HGG compared to LGG, but with considerable overlap due to interindividual differences, hampering predictions on the individual level. In dynamic PET, the dynamic behaviour is generally different between LGG and HGG. In the low-grade patients, the $^{18}$F-FET signal is continuously increasing until the end of acquisition, while in HGG patients the uptake shows an early peak (around 10–20 min p.i.) followed by a decrease. This trend is illustrated in figure 7.2.




**(a)** Low-grade gliomas **(b)** High-grade gliomas

**Figure 7.2:** Examples of the kinetic uptake pattern of $^{18}$F-FET in glioma. Whereas low-grade glioma typically show a continuously increasing signal, high-grade glioma display an early peak followed by tracer wash-out. Reprinted with permission from [124], Copyright Springer 2007.

**Table 7.1:** Overview of the literature on tumour grading of untreated glioma patients based on $^{18}$F-FET PET.

| Author | Static/ dynamic | Patients | Findings |
|---|---|---|---|
| Weckesser et al. (2005) [242] | Static + dynamic | 44 patients, 22 gliomas (5 LGG, 17 HGG) | Significant difference between LGG and HGG in first image frame (0–10 min p.i.), different kinetic behaviour |
| Pöpperl et al. (2007) [124] | Static + dynamic | 54 patients (20 LGG, 34 HGG) | Significant increase in $SUV_{max}$/BG (20–40 min p.i.) from LGG to HGG, but with marked overlap; different kinetic behaviour: 94% sensitivity, 100% specificity |
| Calcagni et al. (2011) [248] | Dynamic | 32 patients (17 LGG, 15 HGG) | Identification of three types of time-activity curves (TACs): always ascending, midway peak followed by plateau or slow descent, early peak followed by steep descent (87% accuracy); identification of two parameters with high accuracy: early-to-middle ratio and Tpeak (both 94%); logistic regression with early SUV and SumOfDifferences shows accuracy of 97% |
| Jansen et al. (2012) [243] | Static + dynamic | 127 patients (71 LGG, 47 HGG, 9 others), non-contrast-enhancing MRI lesions suspicious for LGG | No statistically significant differences in static parameters between LGG and HGG, unless excluding oligodendroglioma; decreasing TAC: PPV 74.1% HGG, increasing TAC: NPV 94.7% HGG |
| Rapp et al. (2013) [249] | Static | 174 patients (77 LGG, 66 HGG, 31 others) | Significant differences between gliomas and non-neoplastic lesions, significant differences between LGG and HGG, but not sufficient to justify clinical decision |
| Dunet et al. (2014) [250] | Static + dynamic | 38 patients (16 LGG, 22 HGG) | Poor accuracies for individual grading based on single parameters (ADC, magnetic resonance spectroscopy (MRS), PET); combination of decreasing TAC and bimodal ADC histogram yields best predictor of HGG |
| Albert et al. (2016) [251] | Static + dynamic | 314 patients (131 LGG, 183 HGG) | Static: best discrimination in 5–15 min time frame (accuracy 77.4%); dynamic: decreasing TACs yields accuracy of 79.7% |
| Pyka et al. (2016) [252] | Static | 113 HGG patients (26 WHO III, 87 WHO IV) | Linear combination of texture parameters and MTV yields grading accuracy of 85%; three texture parameters significantly correlated with OS and progression-free survival (PFS) |
| Verger et al. (2017) [253] | Static + dynamic | 72 patients (22 LGG, 50 HGG) | Significant differences between LGG and HGG based on static (AUC=0.80–0.83) and dynamic parameters (AUC=0.78); similar accuracies on PWI rCBV parameters (AUC=0.80–0.81) |
| Röhrich et al. (2018) [254] | Static + dynamic | 44 patients (10 WHO II, 13 WHO III, 21 GBM) | Better correspondence between kinetic behaviour of $^{18}$F-FET and methylation-based diagnosis than histological tumour classification; SUV significantly higher in GBM compared to lower-grade; no difference in SUV between IDH wildtype or mutant but TTP significantly shorter in IDH-mutant gliomas; significantly different $K1$ between GBM and lower-grade gliomas, both histologically and methylation-based |

The higher uptake in HGG compared to LGG is probably due to a higher regional blood volume, resulting from hypervascularisation, increased angiogenesis and intratumoural microvessel density, and facilitated amino acid transport in malignant glioma [247, 124]. The exact reasons for the difference in tracer kinetics are still not fully understood, and a detailed discussion is outside the scope of this thesis. A disruption in the blood-brain barrier (BBB) might play a role, but is probably not the only factor, since tracer washout is also observed in high-grade patients without contrast-enhancement on MRI [243]. With increasing malignancy, tumour cells could lose their ability to take up and retain $^{18}$F-FET [248]. Another mechanism could be the changed transport between intra- and extracellular amino acids, leading to a loss of unbound, non-metabolised $^{18}$F-FET [124].

Similarly to tumour grading, Jansen et al. [149] showed that oligodendrogliomas (with loss of heterozygosity on chromosomes 1p and 19q) show a significantly higher $^{18}$F-FET uptake compared to astrocytomas, especially in the WHO grade II gliomas. Moreover, they showed that tracer uptake was independent of grade in oligodendroglioma, while in astrocytoma a positive correlation between $^{18}$F-FET uptake and tumour grade was seen. However, there was a significant overlap in uptake values from oligodendroglial tumours and high-grade astrocytomas, and low-grade oligodendroglioma often showed a decreasing uptake curve mimicking high-grade tumours, making it difficult to provide a PET-based prediction of tumour type.

Finally, $^{18}$F-FET PET can be used to detect progression to malignancy of low-grade glioma, as was shown by Galldiks et al. [255]. Compared to MRI, where a change in contrast enhancement is an indication of tumour progression, PET achieves a significantly higher accuracy to identify malignant progression. Therefore, repeated $^{18}$F-FET PET can be considered for low-grade glioma patients for whom a watch-and-wait strategy was opted.

### Treatment planning and prognosis

Due to the specificity of $^{18}$F-FET imaging of glioma, it can play an important role in radiotherapy planning. Several authors compared the

biological tumour volume (BTV) based on $^{18}$F-FET PET with the conventional gross tumour volume (GTV) based on MRI [256, 257], showing the added value of amino acid imaging over MRI. On the other hand, differences between BTVs on early and late $^{18}$F-FET PET images might be an indication of IDH-status, as was shown recently [258].

Next to treatment planning, $^{18}$F-FET imaging can be used for prognosis. Floeth et al. [259] were able to stratify 33 low-grade glioma patients into three prognostic subgroups based on MRI and $^{18}$F-FET PET features: patients with no malignant transformation ($n = 11$), patients with tumour progression but a low probability of death ($n = 13$), and patients with tumour progression to malignancy and a high risk of death ($n = 9$). Jansen et al. [260] evaluated $^{18}$F-FET PET scans of low-grade astrocytoma patients. Their analysis showed that decreasing time-activity curves are highly prognostic for shorter PFS and time to malignant transformation. This finding was confirmed by Suchoska et al. [261] for the overall survival (OS) in glioblastoma patients. Moreover, they showed that a smaller BTV prior to radiochemotherapy is related to a longer PFS and OS in glioblastoma.

## Therapy response assessment and tumour recurrence

Amino acid imaging can also play an important role during and after therapy. A decrease in $^{18}$F-FET uptake in the early days after completion of chemoradiotherapy is a highly significant predictor for a longer PFS and OS in glioblastoma patients [262]. Also in patients with recurrent high-grade gliomas treated with antiangiogenic drugs such as bevacizumab, $^{18}$F-FET PET can be used to assess the treatment response better than MRI, both using static [263] and dynamic [264] parameters.

One of the main problems in the follow-up of glioma patients is distinguishing between treatment-related changes and tumour recurrence or progression. On conventional MRI this distinction is hard to make. Increased mass effect or changing contrast enhancement are typical features of tumour recurrence or progression, but can also occur as a side-effect (e.g. radionecrosis) to radio(chemo)therapy. These symptoms can happen in the early stage (within the first three months, i.e. pseudoprogression), but also later on, even up to several years after finishing the

therapy [265, 266, 125].

In 2004, Pöpperl et al. [267] showed that focally increased uptake of $^{18}$F-FET could confidently distinguish between benign therapy-induced changes and tumour recurrence/progression. However, there was a large signal overlap between the different tumour grades, making a differential diagnosis not possible based on static parameters alone. On a different patient cohort, they achieved a positive predictive value of glioma recurrence after treatment of 84% using the previously established $^{18}$F-FET criteria [268]. The same authors [247] showed that dynamic $^{18}$F-FET PET has an added value in pretreated patients with a suspicious MRI. Similarly as in untreated cases, tumour-free or low-grade tumour patients show an increased PET signal until the end of acquisition. However, in high-grade glioma patients, the uptake showed a peak around 5–15 minutes after injection and then decreased.

In a large study, Galldiks et al. [125] showed that a mean tumour-to-background ratio larger than 2 and a time-to-peak less than 45 minutes achieves an accuracy of 93% in identifying tumour recurrence or progression. A more recent study showed that static $^{18}$F-FET PET parameters can be used in the discrimination of tumour recurrence and treatment-related changes in glioblastoma patients treated with tumour-treating fields [269]. This is a recent technique applying low intensity alternating electric fields to the tumour, thereby inhibiting tumour growth [270].

Finally, Lohmann et al. [158] achieved an increased accuracy in distinguishing radiation injury from tumour recurrence in brain metastases when performing textural analysis. They showed that texture features calculated on standard static scans have an added value over conventional intensity-based and dynamic parameters.

### 7.1.3 Goal

The goal of our study is to accurately predict the tumour grade (low-grade or high-grade) of untreated glioma patients. Therefore, we will apply the methodology developed in the previous chapters to static $^{18}$F-FET PET images. This is preferred over dynamic scanning to maximise the patients' comfort due to the shorter scanning time. Inspired by the texture analysis approach of Pyka et al. [252], we will extract quan-

titative features and analyse them using machine learning. However, instead of delineating the tumour on the PET scan, we will automatically segment the tumour into several compartments on structural MRI images, and apply these masks to the PET scans.

## 7.2 Materials and methods

We start this section with an overview of the patient data that are used for the analyses. Afterwards, the preprocessing steps, automated tumour segmentation and feature extraction are explained. Finally, the quantitative features are analysed using machine learning techniques.

### 7.2.1 Data

We included thirty patients in this retrospective study. Inclusion criteria were the availability of both a $^{18}$F-FET PET and a structural MRI scan prior to surgery or treatment, and a histologically proven primary brain tumour. Fourteen patients were diagnosed with a low-grade glioma, among which 11 diffuse low-grade astrocytoma (WHO grade II), 1 low-grade ependymoma (WHO grade II), and 2 grade I neuronal-glial tumours (1 ganglioglioma and 1 rosette-forming glioneuronal tumour). From the 16 patients in the high-grade glioma class, there were 4 anaplastic astrocytomas (WHO grade III), 4 anaplastic oligodendrogliomas (WHO grade III) and 8 glioblastomas.

All patients fasted for at least 6 hours before an intravenous $^{18}$F-FET bolus injection (dose expressed in MBq) of 2.7–2.8 times the body weight (expressed in kg). Dynamic image acquisition started at the moment of tracer injection and lasted for 50 minutes. A summed image of the last 10 minutes of acquisition is saved as static scan.

For three patients, the PET scans were performed on an Allegro PET imaging system (Philips Co., Cleveland, Ohio, USA), which consists of a gadolinium oxyorthosilicate (GSO) full-ring PET scanner with a spatial resolution of about 5 mm (centre field-of-view (FOV): 4.65 mm vertical, 5.00 mm horizontal full width at half maximum (FWHM); 10 cm transverse offset Y: 5.47 mm vertical, 5.25 mm horizontal; 10 cm trans-

verse offset X: 5.26 mm vertical, 5.66 mm horizontal according to internal acceptance report). The system also includes $^{137}$Cs rods for transmission scanning, used for attenuation and scatter correction purposes. Scans are acquired in high-resolution mode (matrix $128 \times 128 \times 90$ and voxel size $2\,\text{mm} \times 2\,\text{mm} \times 2\,\text{mm}$). Image reconstruction was performed using the row-action maximum likelihood algorithm (RAMLA) in 3D, with two iterations and using generalized Kaiser-Bessel window functions ("blobs") as basis functions. Blob- and reconstruction parameters were optimized for brain PET imaging.

All other patients were scanned on a Biograph mCT Flow system (Siemens Healthcare, Erlangen, Germany), consisting of four lutetium oxyorthosilicate (LSO) detector rings of 842 mm in diameter [271]. The resolution is about 4.5 mm (centre FOV: 4.5 mm axial, 4.4 mm transverse; at 10 cm offset: 5.9 mm axial, 4.9 mm transverse, according to internal acceptance report). The ordered subset expectation maximization (OSEM) algorithm was used for image reconstruction with CT-based attenuation and scatter correction, using two iterations and 21 subsets including time-of-flight information and resolution recovery, and a Gaussian post filter of 3 mm. The resulting image matrix with UHD settings contains $400 \times 400 \times 400$ voxels with a size of $1.018\,\text{mm} \times 1.018\,\text{mm} \times 3\,\text{mm}$.

### 7.2.2 Preprocessing

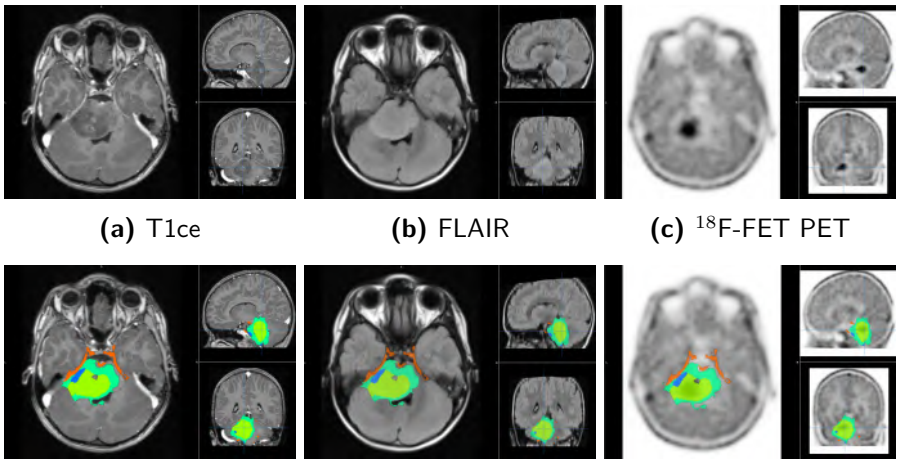
As before, the T1ce scans were normalised to MNI-space, trilinearly interpolated to a $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$ voxel size and bias field corrected using SPM12 (version 6906, Wellcome Trust Centre for Neuroimaging, London) [182], running on MATLAB R2017b (The MathWorks, Inc., Natick, MA). SPM12 was also used to coregister the FLAIR scans, if available, to the T1ce images. The $^{18}$F-FET PET scans were coregistered to the T1ce scans using PMOD (PMOD Technologies, Switzerland, `www.pmod.com`). This software was preferred over SPM12 since it yielded visually better results.

The intensities on the MRI scans were normalised using the white-stripe method [138]. Since the parameters necessary for conversion to standardised uptake value (SUV) were removed during the conversion

of the PET scans to NIfTI-format, we simply divided the intensities by the maximal value to obtain normalised images.

### 7.2.3 Segmentation and feature extraction

The automated Random Forests based segmentation algorithm developed in chapter 4 was applied to the T1ce and FLAIR scans. However, for 5 patients only a presurgical T1ce optimised for neuronavigation was available. Therefore, the workflow of chapter 4 was repeated in order to create a segmentation algorithm able to work on T1ce scans alone. Although achieving a lower performance than the version incorporating both T1ce and FLAIR scans, the method is still able to delineate the contrast-enhancement, tumour core and total abnormal region with reasonable accuracy [191]. The segmentation masks were transferred to the PET scans, as is illustrated in figure 7.3.



**(a)** T1ce **(b)** FLAIR **(c)** $^{18}$F-FET PET

**(d)** T1ce + segmentation **(e)** FLAIR + segmentation **(f)** $^{18}$F-FET PET + segmentation

**Figure 7.3:** Patient with a large lesion in the cerebellum showing high focal $^{18}$F-FET uptake diagnosed as glioblastoma. The automated segmentation method developed in chapter 4 is applied to the T1ce and FLAIR scans, and transferred to the PET scan. The segmentation process was not entirely successful, as the small contrast-enhancing regions in the middle of the tumour are missed, and blood vessels are assigned a contrast-enhancing tumour label.

Next, we calculated the 207 features introduced in chapter 3 in four tumour masks on both T1ce and $^{18}$F-FET PET scans:

- oedema
- non-enhancing, non-necrotic region
- tumour core
- total abnormal region.

These are complemented with 27 features capturing the location of the tumour and the contrast between the tumour intensities and the surrounding tissue. In this way a total of 1683 features were extracted for every patient. Since a FLAIR scan was not available for all patients, we did not calculate features on this scan type.

### 7.2.4 Feature reduction and machine learning

Every feature is first transformed to have zero mean and unit variance. Then, we hold out one patient as test set and use the other patients to train on. A two-sample $t$-test with the assumption of unequal variance is performed to select features with a significantly different mean between the LGG and HGG classes, and the 150 features with smallest $p$-value are retained. Next, sequential forward selection is applied using five-fold cross validation on the training set, assessed using Random Forests. In this way, 15 features with complementary predictive power are selected. This process is repeated for every patient.

For every left out patient, a different set of 15 features can be found, making it difficult to assess the global accuracy. Therefore, we combine the different feature sets in order to yield a ranked set of features valid for all patients. When a feature is selected first in the SFS algorithm, we accredit it with 15 points, when selected second 14 points, and so on. Then the scores for all features are summed up and ranked to yield the most predictive features.

Finally, the previously ranked features are used to train a Random Forests model. Since the class sizes are not equal, we train four different Forests, each time on random subsets of LGG and HGG patients with the

smallest class size. In particular, when evaluating a low-grade patient, we use the remaining 13 LGG samples, and randomly sample 13 of the 16 HGG patients to train a first model. When evaluating a high-grade patient, we use all 14 LGG samples, and sample 14 of the remaining 15 HGG samples. This procedure is repeated four times, as this reduces the probability for a patient not being selected to less than 0.1% ($(3/16)^4$). The output of these four models is then combined to yield the final prediction.

Due to the small sample size, we assess the accuracy of this method using leave-one-out validation. This experiment is repeated three times: using only T1ce-based features, only $^{18}$F-FET PET-based features, or the combination of both.

## 7.3 Results

### 7.3.1 Feature ranking

The best performing features for every model are shown in table 7.2. Many of the most predictive features are based on infiltrating or oedematous tissue rather than the tumour core (7/10 on T1ce, 6/10 on $^{18}$F-FET, 6/10 on the combination). Moreover, the GLCM-based feature "informational measure of correlation 2", a very complex texture parameter (see feature definition in Appendix A), is found seven times in the top-10 most predictive features on the combination of $^{18}$F-FET PET and T1ce.

In the combination list, only one out of ten features is based on T1ce. This feature also ranked first in the T1ce-based list. Moreover, most highly ranked texture features are based on a distance of 2–3 mm on both modalities. This demonstrates that relevant texture occurs at larger spatial scales than the 1 mm$^3$ voxel size.

