Robust Imaging Biomarkers for Brain Tumors

Pálsson, Sveinn

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Robust Imaging Biomarkers for Brain Tumors

Sveinn Pálsson

Kongens Lyngby 2021

Technical University of Denmark
Department of Health Technology
Ørsteds Plads, Building 345C,
2800 Kongens Lyngby, Denmark
healthtech-info@dtu.dk
www.healthtech.dtu.dk/

Summary (English)

The goal of this PhD project is to develop robust imaging biomarkers for building prediction models of brain tumors. Brain tumor biomarkers are used for analysis of the disease, e.g. with respect to tumor grade, tumor recurrence and survival. We focus on magnetic resonance (MR) imaging, which is the most widely used imaging method for studying the brain and its disorders. Various scanning sequences exist for MR imaging that each give different contrasts and are used to visualize different biological properties of the tissues of interest. To study and diagnose brain tumors, several different MR sequences are typically used. However, an MR scan in one clinical center may not look the same as in another, even if the same subject and type of sequence is used. There are many factors that influence the resulting image, which makes development of computational diagnostic tools difficult. In this thesis we emphasize developing methods that overcome these issues to facilitate clinical adoption.

We build generative models of tumor shape and apply them to both prediction and segmentation of brain tumors. We develop interpretable and robust imaging biomarkers based on whole-brain segmentations of MR images, and apply them to survival prediction of glioblastoma. These biomarkers measure the deformation of brain structures surrounding the tumor, and are computed fully automatically, while being sequence-adaptive. Using these biomarkers, we show improvement in survival prediction over models that only consider conventional non-imaging biomarkers.

Summary (Danish)

Dette PhD-projekts primære mål er at udvikle robuste billeddannende biomarkører til opbygning af forudsigelsesmodeller af hjernetumorer. Hjernetumorbiomarkører bruges til analyse af sygdommen, f.eks. med hensyn til tumorgrad, tumorrecidiv og overlevelse. Vi fokuserer på magnetisk resonans (MR) billeddannelse, som er den mest anvendte billeddannelsesmetode til at studere hjernen og dens lidelser. Der findes forskellige scanningssekvenser til MR -billeddannelse, der hver giver forskellige kontraster og bruges til at visualisere forskellige biologiske egenskaber for væv af interesse. For at studere og diagnosticere hjernetumorer bruges typisk flere forskellige MR -sekvenser. Imidlertid kan en MR -scanning i et klinisk center muligvis ikke se det samme ud som i et andet, selvom det samme emne og den type sekvens bruges. Der er mange faktorer, der påvirker det resulterende billede, hvilket gør udviklingen af beregningsdiagnostiske værktøjer vanskelig. I dette speciale lægger vi vægt på at udvikle metoder, der overvinder disse spørgsmål for at lette klinisk adoption.

Vi bygger generative modeller for tumorform og anvender dem til både forudsigelse og segmentering af hjernetumorer. Vi udvikler fortolkelige og robuste billedbiomarkører baseret på helhjernesegmenteringer af MR-billeder og anvender dem til forudsigelse af overlevelse af glioblastom. Disse biomarkører måler deformationen af hjernestrukturer, der omgiver tumoren, og beregnes fuldt automatisk, samtidig med at de er sekvens-adaptive. Ved hjælp af disse biomarkører viser vi forbedring i overlevelse forudsigelse i forhold til modeller, der kun overvejer konventionelle ikke-billeddannende biomarkører.

Preface

This thesis was prepared at the Department of Health Technology at the Technical University of Denmark in partial fulfillment of the Ph.D. degree requirements. Professor Koen Van Leemput, from the Technical University of Denmark and the Athinoula A. Martinos Center for Biomedical Imaging, acted as the main supervisor. The Ph.D. project was performed in collaboration with Rigshospitalet, where professor Ian Law acted as co-supervisor. The work described in this thesis was funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 765148. The thesis focuses on robust imaging biomarkers for brain tumors.

Lyngby, 1-October-2021

Sveinn Pálsson

Sveinn Pálsson

Acknowledgements

Firstly, I would like to thank my main supervisor, Koen Van Leemput, for guiding me through the PhD studies. His vast knowledge and experience in the field has been tremendously helpful, and his integrity and passion for science has been immensely inspiring. I would also like to thank my co-supervisor Ian Law for hosting me at Rigshospitalet and giving me valuable insights into the clinical aspects of the project.

Thanks to my fellow TRABIT network ESRs: Ines Meyer, Francesco La Rosa, Stefano Cerri, Ivan Ezhov, Andrey Zhylka, Daniel Krahulec, Luca Canalini, Lucas Fidon, Thomas Yu, Athena Taymourtash, Suprosanna Shit, Carmen Moreno Genis, Amnah Mahroo and Ezequiel de la Rosa. Thanks for the time we spent together during workshops, summer schools and conferences. I would also like to thank the TRABIT supervisors for organizing these events.

Many thanks to my colleagues at DTU, Chiara Mauri, Stefano Cerri, Jacob Frøsig and Mikael Agn. We had many good times together, and I hope many more will come in the future.

Last but not least, I want to thank my wife, Renfei Liu, and my family for their love and support.

Scientific Contributions

Papers included in this thesis

Paper A: Pálsson S., Cerri S., Dittadi A., Van Leemput K., Semi-supervised Variational Autoencoder for Survival Prediction. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2019. Lecture Notes in Computer Science*, Vol. 11993, pp. 124–134, 2020.

Paper B: Pálsson S., Cerri S., Poulsen H., Urup T., Law I., Van Leemput K., Predicting survival of glioblastoma from automaticwhole-brain and tumor segmentation of MR images. In: *Submitted to Scientific Reports*, 2021.

Paper C: Pálsson S., Cerri S., Van Leemput K., Prediction of MGMT Methylation Status of Glioblastoma using Radiomics and Latent Space Shape Features. (accepted) In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2021.*

Contents

Chapter 1

Introduction

The goal of this PhD project is to develop novel, robust imaging biomarkers of brain tumors. Biomarkers are any relevant biological information which can be used for analysis of disease, such as prediction of treatment outcome. Magnetic resonance (MR) imaging is a very efficient way of gathering information about a patient's brain, and is the most widely used imaging method for studying brain disorders. Automatic computational methods to extract relevant information from brain scans exist and are actively being researched and improved.

Various scanning sequences exist for MR imaging that each give different contrasts and are used to visualize different biological properties of the tissues of interest. To study and diagnose brain tumors, several different sequences are typically used. However, there are many factors that influence the resulting image, making development of computational tools difficult.

Recent years have seen extraordinary advancements in mathematical modeling, especially in the realm of machine learning, where numerous methods for computer-aided diagnostics related to MR imaging have been developed. The main barriers that limit clinical adoption of such methods is their direct dependence on the raw image intensities, making them vulnerable to the many sources of variation inherent in MR imaging. Another commonly overlooked issue is that of interpretability. Models such as neural networks are hard to interpret, making them undesirable by clinicians who need to trust that the model

is making reasonable predictions. In this thesis, we focus on building models that are robust and interpretable.