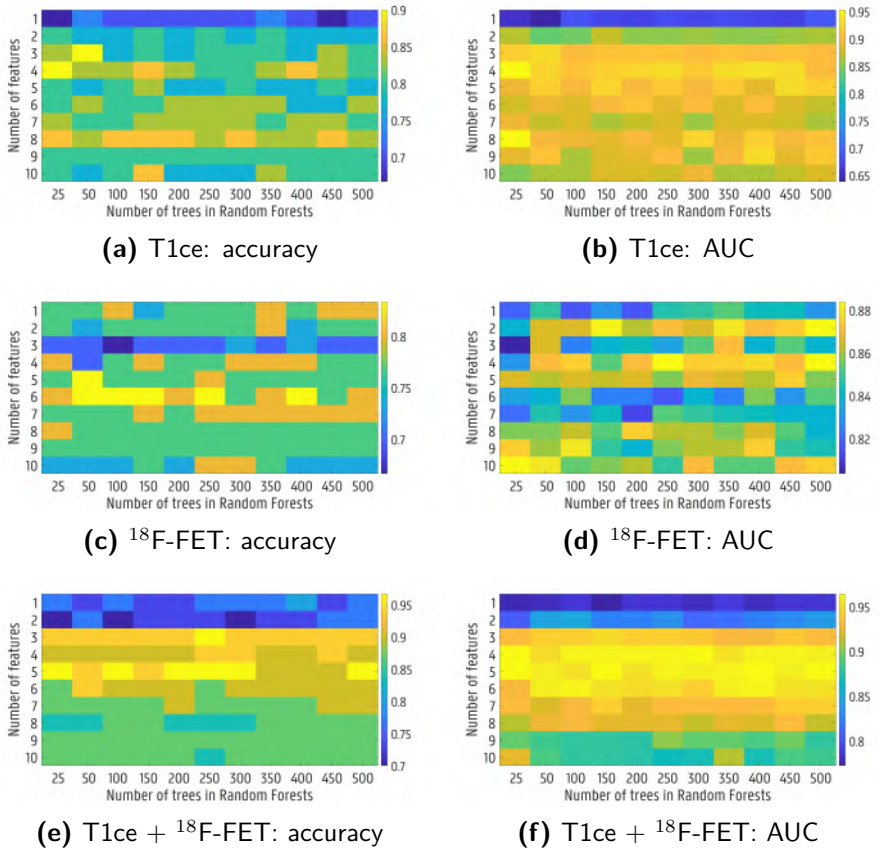
**Table 7.2:** Best feature for distinguishing LGG from HGG patients.

| | based on T1ce | based on $^{18}$F-FET PET | combination |
|---|---|---|---|
| 1 | non-enh/non-necr: GLCM - Informational measure of correlation 2 ($d$=2, std) | non-enh/non-necr: GLCM - Informational measure of correlation 2 ($d$=3, std) | FET - total abnormal: histogram - Range |
| 2 | non-enh/non-necr: GLRLM - High gray level run emphasis (std) | non-enh/non-necr: GLCM - Informational measure of correlation 2 ($d$=2, std) | FET - non-enh/non-necr: GLCM - Informational measure of correlation 2 ($d$=3, std) |
| 3 | core: GLCM - Variance ($d$=2, std) | total abnormal: GLCM - Informational measure of correlation 2 ($d$=2, std) | T1ce - non-enh/non-necr: GLCM - Informational measure of correlation 2 ($d$=2, std) |
| 4 | core: GLSZM - Zonelogl | non-enh/non-necr: GLCM - Correlation2 ($d$=2, mean) | FET - non-enh/non-necr: GLCM - Informational measure of correlation 2 ($d$=2, std) |
| 5 | oedema: GLRLM - Run length non-uniformity (mean) | oedema: GLCM - Informational measure of correlation 2 ($d$=2, std) | FET - total abnormal: GLCM - Informational measure of correlation 2 ($d$=3, std) |
| 6 | oedema: GLCM - Correlation2 ($d$=3, mean) | total abnormal: GLCM - Maximum probability ($d$=1, std) | FET - total abnormal: GLCM - Informational measure of correlation 2 ($d$=2, std) |
| 7 | non-enh/non-necr: Shape - Maximum 3D diameter | non-enh/non-necr: GLRLM - Run percentage (std) | FET - non-enh/non-necr: histogram - Range |
| 8 | non-enh/non-necr: GLRLM - Gray level non-uniformity (std) | core: GLRLM - Short run low gray level emphasis (std) | FET - oedema: GLCM - Informational measure of correlation 2 ($d$=2, std) |
| 9 | non-enh/non-necr: GLSZM - Glnonunif | total abnormal: GLCM - Correlation2 ($d$=2, mean) | FET - core: GLCM - Informational measure of correlation 1 ($d$=1, mean) |
| 10 | total abnormal: GLCM - Autocorrelation ($d$=3, std) | oedema: GLRLM - Long run low gray level emphasis (mean) | FET - oedema: GLCM - Correlation2 ($d$=2, mean) |

### 7.3.2 Model parameters leading to best performance

The optimal model parameters (number of trees and number of features) are determined using grid search, where different combinations of parameters are tested. The accuracies and corresponding AUC values for the different models are visually given in figure 7.4.

From these graphs, it is clear that the best performance is obtained when 3–6 features are included in the model combining $^{18}$F-FET PET and T1ce features. However, as including five features is most robust when changing the number of trees in the model, this setting was chosen. Similarly, we chose the optimal parameters for the individual PET and MRI models. Using T1ce alone, the best accuracy (27/30) is obtained when including 3–4 features and a low number of trees, but since this was




**(a)** T1ce: accuracy **(b)** T1ce: AUC

**(c)** $^{18}$F-FET: accuracy **(d)** $^{18}$F-FET: AUC

**(e)** T1ce + $^{18}$F-FET: accuracy **(f)** T1ce + $^{18}$F-FET: AUC

**Figure 7.4:** Grid search showing the prediction performance, validated by leave-one-out, of Random Forests classification model in function of the number of features and the number of trees. The total number of trees is four times higher, since the model is trained on four slightly different, balanced subsets, and the prediction scores are averaged. Optimal results for every model are given in table 7.3.

not reproducible when increasing the forest size, this result is probably not robust. Therefore, we chose to include 8 features instead. The corresponding accuracies and AUC-values are given in table 7.3.

Using 6 PET-based features, the model correctly classifies 25/30 patients, while using 8 T1ce-based features, the accuracy reaches 26/30. When combining both scans, the accuracy increases to 29/30 patients correctly classified. The one misclassified patient was diagnosed with a

**Table 7.3:** Model parameters leading to the best performance in the discrimination of LGG and HGG patients.

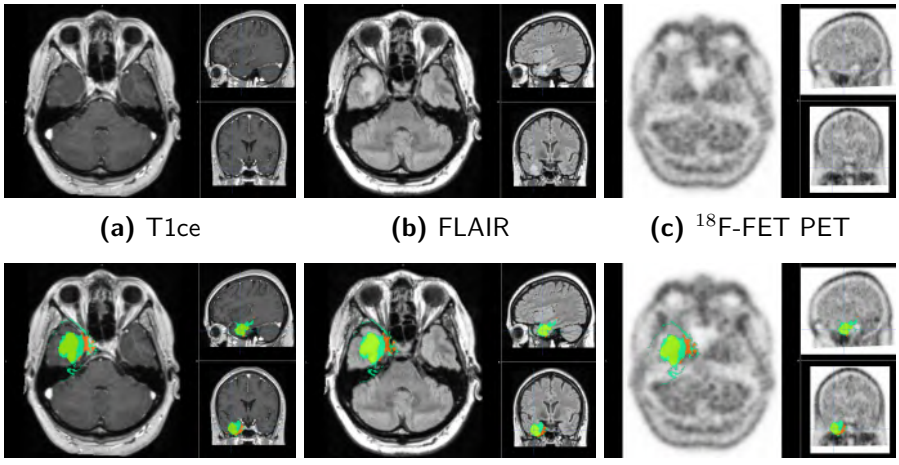
| | no. features | no. trees | accuracy | AUC |
|---|---|---|---|---|
| based on T1ce MRI | 8 | 25 | 86.7% | 0.953 |
| based on $^{18}$F-FET PET | 6 | 50 | 83.3% | 0.839 |
| combination | 5 | 25 | 96.7% | 0.964 |

diffuse low-grade astrocytoma, WHO grade II, but was predicted with 85.5% probability to be high-grade by the model. Scans of this patient are displayed in figure 7.5. Visually, there are no signs of a high-grade glioma, as it is a small temporal lobe lesion showing no contrast-enhancement or increased $^{18}$F-FET uptake. However, errors during the coregistration step (obvious when comparing the size of the cerebellum on the MRI and PET images) might lead to the poor prediction performance of this patient.

## 7.4 Discussion

In this chapter, we conducted a study based on quantitative features calculated on static $^{18}$F-FET PET and T1ce MRI in order to discriminate low-grade glioma from high-grade glioma patients. We achieve a near-perfect prediction of 29/30 correctly classified patients, which is matching the best performance in literature (Calcagni et al. [248]: 97% accuracy using logistic regression on two dynamic features including 32 patients).

The main difference between this study and studies in literature is the delineation of the tumour on the scan. Usually, the tumour is defined by an intensity threshold on the $^{18}$F-FET PET scan, such as a factor (e.g. 1.4) of the background uptake or a percentage of the maximal value in the tumour. Alternatively, a more robust tumour segmentation algorithm on the PET scan can be used (see e.g. the results on the first MICCAI challenge on PET tumour segmentation for a thorough comparison of different methods [272]). However, since gliomas do not always show an increased $^{18}$F-FET uptake compared to the background,



**(a)** T1ce **(b)** FLAIR **(c)** $^{18}$F-FET PET

**(d)** T1ce + segmentation **(e)** FLAIR + segmentation **(f)** $^{18}$F-FET PET + segmentation

**Figure 7.5:** Patient with a lesion in the left temporal lobe, showing hypointensity on T1ce, hyperintensity on FLAIR, but no clear $^{18}$F-FET uptake. Again, the segmentation process was not entirely successful, as blood vessels are assigned a contrast-enhancing tumour label. The patient was diagnosed as diffuse low-grade astrocytoma, WHO grade II, but misclassified as high-grade by the Random Forests model.

such an approach was not possible in our study. We therefore chose to automatically delineate the tumour on MRI scans into several compartments. This has the added advantage of using the high-resolution anatomical information, instead of the much lower resolution obtained in PET imaging. Still, as the GTV seen on MRI does not always agree with the BTV seen on PET [256, 257], relevant information on the tumour visible on PET might not be included. Moreover, good performance of the coregistration step of the PET image to the T1ce scan is required.

The only study up to now evaluating the role of texture analysis in $^{18}$F-FET imaging with the goal of tumour grading is performed by Pyka et al. [252]. They only included grade III and grade IV glioma, and calculated textural features based on the neighbourhood grey-tone difference matrix (NGTDM). Combined with the tumour volume, their model achieved an accuracy of 87%. This is slightly higher than the per-

formance we achieved when only including PET-based features (83.3%). A significantly larger patient population (113 patients: 22 grade III, 87 grade IV) might explain the better result. Although we also included the same NGTDM features, there are only features based on the GLCM in our best performing model. A possible explanation for this diverging finding might be that the NGTDM retains the units of the original data, being SUV in the study of Pyka et al. Since we were lacking the necessary information to transform the data into SUV due to the anonymisation process, we had to apply a different intensity normalisation procedure. This might also cause the histogram features to be less robust. There are indeed no histogram features in the selected list based on $^{18}$F-FET PET alone. However, the feature "range" on the total abnormal region was selected first in the combined feature list.

Our grid search analysis shows that the best results are obtained when using a relatively low number of trees per forest (either $4 \times 25$ or $4 \times 50$), and that increasing the number of trees does not lead to a better performance, but to a more computationally costly procedure. This is in line with the results of Oshiro et al. [273]. They applied Random Forests for classification on 29 different datasets, and concluded that using 64–128 trees in a model shows the optimal trade-off between computational cost and performance.

In chapters 3 and 6 we also performed binary classification between different tumour grades, based on structural MRI parameters. For distinguishing between lower-grade tumours and glioblastoma, we achieved accuracies of 88.0%, while for distinguishing between LGG and HGG the accuracy was 83.0%. From these analyses, it was clear that the best performing features are based on the T1ce scan. In this chapter, we obtain a performance of 86.7% when only including T1ce features in the discrimination of LGG and HGG samples. However, when including $^{18}$F-FET PET features as well, the accuracy increases with 10%, proving the added value of amino acid imaging for glioma grading. To the best of our knowledge, this is the first study to combine MRI and PET features for this purpose.

There are some limitations to our study. First of all, the feature selection procedure is not independent from the test set, which might cause overfitting. This leads to features that are best suited to describe

our dataset, but we have little certainty that this will also be the case on an independent dataset. Therefore, our method should be thoroughly validated on a large and independent dataset in order to gauge the real performance. However, since we only had data of 30 patients, this was not possible.

Secondly, we did not use dynamic data, nor did we have information on the molecular diagnosis of our patients. A recent study by Röhrich et al. [254] showed that dynamic $^{18}$F-FET PET is very well suited to predict the methylation status of gliomas, while a worse prediction performance is obtained when the tumours are classified according to the standard histological parameters. Specifically, a significant difference in time-to-peak was noted when the IDH-wildtype lower-grade glioma were included in the glioblastoma group, whereas this was not the case based on histology. Moreover, all the histological grade II/III gliomas which were clustered in the GBM methylation group showed a HGG-like TAC. As was shown by Jansen et al. [260], patients with a typical high-grade uptake profile show a highly significant worse prognosis. It is therefore not unthinkable that in future updates of the WHO classification of gliomas, both tumour methylation status and dynamic $^{18}$F-FET PET behaviour will be taken into account for the distinction between low-grade and high-grade tumours.

In our dataset, IDH-status is only based on immunohistochemical (IHC) analysis for mutant R132H IDH1 protein. Nine LGG grade II and two grade III patients had a negative IHC result, but according to the WHO guidelines, this is not sufficient for the diagnosis of a IDH-wildtype tumour [54], as sequencing is not available. Moreover, one glioblastoma had a positive IHC and is therefore IDH-mutant. This patient should therefore be rearranged to a high-risk IDH-mutant glioma without 1p/19q codeletion subgroup. More crucially however, the diagnosis of 16 patients was based on a biopsy sample, whereas the remaining 14 patients received a tumour resection. This might lead to several erroneous groundtruth grade labels [274, 89].

In a follow-up study, it would be very interesting to learn how well texture parameters, calculated on a static image and possibly complemented with MRI-features, compare to dynamic $^{18}$F-FET behaviour in predicting the tumour status of patients who are fully stratified accord-

ing to their molecular profile. Since for static images a much shorter scanning protocol is needed (e.g. 10 minutes compared to 40 minutes for dynamic acquisition), this increases the comfort of the patient, and makes it possible to scan more patients per day.

## 7.5 Conclusion

In this chapter, we showed that amino acid imaging by means of $^{18}$F-FET PET can complement MRI for the grading of primary brain tumours. Automatically segmented tumour regions on MRI were transferred to the PET images, and from both modalities we calculated quantitative features. In a Random Forests classification model, 29 out of 30 patients were correctly predicted. A thorough validation on a large and independent dataset with full molecular tumour characterisation is however necessary to confirm this result.

# 8 Conclusion and future perspectives

In this final chapter, we look back at each part of this thesis and summarise the main conclusions that can be drawn from them. Based on these results, we formulate possibilities for future research directions. Finally, we end this book with a conclusion.

## 8.1 Summary

In chapter 2 we introduced five research domains related to this PhD dissertation. Computer-aided diagnosis was defined as a class of computer systems that aim to assist in the detection and/or diagnosis of diseases through a "second opinion". Different CAD systems are already routinely used in clinical practice, such as in screening mammography or chest radiography. Next, we focused on neuro-oncology, and mainly on primary brain tumours. In the 2016 classification of the WHO, for the first time genetic and molecular parameters are integrated, complementing the histological findings. This leads to a more objective diagnosis, which is also better able to describe the patient's prognosis. This section was followed by explaining the working principles and neuro-oncological applications of MRI and PET. These imaging modalities provide structural and functional information on the tumour and its environment. They are however not able to directly map the important genetic and molecular features. The *radiomics* hypothesis is that distinct microscopical patterns might be translated into imaging signals that can be picked

up by dedicated computer algorithms. Therefore, techniques from machine learning are applied to analyse feature vectors extracted from the medical images. We explained several techniques from both supervised and unsupervised learning.

Primary brain tumour grading has important prognostic and therapeutic implications. Therefore, we studied a non-invasive tool in order to distinguish between lower-grade gliomas and glioblastomas in chapter 3. The BraTS 2017 dataset, consisting of 75 lower-grade and 210 high-grade glioma patients, was used for this experiment. For every patient, four different MRI scans and a manual tumour delineation are provided. We explained different types of quantitative features, including histogram, shape, texture and environment parameters. In total, 2097 features were extracted per patient, capturing the appearance on two MRI sequences and in different tumour regions. Since this number is much larger than the amount of patients, we discussed several dimensionality reduction methods. Finally, we compared the performance of five binary classifiers combined with six feature reduction techniques. The best result was obtained when using a Random Forests classifier including 700 features, achieving an accuracy of 88.0%. When combining Random Forests with SFS, an accuracy of 85.7% was obtained with only 24 features.

Brain tumour segmentation on medical images is a time-consuming job and can give rise to inter- and intra-observer variability when performed in a manual fashion. Moreover, results in literature have shown that quantitative features are more robust when calculated on (semi-) automatically segmented tumour masks. Therefore, we examined different automated approaches towards brain tumour segmentation in chapter 4. These can roughly be divided into three categories: generative methods, discriminative methods, and deep learning. We developed a flexible, generative algorithm based on outlier detection. This method is able to accept any number and type of scans. Without requiring a training set, it achieves a median Dice score of 73.3% on the BraTS dataset. However, as it relies on the detection of healthy tissue, an excellent coregistration of healthy tissue probabilities maps is necessary. Moreover, although segmenting the tumour into different classes, only a reliable label for the entire abnormal region is provided.

Therefore, we developed a second, discriminative approach. For every voxel, texture and abnormality features are calculated and classified by a pretrained Random Forests model. This method requires a T1ce and FLAIR MRI scan, and recognises the appearance of different tumour tissues. Especially for high-grade glioma, this method provides good results, with median Dice scores of 74.8%, 75.0% and 80.1% for segmenting the enhancing tumour, tumour core and total abnormal region, respectively. However, discriminating between oedema and non-enhancing tumour proved to be difficult.

This segmentation algorithm was applied to clinical scans in chapter 5. We collected 352 patients from 8 different centres, divided into 6 tumour classes. For every patient, we extracted a total of 2097 features, divided over two MRI scans and five tumour regions. These features were implemented into multiclass Random Forests classification models for grade and tumour type. Using these classifiers, tumour grade can be predicted with an overall accuracy of 60.3%, and tumour type with 65.6%. However, the models show a poor performance for grade III gliomas and oligodendrogliomas, respectively, hampering the use in clinical practice.

Therefore, we investigated 14 specific binary models in chapter 6. These can provide answers when previous knowledge on the tumour is available. Suppose for example that the radiologist suspects a lower-grade glioma based on the images, then the models are able to accurately predict probabilities for binary questions such as "grade II versus grade III", or "astrocytoma versus oligodendroglioma". For every one of the 14 classifiers accuracies exceeding 75% are obtained. In a second part, we designed four decision schemes combining the binary classifiers into a multiclass model. Our best result was obtained when first distinguishing between the lower-grade gliomas and meningiomas/glioblastomas, followed by splitting the lower-grade gliomas according to type, and finally by grade, yielding an overall accuracy of 52.8%. Letting the computer decide how to best combine the scores from the binary models did not improve this result.

Finally, in chapter 7 we were able to optimise the automated distinction between low-grade and high-grade glioma based on features calculated on $^{18}$F-FET PET. In literature, there is sufficient evidence towards

using the dynamic uptake of amino acid radiotracers as a biomarker of tumour grade. In this dissertation, we showed that a Random Forests classifier trained on static $^{18}$F-FET PET and T1ce MRI texture features achieve a near-perfect accuracy of 29/30 correctly predicted patients. This would mean that the scanning time can be significantly reduced, thereby increasing the patient's comfort. However, validation on a larger and independent dataset is necessary to confirm this result.

## 8.2 Research possibilities

First of all, this work should be viewed in the light of computer-aided diagnosis. We have stressed that all our models provide probabilities for different possible classes, which can be interpreted by a radiologist in order to lead to the correct diagnosis. An interesting study is therefore to measure in fact how well the computed models can aid the radiologist. A possible study design to perform this experiment consists of three branches. As a first task, one or preferably more radiologists are asked to look at brain tumour scans, and subsequently write down the diagnosis they deem most probable. In the second step, the same physicians look at scans from different patients, but they simultaneously receive the output probabilities from all machine learning models. In the last part, the radiologists again get a different set of images and are asked to give their most probable diagnosis, after which they can reveal a selection of their choice of the computer predictions and adapt their diagnosis if wanted. This experiment will not only allow to gauge the human performance, but also if this improves when using the computer models. Moreover, the last section will show which classifiers are deemed necessary, and to which extent this influences the decision of the physician.

As a second future perspective, a larger amount of more detailed patient data could significantly impact the performance. Combining datasets from different centres would for example not only yield more patients, but would also further increase the heterogeneity of imaging settings. A careful analysis of the extracted feature vectors will result in features that are able to overcome this heterogeneity, and are therefore stable across imaging systems, increasing the clinical applicability.

Moreover, we could further expand the datasets with, next to $^{18}$F-FET PET, images from diffusion and perfusion weighted MRI. These imaging techniques provide a detailed view on the tumour biology. Results in literature convincingly show the added value for several classification tasks using these advanced MRI methods. When collecting a large database, one should also pay attention to the full molecular and genetic characterisation of the patients. This will allow to non-invasively predict some of these biomarkers, such as IDH-status, MGMT promoter methylation, or proliferation with ki-67. This is the research domain of *radiogenomics*.

A final possible research track is the improvement of the techniques we used. Most of the current applications of AI are based on deep learning. In a first step, the segmentation can be improved using CNNs, as this is the method of choice of the winners of the last few editions of the BraTS competition. A better segmentation result will also result in more robust extracted features, and therefore an improved classification result. In a second phase, the segmentation phase can even by bypassed using deep learning, and CNNs can be learned that are directly able to predict tumour grade, type or molecular parameters based on the scans themselves instead of extracted features. However, when trained from scratch, these models need a very large training set, typically consisting of tens of thousands patients. A transfer learning approach can reduce the need for these large amounts of data. In this case, the network is first learned on a large labelled database such as ImageNet, containing millions of images in thousands of categories such as “balloon” or “strawberry”. In this way, the model learns to extract robust features which contain a large amount of information. Afterwards, the pretrained network parameters are further optimised for the specific tumour classification task.

## 8.3 Conclusion

In this dissertation, we have shown several applications of artificial intelligence that can aid in the diagnosis of primary brain tumours based on medical images. We have developed an automated brain tumour segmentation algorithm using Random Forests classification. The de-

lineated tumour regions can be used to extract quantitative features describing the appearance of the tumour on different scans. Based on these features, we were able to solve 14 binary problems with high accuracy using structural MRI scans. When combined with texture features calculated on static PET scans of the amino acid radiotracer $^{18}$F-FET, we achieve a near-perfect prediction of tumour grade, a finding with important prognostic and therapeutic consequences for brain tumour patients. These techniques can thus aid physicians in non-invasively predicting the diagnosis prior to surgery.

# A
# Radiomics features

In this appendix, we provide formal definitions for the quantitative features that we used in this dissertation. Most formulas are identical to those by Aerts et al. [144], except for the features based on the GLSZM and NGTDM, where we used the definitions from Willaime et al. [209].

## Histogram

Assume that the 3D image is given by the matrix $\mathbf{X}$. The tumour ROI consists of $N$ voxels, all other voxels are set to not-a-number (NaN). We calculate the histogram $\mathbf{P}$ with $N_l$ discrete intensity levels. The histogram features are then given by:

$$\begin{aligned}
\text{energy} \quad & = \sum_{i=1}^{N} \mathbf{X}(i)^2 \\
\text{entropy} \quad & = -\sum_{i=1}^{N_l} \mathbf{P}(i) \log_2 \mathbf{P}(i) \\
\text{mean} \quad & = \bar{X} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}(i)
\end{aligned}$$

| | |
|---|---|
| kurtosis | $= \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}(i)-\bar{X}\right)^4}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}(i)-\bar{X}\right)^2}\right)^4}$ |
| minimum | = minimum intensity value of $\mathbf{X}$ |
| maximum | = maximum intensity value of $\mathbf{X}$ |
| median | = median intensity value of $\mathbf{X}$ |
| mean absolute deviation | $= \frac{1}{N}\sum_{i=1}^{N}\lvert\mathbf{X}(i)-\bar{X}\rvert$ |
| range | = maximum − minimum |
| root mean square | $= \sqrt{\frac{\sum_{i=1}^{N}\mathbf{X}(i)^2}{N}}$ |
| skewness | $= \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}(i)-\bar{X}\right)^3}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}(i)-\bar{X}\right)^2}\right)^3}$ |
| standard deviation | $= \left(\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{X}(i)-\bar{X}\right)^2\right)^{\frac{1}{2}}$ |
| uniformity | $= \sum_{i=1}^{N_l}\mathbf{P}(i)^2$ |
| variance | $= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{X}(i)-\bar{X}\right)^2$ |

# Shape and size

| | |
|---|---|
| volume | $= V$ = number of voxels in the ROI multiplied with the voxel size |
| area | $= A$ = surface area, calculated by triangulation<br>$= \sum_{i=1}^{N}\frac{1}{2}\lvert\mathbf{a}_i\mathbf{b}_i \times \mathbf{a}_i\mathbf{c}_i\rvert$<br>where $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ are edge vectors of the $N$ triangles |
| surface to volume ratio | $= \frac{A}{V}$ |

compactness 1 $= \frac{V}{\sqrt{\pi} A^{\frac{2}{3}}}$

compactness 2 $= 36\pi \frac{V^2}{A^3}$

spherical disproportion $= \frac{A}{4\pi R^2}$
with $R = \left(\frac{3V}{4\pi}\right)^{\frac{1}{3}}$ the radius of a sphere with the same volume as the ROI

sphericity $= \frac{\pi^{\frac{1}{3}} (6V)^{\frac{2}{3}}}{A}$

maximum 3D diameter = largest pairwise Euclidean distance between voxels on the surface of the ROI

# Texture

## GLCM

If $GLCM$ is the grey-level co-occurrence matrix of the ROI with $N$ voxels, in direction $\theta$ with distance $d$ and $N_g$ the number of discrete intensity levels, then we first define:

$$p(i,j) = \frac{GLCM(i,j)}{\sum_{i,j} GLCM(i,j)} \qquad p_x(i) = \sum_{j=1}^{N_g} p(i,j)$$

$$\mu = \frac{1}{N} \sum_{i,j} p(i,j) \qquad p_y(j) = \sum_{i=1}^{N_g} p(i,j)$$

$\mu_x$ = mean of $p_x$ $\qquad$ $\sigma_x$ = standard deviation of $p_x$

$\mu_y$ = mean of $p_y$ $\qquad$ $\sigma_y$ = standard deviation of $p_y$

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)\,;\, i+j=k\,;\, k = 2, ..., 2N_g$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)\,;\, |i-j|=k\,;\, k = 0, ..., N_g - 1$$

$$H_X = -\sum_{i=1}^{N_g} p_x(i) \log_2 \left(p_x(i)\right) \quad (= \text{entropy of } p_x)$$

$$H_Y = -\sum_{j=1}^{N_g} p_y(j) \log_2 \left(p_y(j)\right) \quad (= \text{entropy of } p_y)$$

$$H_{XY} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j) \log_2 \left(p(i,j)\right) \quad (= \text{entropy of } p)$$

$$H_{XY1} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j) \log_2 \left(p_x(i)\, p_y(j)\right)$$

$$H_{XY2} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_x(i)\, p_y(j) \log_2 \left(p_x(i)\, p_y(j)\right)$$

$$\mu_{x2} = \sum_{i=1}^{N_g} i\, p_x(i) \qquad \sigma_{x2} = \sqrt{\sum_{i=1}^{N_g} (i-\mu_{x2})^2 p_x(i)}$$

$$\mu_{y2} = \sum_{j=1}^{N_g} j\, p_y(j) \qquad \sigma_{y2} = \sqrt{\sum_{y=1}^{N_g} (j-\mu_{y2})^2 p_y(j)}$$

We can now calculate the following features. Every GLCM feature is calculated as the mean of the feature calculations for each of the 13 directions.

$$\begin{aligned}
\text{autocorrelation} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ij\, p(i,j) \\
\text{cluster prominence} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \left[i + j - \mu_x - \mu_y\right]^4 p(i,j) \\
\text{cluster shade} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \left[i + j - \mu_x - \mu_y\right]^3 p(i,j) \\
\text{cluster tendency} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \left[i + j - \mu_x - \mu_y\right]^2 p(i,j)
\end{aligned}$$

$$
\begin{aligned}
\text{contrast} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} |i-j|^2\, p(i,j) \\
\text{correlation} &= \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (ij\, p(i,j)) - \mu_x\mu_y}{\sigma_x\sigma_y} \\
\text{correlation 2 [275]} &= \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i-\mu_{x2})(j-\mu_{y2})\, p(i,j)}{\sigma_{x2}\sigma_{y2}} \\
\text{difference entropy} &= -\sum_{i=0}^{N_g} p_{x-y}(i)\log_2\left(p_{x-y}(i)\right) \\
\text{dissimilarity} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} |i-j|\, p(i,j) \\
\text{energy} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \left(p(i,j)\right)^2 \\
\text{entropy} &= -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log_2\left(p(i,j)\right) \\
\text{homogeneity 1} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{1+|i-j|} \\
\text{homogeneity 2} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{1+(i+j)^2} \\
\text{informational measure of correlation 1} &= \frac{H_{XY} - H_{XY1}}{\max[H_X, H_Y]} \\
\text{informational measure of correlation 2} &= \sqrt{1-\exp\left(-2(H_{XY2}-H_{XY})\right)} \\
\text{inverse difference normalised} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{1+\left(\frac{|i-j|}{N}\right)} \\
\text{inverse variance} &= \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p(i,j)}{|i-j|^2}\ ,\ i \neq j \\
\text{maximum probability} &= \max\left[p(i,j)\right] \\
\text{sum average} &= \sum_{i=2}^{2N_g} \left(i\, p_{x+y}(i)\right)
\end{aligned}
$$

$$\begin{aligned}
\text{sum entropy} &= SE = \sum_{i=2}^{2N_g} p_{x+y}(i) \log_2 \left(p_{x+y}(i)\right) \\
\text{sum variance} &= \sum_{i=2}^{2N_g} (i - SE)^2 p_{x+y}(i) \\
\text{variance} &= \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - \mu)^2 \, p(i,j)
\end{aligned}$$

## GLRLM

If $p = GLRLM$ is the grey-level run-length matrix of the ROI consisting of $N$ voxels, in direction $\theta$ with $N_g$ the number of discrete intensity levels and $N_r$ the longest run-length, then we can calculate the following features, and again average them over 13 directions.

$$\begin{aligned}
\text{short run emphasis} &= \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{p(i,j)}{j^2}}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)} \\
\text{long run emphasis} &= \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} j^2 p(i,j)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)} \\
\text{grey-level non-uniformity} &= \frac{\sum_{i=1}^{N_g} \left[\sum_{j=1}^{N_r} p(i,j)\right]^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)} \\
\text{run-length non-uniformity} &= \frac{\sum_{j=1}^{N_r} \left[\sum_{i=1}^{N_g} p(i,j)\right]^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r}} \\
\text{run percentage} &= \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{p(i,j)}{N} \\
\text{low grey-level run emphasis} &= \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{p(i,j)}{i^2}}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)} \\
\text{high grey-level run emphasis} &= \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} i^2 p(i,j)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)} \\
\text{short run low grey-level emphasis} &= \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{p(i,j)}{i^2 j^2}}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)}
\end{aligned}$$

$$\text{short run high grey-level emphasis} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{i^2\,p(i,j)}{j^2}}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j)}$$

$$\text{long run low grey-level emphasis} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{j^2\,p(i,j)}{i^2}}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j)}$$

$$\text{long run high grey-level emphasis} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}i^2j^2\,p(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}p(i,j)}$$

## GLSZM

If $p = GLSZM$ is the grey-level size-zone matrix of the ROI consisting of $N$ voxels, with $N_g$ the number of discrete intensity levels and $N_z$ the largest size-zone, then we first define:

$$\mu_{gl} = \frac{1}{N_g \times N_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} i \times\, p(i,j)$$

$$\mu_{Nz} = \frac{1}{N_g \times N_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} \text{szpcent} \times j \times p(i,j)$$

We can now calculate the following features.

$$\text{szpcent} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}p(i,j)}{N}$$

$$\text{smallzone} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}\left(\frac{\text{szpcent}}{j}\right)^2 p(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}p(i,j)}$$

$$\text{largezone} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}(j \times \text{szpcent})^2\,p(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}p(i,j)}$$

$$\text{glnonunif} = \frac{\sum_{i=1}^{N_g}\left[\sum_{j=1}^{N_z}p(i,j)\right]^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}p(i,j)}$$

$$\text{sznonunif} = \frac{\sum_{j=1}^{N_z}\left[\sum_{i=1}^{N_g}p(i,j)\right]^2}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}p(i,j)}$$

$$\text{zonelogl} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}\frac{p(i,j)}{i^2}}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z}p(i,j)}$$

$$
\begin{aligned}
\text{zonehigl} &= \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} i^2\, p(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} p(i,j)} \\
\text{szonelogl} &= \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} \left(\frac{\text{szpcent}}{i\,j}\right)^2 p(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} p(i,j)} \\
\text{szonehigl} &= \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} \left(\frac{i\times\text{szpcent}}{j}\right)^2 p(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} p(i,j)} \\
\text{lzonelogl} &= \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} \left(\frac{j\times\text{szpcent}}{i}\right)^2 p(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} p(i,j)} \\
\text{lzonehigl} &= \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} i^2 j^2 \times \text{szpcent}^2 \times p(i,j)}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} p(i,j)} \\
\text{glvariance} &= \left(\frac{1}{N_g \times N_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} \left(i \times p(i,j) - \mu_{gl}\right)^2\right)^{1/2} \\
\text{szvariance} &= \left(\frac{1}{N_g \times N_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_z} \left(j \times \text{szpcent} \times p(i,j) - \mu_{Nz}\right)^2\right)^{1/2}
\end{aligned}
$$

## NGTDM

If $M = NGTDM$ is the neighbourhood grey-tone difference matrix of the ROI consisting of $N$ voxels, with $N_g$ the number of discrete intensity levels, $N_i$ the number of voxels of intensity $i$, $p_i = N_i/N$ the probability of occurrence of intensity $i$, $N_t$ the number of different grey-levels present in the ROI, and $\varepsilon$ a small number, then we can calculate the following features.

$$
\begin{aligned}
\text{coarseness} &= \left[\varepsilon + \sum_{i=1}^{N_g} p_i\, M(i)\right]^{-1} \\
\text{contrast} &= \left[\frac{1}{N_t(N_t-1)}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_i p_j (i-j)^2\right]\left[\frac{1}{N^2}\sum_{i=1}^{N_g} M(i)\right]
\end{aligned}
$$

$$\text{complexity} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left[\frac{|i-j|}{N_i+N_j}\right][p_i\,M(i)+p_j\,M(j)]$$

$$\text{strength} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(p_i+p_j)(i-j)^2}{\varepsilon+\sum_{i=1}^{N_g}M(i)}\ ,\ p_i\neq 0, p_j\neq 0$$

## Localisation and environment features

For the tumour core and the total abnormal region, we also calculate features capturing the localisation and the contrast between the ROI and surrounding tissue. These are:

$x$-coordinate centre of mass

$y$-coordinate centre of mass

$z$-coordinate centre of mass

distance to centre coordinate system ($d=\sqrt{x^2+y^2+z^2}$)

ratio average intensity inside ROI and average intensity in region 3 mm outside ROI (on original image)

ratio average intensity inside ROI and average intensity in region 5 mm outside ROI (on original image)

ratio average intensity inside ROI and average intensity in region 3 mm outside ROI (after intensity normalisation)

ratio average intensity inside ROI and average intensity in region 5 mm outside ROI (after intensity normalisation)

# B
# Features used in segmentation algorithm

On the next page, a table is given with all the 275 feature maps calculated for the Random Forests based segmentation algorithm. The 52 features selected for the final model are depicted in bold. (DS = downsizing level)

**Table B.1:** Overview of the features used in Random Forests based segmentation method (DS = downsizing level).

| | | **T1ce** | | | | **FLAIR** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | DS 1 | DS 2 | DS 4 | DS 8 | DS 1 | DS 2 | DS 4 | DS 8 |
| texture features | | $I_{\text{disrc}}$ | $I_{\text{disrc}}$ | $I_{\text{disrc}}$ | $I_{\text{disrc}}$ | $I_{\text{disrc}}$ | $I_{\text{disrc}}$ | $I_{\text{disrc}}$ | $I_{\text{disrc}}$ |
| | GLCM | **autoc** | autoc | **autoc** | **autoc** | autoc | autoc | **autoc** | **autoc** |
| | | **ClT** | ClT | **ClT** | **ClT** | ClT | ClT | **ClT** | ClT |
| | | **corr** | corr | corr | corr | corr | corr | corr | corr |
| | | dissim | dissim | dissim | **dissim** | dissim | dissim | dissim | dissim |
| | | energy | energy | energy | energy | energy | energy | energy | energy |
| | | homog | homog | **homog** | **homog** | homog | homog | **homog** | **homog** |
| | | MaxP | MaxP | MaxP | MaxP | MaxP | MaxP | MaxP | MaxP |
| | | SumAvg | SumAvg | SumAvg | **SumAvg** | SumAvg | SumAvg | SumAvg | SumAvg |
| | | var | var | var | var | var | var | **var** | var |
| | GLRLM | SRE | SRE | SRE | SRE | SRE | SRE | SRE | **SRE** |
| | | LRE | LRE | LRE | **LRE** | LRE | LRE | LRE | LRE |
| | | GLN | GLN | GLN | GLN | GLN | GLN | GLN | GLN |
| | | RLN | RLN | RLN | RLN | RLN | RLN | RLN | **RLN** |
| | | **LGLRE** | LGLRE | LGLRE | LGLRE | **LGLRE** | LGLRE | **LGLRE** | LGLRE |
| | | HGLRE | HGLRE | HGLRE | HGLRE | HGLRE | HGLRE | **HGLRE** | HGLRE |
| | | SRLGLE | SRLGLE | SRLGLE | SRLGLE | SRLGLE | SRLGLE | SRLGLE | SRLGLE |
| | | SRHGLE | SRHGLE | **SRHGLE** | **SRHGLE** | SRHGLE | SRHGLE | **SRHGLE** | **SRHGLE** |
| | | **LRLGLE** | LRLGLE | LRLGLE | LRLGLE | LRLGLE | LRLGLE | **LRLGLE** | LRLGLE |
| | | LRHGLE | LRHGLE | LRHGLE | LRHGLE | LRHGLE | LRHGLE | LRHGLE | LRHGLE |
| | GLSZM | SmZone | SmZone | SmZone | SmZone | SmZone | SmZone | SmZone | SmZone |
| | | LgZone | LgZone | LgZone | LgZone | LgZone | LgZone | LgZone | LgZone |
| | | GLnonu | GLnonu | GLnonu | GLnonu | GLnonu | GLnonu | GLnonu | GLnonu |
| | | SZnonu | SZnonu | SZnonu | SZnonu | SZnonu | SZnonu | SZnonu | SZnonu |
| | | **ZnLoGL** | ZnLoGL | ZnLoGL | ZnLoGL | ZnLoGL | ZnLoGL | ZnLoGL | ZnLoGL |
| | | ZnHiGL | ZnHiGL | ZnHiGL | **ZnHiGL** | ZnHiGL | ZnHiGL | ZnHiGL | **ZnHiGL** |
| | | SZLoGL | SZLoGL | SZLoGL | SZLoGL | SZLoGL | SZLoGL | SZLoGL | SZLoGL |
| | | SZHiGL | SZHiGL | SZHiGL | **SZHiGL** | SZHiGL | SZHiGL | SZHiGL | SZHiGL |
| | | LZLoGL | LZLoGL | LZLoGL | LZLoGL | LZLoGL | LZLoGL | LZLoGL | LZLoGL |
| | | LZHiGL | LZHiGL | LZHiGL | LZHiGL | LZHiGL | LZHiGL | LZHiGL | **LZHiGL** |
| (ab)normality | | **TPM1** TPM2 **TPM3** **TPM4** TPM5 | | | | | | | |
| | | $P_{GM}$ $P_{WM}$ $P_{CSF}$ $P_{\text{non-brain}}$ $P_{\text{tumour}}$ | | | | | | | |
| | | **aZnLoGL$_{GM}$** **aZnLoGL$_{WM}$** | | | | **aZnHiGL$_{GM}$** **aZnHiGL$_{WM}$** **aZnLoGL$_{GM}$** **aZnLoGL$_{WM}$** | | | |
| | | **Z-map** symmetry | | | | **Z-map** symmetry | | | |
| | | **symmetry$_{discr}$ (DS 1)** **symmetry$_{discr}$ (DS 4)** | | | | **symmetry$_{discr}$ (DS 1)** **symmetry$_{discr}$ (DS 4)** | | | |
| | | $P_{GM}$ $P_{WM}$ $P_{CSF}$ | | | | $P_{\text{non-brain}}$ **$P_{tumour}$** **Z-map** | | | |

# C

# Optimal parameters for multiple binary classification

## Grid search

We assessed the optimal hyperparameters for the multiple binary classification algorithm by calculating the accuracy of decision scheme 4 (the best method) for different combinations of the `MinLeafSize` parameter and the number of features included in the Random Forests models, averaged over 200 iterations with 80 trees per Random Forest. Optimal accuracy is obtained when 15 features are included with `MinLeafSize` = 2. These settings are also applied for the other decision schemes. This is illustrated in figure C.1.