1.1 Contributions

In paper A, we describe a semi-supervised variational autoencoder to predict survival of glioblastoma patients from the tumor's shape. We submitted our method to the BraTS challenge of the BrainLes workshop in the MICCAI conference of 2019.

For paper B, we developed interpretable and robust imaging biomarkers based on segmentations of MR images, and used them for survival prediction of glioblastoma patients. The method is based on using a whole-brain and tumor segmentation method (also described in this thesis) to measure deformation of various brain structures, caused by the tumor. We show that the severity of deformation of some structures has prognostic value in terms of overall-survival and progression-free survival.

In paper C, we predict MGMT promoter status of glioblastoma patients from MR images, using a combination of shape features and radiomics features. We submitted the method to the BraTS 2021 challenge.

1.2 Overview of the thesis

The remainder of this thesis is structured as follows:

- Chapter 2 gives on overview of the topics covered in this thesis. We discuss brain tumor imaging and how mathematical models can be used for various clinical tasks.
- Chapter 3 describes the tumor shape model developed in paper A, which we also used for the segmentation method of paper B and feature extraction in paper C.
- Chapter 4 describes the segmentation method applied in paper B. We first describe a brain segmentation method, and then how we extend it to handle brain tumors.

- Chapter 5 describes the research related to paper B, where we develop interpretable and robust imaging biomarkers for brain tumor patients and show that they can be used to predict overall survival and progression-free survival.
- Chapter 6 describes the research related to paper C, where we predict MGMT promoter status of glioblastoma using radiomics and shape features extracted using a deep generative model of tumor shape.
- Finally, in Chapter 7, we conclude the thesis and discuss ideas for future work based on the contributions of this thesis.

Chapter 2

Brain tumor imaging

This chapter describes how and why images of brain tumors are acquired. The chapter is structured as follows:

- In the first section, we discuss brain imaging and describe how images are acquired.
- In the second section, we describe brain tumors and brain tumor treatment.
- Finally, in the third section, we describe modeling approaches for brain images and brain tumors and how such models are used.

2.1 Brain image acquisition

Acquiring images of the brain is done with several different techniques. The most commonly used techniques are magnetic resonance (MR) imaging, computed tomography (CT) and positron emission tomography (PET). MR images provide the best soft-tissue contrast and are therefore the most commonly used for segmentation of brain anatomy and pathologies. CT and PET have comparably

poor soft-tissue contrast, but are useful for other things, such as radiotherapy planning. CT contains information about attenuation and scatter of high-energy photon radiation in various parts of the head and brain. PET is used to visualize metabolic activity of a tumor by using a radioactive substance that is injected into the blood stream of the patient.

In this thesis, we will mainly focus on MR images. MR imaging exploits the fact that hydrogen atoms in the body produce an electromagnetic signal when placed in a strong constant magnetic field and then perturbed with a weak oscillating magnetic field. By algorithmically perturbing the hydrogen atoms and measuring their response, an image can be constructed. Differences in the local environment of the hydrogen atoms between different soft-tissues provides a mechanism for image contrast. This image acquisition procedure has many parameters which are varied to obtain specific desired contrasts. Specific configurations of these parameters are referred to as MR sequences.

The most commonly used MR sequences for brain imaging are

- **T1w**: T1-weighted images are the most commonly used for visualization and segmentation of the structures in the brain as they provide particularly good contrast between white and gray matter tissue. MPRAGE (Brant-Zawadzki et al., 1992) is one 3D T1-weighted imaging sequence, commonly used for brain tumor imaging. An example of an MPRAGE image is shown in Fig. 2.1 (A).
- **T1w-c**: A trick, commonly used to highlight lesions in a T1-weighted image, is to inject a contrast-enhancing substance, a so-called "contrast agent", into the blood stream of the patient before scanning. The most commonly used agent for brain tumor imaging is the rare-earth metal, gadolinium. The contrast agent accumulates near damage in the blood-brain barrier, which may indicate presence of an aggressive brain tumor, such as glioblastoma. An example of a T1w-c (using gadolinium) image is shown in Fig. 2.1 (B).
- **T2w**: T2-weighted images have good contrast for free fluids such as cerebrospinal fluid (CSF) and blood. Abnormalities such as edema and tumor tissue can be detected from signal change in T2w images. An example of a T2-weighted image is shown in Fig. 2.1 (C).
- **T2w-FLAIR**: T2-weighted fluid attenuated inversion recovery (FLAIR) images are also used to detect abnormalities similarly to T2w images. However, abnormalities remain bright in FLAIR but CSF is attenuated and appears dark. An example of a FLAIR image is shown in Fig. 2.1 (D).

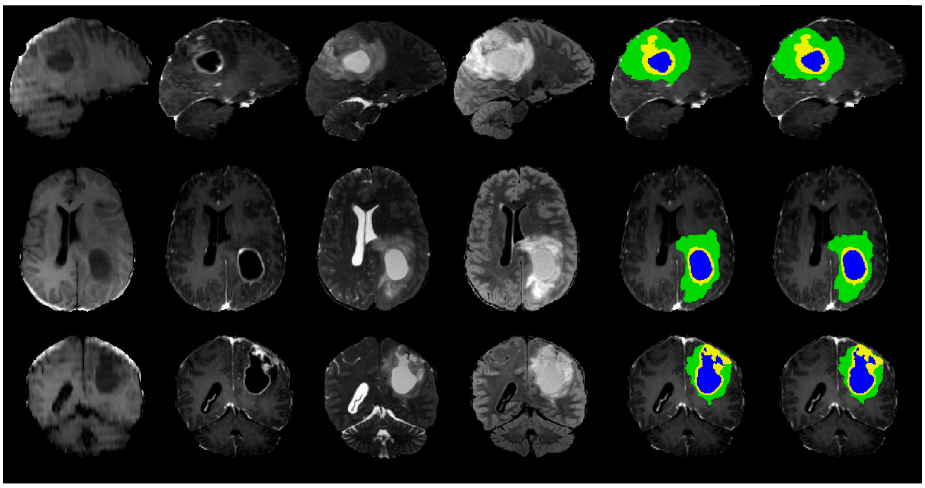

Fig. 2.1: MR images of a brain tumor patient. From top to bottom: sagittal, axial and coronal view. The columns show (A) T1w, (B) T1w-c, (C) T2w, (D) T2w-FLAIR, (E) manual tumor segmentation, (F) automatic tumor segmentation.

There is a large number of parameters involved in MR imaging, and while standardization is achieved to some extent by defining imaging sequences, there remain many parameters in the process that can differ, e.g. between clinical centers, equipment, and software. Some of these differences are visible to humans but some are subtle and numerical, and only noticed when an image processing algorithms fails to correctly process the image. Other sources of variation include image artifacts caused by motion, intensity inhomogeneity (bias-field), and partial volume effect. Bias-field appears as a smoothly varying intensity artifact and is more pronounced in images acquired using stronger magnetic fields. Partial volume effect arises when more than one tissue type occurs in a voxel. All of these issues have to be considered when processing MR images.

2.2 Brain Tumors

A brain tumor is a cluster of abnormal cells in the brain. In 2015, global prevalence of brain tumors was estimated to be 1.2 million (Vos et al., 2016), with deaths estimated at 228,000 (Wang et al., 2016). Brain tumors may be benign or malignant, can occur anywhere within the brain, and are classified into

either primary or secondary tumors, based on whether the tumor growth started within the brain (primary) or spread from other parts of the body (secondary).

In this thesis, we will primarily focus on gliomas. Gliomas are the most common primary brain tumors and comprise about 80 percent of all malignant brain tumours (Goodenberger and Jenkins, 2012). Gliomas begin in glial cells, support cells that surround neurons and help them function (Jessen and Mirsky, 1980). Gliomas are classified into three types, by the type of glial cells that produced them: astrocytomas, ependymomas and oligodendrogliomas. Gliomas are further classified by their grade, from I to IV (Wesseling and Capper, 2018), with grade IV being the most aggressive and malignant. Grade I gliomas are benign and can often be removed by surgery. Grade II gliomas are often benign tumors and are referred to as low-grade gliomas, whereas grades III and IV are malignant and are called high-grade gliomas. Low-grade gliomas often increase in grade over time.

Glioblastoma is a high-grade glioma with particularly poor prognosis and is estimated to occur in 3 out of 100,000 people per year (Gallego, 2015). Median overall survival (OS) of glioblastoma patients is less than 15 months, and the 5-year OS rate is only about 10%, even when aggressively treated (Louis et al., 2007; Gutman et al., 2013; Stupp et al., 2009; Poulsen et al., 2017). The standard treatment consists of maximal surgical resection followed by radiation therapy and chemotherapy with temozolomide (Stupp et al., 2009). An example of MR scans of a glioblastoma patient is shown in Fig. 2.1. In Fig. 2.1 (E), the tumor areas have been drawn by a clinical expert. The yellow area shows contrast enhancing tumor core, which is defined as the part of the tumor that lights up in the T1w-c image (Fig. 2.1 (B)). Such contrast enhancing areas are characteristic of high-grade gliomas. The blue area is non-enhancing tumor core and the green area is edema.

2.3 Models of brain images and brain tumors

MR imaging is the primary tool for detecting brain tumors. While detecting the presence of a brain tumor is fairly simple using the MR sequences described in Section 2.1, mathematical models of the images are needed for more complex tasks, such as treatment planning and analysis of the disease.

2.3.1 Segmentation

Segmentation of brain images is useful for purposes such as studying and diagnosing disorders of the brain. Manually drawing tumor segmentations, such as the one shown in Fig. 2.1 (E), is a tedious task and one that requires an expert to complete. An even more tedious task is to segment the *whole brain*, which is useful for planning radiation therapy, for example. The goal of radiation therapy is to target the tumor with radiation while minimizing radiation of sensitive healthy brain structures (Shaffer et al., 2010). Fortunately, automated segmentation methods exist (see Fig. 2.1 (F)), which we will now discuss.

Say we are given MR image data $\mathbf{D}$ of a subject and we aim to infer a label map $\mathbf{l}$ from the given image data. We will discuss probabilistic segmentation models, which are divided into two categories: *discriminative* and *generative*. Discriminative models aim to model the conditional distribution $p(\mathbf{l}|\mathbf{D})$ directly while generative models aim to model the joint distribution $p(\mathbf{D}, \mathbf{l})$. In the generative approach, the generative process of the data, $p(\mathbf{D}|\mathbf{l})$, and prior knowledge of the target variable, $p(\mathbf{l})$, are both modeled and inference of $\mathbf{l}$ is achieved by "inverting" the model using Bayes rule: $p(\mathbf{l}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{l})p(\mathbf{l})$.

Parameters of discriminative models are typically learned from training data $\{\mathbf{D}_n, \mathbf{l}_n\}_n^N$ of N pairs. In this setting, the image data has usually been labelled by human experts and the model, f, is trained such that (on average) the estimated label maps, $\hat{\mathbf{l}}_n = f(\mathbf{D}_n)$, are close to the human-labelled ones. Recently, discriminative segmentation models are usually implemented as neural networks containing a large number of parameters, requiring a large amount of training data to fit. For medical image segmentation, convolutional neural networks (CNNs) have been very successful, especially so-called "U-Nets" (Ronneberger et al., 2015). Isensee et al. (2021) created a self-configuring U-Net framework, winning medical image segmentation challenges including the MICCAI-BraTS 2020 tumor segmentation challenge (Isensee and Maier-Hein, 2021; Bakas et al., 2018; Menze et al., 2014). However, the performance of discriminative models tends to degrade when presented with data that differs from the training data. This problem is inherent in medical imaging, where images tend to differ significantly between clinical centers, types of scanners used and sequences used. A discriminative model trained on image data from one center may completely fail when presented with data from other centers. Furthermore, discriminative models require a lot of training data, which is a particularly scarce resource in the case of medical images. The scarcity of medical training data is a consequence of factors such as data privacy laws and the high cost of generating high quality data with expert annotations.

Generative models, as we mentioned, model the join distribution, $p(\mathbf{D}|\mathbf{l})p(\mathbf{l})$.

In the case of brain segmentation, the prior $p(\mathbf{l})$ encodes knowledge about the segmentation labels that we have before seeing the data, e.g. that brain tissue can only be inside the head or that the thalamus is located near the center of the brain. The likelihood function $p(\mathbf{D}|\mathbf{l})$ models the image intensities given the labels and can chosen carefully based on knowledge of the problem, e.g. knowing that image intensities of a particular structure follow a Gaussian distribution. Without knowledge of intensity distribution given the labels, the likelihood function can be chosen to be more expressive e.g. mixture of Gaussians (Ashburner and Friston, 2005). Generative models can be made more adaptive to differences between subjects than discriminative methods because the parameters of the likelihood function can be estimated directly from the image to be segmented ("unsupervised"). The prior $p(\mathbf{l})$ can be learned from training data, e.g. in the form of a *probabilistic atlas* (Van Leemput et al., 1999), obtained by estimating the frequency of brain tissues at each location. The combination of an unsupervised intensity model and a probabilistic atlas has been used to segment brain tissue (Ashburner and Friston, 1997; Van Leemput et al., 1999; Ashburner and Friston, 2005), and to segment the whole-brain (Puonti et al., 2016). With a similar approach, segmenting the whole-brain and tumor simultaneously has been done in Agn et al. (2019), where the likelihood function is not entirely unsupervised, but has several constraints on it that are learned from training data. The two main problems we discussed for discriminative models, lack of data and variability, are somewhat avoided by the generative approach. Lack of data is less of an issue because a rather small set of training data is required for building a probabilistic atlas, compared to what is needed for estimating (thousands or millions of) parameters of neural networks. Being unsupervised allows the image intensity model to adapt to variability of the data, such as the types of MR sequences used.