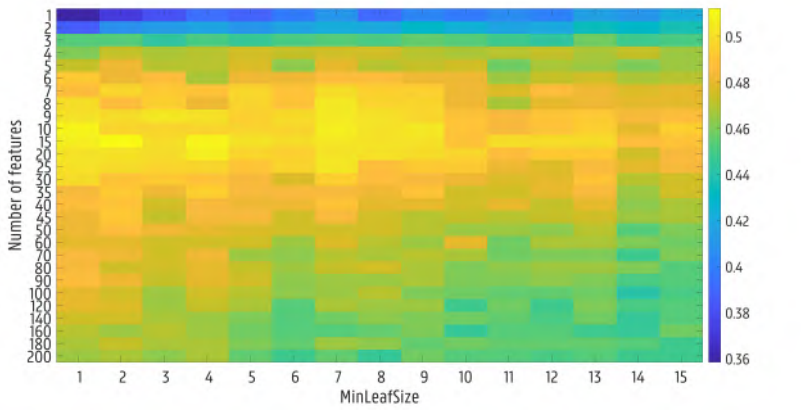



**Figure C.1:** Multiple binary classification: grid search.

# Optimal features for binary problems

In the following tables, the best 15 features for all binary problems are given. Also included are the $p$-values for a two-sample $t$-test with unequal variance and AUC values for the individual features.

**Table C.1:** Meningioma versus glioblastoma. Model performance: accuracy = 95.0%, AUC = 0.989.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | ratio core / surrounding total T1ce (5 voxels) | 3.7e-18 | 0.899 |
| 2 | T1ce: enhancing tumour - histogram: Median | 3.2e-13 | 0.840 |
| 3 | T1ce: tumour core - GLRLM: Short run high gray level emphasis (std) | 2.2e-17 | 0.879 |
| 4 | T1ce: tumour core - GLCM: Autocorrelation ($d$=3, mean) | 4.2e-13 | 0.888 |
| 5 | T1ce: oedema - GLCM: Correlation2 ($d$=1, mean) | 1.1e-13 | 0.889 |
| 6 | ratio core / surrounding core T1ce (5 voxels) | 1.2e-11 | 0.828 |
| 7 | T1ce: total abnormal - GLCM: Informational measure of correlation 2 ($d$=3, std) | 1.6e-11 | 0.828 |
| 8 | T1ce: tumour core - GLCM: Autocorrelation ($d$=1, mean) | 2.2e-12 | 0.835 |
| 9 | T1ce: enhancing tumour - GLSZM: Glvariance | 5.2e-06 | 0.851 |
| 10 | T1ce: enhancing tumour - histogram: Skewness | 7.6e-17 | 0.876 |
| 11 | ratio core / surrounding total T1ce (3 voxels) | 2.5e-15 | 0.811 |
| 12 | T1ce: total abnormal - GLCM: Informational measure of correlation 2 ($d$=2, std) | 3.8e-10 | 0.758 |
| 13 | T1ce: total abnormal - histogram: Mean absolute deviation | 1.9e-09 | 0.858 |
| 14 | T1ce: tumour core - GLRLM: Long run high gray level emphasis (mean) | 1.1e-14 | 0.718 |
| 15 | FLAIR: tumour core - GLCM: Correlation2 ($d$=3, mean) | 2.9e-08 | 0.730 |

**Table C.2:** Meningioma versus glioma. Model performance: accuracy = 92.6%, AUC = 0.976.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | ratio core / surrounding total T1ce (5 voxels) | 1.4e-36 | 0.949 |
| 2 | T1ce: total abnormal - GLCM: Informational measure of correlation 2 ($d$=3, std) | 1.2e-19 | 0.873 |
| 3 | T1ce: enhancing tumour - histogram: Skewness | 6.9e-22 | 0.877 |
| 4 | ratio core / surrounding total T1ce (3 voxels) | 8.1e-33 | 0.946 |
| 5 | T1ce: total abnormal - histogram: Mean | 8.6e-34 | 0.943 |
| 6 | FLAIR: enhancing tumour - NGTDM: Complexity ($d$=1) | 2.0e-30 | 0.881 |
| 7 | T1ce: enhancing tumour - NGTDM: Complexity ($d$=1) | 1.7e-29 | 0.881 |
| 8 | T1ce: total abnormal - GLCM: Informational measure of correlation 2 ($d$=2, std) | 3.6e-16 | 0.870 |
| 9 | T1ce: tumour core - histogram: Mean | 2.7e-19 | 0.881 |
| 10 | T1ce: enhancing tumour - GLRLM: Short run high gray level emphasis (std) | 3.1e-35 | 0.890 |
| 11 | ratio core / surrounding core T1ce (3 voxels) | 2.7e-19 | 0.885 |
| 12 | FLAIR: enhancing tumour - NGTDM: Complexity ($d$=3) | 1.5e-16 | 0.843 |
| 13 | FLAIR: enhancing tumour - Shape: Sphericity | 3.6e-20 | 0.870 |
| 14 | T1ce: total abnormal - GLCM: Dissimilarity ($d$=3, mean) | 5.1e-19 | 0.788 |
| 15 | T1ce: total abnormal - GLCM: Contrast ($d$=2, std) | 9.8e-24 | 0.823 |

**Table C.3:** Lower-grade glioma versus meningioma/glioblastoma. Model performance: accuracy = 90.7%, AUC = 0.952.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | T1ce: total abnormal - histogram: Mean | 2.3e-23 | 0.848 |
| 2 | ratio core / surrounding core T1ce (5 voxels) | 1.6e-21 | 0.684 |
| 3 | ratio core / surrounding core T1ce (3 voxels) | 5.0e-24 | 0.677 |
| 4 | T1ce: enhancing tumour - Shape: Surface to volume ratio | 2.1e-23 | 0.842 |
| 5 | ratio core / surrounding total T1ce (5 voxels) | 1.3e-29 | 0.875 |
| 6 | FLAIR: enhancing tumour - GLRLM: Long run emphasis (std) | 2.4e-16 | 0.769 |
| 7 | FLAIR: enhancing tumour - GLRLM: Long run emphasis (mean) | 8.7e-18 | 0.759 |
| 8 | FLAIR: enhancing tumour - Shape: Surface to volume ratio | 2.1e-23 | 0.695 |
| 9 | T1ce: total abnormal - GLCM: Inverse difference moment normalized ($d$=3, std) | 2.3e-15 | 0.772 |
| 10 | T1ce: total abnormal - GLCM: Dissimilarity ($d$=3, std) | 2.4e-16 | 0.678 |
| 11 | T1ce: total abnormal - GLCM: Difference entropy ($d$=3, mean) | 3.9e-12 | 0.824 |
| 12 | T1ce: total abnormal - GLCM: Inverse difference moment normalized ($d$=2, std) | 6.7e-15 | 0.782 |
| 13 | T1ce: enhancing tumour - GLRLM: Run length non-uniformity (mean) | 3.2e-13 | 0.659 |
| 14 | FLAIR: tumour core - Shape: Surface to volume ratio | 6.8e-15 | 0.708 |
| 15 | FLAIR: enhancing tumour - NGTDM: Complexity ($d$=2) | 6.2e-18 | 0.759 |

**Table C.4:** Grade III versus grade IV. Model performance: accuracy = 88.8%, AUC = 0.944.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | ratio core / surrounding core T1ce (5 voxels) | 8.7e-07 | 0.724 |
| 2 | T1ce: total abnormal - histogram: Mean | 1.2e-08 | 0.578 |
| 3 | ratio core / surrounding core T1ce (3 voxels) | 1.3e-07 | 0.587 |
| 4 | T1ce: total abnormal - histogram: Median | 5.4e-08 | 0.721 |
| 5 | FLAIR: enhancing tumour - GLSZM: Szpcent | 6.2e-08 | 0.775 |
| 6 | T1ce: total abnormal - GLRLM: High gray level run emphasis (std) | 3.7e-09 | 0.680 |
| 7 | T1ce: total abnormal - GLCM: Entropy ($d$=3, std) | 4.1e-05 | 0.659 |
| 8 | T1ce: total abnormal - GLCM: Difference entropy ($d$=2, std) | 6.5e-06 | 0.594 |
| 9 | ratio core / surrounding total T1ce (5 voxels) | 3.3e-08 | 0.699 |
| 10 | FLAIR: enhancing tumour - GLRLM: Long run emphasis (std) | 3.5e-07 | 0.584 |
| 11 | FLAIR: enhancing tumour - GLCM: Difference entropy ($d$=2, mean) | 0.00023 | 0.739 |
| 12 | FLAIR: enhancing tumour - GLCM: Homogeneity 1 ($d$=1, mean) | 8.3e-06 | 0.697 |
| 13 | FLAIR: enhancing tumour - GLCM: Correlation2 ($d$=1, std) | 6.5e-05 | 0.575 |
| 14 | FLAIR: enhancing tumour - Shape: Surface to volume ratio | 1.6e-08 | 0.635 |
| 15 | T1ce: total abnormal - GLCM: Informational measure of correlation 1 ($d$=2, std) | 0.018 | 0.691 |

**Table C.5:** Grade III versus meningioma/glioblastoma. Model performance: accuracy = 88.5%, AUC = 0.937.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | T1ce: total abnormal - histogram: Mean | 2.6e-15 | 0.801 |
| 2 | ratio core / surrounding total T1ce (5 voxels) | 4.5e-17 | 0.677 |
| 3 | FLAIR: enhancing tumour - Shape: Surface to volume ratio | 5.8e-14 | 0.682 |
| 4 | T1ce: enhancing tumour - NGTDM: Complexity ($d$=3) | 8.0e-10 | 0.799 |
| 5 | T1ce: total abnormal - GLRLM: High gray level run emphasis (std) | 1.7e-09 | 0.835 |
| 6 | T1ce: enhancing tumour - GLSZM: Szpcent | 1.0e-10 | 0.750 |
| 7 | ratio core / surrounding core T1ce (5 voxels) | 8.7e-14 | 0.736 |
| 8 | T1ce: total abnormal - GLCM: Contrast ($d$=3, std) | 4.6e-08 | 0.688 |
| 9 | T1ce: tumour core - histogram: Median | 8.2e-11 | 0.763 |
| 10 | T1ce: enhancing tumour - Shape: Surface to volume ratio | 5.8e-14 | 0.688 |
| 11 | ratio core / surrounding core T1ce (3 voxels) | 1.1e-14 | 0.790 |
| 12 | T1ce: total abnormal - histogram: Median | 2.4e-12 | 0.749 |
| 13 | T1ce: total abnormal - histogram: Kurtosis | 4.6e-09 | 0.675 |
| 14 | T1ce: enhancing tumour - GLRLM: Short run high gray level emphasis (std) | 1.5e-06 | 0.690 |
| 15 | T1ce: enhancing tumour - GLRLM: Run percentage (mean) | 2.4e-09 | 0.742 |

**Table C.6:** Grade II/III versus glioblastoma. Model performance: accuracy = 88.0%, AUC = 0.934.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | ratio core / surrounding core T1ce (3 voxels) | 3.4e-12 | 0.790 |
| 2 | FLAIR: enhancing tumour - GLRLM: Long run emphasis (mean) | 5.6e-16 | 0.588 |
| 3 | T1ce: total abnormal - histogram: Mean | 3.0e-10 | 0.581 |
| 4 | ratio core / surrounding core T1ce (5 voxels) | 2.9e-10 | 0.782 |
| 5 | T1ce: total abnormal - GLRLM: High gray level run emphasis (std) | 7.1e-16 | 0.833 |
| 6 | T1ce: total abnormal - histogram: Median | 5.3e-09 | 0.704 |
| 7 | T1ce: tumour core - Shape: Surface to volume ratio | 1.1e-10 | 0.689 |
| 8 | T1ce: enhancing tumour - Shape: Surface area | 4.5e-09 | 0.605 |
| 9 | FLAIR: tumour core - Shape: Surface to volume ratio | 1.1e-10 | 0.715 |
| 10 | FLAIR: enhancing tumour - GLRLM: Run percentage (std) | 1.5e-11 | 0.574 |
| 11 | FLAIR: enhancing tumour - GLRLM: Long run emphasis (std) | 6.4e-15 | 0.780 |
| 12 | FLAIR: enhancing tumour - GLCM: Autocorrelation ($d$=3, std) | 1.3e-06 | 0.736 |
| 13 | FLAIR: enhancing tumour - Shape: Compactness 1 | 9.8e-21 | 0.555 |
| 14 | T1ce: total abnormal - GLCM: Contrast ($d$=3, std) | 7.2e-08 | 0.656 |
| 15 | T1ce: tumour core - Shape: Compactness 1 | 5.9e-11 | 0.710 |

**Table C.7:** Oligodendroglioma grade II versus oligodendroglioma grade III. Model performance: accuracy = 87.8%, AUC = 0.937.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | T1ce: oedema - histogram: Median | 2.0e-05 | 0.706 |
| 2 | T1ce: oedema + non-enhancing tumour - histogram: Mean | 2.0e-05 | 0.502 |
| 3 | FLAIR: oedema - GLCM: Correlation2 ($d$=3, mean) | 0.0018 | 0.617 |
| 4 | T1ce: total abnormal - histogram: Median | 1.1e-05 | 0.677 |
| 5 | T1ce: oedema - histogram: Mean | 1.9e-05 | 0.834 |
| 6 | T1ce: total abnormal - histogram: Mean | 7.3e-06 | 0.504 |
| 7 | FLAIR: oedema - GLCM: Informational measure of correlation 1 ($d$=2, mean) | 0.0043 | 0.496 |
| 8 | FLAIR: enhancing tumour - GLCM: Homogeneity 2 ($d$=3, mean) | 0.0008 | 0.492 |
| 9 | FLAIR: enhancing tumour - GLCM: Homogeneity 2 ($d$=2, mean) | 0.00061 | 0.663 |
| 10 | FLAIR: enhancing tumour - GLCM: Difference entropy ($d$=2, mean) | 0.00046 | 0.644 |
| 11 | FLAIR: enhancing tumour - GLCM: Difference entropy ($d$=1, mean) | 0.00044 | 0.625 |
| 12 | FLAIR: oedema - GLCM: Correlation2 ($d$=2, mean) | 0.00079 | 0.575 |
| 13 | ratio core / surrounding total T1ce (5 voxels) | 0.002 | 0.595 |
| 14 | FLAIR: total abnormal - GLCM: Autocorrelation ($d$=3, mean) | 0.0013 | 0.572 |
| 15 | FLAIR: total abnormal - GLCM: Sum variance ($d$=1, mean) | 0.0022 | 0.548 |

**Table C.8:** Astrocytoma grade II versus astrocytoma grade III. Model performance: accuracy = 84.3%, AUC = 0.919.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | T1ce: tumour core - histogram: Uniformity | 0.0032 | 0.610 |
| 2 | T1ce: oedema + non-enhancing tumour - histogram: Kurtosis | 0.00029 | 0.533 |
| 3 | FLAIR: non-enhancing tumour - GLCM: Inverse difference normalized ($d$=2, mean) | 0.00077 | 0.545 |
| 4 | z-coordinate center of mass | 0.005 | 0.613 |
| 5 | FLAIR: enhancing tumour - GLCM: Inverse difference moment normalized ($d$=3, std) | 0.00093 | 0.611 |
| 6 | FLAIR: enhancing tumour - GLCM: Difference entropy ($d$=2, mean) | 3.2e-05 | 0.565 |
| 7 | FLAIR: non-enhancing tumour - GLCM: Inverse difference moment normalized ($d$=3, mean) | 0.00042 | 0.570 |
| 8 | T1ce: total abnormal - Shape: Spherical disproportion | 0.035 | 0.537 |
| 9 | T1ce: tumour core - GLCM: Energy ($d$=1, std) | 0.0057 | 0.559 |
| 10 | T1ce: tumour core - histogram: Entropy | 0.0069 | 0.534 |
| 11 | FLAIR: tumour core - GLRLM: Long run emphasis (mean) | 0.0091 | 0.610 |
| 12 | FLAIR: tumour core - GLCM: Inverse variance ($d$=3, mean) | 0.033 | 0.564 |
| 13 | FLAIR: tumour core - GLCM: Difference entropy ($d$=3, mean) | 0.0054 | 0.478 |
| 14 | FLAIR: tumour core - GLCM: Difference entropy ($d$=2, mean) | 0.0091 | 0.533 |
| 15 | FLAIR: enhancing tumour - GLCM: Inverse difference moment normalized ($d$=3, mean) | 8.1e-05 | 0.536 |

**Table C.9:** Grade II versus grade III/glioblastoma/meningioma. Model performance: accuracy = 83.5%, AUC = 0.901.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | ratio core / surrounding total T1ce (5 voxels) | 3.6e-13 | 0.778 |
| 2 | FLAIR: enhancing tumour - Shape: Volume | 1.1e-09 | 0.616 |
| 3 | T1ce: enhancing tumour - GLSZM: Sznonunif | 6.8e-11 | 0.594 |
| 4 | ratio core / surrounding core T1ce (5 voxels) | 9.2e-11 | 0.771 |
| 5 | ratio core / surrounding total T1ce (3 voxels) | 1.5e-12 | 0.805 |
| 6 | FLAIR: enhancing tumour - GLCM: Homogeneity 2 ($d$=1, mean) | 3.5e-13 | 0.681 |
| 7 | T1ce: enhancing tumour - GLCM: Informational measure of correlation 2 ($d$=3, mean) | 4.0e-13 | 0.680 |
| 8 | T1ce: enhancing tumour - GLCM: Informational measure of correlation 2 ($d$=2, mean) | 1.1e-10 | 0.625 |
| 9 | T1ce: enhancing tumour - Shape: Surface to volume ratio | 8.6e-11 | 0.692 |
| 10 | FLAIR: enhancing tumour - GLRLM: Run length non-uniformity (std) | 1.9e-09 | 0.589 |
| 11 | FLAIR: enhancing tumour - GLCM: Difference entropy ($d$=1, mean) | 1.4e-14 | 0.748 |
| 12 | FLAIR: enhancing tumour - Shape: Compactness 1 | 2.5e-13 | 0.707 |
| 13 | T1ce: total abnormal - histogram: Mean | 5.7e-10 | 0.561 |
| 14 | T1ce: enhancing tumour - Shape: Compactness 1 | 2.5e-13 | 0.644 |
| 15 | FLAIR: enhancing tumour - GLSZM: Smallzone | 3.0e-08 | 0.669 |

**Table C.10:** Grade II versus grade III. Model performance: accuracy = 83.1%, AUC = 0.907.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | FLAIR: enhancing tumour - GLCM: Difference entropy ($d$=1, mean) | 3.7e-06 | 0.635 |
| 2 | FLAIR: enhancing tumour - histogram: Entropy | 4.7e-05 | 0.524 |
| 3 | T1ce: oedema + non-enhancing tumour - histogram: Kurtosis | 0.0027 | 0.504 |
| 4 | FLAIR: enhancing tumour - GLCM: Inverse difference normalized ($d$=3, mean) | 1.9e-07 | 0.629 |
| 5 | FLAIR: enhancing tumour - GLCM: Difference entropy ($d$=2, mean) | 1.0e-07 | 0.668 |
| 6 | FLAIR: enhancing tumour - Shape: Compactness 1 | 0.0002 | 0.545 |
| 7 | FLAIR: total abnormal - GLCM: Inverse difference normalized ($d$=1, std) | 0.0063 | 0.552 |
| 8 | T1ce: tumour core - GLCM: Sum entropy ($d$=3, mean) | 0.015 | 0.530 |
| 9 | T1ce: oedema + non-enhancing tumour - GLRLM: Gray level non-uniformity (mean) | 0.00047 | 0.583 |
| 10 | T1ce: enhancing tumour - GLCM: Informational measure of correlation 2 ($d$=3, mean) | 1.6e-06 | 0.513 |
| 11 | T1ce: enhancing tumour - Shape: Compactness 1 | 0.0002 | 0.613 |
| 12 | T1ce: non-enhancing tumour - GLCM: Dissimilarity ($d$=2, std) | 0.00051 | 0.568 |
| 13 | FLAIR: total abnormal - Shape: Surface area | 0.0012 | 0.544 |
| 14 | FLAIR: tumour core - GLCM: Difference entropy ($d$=2, mean) | 0.0027 | 0.543 |
| 15 | FLAIR: tumour core - GLCM: Homogeneity 1 ($d$=1, mean) | 0.0089 | 0.537 |

**Table C.11:** Low-grade glioma versus high-grade glioma. Model performance: accuracy = 83.0%, AUC = 0.906.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | ratio core / surrounding core T1ce (3 voxels) | 8.9e-08 | 0.736 |
| 2 | T1ce: total abnormal - histogram: Mean | 5.9e-06 | 0.558 |
| 3 | T1ce: enhancing tumour - Shape: Volume | 1.1e-09 | 0.532 |
| 4 | ratio core / surrounding total T1ce (3 voxels) | 1.2e-07 | 0.726 |
| 5 | FLAIR: enhancing tumour - GLRLM: Run length non-uniformity (std) | 1.7e-09 | 0.769 |
| 6 | FLAIR: enhancing tumour - histogram: Uniformity | 1.9e-06 | 0.629 |
| 7 | T1ce: enhancing tumour - GLCM: Entropy ($d$=2, mean) | 1.9e-06 | 0.629 |
| 8 | FLAIR: enhancing tumour - GLSZM: Szpcent | 9.3e-12 | 0.569 |
| 9 | FLAIR: enhancing tumour - GLRLM: Long run emphasis (mean) | 5.6e-12 | 0.650 |
| 10 | FLAIR: enhancing tumour - GLCM: Inverse difference normalized ($d$=3, mean) | 1.4e-15 | 0.478 |
| 11 | FLAIR: enhancing tumour - GLCM: Homogeneity 1 ($d$=3, mean) | 1.4e-13 | 0.708 |
| 12 | FLAIR: enhancing tumour - GLCM: Informational measure of correlation 2 ($d$=2, std) | 5.2e-08 | 0.663 |
| 13 | FLAIR: enhancing tumour - GLCM: Difference entropy ($d$=2, mean) | 1.4e-15 | 0.508 |
| 14 | T1ce: total abnormal - GLRLM: High gray level run emphasis (std) | 2.2e-06 | 0.604 |
| 15 | T1ce: tumour core - GLCM: Sum entropy ($d$=3, mean) | 2.1e-06 | 0.625 |

**Table C.12:** Astrocytoma grade II versus oligodendroglioma grade II. Model performance: accuracy = 80.5%, AUC = 0.906.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | FLAIR: total abnormal - GLCM: Difference entropy ($d$=3, std) | 0.0052 | 0.592 |
| 2 | T1ce: tumour core - GLRLM: Gray level non-uniformity (std) | 0.016 | 0.582 |
| 3 | z-coordinate center of mass | 0.0011 | 0.646 |
| 4 | FLAIR: oedema - GLCM: Autocorrelation ($d$=3, mean) | 0.065 | 0.620 |
| 5 | FLAIR: oedema - GLCM: Correlation2 ($d$=2, mean) | 0.0013 | 0.543 |
| 6 | T1ce: tumour core - GLSZM: Glvariance | 0.023 | 0.526 |
| 7 | T1ce: tumour core - GLSZM: Largezone | 0.0078 | 0.563 |
| 8 | T1ce: tumour core - GLCM: Energy ($d$=2, std) | 0.0089 | 0.590 |
| 9 | T1ce: enhancing tumour - GLCM: Sum average ($d$=1, mean) | 0.021 | 0.514 |
| 10 | FLAIR: total abnormal - GLCM: Sum variance ($d$=2, std) | 0.0066 | 0.586 |
| 11 | FLAIR: total abnormal - GLCM: Correlation ($d$=2, mean) | 0.73 | 0.477 |
| 12 | FLAIR: total abnormal - GLCM: Sum variance ($d$=1, std) | 0.0079 | 0.464 |
| 13 | FLAIR: oedema + non-enhancing tumour - GLCM: Difference entropy ($d$=3, std) | 0.011 | 0.531 |
| 14 | FLAIR: oedema + non-enhancing tumour - GLCM: Sum variance ($d$=2, mean) | 0.27 | 0.506 |
| 15 | FLAIR: oedema + non-enhancing tumour - GLCM: Inverse variance ($d$=2, std) | 0.099 | 0.489 |

**Table C.13:** Astrocytoma versus oligodendroglioma. Model performance: accuracy = 80.3%, AUC = 0.892.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | y-coordinate center of mass | 0.0068 | 0.599 |
| 2 | FLAIR: non-enhancing tumour - GLCM: Homogeneity 2 ($d$=3, std) | 0.072 | 0.609 |
| 3 | FLAIR: total abnormal - GLSZM: Zonehigl | 0.062 | 0.573 |
| 4 | T1ce: tumour core - GLSZM: Largezone | 0.0016 | 0.612 |
| 5 | T1ce: oedema + non-enhancing tumour - GLCM: Homogeneity 1 ($d$=1, std) | 0.028 | 0.548 |
| 6 | FLAIR: total abnormal - GLCM: Difference entropy ($d$=3, std) | 0.022 | 0.493 |
| 7 | FLAIR: total abnormal - histogram: Entropy | 0.11 | 0.527 |
| 8 | FLAIR: oedema - GLCM: Correlation2 ($d$=3, mean) | 0.17 | 0.607 |
| 9 | T1ce: total abnormal - GLCM: Informational measure of correlation 2 ($d$=2, std) | 0.0045 | 0.483 |
| 10 | T1ce: tumour core - GLCM: Informational measure of correlation 1 ($d$=3, std) | 0.022 | 0.504 |
| 11 | T1ce: tumour core - GLCM: Informational measure of correlation 1 ($d$=2, mean) | 0.041 | 0.492 |
| 12 | T1ce: tumour core - GLCM: Entropy ($d$=1, mean) | 0.1 | 0.518 |
| 13 | T1ce: oedema + non-enhancing tumour - GLCM: Homogeneity 1 ($d$=2, std) | 0.036 | 0.498 |
| 14 | T1ce: enhancing tumour - GLCM: Homogeneity 1 ($d$=3, std) | 0.011 | 0.516 |
| 15 | T1ce: oedema - GLCM: Homogeneity 1 ($d$=2, std) | 0.077 | 0.521 |

**Table C.14:** Astrocytoma grade III versus oligodendroglioma grade III. Model performance: accuracy = 75.1%, AUC = 0.829.

| | Feature name | $p$-value | AUC |
|---|---|---|---|
| 1 | T1ce: oedema - histogram: Mean | 0.0076 | 0.614 |
| 2 | FLAIR: tumour core - GLCM: Entropy ($d$=3, std) | 0.15 | 0.629 |
| 3 | FLAIR: necrosis - GLCM: Correlation ($d$=3, mean) | 0.086 | 0.487 |
| 4 | FLAIR: total abnormal - GLCM: Dissimilarity ($d$=1, mean) | 0.12 | 0.612 |
| 5 | FLAIR: tumour core - GLCM: Entropy ($d$=2, std) | 0.2 | 0.651 |
| 6 | FLAIR: oedema + non-enhancing tumour - GLSZM: Zonehigl | 0.046 | 0.538 |
| 7 | FLAIR: oedema + non-enhancing tumour - GLRLM: High gray level run emphasis (mean) | 0.021 | 0.511 |
| 8 | FLAIR: oedema + non-enhancing tumour - GLCM: Cluster Prominence ($d$=3, mean) | 0.034 | 0.617 |
| 9 | FLAIR: oedema + non-enhancing tumour - GLCM: Cluster Tendency ($d$=2, mean) | 0.028 | 0.554 |
| 10 | FLAIR: enhancing tumour - GLCM: Variance ($d$=1, mean) | 0.097 | 0.587 |
| 11 | FLAIR: non-enhancing tumour - GLCM: Correlation2 ($d$=3, mean) | 0.0029 | 0.502 |
| 12 | FLAIR: oedema - GLCM: Cluster Prominence ($d$=3, mean) | 0.014 | 0.507 |
| 13 | FLAIR: oedema - GLCM: Cluster Shade ($d$=2, mean) | 0.015 | 0.538 |
| 14 | FLAIR: oedema - GLCM: Informational measure of correlation 1 ($d$=1, mean) | 0.14 | 0.538 |
| 15 | FLAIR: oedema - GLCM: Cluster Prominence ($d$=1, mean) | 0.01 | 0.528 |

# Bibliography


[1] XinhuaNet.com. China Focus: AI beats human doctors in neuroimaging recognition contest. www.xinhuanet.com/english/2018-06/30/c_137292451.htm, 2018. Accessed: 2018-08-09.

[2] Paul M. Parizel. I've seen the future: a competition between physicians and AI. www.myesr.org/article/1840, 2018. Accessed: 2018-08-09.

[3] Memorial Sloan Kettering Cancer Center. Watson oncology. www.mskcc.org/about/innovative-collaborations/watson-oncology, 2017. Accessed: 2018-10-18.

[4] B.J. Copeland. Artificial intelligence. www.britannica.com/technology/artificial-intelligence, 2018. Accessed: 2018-04-06.

[5] Macedo Firmino, Giovani Angelo, Higor Morais, Marcel R Dantas, and Ricardo Valentim. Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. *Biomedical engineering online*, 15(1):2, 2016.

[6] Martin Lipkin and James D Hardy. Mechanical correlation of data in differential diagnosis of hematological diseases. *Journal of the American Medical Association*, 166(2):113–125, 1958.

[7] Wallace Feurzeig, Preston Munter, John Swets, and Myra Breen. Computer-aided teaching in medical diagnosis. *Academic Medicine*, 39(8):746–754, 1964.

[8] Gwilym S Lodwick. Computer-aided diagnosis in radiology: A research plan. *Investigative Radiology*, 1(1):72–80, 1966.

[9] G Anthony Gorry. Strategies for computer-aided diagnosis. *Mathematical Biosciences*, 2(3-4):293–318, 1968.

[10] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007.

[11] Heang-Ping Chan, Simranjit Galhotra, Carl J Vyborny, Heber MacMahon, Peter M Jokich, et al. Image feature analysis and computer-aided diagnosis in digital radiography. i. automated detection of microcalcifications in mammography. *Medical physics*, 14(4):538–548, 1987.

[12] DH Davies and DR Dance. Automatic computer detection of clustered calcifications in digital mammograms. *Physics in Medicine & Biology*, 35(8):1111, 1990.

[13] Heang-Ping Chan, Kunio Doi, Carl J Vyborny, Robert A Schmidt, Charles E Metz, Kwok Leung Lam, Toshihiro Ogura, YZ Wu, and Heber MacMahon. Improvement in radiologists' detection of clustered microcalcifications on mammograms. the potential of computer-aided diagnosis. *Investigative radiology*, 25(10):1102–1110, 1990.

[14] Maryellen Lissak Giger, Heber MacMahon, et al. Image feature analysis and computer-aided diagnosis in digital radiography. 3. automated detection of nodules in peripheral lung fields. *Medical Physics*, 15(2):158–166, 1988.

[15] Shigehiko Katsuragawa, Heber MacMahon, et al. Image feature analysis and computer-aided diagnosis in digital radiography: Detection and characterization of interstitial lung disease in digital chest radiographs. *Medical Physics*, 15(3):311–319, 1988.

[16] Maryellen L Giger, Kunio Doi, Heber MacMahon, Charles E Metz, and Fang-Fang Yin. Pulmonary nodules: computer-aided detection in digital chest images. *Radiographics*, 10(1):41–51, 1990.

[17] Rebecca Smith-Bindman, Diana L Miglioretti, and Eric B Larson. Rising use of diagnostic medical imaging in a large integrated health system. *Health affairs*, 27(6):1491–1502, 2008.

[18] Fred A Mettler Jr, Mythreyi Bhargavan, Keith Faulkner, Debbie B Gilley, Joel E Gray, Geoffrey S Ibbott, Jill A Lipoti, Mahadevappa Mahesh, John L McCrohan, Michael G Stabin, et al. Radiologic and nuclear medicine studies in the united states and worldwide: frequency, radiation dose, and comparison with other radiation sources - 1950–2007. *Radiology*, 253(2):520–531, 2009.

[19] Mythreyi Bhargavan, Adam H Kaye, Howard P Forman, and Jonathan H Sunshine. Workload of radiologists in united states in 2006–2007 and trends since 1991–1992. *Radiology*, 252(2):458–467, 2009.

[20] Robert M Nishikawa and Kyongtae T Bae. Importance of better human-computer interaction in the era of deep learning: Mammography computer-aided diagnosis as a use case. *Journal of the American College of Radiology*, 15(1):49–52, 2018.

[21] Vijay M Rao, David C Levin, Laurence Parker, Barbara Cavanaugh, Andrea J Frangos, and Jonathan H Sunshine. How widely is computer-aided detection used in screening and diagnostic mammography? *Journal of the American College of Radiology*, 7(10):802–805, 2010.

[22] Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, and Diana L Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11):1828–1837, 2015.

[23] Ting-Wei Lin, Po-Yu Huang, and Claire Wan-Chiung Cheng. Computer-aided diagnosis in medical imaging: Review of legal

barriers to entry for the commercial systems. In *e-Health Networking, Applications and Services (Healthcom), 2016 IEEE 18th International Conference on*, pages 1–5. IEEE, 2016.

[24] Afsaneh Jalalian, Syamsiah BT Mashohor, Hajjah Rozi Mahmud, M Iqbal B Saripan, Abdul Rahman B Ramli, and Babak Karasfi. Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clinical imaging*, 37(3):420–426, 2013.

[25] Karthikeyan Ganesan, U Rajendra Acharya, Chua Kuang Chua, Lim Choo Min, K Thomas Abraham, and Kwan-Hoong Ng. Computer-aided breast cancer detection using mammograms: a review. *IEEE Reviews in biomedical engineering*, 6:77–98, 2013.

[26] Leila H Eadie, Paul Taylor, and Adam P Gibson. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *European journal of radiology*, 81(1):e70–e76, 2012.

[27] Thijs Kooi, Geert Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*, 35:303–312, 2017.

[28] Azam Hamidinekoo, Erika Denton, Andrik Rampun, Kate Honnor, and Reyer Zwiggelaar. Deep learning in mammography and breast histology, an overview and future trends. *Medical Image Analysis*, 2018.

[29] Bram Van Ginneken, BM Ter Haar Romeny, and Max A Viergever. Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on medical imaging*, 20(12):1228–1241, 2001.

[30] Feng Li, Roger Engelmann, Samuel G Armato, and Heber MacMahon. Computer-aided nodule detection system: results in an unselected series of consecutive chest radiographs. *Academic radiology*, 22(4):475–480, 2015.

[31] Nikolaos Dellios, Ulf Teichgraeber, Robert Chelaru, Ansgar Malich, and Ismini E Papageorgiou. Computer-aided detection fidelity of pulmonary nodules in chest radiograph. *Journal of clinical imaging science*, 7, 2017.

[32] World Health Organization. Tuberculosis. `www.who.int/mediacentre/factsheets/fs104/en/`, 2018. Accessed: 2018-04-11.

[33] Marianne Breuninger, Bram van Ginneken, Rick HHM Philipsen, Francis Mhimbira, Jerry J Hella, Fred Lwilla, Jan van den Hombergh, Amanda Ross, Levan Jugheli, Dirk Wagner, et al. Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from sub-saharan africa. *PLoS One*, 9(9):e106381, 2014.

[34] Jaime Melendez, Bram van Ginneken, Pragnya Maduskar, Rick HHM Philipsen, Klaus Reither, Marianne Breuninger, Ifedayo MO Adetifa, Rahmatulai Maane, Helen Ayles, and Clara I Sánchez. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays. *IEEE transactions on medical imaging*, 34(1):179–192, 2015.

[35] Zishan Sheikh and Caron Parsons. Bone age assessment. `radiopaedia.org/articles/bone-age-assessment`, 2018. Accessed: 2018-04-11.

[36] William Walter Greulich and S Idell Pyle. Radiographic atlas of skeletal development of the hand and wrist. *The American Journal of the Medical Sciences*, 238(3):393, 1959.

[37] Arsalan Manzoor Mughal, Nuzhat Hassan, and Anwar Ahmed. Bone age assessment methods: A critical review. *Pakistan journal of medical sciences*, 30(1):211, 2014.

[38] Mari Satoh. Bone age: assessment methods and clinical applications. *Clinical Pediatric Endocrinology*, 24(4):143–152, 2015.

[39] James Mourilyan Tanner. *Assessment of skeletal maturity and prediction of adult height (TW2 method)*. Academic Press, 1983.

[40] RK Bull, PD Edwards, PM Kemp, S Fry, and IA Hughes. Bone age assessment: a large scale comparison of the greulich and pyle, and tanner and whitehouse (tw2) methods. *Archives of disease in childhood*, 81(2):172–173, 1999.

[41] Hans Henrik Thodberg, Sven Kreiborg, Anders Juul, and Karen Damgaard Pedersen. The BoneXpert method for automated determination of skeletal maturity. *IEEE transactions on medical imaging*, 28(1):52–66, 2009.

[42] Hans Henrik Thodberg, Julia Neuhof, Michael B Ranke, Oskar G Jenni, and David D Martin. Validation of bone age methods by their ability to predict adult height. *Hormone research in paediatrics*, 74(1):15–22, 2010.

[43] David D Martin, Jan M Wit, Ze'ev Hochberg, Lars Sävendahl, Rick R Van Rijn, Oliver Fricke, Noël Cameron, Janina Caliebe, Thomas Hertel, Daniela Kiepe, et al. The use of bone age in clinical practice–part 1. *Hormone research in paediatrics*, 76(1):1–9, 2011.

[44] Icometrix. Transforming patient care through imaging ai. `www.icometrix.com`, 2018. Accessed: 2018-04-17.

[45] Saurabh Jain, Diana M Sima, Annemie Ribbens, Melissa Cambron, Anke Maertens, Wim Van Hecke, Johan De Mey, Frederik Barkhof, Martijn D Steenwijk, Marita Daams, et al. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *NeuroImage: Clinical*, 8:367–375, 2015.

[46] Dirk Smeets, Annemie Ribbens, Diana M Sima, Melissa Cambron, Dana Horakova, Saurabh Jain, Anke Maertens, Eline Van Vlierberghe, Vasilis Terzopoulos, Anne-Marie Van Binst, et al. Reliable measurements of brain atrophy in individual patients with multiple sclerosis. *Brain and behavior*, 6(9), 2016.

[47] Annemie Ribbens, Dirk Smeets, Vasilis Terzopoulos, Saurabh Jain, Diana M Sima, Hanne Struyfs, Sebastiaan Engelborghs, and

Wim Van Hecke. Admetrix: A new method for atrophy quantification in alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 11(7):P876, 2015.