2.3.2 Prediction

Brain MR images of tumor patients contain vast amounts of information about the disease that is not readily visible to the human eye. Without mathematical models we are limited to only rudimentary estimates of size, location and tumor type. Models to predict various attributes of disease have been created, for example, to stratify patients for clinical trials and to guide treatment based on expected outcome (Gorlia et al., 2008; Katzman et al., 2018), to study tumor recurrence (Rathore et al., 2018; Lundemann et al., 2019), and predicting tumor grade (Wang et al., 2019) . Prediction models can also be built with the purpose of discovering patterns that lead to a better understanding of a disease. For example, consider the task of predicting glioma grade using variables such as the size of edema, enhancing-core and non-enhancing core (cf. Fig. 2.1 (E)). A pattern one might find with a simple model is that non-zero size of enhancing

core is a good indicator of high tumor grade. Now for the same task, say we instead used a deep neural network that takes the image volume as input and train it to predict tumor grade; its accuracy may reach higher than in the previous case, but interpreting the model predictions is difficult. Interpretability is also important for the users (e.g. clinicians) to trust the model (Shortliffe and Sepúlveda, 2018), as accuracy in one study can be insufficient to guarantee generalization to their data.

Radiomics (Lambin et al., 2012), is a very popular approach to extracting biomarkers or "features" from medical images, to use for prediction tasks related to tumors (Yip and Aerts, 2016), including analysis of glioblastoma (Narang et al., 2016). The method has many implementations but essentially involves extracting a large number of features from the tumor region of the available images. The features, called "radiomic" features are a variety of statistical, shape and texture features. The standard workflow then consists of selecting a subset of these features and training a discriminative classifier or regression model. Radiomics has been rather successfully applied to the prediction tasks that we consider in this thesis, but they do face challenges with interpretability and generalizability. We will discuss these issues later in this thesis, in Chapter 5, and propose alternative features.

Chapter 3

A deep generative model for tumor shape

In this chapter, we will describe a generative model for tumor shape. This model is used in paper B as part of the image segmentation method described in detail in Chapter 4. The model is also used in paper C, to extract shape features from glioblastoma images, which are then used in a discriminative classifier to classify MGMT methylation (Chapter 6). In paper A, a similar model is used in a semi-supervised setting to classify glioblastoma patients into different survival groups. The generative model is a variational autoencoder (VAE), a nonlinear latent variable model trained with a gradient-based procedure based on variational principles. In this chapter:

- We begin with an overview of the theory of VAE
- Next, we describe the implementation of a VAE for tumor shape
- Finally, we describe the modelling approach of paper A, applying the VAE model in a semi-supervised setting for survival prediction.

3.1 Variational autoencoder

A variational autoencoder (Kingma and Welling, 2013) is a latent variable model, meaning that it assumes an observed data sample $\mathbf{x}$ is generated by a random process involving latent variables $\mathbf{z}$, where the number of latent variables is typically much lower than the dimensionality of the data.

The purpose of the model is to learn the generative process $p(\mathbf{x}|\mathbf{z})$ while assuming some known prior $p(\mathbf{z})$ of the latent variables. The idea is to introduce a function p_θ with parameters θ to approximate the generative distribution by maximizing the probability of observed data under the model. Assuming the observed training data are N samples $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$, the problem is to maximize the probability the model assigns to these samples. Formally, the problem is to maximize

$$\sum_{i=1}^{N} \log p_\theta(\mathbf{x}^{(i)}) = \sum_{i=1}^{N} \log \int_{\mathbf{z}} p_\theta(\mathbf{x}^{(i)}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \tag{3.1}$$

with respect to θ. However, the optimization problem is difficult, as the probability of observed data involves an intractable integral over the latent variables. To avoid this problem, the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is used to exploit the fact that the optimization would be easier if the latent variables were known. The EM algorithm iteratively constructs and maximizes a lower bound to $\log p_\theta(\mathbf{x})$ [1] in a process that involves "filling in" the missing latent variables using their posterior distribution, $p_\theta(\mathbf{z}|\mathbf{x})$. Since this posterior is intractable, an approximation to it, $q_\phi(\mathbf{z}|\mathbf{x})$ is introduced. This approximation is defined as a multivariate Gaussian distribution with diagonal covariance matrix where the mean and covariance are functions of $\mathbf{x}$, parameterized by ϕ.

The lower bound to $\log p_\theta(\mathbf{x})$ is derived in the following way:

[1] We will omit the index i when it is irrelevant

$$
\begin{aligned}
\log p_\theta(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x})] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{z}|\mathbf{x})}\right]\right] \\
&= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right]}_{=\mathcal{L}_{\theta,\phi}(\mathbf{x})} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]\right]}_{=D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x}))}
\end{aligned}
$$

where D_{KL} denotes the Kullback-Leibler (KL) divergence. Since the KL divergence is always non-negative, we have that

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}_{\theta,\phi}(\mathbf{x}) \tag{3.2}$$

The objective of the VAE is then to maximize the lower bound to Eq. 3.1:

$$\sum_{i=1}^{N} \mathcal{L}_{\theta,\phi}(\mathbf{x}^{(i)}) \tag{3.3}$$

with respect to both the variational parameters ϕ and the generative parameters θ, which can be achieved by stochastic gradient ascent. However, the gradient of $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ with respect to ϕ is difficult to estimate but the problem is avoided by the reparameterization trick described in Kingma and Welling (2013). The idea is to reparameterize the multivariate Gaussian $q_\phi(\mathbf{z}|\mathbf{x})$ using a noise distribution $p(\epsilon) = \mathcal{N}(0, \mathbf{I})$ and let $\mathbf{z} = g_\phi(\epsilon, \mathbf{x}) = \mu_\phi(x) + \epsilon \odot \sigma_\phi(x)$ which results in $\mathbf{z}$ having the desired distribution $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu_\phi(x), \sigma_\phi(x))$. Here we used $\odot$ to denote element-wise multiplication.

The lower bound $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ can be written as

$$
\begin{aligned}
\mathcal{L}_{\theta,\phi}(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]\right] + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] \\
&= -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]
\end{aligned} \tag{3.4}
$$

As both $q_\phi(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$ are Gaussian, the KL-divergence has a closed form expression. The lower bound can be approximated with a set of samples $\epsilon^{(l)}$, $l = 1, \ldots, L$ as

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) \approx -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \sum_{l=1}^{L} \log p_\theta(\mathbf{x}|\mathbf{z}^{(l)}) \tag{3.5}$$

where $\mathbf{z}^{(l)} = \mu_\phi(x) + \epsilon^{(l)} \odot \sigma_\phi(x)$ and $\epsilon^{(l)} \sim \mathcal{N}(\epsilon|0, \mathbf{I})$

From an information theory point of view, the latent variables can be seen as a code and therefore the distributions $q_\phi(\mathbf{z}|\mathbf{x})$ and $p_\theta(\mathbf{x}|\mathbf{z})$ can be seen as a probabilistic encoder and decoder, respectively. The encoder and decoder are typically implemented with deep neural networks, with architecture depending on the data structure. In the next section we will discuss how we apply the VAE framework to tumor segmentations.

3.2 Implementation of VAE for tumor shape

As we mentioned in the beginning of this chapter, we apply a generative tumor shape model in a couple of different scenarios in this thesis. In this section, we will describe how the model is implemented in the VAE framework we described in Section 3.1.

The data we aim to model are brain tumor segmentations and we have training data from N subjects, $\{\mathbf{x}^{(1)}, \ldots \mathbf{x}^{(N)}\}$ where $\mathbf{x}^{(i)} \in \{1, ..., K\}^D$ is the i-th subject's segmentation data in the form of a segmentation map with D voxels. In our case we have the segmentation of $K = 4$ different tumor classes as input to the model. One of the tumor classes represents absence of tumor while the other three are non-enhancing core, edema and enhancing core.

We implement the encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and the decoder $p_\theta(\mathbf{x}|\mathbf{z})$ as deep convolutional neural networks (CNNs) using Tensorflow (Abadi et al., 2015). An overview of the encoder and decoder network architectures is given in Table 3.1 and Table 3.2, respectively. The encoder network consists of 3 convolutional network blocks, followed by two fully connected layers. Each block consists of 2 convolutional layers followed by a max pooling layer. The decoder network has a symmetrical architecture to the encoder, where the convolutional layers are replaced with transposed convolutional (deconvolutional) layers (Dosovitskiy et al., 2015). After each convolutional layer in both networks, a leaky

Layer	Output Shape	Number of Parameters
Input	(240, 240, 155, 4)	0
Conv3D	(120, 120, 78, 32)	3488
Conv3D	(60, 60, 39, 32)	27680
MaxPooling3D	(30, 30, 20, 32)	0
Conv3D	(15, 15, 10, 32)	27680
Conv3D	(15, 15, 10, 32)	27680
MaxPooling3D	(8, 8, 5, 32)	0
Conv3D	(8, 8, 5, 32)	27680
Conv3D	(8, 8, 5, 32)	27680
MaxPooling3D	(4, 4, 3, 32)	0
Flatten	(1536)	0
Dense	(256)	393472
Dense	(64)	16448

Table 3.1: Overview of the layers in the encoder network. The network has 551,808 parameters in total

ReLU (Nair and Hinton, 2010) activation is applied, except at the last layer of the decoder whose output is interpreted as logits that are passed through a soft-max layer to produce tumor class probabilities in each voxel. The total number of parameters in the encoder and decoder are 551,808 and 600,420, respectively. The input to the encoder is 4-dimensional in each voxel, where the 4-dimensional vector is a one-hot encoding of the tumor class.

We choose a functional form for the decoder to be voxel-wise categorical distributions

$$p_\theta(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{D} \prod_{k=1}^{K} f_{ik}(\mathbf{z})^{I[x_i=k]}$$

where $I[x_i = k]$ evaluates to 1 if the i-th voxel of $\mathbf{x}$ (x_i) has tumor class k, else 0, and $f_{ik}(\mathbf{z})$ is the output of the decoder network at voxel i for tumor class k, and $\sum_{k=1}^{K} f_{ik}(\mathbf{z}) = 1$.

To train the model, we use the training data from the BraTS2020 dataset, which is publicly available and was released for the BraTS2020 segmentation challenge (Bakas et al., 2018). The dataset consists of 369 manual segmentations of the tumor regions in scans of grade II-IV glioma patients, where all data have been co-registered to a template of size (240,240,155). 75 of the subjects were used to measure error on unseen data while 294 were used for training the model.

The VAE is trained using the ADAM optimization algorithm (Kingma and Ba,

Layer	Output Shape	Number of Parameters
Input	(32)	0
Dense	(256)	8448
Dense	(1536)	394752
Reshape	(4, 4, 3, 32)	0
Conv3DTranspose	(4, 4, 3, 32)	27680
Conv3DTranspose	(4, 4, 3, 32)	27680
Conv3DTranspose	(8, 8, 6, 32)	27680
Conv3DTranspose	(16, 16, 12, 32)	27680
Conv3DTranspose	(32, 32, 24, 32)	27680
Conv3DTranspose	(64, 64, 48, 32)	27680
Conv3DTranspose	(128, 128, 96, 32)	27680
Conv3DTranspose	(240, 240, 155, 4)	3460

Table 3.2: Overview of the layers in the decoder network. The network has 600,420 parameters in total

2014) with learning rate of 10^{-4} and exponential decay rates for the 1st and 2nd moments of 0.9 and 0.999, respectively. The number of latent variables is chosen to be 32. This choice results in the output dimension of the encoder being 64 (see Table 3.1), since the posterior over latent variables has two parameters for each latent variable. We use data augmentation where the input is reversed along a randomly chosen set of axes where all possible sets of axes have equal probability of being chosen. Training is stopped when error on the unseen data does not decrease for 10 epochs.

We use this model in Paper B, where the tumor VAE model acts as regularization for tumor shape in a segmentation model. This model is again used in paper C, where we use the encoder output as shape features of the tumor for a disease classification task. In the next section, we discuss how we apply this model in a semi-supervised setting to solve a classification task.

3.3 Semi-supervised VAE for survival prediction

In this section, we describe the method we developed in paper A for survival prediction of glioblastoma patients.

In recent years there has been an increased interest in survival prediction of brain tumors based on MR images, mostly using discriminative models that directly encode the relationship between image data and prediction labels (Bakas

et al., 2018). As we discussed in Section 2.3, the flexibility of MR imaging makes it difficult for these models to generalize across scanners and clinical centers, limiting their potential applicability in clinical settings. In paper A, we explore whether these issues with supervised intensity-based methods can be ameliorated by using a semi-supervised approach instead, using only segmentation masks as input. In particular, we adapt a semi-supervised variational autoencoder model (Kingma et al., 2014) to predict overall survival from a small amount of labeled training subjects, augmented with *unlabeled* subjects in which only imaging data is available. Because the model only takes segmentation masks as input, all assumptions on the image modalities and scanners used are removed.