[48] Andrew IR Maas, David K Menon, P David Adelson, Nada Andelic, Michael J Bell, Antonio Belli, Peter Bragge, Alexandra Brazinova, András Büki, Randall M Chesnut, et al. Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. *The Lancet Neurology*, 2017.

[49] Nicolas Henri Jacob. Bourgery neuroanatomie. `commons.wikimedia.org/wiki/File:Bourgery_Neuroanatomie.jpg`, 1844. Accessed: 2018-04-19.

[50] Henry Gray. Anatomy of the human body, diagrammatic section of the scalp. `commons.wikimedia.org/wiki/File:Gray1196.png`, before 1858. Accessed: 2018-04-19.

[51] John A Beal. Human brain frontal (coronal) section. `commons.wikimedia.org/wiki/File:Human_brain_frontal_(coronal)_section.JPG`, 2005. Accessed: 2018-04-19.

[52] Holly Fischer. Glial cell types. `commons.wikimedia.org/wiki/File:Glial_Cell_Types.png`, 2013. Accessed: 2018-04-18.

[53] D Purves, GJ Augustine, D Fitzpatrick, et al. Neuroglial cells. `www.ncbi.nlm.nih.gov/books/NBK10869/`, 2018. Accessed: 2018-04-18.

[54] David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.

[55] Sebastian Brandner. Macroscopic pathology of glioblastoma multiforme. commons.wikimedia.org/wiki/File:Glioblastoma_macro.jpg, 2007. Accessed: 2018-04-24.

[56] The Armed Forces Institute of Pathology. CNS: Meningioma. commons.wikimedia.org/wiki/File:Meningioma.jpg, 2018. Accessed: 2018-04-24.

[57] Amir Reazee et al. WHO grading of CNS tumours. radiopaedia.org/articles/who-grading-of-cns-tumours, 2018. Accessed: 2018-04-24.

[58] David N Louis, Hiroko Ohgaki, Otmar D Wiestler, Webster K Cavenee, Peter C Burger, Anne Jouvet, Bernd W Scheithauer, and Paul Kleihues. The 2007 WHO classification of tumours of the central nervous system. *Acta neuropathologica*, 114(2):97–109, 2007.

[59] Mark A Mittler, Beverly C Walters, and Edward G Stopa. Observer reliability in histological grading of astrocytoma stereotactic biopsies. *Journal of neurosurgery*, 85(6):1091–1094, 1996.

[60] Janet M Bruner, Lila Inouye, Gregory N Fuller, and Lauren A Langford. Diagnostic discrepancies and their clinical impact in a neuropathology referral practice. *Cancer*, 79(4):796–803, 1997.

[61] Martin J van den Bent. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta neuropathologica*, 120(3):297–304, 2010.

[62] Catherine L Nutt, DR Mani, Rebecca A Betensky, Pablo Tamayo, J Gregory Cairncross, Christine Ladd, Ute Pohl, Christian Hartmann, Margaret E McLaughlin, Tracy T Batchelor, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer research*, 63(7):1602–1607, 2003.

[63] Lonneke AM Gravendeel, Mathilde CM Kouwenhoven, Olivier Gevaert, Johan J de Rooi, Andrew P Stubbs, J Elza Duijm, Anneleen Daemen, Fonnet E Bleeker, Linda BC Bralten, Nanne K Kloosterhof, et al. Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer research*, 69(23):9065–9072, 2009.

[64] Zachary J Reitman and Hai Yan. Isocitrate dehydrogenase 1 and 2 mutations in cancer: alterations at a crossroads of cellular metabolism. *Journal of the National Cancer Institute*, 102(13):932–941, 2010.

[65] Lenny Dang, Shengfang Jin, and Shinsan M Su. IDH mutations in glioma and acute myeloid leukemia. *Trends in molecular medicine*, 16(9):387–397, 2010.

[66] M Preusser, D Capper, and C Hartmann. IDH testing in diagnostic neuropathology: review and practical guideline article invited by the Euro-CNS research committee. *Clinical neuropathology*, 30(5):217–230, 2011.

[67] Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015.

[68] David E Reuss, Annekathrin Kratz, Felix Sahm, David Capper, Daniel Schrimpf, Christian Koelsche, Volker Hovestadt, Melanie Bewerunge-Hudler, David TW Jones, Jens Schittenhelm, et al. Adult IDH wild type astrocytomas biologically and clinically resolve into other tumor entities. *Acta neuropathologica*, 130(3):407–417, 2015.

[69] Adriana Olar, Khalida M Wani, Kristin D Alfaro-Munoz, Lindsey E Heathcock, Hinke F van Thuijl, Mark R Gilbert, Terri S Armstrong, Erik P Sulman, Daniel P Cahill, Elizabeth Vera-Bolanos, et al. Idh mutation status and role of who grade and mitotic index in overall survival in grade ii–iii diffuse gliomas. *Acta neuropathologica*, 129(4):585–596, 2015.

[70] David E Reuss, Yasin Mamatjan, Daniel Schrimpf, David Capper, Volker Hovestadt, Annekathrin Kratz, Felix Sahm, Christian Koelsche, Andrey Korshunov, Adriana Olar, et al. IDH mutant diffuse and anaplastic astrocytomas have similar age at presentation and little difference in survival: a grading problem for WHO. *Acta neuropathologica*, 129(6):867–873, 2015.

[71] Martin J Van Den Bent, Christian Hartmann, Matthias Preusser, Thomas Ströbel, Hendrikus J Dubbink, Johan M Kros, Andreas Von Deimling, Blandine Boisselier, Marc Sanson, Kevin C Halling, et al. Interlaboratory comparison of IDH mutation detection. *Journal of neuro-oncology*, 112(2):173–178, 2013.

[72] C Houillier, X Wang, G Kaloshi, K Mokhtari, R Guillevin, J Laffaire, S Paris, B Boisselier, A Idbaih, F Laigle-Donadey, et al. IDH1 or IDH2 mutations predict longer survival and response to temozolomide in low-grade gliomas. *Neurology*, 75(17):1560–1566, 2010.

[73] Michael Weller, Roger Stupp, Monika E Hegi, Martin van den Bent, Joerg C Tonn, Marc Sanson, Wolfgang Wick, and Guido Reifenberger. Personalized care in neuro-oncology coming of age: why we need MGMT and 1p/19q testing for malignant glioma patients in clinical practice. *Neuro-oncology*, 14(suppl_4):iv100–iv108, 2012.

[74] Jeanette E Eckel-Passow, Daniel H Lachance, Annette M Molinaro, Kyle M Walsh, Paul A Decker, Hugues Sicotte, Melike Pekmezci, Terri Rice, Matt L Kosel, Ivan V Smirnov, et al. Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. *New England Journal of Medicine*, 372(26):2499–2508, 2015.

[75] Paula de Robles, Kirsten M Fiest, Alexandra D Frolkis, Tamara Pringsheim, Callie Atta, Christine St. Germaine-Smith, Lundy Day, Darren Lam, and Nathalie Jette. The worldwide incidence and prevalence of primary brain tumors: a systematic review and meta-analysis. *Neuro-oncology*, 17(6):776–783, 2014.

[76] Katharine A McNeill. Epidemiology of brain tumors. *Neurologic clinics*, 34(4):981–998, 2016.

[77] Quinn T Ostrom, Haley Gittleman, Jordan Xu, Courtney Kromer, Yingli Wolinsky, Carol Kruchko, and Jill S Barnholtz-Sloan. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2009–2013. *Neuro-oncology*, 18(suppl_5):v1–v75, 2016.

[78] Lisa M DeAngelis. Brain tumors. *New England Journal of Medicine*, 344(2):114–123, 2001.

[79] INTERPHONE Study Group. Brain tumour risk in relation to mobile telephone use: results of the interphone international case–control study. *International journal of epidemiology*, 39(3):675–694, 2010.

[80] Anthony Behin, Khe Hoang-Xuan, Antoine F Carpentier, and Jean-Yves Delattre. Primary brain tumours in adults. *The Lancet*, 361(9354):323–331, 2003.

[81] Maarten MJ Wijnenga, Tariq Mattni, Pim J French, Geert-Jan Rutten, Sieger Leenstra, Fred Kloet, Martin JB Taphoorn, Martin J van den Bent, Clemens MF Dirven, Marie-Lise van Veelen, et al. Does early resection of presumed low-grade glioma improve survival? A clinical perspective. *Journal of neuro-oncology*, 133(1):137–146, 2017.

[82] Roland Goldbrunner, Giuseppe Minniti, Matthias Preusser, Michael D Jenkinson, Kita Sallabanda, Emmanuel Houdart, Andreas von Deimling, Pantelis Stavrinou, Florence Lefranc, Morten Lund-Johansen, et al. EANO guidelines for the diagnosis and treatment of meningiomas. *The Lancet Oncology*, 17(9):e383–e391, 2016.

[83] Michael Weller, Martin Van Den Bent, Jörg C Tonn, Roger Stupp, Matthias Preusser, Elizabeth Cohen-Jonathan-Moyal, Roger Henriksson, Emilie Le Rhun, Carmen Balana, Olivier Chinot, et al. European association for neuro-oncology (eano) guideline on the

diagnosis and treatment of adult astrocytic and oligodendroglial gliomas. *The lancet oncology*, 18(6):e315–e329, 2017.

[84] Michel Lacroix and Steven A Toms. Maximum safe resection of glioblastoma multiforme. *Journal of Clinical Oncology*, 32(8):727–728, 2014.

[85] SL Hervey-Jumper and MS Berger. Technical nuances of awake brain tumor surgery and the role of maximum safe resection. *Journal of neurosurgical sciences*, 59(4):351–360, 2015.

[86] Brian T Ragel, Timothy C Ryken, Steven N Kalkanis, Mateo Ziu, Daniel Cahill, and Jeffrey J Olson. The role of biopsy in the management of patients with presumed diffuse low grade glioma. *Journal of neuro-oncology*, 125(3):481–501, 2015.

[87] Parakrama T Chandrasoma, Maurice M Smith, and Michael LJ Apuzzo. Stereotactic biopsy in the diagnosis of brain masses: comparison of results of biopsy and resected surgical specimen. *Neurosurgery*, 24(2):160–165, 1989.

[88] Wolfgang Feiden, Ulrich Steude, Karl Bise, and Ortrun Gündisch. Accuracy of stereotactic brain tumor biopsy: comparison of the histologic findings in biopsy cylinders and resected tumor tissue. *Neurosurgical review*, 14(1):51–56, 1991.

[89] Graeme Woodworth, Matthew J McGirt, Amer Samdani, Ira Garonzik, Alessandro Olivi, and Jon D Weingart. Accuracy of frameless and frame-based image-guided stereotactic brain biopsy in the diagnosis of glioma: comparison of biopsy and open resection specimen. *Neurological research*, 27(4):358–362, 2005.

[90] University of Kansas Medical Center Department of Neurosurgery. Stereotactic radiosurgergy program. `www.kumc.edu/school-of-medicine/neurosurgery/clinical-specialties/stereotactic-radiosurgery-program.html`, 2018. Accessed: 2018-05-03.

[91] Eric Barbarite, Justin T Sick, Emmanuel Berchmans, Amade Bregy, Ashish H Shah, Nagy Elsayyad, and Ricardo J Komotar. The role of brachytherapy in the treatment of glioblastoma multiforme. *Neurosurgical review*, 40(2):195–211, 2017.

[92] Greg Freiherr. The eclectic history of medical imaging. `www.itnonline.com/article/eclectic-history-medical-imaging`, 2014. Accessed: 2018-05-07.

[93] Wilhelm Röntgen. Hand mit Ringen: a print of one of the first X-rays by Wilhelm Röntgen. `commons.wikimedia.org/wiki/File:First_medical_X-ray_by_Wilhelm_Rontgen_of_his_wife_Anna_Bertha_Ludwig%27s_hand_-_18951222.gif`, 1895. Accessed: 2018-05-07.

[94] Johann Radon. 1.1 Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Classic papers in modern diagnostic radiology*, 5:21, 2005.

[95] impactscan.org. A brief history of CT. `www.impactscan.org/CThistory.htm`, 2013. Accessed: 2018-05-08.

[96] Dieter Schellinger, Giovanni Di Chiro, Stewart P Axelbaum, Homer L Twigg, and Robert S Ledley. Early clinical experience with the ACTA scanner. *Radiology*, 114(2):257–261, 1975.

[97] Society of Nuclear Medicine and Molecular Imaging. Historical timeline. `www.snmmi.org/AboutSNMMI/Content.aspx?ItemNumber=4175`, 2018. Accessed: 2018-05-08.

[98] SM Seidlin, LD Marinelli, and Eleanor Oshry. Radioactive iodine therapy: effect on functioning metastases of adenocarcinoma of the thyroid. *Journal of the American Medical Association*, 132(14):838–847, 1946.

[99] David E Kuhl and Roy Q Edwards. Image separation radioisotope scanning. *Radiology*, 80(4):653–662, 1963.

[100] Michael E Phelps, Edward J Hoffman, Nizar A Mullani, and Michel M Ter-Pogossian. Application of annihilation coincidence detection to transaxial reconstruction tomography. *Journal of Nuclear Medicine*, 16(3):210–224, 1975.

[101] Michel M Ter-Pogossian, Michael E Phelps, Edward J Hoffman, and Nizar A Mullani. A positron-emission transaxial tomograph for nuclear imaging (PETT). *Radiology*, 114(1):89–98, 1975.

[102] Martin Reivich, D Kuhl, A Wolf, J Greenberg, M Phelps, T Ido, V Casella, J Fowler, E Hoffman, A Alavi, et al. The [18F] fluorodeoxyglucose method for the measurement of local cerebral glucose utilization in man. *Circulation research*, 44(1):127–137, 1979.

[103] Thomas Beyer, David W Townsend, Tony Brun, Paul E Kinahan, et al. A combined PET/CT scanner for clinical oncology. *The Journal of nuclear medicine*, 41(8):1369, 2000.

[104] I I Rabi, J R Zacharias, S Millman, and P Kusch. A new method of measuring nuclear magnetic moment. *Phys. Rev.*, 53:318, 1938.

[105] John R Mallard. Magnetic resonance imaging —- the aberdeen perspective on developments in the early years. *Physics in Medicine & Biology*, 51(13):R45, 2006.

[106] David S Hersh, Anthony J Kim, Jeffrey A Winkles, Howard M Eisenberg, Graeme F Woodworth, and Victor Frenkel. Emerging applications of therapeutic ultrasound in neuro-oncology: moving beyond tumor ablation. *Neurosurgery*, 79(5):643–654, 2016.

[107] Aliasgar V Moiyadi. Intraoperative ultrasound technology in neuro-oncology practice – current role and future applications. *World neurosurgery*, 93:81–93, 2016.

[108] Christine Runyon. Understanding MRI gradients & slew rates. `www.amberusa.com/blog/understanding-mri-gradients-and-slew-rates`, 2016. Accessed: 2018-05-12.

[109] Siemens Healthineers. 32-channel head coil. `www.healthcare.siemens.com/magnetic-resonance-imaging/options-and-upgrades/coils/32-channel-head-coil`, 2018. Accessed: 2018-05-12.

[110] Steren Giannini. Visual representation of the spin of a proton in a constant magnetic field B0 and then under a magnetic wave B1. Visualization of T1 and T2 times. `commons.wikimedia.org/w/index.php?title=File%3AProton_spin_MRI.webm`, 2013. Accessed: 2018-05-14.

[111] Kinuko Kono, Yuichi Inoue, Keiko Nakayama, Miyuki Shakudo, Michiharu Morino, Kenji Ohata, Kenichi Wakasa, and Ryusaku Yamada. The role of diffusion-weighted imaging in patients with brain tumors. *American Journal of Neuroradiology*, 22(6):1081–1088, 2001.

[112] Thomas L Chenevert, Lauren D Stegman, Jeremy MG Taylor, Patricia L Robertson, Harry S Greenberg, Alnawaz Rehemtulla, and Brian D Ross. Diffusion magnetic resonance imaging: an early surrogate marker of therapeutic efficacy in brain tumors. *JNCI: Journal of the National Cancer Institute*, 92(24):2029–2036, 2000.

[113] Bradford A Moffat, Thomas L Chenevert, Theodore S Lawrence, Charles R Meyer, Timothy D Johnson, Qian Dong, Christina Tsien, Suresh Mukherji, Douglas J Quint, Stephen S Gebarski, et al. Functional diffusion map: a noninvasive MRI biomarker for early stratification of clinical brain tumor response. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):5524–5529, 2005.

[114] Hanna Järnum, Elena G Steffensen, Linda Knutsson, Ernst-Torben Frund, Carsten Wiberg Simonsen, Søren Lundbye-Christensen, Ajit Shankaranarayanan, David C Alsop, Finn Taagehøj Jensen, and Elna-Marie Larsson. Perfusion MRI of brain tumours: a comparative study of pseudo-continuous arterial spin labelling and dynamic susceptibility contrast imaging. *Neuroradiology*, 52(4):307–317, 2010.

[115] Henrik Thomsen, E Steffensen, and Elna-Marie Larsson. Perfusion MRI (dynamic susceptibility contrast imaging) with different measurement approaches for the evaluation of blood flow and blood volume in human gliomas. *Acta Radiologica*, 53(1):95–101, 2012.

[116] Neetu Soni, Devender Pal S Dhanota, Sunil Kumar, Awadhesh K Jaiswal, Arun K Srivastava, et al. Perfusion MR imaging of enhancing brain tumors: Comparison of arterial spin labeling technique with dynamic susceptibility contrast technique. *Neurology India*, 65(5):1046, 2017.

[117] Jens Maus. PET scheme. `commons.wikimedia.org/wiki/File:PET-schema.png`, 2003. Accessed: 2018-05-15.

[118] Vincent Keereman, Yves Fierens, Tom Broux, Yves De Deene, Max Lonneux, and Stefaan Vandenberghe. MRI-based attenuation correction for PET/MRI using ultrashort echo time sequences. *Journal of nuclear medicine*, 51(5):812–818, 2010.

[119] Gaspar Delso and Johan Nuyts. PET/MRI: Attenuation correction. In *PET/MRI in Oncology*, pages 53–75. Springer, 2018.

[120] Nobel Prize. The nobel prize in chemistry 1943: George de Hevesy. `www.nobelprize.org/nobel_prizes/chemistry/laureates/1943/`, 2018. Accessed: 2018-05-16.

[121] Vasilis Ntziachristos, Anne Leroy-Willig, and Bertrand Tavitian. *Textbook of in vivo Imaging in Vertebrates*. John Wiley & Sons, 2007.

[122] Kam Leung. O-(2-[18f]fluoroethyl)-l-tyrosine. in: Molecular imaging and contrast agent database (micad). `www.ncbi.nlm.nih.gov/books/NBK23454/`, 2005. Accessed: 2018-05-16.

[123] Nathalie L Albert, Michael Weller, Bogdana Suchorska, Norbert Galldiks, Riccardo Soffietti, Michelle M Kim, Christian La Fougère, Whitney Pope, Ian Law, Javier Arbizu, et al. Response Assessment in Neuro-Oncology working group and European Association for Neuro-Oncology recommendations for

the clinical use of PET imaging in gliomas. *Neuro-oncology*, 18(9):1199–1208, 2016.

[124] Gabriele Pöpperl, Friedrich W Kreth, Jan H Mehrkens, Jochen Herms, Klaus Seelos, Walter Koch, Franz J Gildehaus, Hans A Kretzschmar, Jörg C Tonn, and Klaus Tatsch. FET PET for the evaluation of untreated gliomas: correlation of FET uptake and uptake kinetics with tumour grading. *European journal of nuclear medicine and molecular imaging*, 34(12):1933–1942, 2007.

[125] Norbert Galldiks, Gabriele Stoffels, Christian Filss, Marion Rapp, Tobias Blau, Caroline Tscherpel, Garry Ceccon, Veronika Dunkl, Martin Weinzierl, Michael Stoffel, et al. The use of dynamic O-(2-18F-fluoroethyl)-l-tyrosine PET in the diagnosis of patients with progressive and recurrent glioma. *Neuro-oncology*, 17(9):1293–1300, 2015.