Although survival prediction is usually formulated as a regression problem, in paper A we divided the subjects into three categories and formulate the problem as a classification task. We aim to classify patients into three prognosis groups: **long-survivors** (>15 months), **short-survivors** (<10 months), and **mid-survivors** (between 10 and 15 months), all relative to the time of diagnosis.

3.3.1 Model

In the previous section, we already described the variational autoencoder to model the data distribution. We now show how we simultaneously model the input data and the target class variable.

The available training data $\{(\boldsymbol{x}^{(1)}, y^{(1)}), ..., (\boldsymbol{x}^{(N_l)}, y^{(N_l)})\}$ consists of a set of N_l labeled pairs, possibly augmented with a set of N_u *unlabeled* data points $\{\boldsymbol{x}^{(N_l+1)}, ..., \boldsymbol{x}^{(N_l+N_u)}\}$, where $\boldsymbol{x}^{(i)} \in \{1, ..., M_x\}^D$ is the i-th subject's image data in the form of a segmentation map with D voxels, and the target variable $y^{(i)} \in \{1, ..., M_y\}$ denotes the survival group the subject belongs to. In our case we have the segmentation of $M_x = 4$ different tumor structures as input to the model, and $M_y = 3$ different survival groups. For convenience, we will omit the index i when possible in the remainder.

We assume that the data is generated by a random process, illustrated in Figure 3.1, that involves latent variables $\boldsymbol{z} \in \mathcal{R}^L$, assumed to be independent of y, where $L \ll D$. As in the previous section, the latent variables' purpose is to encode high-level tumor shape and location features. Specifically, we assume a generative model of the form

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|y, \boldsymbol{z})p(\boldsymbol{z})p(y), \tag{3.6}$$

where $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \mathbf{I})$ is a zero-mean isotropic multivariate Gaussian, $p(y) \propto 1$

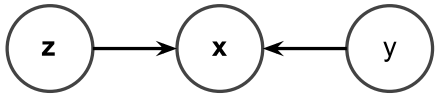

Fig. 3.1: Probabilistic graphical model of the generative process.

is a flat categorical prior distribution over y, and $p_{\boldsymbol{\theta}}(\boldsymbol{x}|y,\boldsymbol{z})$ is a conditional distribution parameterized by $\boldsymbol{\theta}$.

The task is to find the maximum likelihood parameters, i.e., the parameter values $\boldsymbol{\theta}$ that maximize the probability of the training data under the model. This is equivalent to maximizing

$$\sum_{i=1}^{N_l} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}, y^{(i)}) + \sum_{i=N_l+1}^{N_l+N_u} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) \tag{3.7}$$

with respect to $\boldsymbol{\theta}$, where

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = \int_{\boldsymbol{z}} p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z}) d\boldsymbol{z} \tag{3.8}$$

and

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{y} p_{\boldsymbol{\theta}}(\boldsymbol{x}, y). \tag{3.9}$$

Once suitable parameter values are found, the survival group of a new subject with image data $\boldsymbol{x}$ can be predicted by assessing $p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}, y)/p_{\boldsymbol{\theta}}(\boldsymbol{x})$.

Semi-supervised variational autoencoder

Maximizing Eq. (3.7) for $\boldsymbol{\theta}$ directly is not feasible due to intractability of the integral over the latent variables in Eq. (3.8). We take the same approach to solving this problem as we described for the VAE model in Section 3.1. We use the EM algorithm where we iteratively construct and maximize a lower bound to Eq. (3.7). We approximate $p_{\boldsymbol{\theta}}(\boldsymbol{z}, y|\boldsymbol{x})$ with $q_{\boldsymbol{\phi}}(\boldsymbol{z}, y|\boldsymbol{x})$ with parameters $\boldsymbol{\phi}$, which factorizes as:

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}, y|\boldsymbol{x}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y) q_{\boldsymbol{\phi}}(y|\boldsymbol{x}),$$

where $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)$ is a multivariate Gaussian distribution with diagonal covariance matrix, and $q_{\boldsymbol{\phi}}(y|\boldsymbol{x})$ is a categorical distribution. This approximation can be used to obtain a lower bound to Eq. (3.7) as follows. The probability of each

labeled data point (first term in Eq. (3.7)) can be rewritten as:

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, y)] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log\left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, y)}\right]\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log\left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)}\frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, y)}\right]\right] \\
&= \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log\left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)}\right]\right]}_{=\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x},y)} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log\left[\frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, y)}\right]\right]}_{=D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x},y))}
\end{aligned}$$

where D_{KL} denotes the Kullback-Leibler (KL) divergence. Since the KL divergence is always non-negative, we have that

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) \geq \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}, y). \tag{3.10}$$

Using a similar derivation, the probability of each *unlabeled* data point can be bounded as follows:

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) &\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(y,\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|y, \boldsymbol{x})} - \log q_{\boldsymbol{\phi}}(y|\boldsymbol{x})\right] \\
&= \sum_y q_{\boldsymbol{\phi}}(y|\boldsymbol{x})(\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}, y)) + \mathcal{H}(q_{\boldsymbol{\phi}}(y|\boldsymbol{x})) = \mathcal{U}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}),
\end{aligned} \tag{3.11}$$

where $\mathcal{H}(\cdot)$ denotes the entropy of a probability distribution.

By combining (3.10) and (3.11), a lower bound to Eq. (3.7) is finally obtained as:

$$\mathcal{J}_{\boldsymbol{\theta},\boldsymbol{\phi}} = \sum_{i=1}^{N_l} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}_i, y_i) + \sum_{i=N_l+1}^{N_l+N_u} \mathcal{U}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}_i), \tag{3.12}$$

which we optimize with respect to both the variational parameters $\boldsymbol{\phi}$ and the generative parameters $\boldsymbol{\theta}$. We use stochastic gradient ascent for the optimization, approximating gradients of the expectations in (3.12) as described in Kingma and Welling (2013).

The label predictive distribution $q_{\boldsymbol{\phi}}(y|\boldsymbol{x})$ has the form of a discriminative *classifier*, and can be used as an approximation to $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ for classifying new cases after training.

In paper A, we describe a few model modifications that make the parameter learning process faster and less prone to overfitting. The modifications are

- Adding a weak classification loss to the objective function, as in Kingma et al. (2014), to let the label predictive distribution $q_{\boldsymbol{\phi}}(y|\boldsymbol{x})$ also learn from labeled data.

- Using gumbel-softmax (Jang et al., 2016; Maddison et al., 2016), to avoid marginalization over $q_{\boldsymbol{\phi}}(y|\boldsymbol{x})$ in Eq. (3.11), making the training computationally more efficient.

- Controlling the trade of between accurate reconstruction and constraint of the latent space by scaling the KL (regularization) term as proposed in Higgins et al. (2016). Similarly scaling the entropy of the label predictive distribution in Eq. (3.11) helps to prevent overfitting of the classifier.

3.3.2 Data

The BraTS2019 training dataset consists of scans of 335 subjects for which manual tumor delineation is provided, out of which 210 have known survival times. The remaining 125 subjects were used as part of our unlabelled training data. The challenge used an online evaluation platform where survival predictions for 29 subjects were compared to ground-truth. Note that while the labelled data consist mostly of glioblastoma subjects, the unlabelled data contains more lower grade gliomas, which is not ideal for our set-up.

In all our experiments we performed 3-fold cross-validation by randomly splitting the BraTS 2019 training set with survival labels into a training (75%) and validation set (25%) in each fold, in order to have an alternative to the online evaluation platform, which only validates on 29 subjects. With this set-up, which we call **S0** in the remainder, we effectively trained the model on a training set of $\mathbf{N_l = 157}$ and $\mathbf{N_u = 125}$ for each of the three cross-validation folds. These models were subsequently tested on their corresponding validation sets of 53 subjects, as well as on the standard BraTS 2019 validation set of 29 subjects.

The reason we took the semi-supervised approach was to allow the method learn from all the available unlabeled data, which usually needs to be substantially more than the labelled data. We therefore attempted to generate more unlabelled data by using three open-source methods (Wang et al., 2017; Nuechterlein and Mehta, 2018; Isensee et al., 2017) to automatically segment both the entire BraTS 2019 training and validation sets in order to have many more unlabeled training subjects available. We further augmented these unlabeled data by flipping the images in the coronal plane. With this new set-up, which we call **S1**, we then trained the model on an "augmented" set of $\mathbf{N_l = 157}$ and $\mathbf{N_u = 2268}$ for each of the three cross-validation folds. Ideally, dramatically increasing the set

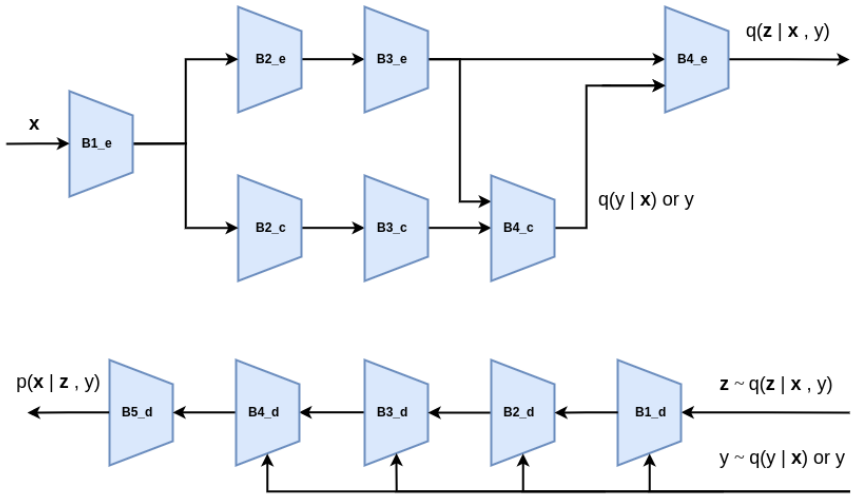

Fig. 3.2: Networks architectures in papar A: encoder, decoder and classifier architectures.

of unlabeled data points this way should help the model learn to better encode tumor representations, thereby increasing classification accuracy.

3.3.3 Implementation

We implemented the encoder $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)$, the decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, y)$ and the classifier $q_{\boldsymbol{\phi}}(y|\boldsymbol{x})$ all as deep convolutional networks using PyTorch (Paszke et al., 2017). The segmentation volumes provided in the BraTS challenge have size 240×240×155, but since large parts of the volume are always zero, we cropped the volumes to 146×188×128 without losing any tumor voxels. We further downsampled the volume by a factor of 2 in all dimensions, resulting in a shape of $73 \times 94 \times 64$, roughly a 95% overall reduction in input image size. This leads to much faster training and larger batches fitting in memory, while losing minimal information.

We refer the reader to paper A and our code repository for a detailed description of the implementation, parameter settings and the network architecture, which we only briefly describe here. The code is available at `https://github.com/sveinnpalsson/semivaebrats`. The three networks consist of 3D convolutional layers, with the exception of a few fully connected layers in the classifier. There are nonlinearities (Scaled Exponential Linear Units, (Klambauer et al., 2017)) and dropout (Srivastava et al., 2014) after each layer, except when noted. What

follows is a high-level description of the network architecture, represented in diagrams in Figure 3.2.

The inference network consists of a convolutional layer (B1_e) with large kernel size and stride (7 and 4, respectively), followed by two residual blocks (He et al., 2016) (B2_e and B3_e). The input to each block is processed in parallel in two branches, one consisting of two convolutional layers, the other of average pooling followed by a linear transformation (without nonlinearities). The results of the two branches are added together. The output of the first layer is also fed into the classifier network, which outputs the class scores (these will be used to compute the classification loss for labeled data). A categorical sample from $q_\phi(y|\mathbf{x})$ is drawn using the Gumbel-Softmax reparameterization given the class scores, and is embedded by a fully connected layer into a real vector space. Such embedding is then concatenated to the output of the two encoder blocks, so that the means and variances of the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x}, y)$, that are computed by a final convolutional layer, are conditioned on the sampled label. The classifier consists of two residual blocks similar to the ones in the encoder (B2_c and B3_c), followed by two fully connected layers (B4_c).

The decoder network consists of two convolutional layers (B1_d and B2_d), two residual blocks similar to those in the encoder (B3_d and B4_d), and a final convolution followed by a sigmoid nonlinearity (B5_d). In the decoder, most convolutions are replaced by transposed convolutions (for upsampling), and pooling in the residual connections is replaced by nearest neighbour interpolation. The input to the decoder network is a latent vector $\mathbf{z}$ sampled from the approximate posterior. The embedding of y, computed as in the final stage of the inference network, is also concatenated to the input of each layer (except the ones in the middle of a block) to express the conditioning of the likelihood function on the label. Here, the label is either the ground truth (for labeled examples) or a sample from the inferred posterior (for unlabeled examples).