[126] Norbert Galldiks, Ian Law, Whitney B Pope, Javier Arbizu, and Karl-Josef Langen. The use of amino acid PET and conventional MRI for monitoring of brain tumor therapy. *NeuroImage: Clinical*, 13:386–394, 2017.

[127] Julie Bolcaen, Marjan Acou, Benedicte Descamps, Ken Kersemans, Karel Deblaere, Christian Vanhove, and Ingeborg Goethals. *PET for therapy response assessment in glioblastoma*, chapter 10. Codon Publications, 2017.

[128] Anca-Ligia Grosu, Sabrina T Astner, Eva Riedel, Carsten Nieder, Nicole Wiedenmann, Felix Heinemann, Markus Schwaiger, Michael Molls, Hans-Jürgen Wester, and Wolfgang A Weber. An Interindividual Comparison of O-(2-[18F] Fluoroethyl)-L-Tyrosine (FET)–and L-[Methyl-11C] Methionine (MET)–PET in Patients With Brain Gliomas and Metastases. *International Journal of Radiation Oncology – Biology – Physics*, 81(4):1049–1058, 2011.

[129] Clemens Kratochwil, Stephanie E Combs, Karin Leotta, Ali Afshar-Oromieh, Stefan Rieken, Jürgen Debus, Uwe Haberkorn, and Frederik L Giesel. Intra-individual comparison of 18F-FET

and 18F-DOPA in PET imaging of recurrent brain tumors. *Neuro-oncology*, 16(3):434–440, 2013.

[130] Nathalie L Albert, M Unterrainer, DF Fleischmann, S Lindner, F Vettermann, A Brunegraf, L Vomacka, M Brendel, V Wenter, C Wetzel, et al. TSPO PET for glioma imaging using the novel ligand 18F-GE-180: first results in patients with glioblastoma. *European journal of nuclear medicine and molecular imaging*, 44(13):2230–2238, 2017.

[131] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2015.

[132] Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A Eschrich, Matthew B Schabath, Kenneth Forster, Hugo JWL Aerts, Andre Dekker, David Fenstermacher, et al. Radiomics: the process and the challenges. *Magnetic resonance imaging*, 30(9):1234–1248, 2012.

[133] Computation imaging & bioinformatics lab Harvard Medical School. radiomics.io. `www.radiomics.io`, 2017. Accessed: 2018-05-17.

[134] Hugo JWL Aerts. The potential of radiomic-based phenotyping in precision medicine: a review. *JAMA oncology*, 2(12):1636–1642, 2016.

[135] E Sala, E Mema, Y Himoto, H Veeraraghavan, JD Brenton, A Snyder, B Weigelt, and HA Vargas. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clinical radiology*, 72(1):3–10, 2017.

[136] Andriy Marusyk and Kornelia Polyak. Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA) – Reviews on Cancer*, 1805(1):105–117, 2010.

[137] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker,

et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

[138] Russell T Shinohara, Elizabeth M Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J Mateen, Peter A Calabresi, Samson Jarso, Dzung L Pham, Daniel S Reich, Ciprian M Crainiceanu, et al. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19, 2014.

[139] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2015.

[140] Rivka R Colen, Takeo Fujii, Mehmet Asim Bilen, Aikaterini Kotrotsou, Srishti Abrol, Kenneth R Hess, Joud Hajjar, Maria E Suarez-Almazor, Anas Alshawa, David S Hong, et al. Radiomics to predict immunotherapy-induced pneumonitis: proof of concept. *Investigational new drugs*, pages 1–7, 2017.

[141] Ahmad Chaddad, Christian Desrosiers, Lama Hassan, and Camel Tanougast. Hippocampus and amygdala radiomic biomarkers for the study of autism spectrum disorder. *BMC neuroscience*, 18(1):52, 2017.

[142] Peng Huang, Nikolay Shenkov, Sima Fotouhi, Esmaeil Davoodi-Bojd, Lijun Lu, Zoltan Mari, Hamid Soltanian-Zadeh, Vesna Sossi, and Arman Rahmim. Radiomics analysis of longitudinal datscan images for improved progression tracking in Parkinson‘s disease. *Journal of Nuclear Medicine*, 58(supplement 1):412–412, 2017.

[143] Kun Zhao, Yanhui Ding, Pan Wang, Xuejiao Dou, Bo Zhou, Hongxiang Yao, Ningyu An, Yongxin Zhang, Xi Zhang, and Yong Liu. Early classification of Alzheimer‘s disease using hippocampal texture from structural MRI. In *Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging*,

volume 10137, page 101372E. International Society for Optics and Photonics, 2017.

[144] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5:4006, 2014.

[145] M Zhou, J Scott, B Chaudhury, L Hall, D Goldgof, KW Yeom, M Iv, Y Ou, J Kalpathy-Cramer, S Napel, et al. Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches. *American Journal of Neuroradiology*, 39(2):208–216, 2018.

[146] Mu Zhou, Baishali Chaudhury, Lawrence O Hall, Dmitry B Goldgof, Robert J Gillies, and Robert A Gatenby. Identifying spatial imaging biomarkers of glioblastoma multiforme for survival group prediction. *Journal of Magnetic Resonance Imaging*, 46(1):115–123, 2017.

[147] Whitney B Pope, Hyun J Kim, Jing Huo, Jeffry Alger, Matthew S Brown, David Gjertson, Victor Sai, Jonathan R Young, Leena Tekchandani, Timothy Cloughesy, et al. Recurrent glioblastoma multiforme: ADC histogram analysis predicts response to bevacizumab treatment. *Radiology*, 252(1):182–189, 2009.

[148] Philipp Kickingereder, Michael Götz, John Muschelli, Antje Wick, Ulf Neuberger, Russell T Shinohara, Martin Sill, Martha Nowosielski, Heinz-Peter Schlemmer, Alexander Radbruch, et al. Large-scale radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment response. *Clinical Cancer Research*, 22(23):5765–5771, 2016.

[149] Nathalie L Jansen, Christoph Schwartz, Vera Graute, Sabina Eigenbrod, Jürgen Lutz, Rupert Egensperger, Gabriele Pöpperl, Hans A Kretzschmar, Paul Cumming, Peter Bartenstein, et al.

Prediction of oligodendroglial histology and LOH 1p/19q using dynamic [18F] FET-PET imaging in intracranial WHO grade II and III gliomas. *Neuro-oncology*, 14(12):1473–1480, 2012.

[150] Zeynettin Akkus, Issa Ali, Jiří Sedlář, Jay P Agrawal, Ian F Parney, Caterina Giannini, and Bradley J Erickson. Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence. *Journal of digital imaging*, 30(4):469–476, 2017.

[151] Hao Zhou, Martin Vallières, Harrison X Bai, Chang Su, Haiyun Tang, Derek Oldridge, Zishu Zhang, Bo Xiao, Weihua Liao, Yongguang Tao, et al. MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro-oncology*, 19(6):862–870, 2017.

[152] Biqi Zhang, Ken Chang, Shakti Ramkissoon, Shyam Tanguturi, Wenya Linda Bi, David A Reardon, Keith L Ligon, Brian M Alexander, Patrick Y Wen, and Raymond Y Huang. Multimodal MRI features predict isocitrate dehydrogenase genotype in high-grade gliomas. *Neuro-oncology*, 19(1):109–117, 2016.

[153] Jinhua Yu, Zhifeng Shi, Yuxi Lian, Zeju Li, Tongtong Liu, Yuan Gao, Yuanyuan Wang, Liang Chen, and Ying Mao. Noninvasive IDH1 mutation estimation based on a quantitative radiomics approach for grade II glioma. *European radiology*, 27(8):3509–3522, 2017.

[154] Panagiotis Korfiatis, Timothy L Kline, Lucie Coufalova, Daniel H Lachance, Ian F Parney, Rickey E Carter, Jan C Buckner, and Bradley J Erickson. MRI texture features as biomarkers to predict MGMT methylation status in glioblastomas. *Medical physics*, 43(6Part1):2835–2844, 2016.

[155] Yi-bin Xi, Fan Guo, Zi-liang Xu, Chen Li, Wei Wei, Ping Tian, Ting-ting Liu, Lin Liu, Gang Chen, Jing Ye, et al. Radiomics signature: a potential biomarker for the prediction of MGMT promoter methylation in glioblastoma. *Journal of Magnetic Resonance Imaging*, 47(5):1380–1387, 2018.

[156] Xintao Hu, Kelvin K Wong, Geoffrey S Young, Lei Guo, and Stephen T Wong. Support vector machine multiparametric MRI identification of pseudoprogression from tumor recurrence in patients with resected glioblastoma. *Journal of Magnetic Resonance Imaging*, 33(2):296–305, 2011.

[157] Pallavi Tiwari, Prateek Prasanna, Leo Wolansky, Marco Pinho, Mark Cohen, AP Nayate, Amit Gupta, Gagandeep Singh, KJ Hatanpaa, Andrew Sloan, et al. Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric MRI: a feasibility study. *American Journal of Neuroradiology*, 37(12):2231–2236, 2016.

[158] Philipp Lohmann, Gabriele Stoffels, Garry Ceccon, Marion Rapp, Michael Sabel, Christian P Filss, Marcel A Kamp, Carina Stegmayr, Bernd Neumaier, Nadim J Shah, et al. Radiation injury vs. recurrent brain metastasis: combining textural feature radiomics analysis and standard parameters may increase 18F-FET PET accuracy without dynamic scans. *European radiology*, 27(7):2916–2927, 2017.

[159] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[160] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[161] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[162] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[163] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[164] Stijn Bonte, Ingeborg Goethals, and Roel Van Holen. Automated grade prediction of glioma patients based on magnetic resonance imaging and a random forests approach. In *MISS 2016 : poster session booklet*, pages 11–11, 2016.

[165] Stijn Bonte, Ingeborg Goethals, and Roel Van Holen. Automated grade prediction of glioma patients based on magnetic resonance imaging and a random forests approach. In *Neuro-Oncology*, volume 18, pages iv38–iv38. Society for Neuro-Oncology, 2016.

[166] Roger Stupp, Warren P Mason, Martin J Van Den Bent, Michael Weller, Barbara Fisher, Martin JB Taphoorn, Karl Belanger, Alba A Brandes, Christine Marosi, Ulrich Bogdahn, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10):987–996, 2005.

[167] Sarah Ironside, Sunit Das, Arjun Sahgal, Claire Moroney, Todd Mainprize, and James R Perry. Optimal therapies for newly diagnosed elderly patients with glioblastoma. *Current treatment options in oncology*, 18(11):66, 2017.

[168] François-Xavier Ferracci and Hugues Duffau. Improving surgical outcome for gliomas with intraoperative mapping. *Expert review of neurotherapeutics*, 18(4):333–341, 2018.

[169] Hugues Duffau. Is non-awake surgery for supratentorial adult low-grade glioma treatment still feasible? *Neurosurgical review*, 41(1):133–139, 2018.

[170] Fred G Barker, Susan M Chang, Stephen L Huhn, Richard L Davis, Philip H Gutin, Michael W McDermott, Charles B Wilson, and Michael D Prados. Age and the risk of anaplasia in magnetic resonance-nonenhancing supratentorial cerebral tumors. *Cancer*, 80(5):936–941, 1997.

[171] JN Scott, PMA Brasher, RJ Sevick, NB Rewcastle, and PA Forsyth. How often are nonenhancing supratentorial gliomas malignant? a population study. *Neurology*, 59(6):947–949, 2002.

[172] Johan Pallud, Laurent Capelle, Luc Taillandier, Denys Fontaine, Emmanuel Mandonnet, Rémy Guillevin, Luc Bauchet, Philippe Peruzzi, Florence Laigle-Donadey, Michèle Kujas, et al. Prognostic significance of imaging contrast enhancement for WHO grade II gliomas. *Neuro-oncology*, 11(2):176–182, 2009.

[173] Matthew L White, Yan Zhang, Patricia Kirby, and Timothy C Ryken. Can tumor contrast enhancement be used as a criterion for differentiating tumor grades of oligodendrogliomas? *American journal of neuroradiology*, 26(4):784–790, 2005.

[174] L Khalid, M Carone, N Dumrongpisutikul, J Intrapiromkul, D Bonekamp, PB Barker, and David M Yousem. Imaging characteristics of oligodendrogliomas that predict grade. *American Journal of Neuroradiology*, 33(5):852–857, 2012.

[175] Evangelia I Zacharaki, Sumei Wang, Sanjeev Chawla, Dong Soo Yoo, Ronald Wolf, Elias R Melhem, and Christos Davatzikos. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magnetic resonance in medicine*, 62(6):1609–1618, 2009.

[176] Frank G Zöllner, Kyrre E Emblem, and Lothar R Schad. SVM-based glioma grading: optimization by feature reduction analysis. *Zeitschrift für medizinische Physik*, 22(3):205–214, 2012.

[177] Karoline Skogen, Anselm Schulz, Johann Baptist Dormagen, Balaji Ganeshan, Eirik Helseth, and Andrès Server. Diagnostic performance of texture analysis on MRI in grading cerebral gliomas. *European journal of radiology*, 85(4):824–829, 2016.

[178] Kevin Li-Chun Hsieh, Chung-Ming Lo, and Chih-Jou Hsiao. Computer-aided grading of gliomas based on local and global MRI features. *Computer methods and programs in biomedicine*, 139:31–38, 2017.

[179] Kevin Li-Chun Hsieh, Ruei-Je Tsai, Yu-Chuan Teng, and Chung-Ming Lo. Effect of a computer-aided diagnosis system on radiologists' performance in grading gliomas with MRI. *PloS one*, 12(2):e0171342, 2017.

[180] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data*, 4:170117, 2017.

[181] Neuroimaging Informatics Technology Initiative et al. NIfTI-1 data format. `nifti.nimh.nih.gov/`, 2015. Accessed: 2018-08-07.

[182] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[183] Robert M Haralick, Karthikeyan Shanmugam, et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[184] Mary M Galloway. Texture analysis using grey level run lengths. *NASA STI/Recon Technical Report N*, 75, 1974.

[185] Guillaume Thibault, Bernard Fertil, Claire Navarro, Sandrine Pereira, Pierre Cau, Nicolas Levy, Jean Sequeira, and Jean-Luc Mari. Shape and texture indexes application to cell nuclei classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 27(01):1357002, 2013.

[186] Moses Amadasun and Robert King. Textural features corresponding to textural properties. *IEEE Transactions on systems, man, and Cybernetics*, 19(5):1264–1274, 1989.

[187] Kenji Kira and Larry A Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Aaai*, volume 2, pages 129–134, 1992.

[188] Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, 7(1):39–55, 1997.

[189] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[190] Stijn Bonte, Ingeborg Goethals, and Roel Van Holen. An automatic and flexible brain tumour segmentation algorithm using abnormality detection for radiomics applications. In *18th Symposium of the Belgian Society of Nuclear Medicine, 5th to 7th May 2017, Ghent*, pages abstract PHYS P1:127–abstract PHYS P1:128. Belgian Society of Nuclear Medicine (BELNUC), 2017.

[191] Stijn Bonte, Roel Van Holen, and Ingeborg Goethals. Brain tumour segmentation on contrast enhanced T1w MRI using local texture and Random Forests. In *European Conference of Radiology*. European Congress of Radiology, 2018.

[192] Nelly Gordillo, Eduard Montseny, and Pilar Sobrevilla. State of the art survey on mri brain tumor segmentation. *Magnetic resonance imaging*, 31(8):1426–1438, 2013.

[193] Hassan Khotanlou, Olivier Colliot, Jamal Atif, and Isabelle Bloch. 3d brain tumor segmentation in mri using fuzzy classification, symmetry analysis and spatially constrained deformable models. *Fuzzy sets and systems*, 160(10):1457–1473, 2009.

[194] Marcel Prastawa, Elizabeth Bullitt, Sean Ho, and Guido Gerig. A brain tumor segmentation framework based on outlier detection. *Medical image analysis*, 8(3):275–283, 2004.

[195] D Zikic, B Glocker, E Konukoglu, J Shotton, A Criminisi, D Ye, C Demiralp, OM Thomas, T Das, R Jena, et al. Context-sensitive classification forests for segmentation of brain tumor tissues. *Proceedings of MICCAI-BRATS (Multimodal Brain Tumor Segmentation Challenge)*, pages 1–9, 2012.

[196] Bjoern H Menze, Ezequiel Geremia, Nicholas Ayache, and Gabor Szekely. Segmenting glioma in multi-modal images using a generative-discriminative model for brain lesion segmentation. *Proceedings of MICCAI-BRATS (Multimodal Brain Tumor Segmentation Challenge)*, 8, 2012.

[197] Joana Festa, Sergio Pereira, J Antonio Mariz, Nuno Sousa, and Carlos A Silva. Automatic brain tumor segmentation of multi-sequence mr images using random decision forests. *Proceedings of MICCAI-BRATS (Multimodal Brain Tumor Segmentation Challenge)*, 1:23–26, 2013.

[198] S Reza and KM Iftekharuddin. Multi-class abnormal brain tissue segmentation using texture. *Multimodal Brain Tumor Segmentation*, page 38, 2013.

[199] Peter D. Chang. Fully convolutional neural networks with hyperlocal features for brain tumor segmentation. In *Proceedings of MICCAI-BRATS (Multimodal Brain Tumor Segmentation Challenge)*, pages 4–9. MICCAI, 2016.

[200] Konstantinos Kamnitsas, Wenjia Bai, Enzo Ferrante, Steven McDonagh, Matthew Sinclair, Nick Pawlowski, Martin Rajchl, Matthew Lee, Bernhard Kainz, Daniel Rueckert, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. *arXiv preprint arXiv:1711.01468*, 2017.

[201] Andriy Myronenko. Non-rigid image registration: regularization, algorithms and applications. *PhD thesis, Oregon Health & Science University*, 2010.

[202] Mathieu Hatt, Catherine Cheze Le Rest, Alexandre Turzo, Christian Roux, and Dimitris Visvikis. A fuzzy locally adaptive bayesian segmentation approach for volume determination in PET. *IEEE Transactions on Medical Imaging*, 28(6):881–893, 2009.

[203] S Doyle, F Vasseur, M Dojat, and F Forbes. Fully automatic brain tumor segmentation from multiple MR sequences using hidden Markov fields and variational EM. *Proceedings of*

*MICCAI-BRATS (Multimodal Brain Tumor Segmentation Challenge)*, pages 18–22, 2013.

[204] Mikael Agn, Oula Puonti, Ian Law, PM af Rosenschöld, and K van Leemput. Brain tumor segmentation by a generative model with a prior on tumor shape. *Proceeding of the Multimodal Brain Tumor Image Segmentation Challenge*, pages 1–4, 2015.

[205] Tom Haeck, Frederik Maes, and Paul Suetens. Automated model-based segmentation of brain tumors in MR images. In *Proceedings of MICCAI-BRATS (Multimodal Brain Tumor Segmentation Challenge)*, pages 25–28, 2015.

[206] Stijn Bonte, Ingeborg Goethals, and Roel Van Holen. Machine learning based brain tumour segmentation on limited data using local texture and abnormality. *Computers in biology and medicine*, 98:39–47, 2018.

[207] Michael Kistler, Serena Bonaretti, Marcel Pfahrer, Roman Niklaus, and Philippe Büchler. The virtual skeleton database: An open access repository for biomedical research and collaboration. *J Med Internet Res*, 15(11):e245, Nov 2013.

[208] Alan C Evans, D Louis Collins, SR Mills, ED Brown, RL Kelly, and Terry M Peters. 3D statistical neuroanatomical models from 305 MRI volumes. In *Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record.*, pages 1813–1817. IEEE, 1993.

[209] JMY Willaime, FE Turkheimer, LM Kenny, and EO Aboagye. Quantification of intra-tumour cell proliferation heterogeneity using imaging descriptors of 18F fluorothymidine-positron emission tomography. *Physics in medicine and biology*, 58(2):187, 2012.

[210] Anthony Bianchi, James V Miller, Ek Tsoon Tan, and Albert Montillo. Brain tumor segmentation with symmetric texture and symmetric intensity-based decision forests. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 748–751. IEEE, 2013.

[211] Surani Anuradha Jayasuriya and Alan Wee-Chung Liew. Symmetry plane detection in neuroimages based on intensity profile analysis. In *Information Technology in Medicine and Education (ITME), 2012 International Symposium on*, volume 2, pages 599–603. IEEE, 2012.

[212] Jianguo Liu, Jayaram K Udupa, Dewey Odhner, David Hackney, and Gul Moonis. A system for brain tumor volume estimation via MR imaging and fuzzy connectedness. *Computerized Medical Imaging and Graphics*, 29(1):21–34, 2005.

[213] Khan M Iftekharuddin, Jing Zheng, Mohammad A Islam, and Robert J Ogg. Fractal-based brain tumor detection in multimodal mri. *Applied Mathematics and Computation*, 207(1):23–41, 2009.

[214] Dakai Jin, Ziyue Xu, Adam P Harrison, and Daniel J Mollura. White matter hyperintensity segmentation from T1 and FLAIR images using fully convolutional neural networks enhanced with residual connections. *arXiv preprint arXiv:1803.06782*, 2018.

[215] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

[216] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[217] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[218] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medi-*

*cal Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015.

[219] Stijn Bonte, Ingeborg Goethals, and Roel Van Holen. Individual prediction of brain tumor histological grading using radiomics on structural MRI. In *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference*, 2017.

[220] S Herlidou-Meme, JM Constans, B Carsin, D Olivie, PA Eliat, L Nadal-Desbarats, C Gondry, E Le Rumeur, I Idy-Peretti, and JD De Certaines. MRI texture analysis on texture test objects, normal brain and intracranial tumors. *Magnetic resonance imaging*, 21(9):989–993, 2003.

[221] Pantelis Georgiadis, Dionisis Cavouras, Ioannis Kalatzis, Antonis Daskalakis, George C Kagadis, Koralia Sifaki, Menelaos Malamas, George Nikiforidis, and Ekaterini Solomou. Improving brain tumor characterization on MRI by probabilistic neural networks and non-linear transformation of textural features. *Computer methods and programs in biomedicine*, 89(1):24–32, 2008.

[222] Evangelia I Zacharaki, Vasileios G Kanas, and Christos Davatzikos. Investigating machine learning techniques for MRI-based classification of brain neoplasms. *International journal of computer assisted radiology and surgery*, 6(6):821–828, 2011.

[223] Jainy Sachdeva, Vinod Kumar, Indra Gupta, Niranjan Khandelwal, and Chirag Kamal Ahuja. A package-SFERCB-"Segmentation, feature extraction, reduction and classification analysis by both SVM and ANN for brain tumors". *Applied Soft Computing*, 47:151–167, 2016.

[224] Jan Luts, Arend Heerschap, Johan AK Suykens, and Sabine Van Huffel. A combined MRI and MRSI based multiclass system for brain tumour recognition using LS-SVMs with class probabilities and feature selection. *Artificial intelligence in medicine*, 40(2):87–102, 2007.

[225] Juan M García-Gómez, Jan Luts, Margarida Julià-Sapé, Patrick Krooshof, Salvador Tortajada, Javier Vicente Robledo, Willem Melssen, Elies Fuster-García, Iván Olier, Geert Postma, et al. Multiproject–multicenter evaluation of automatic brain tumor classification by magnetic resonance spectroscopy. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 22(1):5, 2009.

[226] Chintan Parmar, Emmanuel Rios Velazquez, Ralph Leijenaar, Mohammed Jermoumi, Sara Carvalho, Raymond H Mak, Sushmita Mitra, B Uma Shankar, Ron Kikinis, Benjamin Haibe-Kains, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PloS one*, 9(7):e102107, 2014.

[227] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

[228] Subha Madhavan, Jean-Claude Zenklusen, Yuri Kotliarov, Himanso Sahni, Howard A Fine, and Kenneth Buetow. Rembrandt: helping personalized medicine become a reality through integrative translational research. *Molecular cancer research*, 7(2):157–167, 2009.

[229] Lisa Scarpace, AE Flanders, R Jain, T Mikkelsen, and DW Andrews. Data from REMBRANDT. The Cancer Imaging Archive, 2015.

[230] David A Gutman, Lee AD Cooper, Scott N Hwang, Chad A Holder, JingJing Gao, Tarun D Aurora, William D Dunn Jr, Lisa Scarpace, Tom Mikkelsen, Rajan Jain, et al. MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. *Radiology*, 267(2):560–569, 2013.

[231] The Cancer Imaging Archive. VASARI research project. `wiki.cancerimagingarchive.net/display/Public/VASARI+Research+Project`, 2015. Accessed: 2018-07-10.

[232] TCGA Research Network. The cancer genome atlas. `cancergenome.nih.gov/`, 2018. Accessed: 2018-07-10.

[233] N Pedano, AE Flanders, L Scarpace, et al. Radiology data from The Cancer Genome Atlas Low Grade Glioma [TCGA-LGG] collection. *Cancer Imaging Arch*, 2016.

[234] Binsheng Zhao, Yongqiang Tan, Wei-Yann Tsai, Jing Qi, Chuanmiao Xie, Lin Lu, and Lawrence H Schwartz. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific reports*, 6:23428, 2016.

[235] Stephen SF Yip and Hugo JWL Aerts. Applications and limitations of radiomics. *Physics in Medicine & Biology*, 61(13):R150, 2016.

[236] Stijn Bonte, Roel Van Holen, and Ingeborg Goethals. Towards computer aided diagnosis of primary brain tumors : a radiomics approach. In *7th European conference on clinical neuroimaging : final program*, pages 51–51, 2018.

[237] Stijn Bonte, Sam Donche, Michaël Henrotte, Roel Van Holen, and Ingeborg Goethals. Radiomics and machine learning on [18f]fet pet and t1ce mri discriminate between low-grade and high-grade glioma. In *NEURO-ONCOLOGY*, volume 20, pages abstract OS6.3:iii226–abstract OS6.3:iii226, 2018.

[238] Hans J Wester, Michael Herz, Wolfgang Weber, Peter Heiss, Reingard Senekowitsch-Schmidtke, Markus Schwaiger, and Gerhard Stocklin. Synthesis and radiopharmacology of O-(2-[18F] fluoroethyl)-L-tyrosine for tumor imaging. *Journal of Nuclear Medicine*, 40(1):205, 1999.

[239] Karl-Josef Langen, Gabriele Stoffels, Christian Filss, Alexander Heinzel, Carina Stegmayr, Philipp Lohmann, Antje Willuweit, Bernd Neumaier, Felix M Mottaghy, and Norbert Galldiks. Imaging of amino acid transport in brain tumours: Positron emission tomography with O-(2-[18F] fluoroethyl)-L-tyrosine (FET). *Methods*, 2017.

[240] Wolfgang A Weber, Hans-Jürgen Wester, Anca L Grosu, Michael Herz, Brigitte Dzewas, Horst-Jürgen Feldmann, Michael Molls, Gerhard Stöcklin, and Markus Schwaiger. O-(2-[18 F] fluoroethyl)-L-tyrosine and L-[methyl-11 C] methionine uptake in brain tumours: initial results of a comparative study. *European journal of nuclear medicine*, 27(5):542–549, 2000.

[241] A Habermeier, J Graf, BF Sandhöfer, J-P Boissel, F Roesch, and Ellen I Closs. System L amino acid transporter LAT1 accumulates O-(2-fluoroethyl)-L-tyrosine (FET). *Amino acids*, 47(2):335–344, 2015.

[242] M Weckesser, KJ Langen, CH Rickert, S Kloska, R Straeter, K Hamacher, G Kurlemann, H Wassmann, HH Coenen, and O Schober. O-(2-[18 F] fluorethyl)-L-tyrosine PET in the clinical evaluation of primary brain tumours. *European journal of nuclear medicine and molecular imaging*, 32(4):422–429, 2005.

[243] Nathalie L Jansen, Vera Graute, Lena Armbruster, Bogdana Suchorska, Juergen Lutz, Sabina Eigenbrod, Paul Cumming, Peter Bartenstein, Jörg-Christian Tonn, Friedrich Wilhelm Kreth, et al. MRI-suspected low-grade glioma: is there a need to perform dynamic FET PET? *European journal of nuclear medicine and molecular imaging*, 39(6):1021–1029, 2012.

[244] Dirk Pauleit, Frank Floeth, Kurt Hamacher, Markus J Riemenschneider, Guido Reifenberger, Hans-Wilhelm Müller, Karl Zilles, Heinz H Coenen, and Karl-Josef Langen. O-(2-[18F] fluoroethyl)-L-tyrosine PET combined with MRI improves the diagnostic assessment of cerebral gliomas. *Brain*, 128(3):678–687, 2005.

[245] M Kunz, N Thon, S Eigenbrod, C Hartmann, R Egensperger, J Herms, J Geisler, C La Fougere, J Lutz, J Linn, et al. Hot spots in dynamic 18FET-PET delineate malignant tumor parts within suspected WHO grade II gliomas. *Neuro-oncology*, 13(3):307–316, 2011.

[246] Dirk Pauleit, Gabriele Stoffels, Ansgar Bachofner, Frank W Floeth, Michael Sabel, Hans Herzog, Lutz Tellmann, Paul Jansen,

Guido Reifenberger, Kurt Hamacher, et al. Comparison of 18F-FET and 18F-FDG PET in brain tumors. *Nuclear medicine and biology*, 36(7):779–787, 2009.

[247] Gabriele Pöpperl, Friedrich W Kreth, Jochen Herms, Walter Koch, Jan H Mehrkens, Franz J Gildehaus, Hans A Kretzschmar, Jorg C Tonn, and Klaus Tatsch. Analysis of 18F-FET PET for grading of recurrent gliomas: Is evaluation of uptake kinetics superior to standard methods? *Journal of Nuclear Medicine*, 47(3):393, 2006.

[248] Maria Lucia Calcagni, Guido Galli, Alessandro Giordano, Silvia Taralli, Carmelo Anile, Andreas Niesen, and Richard Paul Baum. Dynamic O-(2-[18F] fluoroethyl)-L-tyrosine (F-18 FET) PET for glioma grading: assessment of individual probability of malignancy. *Clinical nuclear medicine*, 36(10):841–847, 2011.

[249] Marion Rapp, Alexander Heinzel, Norbert Galldiks, Gabriele Stoffels, Jörg Felsberg, Christian Ewelt, Michael Sabel, Hans J Steiger, Guido Reifenberger, Thomas Beez, et al. Diagnostic performance of 18F-FET PET in newly diagnosed cerebral lesions suggestive of glioma. *Journal of Nuclear Medicine*, 54(2):229–235, 2013.

[250] V Dunet, P Maeder, M Nicod-Lalonde, B Lhermitte, C Pollo, J Bloch, R Stupp, R Meuli, and JO Prior. Combination of MRI and dynamic FET PET for initial glioma grading. *Nuklearmedizin*, 53(4):155–161, 2014.

[251] Nathalie L Albert, Isabel Winkelmann, Bogdana Suchorska, Vera Wenter, Christine Schmid-Tannwald, Erik Mille, Andrei Todica, Matthias Brendel, Jörg-Christian Tonn, Peter Bartenstein, et al. Early static 18F-FET-PET scans have a higher accuracy for glioma grading than the standard 20–40 min scans. *European journal of nuclear medicine and molecular imaging*, 43(6):1105–1114, 2016.

[252] Thomas Pyka, Jens Gempt, Daniela Hiob, Florian Ringel, Jürgen Schlegel, Stefanie Bette, Hans-Jürgen Wester, Bernhard Meyer, and Stefan Förster. Textural analysis of pre-therapeutic [18F]-FET-PET and its correlation with tumor grade and patient sur-

vival in high-grade gliomas. *European journal of nuclear medicine and molecular imaging*, 43(1):133–141, 2016.

[253] Antoine Verger, Christian P Filss, Philipp Lohmann, Gabriele Stoffels, Michael Sabel, Hans J Wittsack, Elena Rota Kops, Norbert Galldiks, Gereon R Fink, Nadim J Shah, et al. Comparison of 18F-FET PET and perfusion-weighted MRI for glioma grading: a hybrid PET/MR study. *European Journal of Nuclear Medicine and Molecular Imaging*, 44(13):2257–2265, 2017.

[254] Manuel Röhrich, Kristin Huang, Daniel Schrimpf, Nathalie L Albert, Thomas Hielscher, Andreas von Deimling, Ulrich Schüller, Antonia Dimitrakopoulou-Strauss, and Uwe Haberkorn. Integrated analysis of dynamic FET PET/CT parameters, histology, and methylation profiling of 44 gliomas. *European journal of nuclear medicine and molecular imaging*, pages 1–12, 2018.

[255] Norbert Galldiks, Gabriele Stoffels, Maximilian I Ruge, Marion Rapp, Michael Sabel, Guido Reifenberger, Zuhal Erdem, Nadim J Shah, Gereon R Fink, Heinz H Coenen, et al. Role of O-(2-18F-fluoroethyl)-L-tyrosine PET as a diagnostic tool for detection of malignant progression in patients with low-grade glioma. *J Nucl Med*, 54(12):2046–2054, 2013.

[256] Hansjörg Vees, Srinivasan Senthamizhchelvan, Raymond Miralbell, Damien C Weber, Osman Ratib, and Habib Zaidi. Assessment of various strategies for 18 F-FET PET-guided delineation of target volumes in high-grade glioma patients. *European journal of nuclear medicine and molecular imaging*, 36(2):182–193, 2009.

[257] Maximilian Niyazi, Julia Geisler, Axel Siefert, Silke Birgit Schwarz, Ute Ganswindt, Sylvia Garny, Oliver Schnell, Bogdana Suchorska, Friedrich-Wilhelm Kreth, Jörg-Christian Tonn, et al. FET–PET for malignant glioma treatment planning. *Radiotherapy and Oncology*, 99(1):44–48, 2011.

[258] M Unterrainer, I Winkelmann, B Suchorska, A Giese, V Wenter, FW Kreth, J Herms, P Bartenstein, JC Tonn, and NL Albert.

Biological tumour volumes of gliomas in early and standard 20–40 min 18 F-FET PET images differ according to IDH mutation status. *European journal of nuclear medicine and molecular imaging*, 45(7):1242–1249, 2018.

[259] Frank W Floeth, Dirk Pauleit, Michael Sabel, Gabriele Stoffels, Guido Reifenberger, Markus J Riemenschneider, Paul Jansen, Heinz H Coenen, Hans-Jakob Steiger, and Karl-Josef Langen. Prognostic value of O-(2-18F-fluoroethyl)-L-tyrosine PET and MRI in low-grade glioma. *Journal of nuclear medicine*, 48(4):519–527, 2007.

[260] Nathalie L Jansen, Bogdana Suchorska, Vera Wenter, Sabina Eigenbrod, Christine Schmid-Tannwald, Andreas Zwergal, Maximilian Niyazi, Mark Drexler, Peter Bartenstein, Oliver Schnell, et al. Dynamic 18F-FET PET in newly diagnosed astrocytic low-grade glioma identifies high-risk patients. *Journal of Nuclear Medicine*, pages jnumed–113, 2014.

[261] Bogdana Suchorska, Nathalie L Jansen, Jennifer Linn, Hans Kretzschmar, Hendrik Janssen, Sabina Eigenbrod, Matthias Simon, Gabriele Pöpperl, Friedrich W Kreth, Christian la Fougere, et al. Biological tumor volume in 18FET-PET before radiochemotherapy correlates with survival in GBM. *Neurology*, 84(7):710–719, 2015.

[262] Norbert Galldiks, Karl-Josef Langen, Richard Holy, Michael Pinkawa, Gabriele Stoffels, Kay W Nolte, Hans J Kaiser, Christan P Filss, Gereon R Fink, Heinz H Coenen, et al. Assessment of treatment response in patients with glioblastoma using O-(2-18F-fluoroethyl)-L-tyrosine PET in comparison to MRI. *Journal of Nuclear Medicine*, 53(7):1048, 2012.

[263] Markus Hutterer, Martha Nowosielski, Daniel Putzer, Dietmar Waitz, Gerd Tinkhauser, Herwig Kostron, Armin Muigg, Irene J Virgolini, Wolfgang Staffen, Eugen Trinka, et al. O-(2-18F-fluoroethyl)-L-tyrosine PET predicts failure of antiangiogenic

treatment in patients with recurrent high-grade glioma. *Journal of Nuclear Medicine*, 52(6):856, 2011.

[264] Norbert Galldiks, Marion Rapp, Gabriele Stoffels, Gereon R Fink, Nadim J Shah, Heinz H Coenen, Michael Sabel, and Karl-Josef Langen. Response assessment of bevacizumab in patients with recurrent malignant glioma using [18 F] Fluoroethyl-L-tyrosine PET in comparison to MRI. *European journal of nuclear medicine and molecular imaging*, 40(1):22–33, 2013.

[265] Ashish H Shah, Brian Snelling, Amade Bregy, Payal R Patel, Danoushka Tememe, Rita Bhatia, Evelyn Sklar, and Ricardo J Komotar. Discriminating radiation necrosis from tumor progression in gliomas: a systematic review what is the best imaging modality? *Journal of neuro-oncology*, 112(2):141–152, 2013.

[266] Alexander Radbruch, Joachim Fladt, Philipp Kickingereder, Benedikt Wiestler, Martha Nowosielski, Philipp Bäumer, Heinz-Peter Schlemmer, Antje Wick, Sabine Heiland, Wolfgang Wick, et al. Pseudoprogression in patients with glioblastoma: clinical relevance despite low incidence. *Neuro-oncology*, 17(1):151–159, 2014.

[267] Gabriele Pöpperl, Claudia Götz, Walter Rachinger, Franz-Josef Gildehaus, Jörg-Christian Tonn, and Klaus Tatsch. Value of O-(2-[18 F] fluoroethyl)-L-tyrosine PET for the diagnosis of recurrent glioma. *European journal of nuclear medicine and molecular imaging*, 31(11):1464–1470, 2004.

[268] JH Mehrkens, G Pöpperl, W Rachinger, J Herms, K Seelos, K Tatsch, JC Tonn, and FW Kreth. The positive predictive value of O-(2-[18 F] fluoroethyl)-L-tyrosine (FET) PET in the diagnosis of a glioma recurrence after multimodal treatment. *Journal of neuro-oncology*, 88(1):27–35, 2008.

[269] Garry Ceccon, Lazaros Lazaridis, Gabriele Stoffels, Marion Rapp, Manuel Weber, Tobias Blau, Phillip Lohmann, Sied Kebir, Ken Herrmann, Gereon R Fink, et al. Use of FET PET in glioblastoma

patients undergoing neurooncological treatment including tumour-treating fields: initial experience. *European journal of nuclear medicine and molecular imaging*, pages 1–10, 2018.

[270] Miklos Pless and Uri Weinberg. Tumor treating fields: concept, evidence and future. *Expert opinion on investigational drugs*, 20(8):1099–1106, 2011.

[271] Ivo Rausch, Jacobo Cal-González, David Dapra, Hans Jürgen Gallowitsch, Peter Lind, Thomas Beyer, and Gregory Minear. Performance evaluation of the Biograph mCT Flow PET/CT system according to the NEMA NU2-2012 standard. *EJNMMI physics*, 2(1):26, 2015.

[272] Mathieu Hatt, Baptiste Laurent, Anouar Ouahabi, Hadi Fayad, Shan Tan, Laquan Li, Wei Lu, Vincent Jaouen, Clovis Tauber, Jakub Czakon, et al. The first miccai challenge on pet tumor segmentation. *Medical image analysis*, 44:177–195, 2018.

[273] Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas. How many trees in a random forest? In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 154–168. Springer, 2012.

[274] Robert J Jackson, Gregory N Fuller, Dima Abi-Said, Frederick F Lang, Ziya L Gokaslan, Wei Ming Shi, David M Wildrick, and Raymond Sawaya. Limitations of stereotactic biopsy in the initial management of gliomas. *Neuro-oncology*, 3(3):193–200, 2001.

[275] Matjaz Bevk and Igor Kononenko. A statistical approach to texture description of medical images: a preliminary study. In *Computer-Based Medical Systems, 2002.(CBMS 2002). Proceedings of the 15th IEEE Symposium on*, pages 239–244. IEEE, 2002.