3.3.4 Results

Conditional generation

We visually tested whether the decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x}|y, \boldsymbol{z})$ is able to generate tumor-like images after training, and whether it can disentangle the classes. For this purpose we sampled $\boldsymbol{z}$ from $\mathcal{N}(\boldsymbol{z}|\mathbf{0}, \mathbf{I})$ and varied y between the three classes, namely, short survivor, mid survivor and long survivor. Figure 3.3 shows the three shapes generated accordingly by one of the models trained in set-up **S0**. From the images we can see that the generated tumor for the short survivor

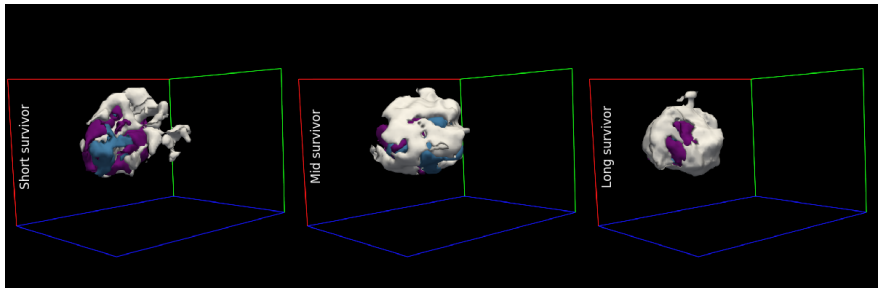

Fig. 3.3: Generated tumor from $p_\theta(\mathbf{x}|y,\mathbf{z})$ where we sampled $\mathbf{z}$ from $\mathcal{N}(\mathbf{z}|\mathbf{0},\mathbf{I})$ and we varied y between short survivor, mid survivor and long survivor. The figure is borrowed from paper A.

Table 3.3: Classification accuracies [%] for both set-ups on the validation set for each of the three cross-validation folds.

Set-up	Fold 1	Fold 2	Fold 3	Avg
S0	42.18 ± 13.30	35.90 ± 12.91	39.53 ± 13.16	39.20 ± 7.59
S1	47.55 ± 13.45	41.13 ± 13.40	42.91 ± 13.32	43.86 ± 7.71

class has an irregular shape with jagged edges while the long survivor generated tumor has a more compact shape with rounded edges. This is in line with findings in related work (Pérez Beteta et al., 2018).

Quantitative evaluation

The classification accuracy is reported here both on the validation set within each fold of cross-validation (Table 3.3) and on the validation data (29 subjects) on the online platform (Table 3.4). All the classification accuracies are reported with binomial confidence interval with normal approximation (Brown et al., 2001).

Table 3.4: Classification accuracies [%] for both set-ups on the BraTS 2019 online evaluation platform.

Set-up	Majority voting
S0	37.90 ± 17.57
S1	31.00 ± 16.83

The results show that in none of the experiments our model achieved a significant improvement over always predicting the largest class, which constitutes around 40% of the labeled cases.

3.3.5 Discussion and conclusions

We described the theory and experiments we did in paper A, where we implemented and evaluated the potential of a semi-supervised deep generative model for classifying brain tumor patients into three overall survival groups, based only on tumor segmentation masks. The main potential advantages of this approach are (1) its in-built invariance to MR intensity variations when different scanners and protocols are used, enabling wide applicability across clinics; and (2) its ability to learn from unlabeled data, which is much more widely available than fully-labeled data.

We compared two different set-ups: one where fewer unlabeled subjects were available for training, and one where their number was (largely artificially) increased using automatic segmentation and data augmentation. Although the latter set-up increased classification performance in our experiments, this increase did not reach statistically significant levels and was not replicated on the small BraTS 2019 validation set. We demonstrated visually that the proposed model effectively learned class-specific information, but overall failed to achieve classification accuracy significantly higher than predicting always the largest class.

Although irregular tumor shape has been previously shown to be an indicator of poor survival (Pérez Beteta et al., 2018), its effect may not be strong enough for reliable subject-specific classification. Most previous work on glioblastoma survival prediction from MR images has been done with textural features, such as radiomics. Since tumor texture may predict tumor sub-type, it can be a better indicator of overall survival than shape alone. Only considering tumor shape is potentially the main drawback of our approach. Therefore, future work could involve implementing a generative model of the tumor intensities, allowing the method to stay agnostic to specific sequences and scanners used while still taking intensity information into account.

Although we attempted to solve the lack-of-data problem by taking the semi-supervised approach, we essentially still run into it because the unlabelled data was somewhat artificial. In the **S0** setup, we used the part of the dataset that didn't have recorded survival times. However, these subjects are of different tumor type, actually mostly belonging to grade II gliomas. Future work may find better results augmenting the dataset with a large set of unlabelled glioblastoma

subjects and, of course, using more data.

The code of our method is publicly available. Although we didn't see the improvement of using more unlabelled data, we are confident that it's not due to the implementation but to the lack of correlation between tumor shape and survival. We tested the code in more controlled settings, for example by generating a 3D version of the MNIST where we could clearly see benefit of increased number of unlabelled data and high accuracy achieved with only a few labelled samples.

Chapter 4

SAMSEG-Tumor: Automatic whole-brain and tumor segmentation

Sequence Adaptive Multimodal SEGmentation (SAMSEG) (Puonti et al., 2016) is a segmentation method that segments dozens of neuroanatomical structures using a generative modeling approach. In this chapter we describe SAMSEG-Tumor, the whole-brain and tumor segmentation method used in paper B. The chapter is structured as follows:

- We begin by describing SAMSEG, the contrast adaptive whole-brain segmentation the method is based on.
- We discuss the previous work on extending SAMSEG to model pathology.
- Next, we describe how SAMSEG-Tumor combines SAMSEG with the tumor shape model discussed in Chapter 3.
- Finally, we discuss the advantages and limitations of SAMSEG-Tumor and future work.

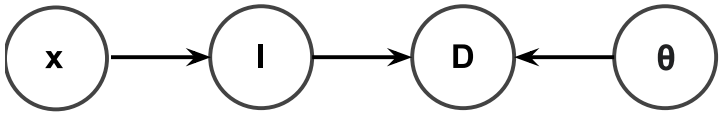

Fig. 4.1: Graphical representation of SAMSEG (Puonti et al., 2016). The observed data $\mathbf{D}$ is assumed to depend on the likelihood parameters $\boldsymbol{\theta}$ and the segmentation labels $\mathbf{l}$, which depend on the deformable atlas configuration variables $\mathbf{x}$.

4.1 SAMSEG: Modality adaptive whole-brain segmentation

SAMSEG is a method that segments multi-contrast MR images of the brain using a generative image model and spatial prior encoded in a neuroanatomical atlas. This approach allows for minimal assumptions made on the scanning platform and pulse sequences used for image acquisition, making it robust and generally applicable to MR data from different centers. The following is a technical description of the modeling approach.

Given N MR scans of the same subject, let $\mathbf{D} = (\mathbf{d}_1, \cdots, \mathbf{d}_I)$ denote MR image intensities where $\mathbf{d}_i \in \mathcal{R}^N$ denotes the intensity values at the i-th voxel and I is the number of voxels. We aim to infer a segmentation $\mathbf{l} = (\mathbf{l}_1, \cdots, \mathbf{l}_I)$, where $\mathbf{l}_i \in \{1, \cdots K\}$ denotes one of K neuroanatomical structures.

The generative model (Fig. 4.1) comprises two parts: a segmentation prior $p(\mathbf{l})$ and a likelihood function $p(\mathbf{D}|\mathbf{l})$. The following is a description of these two parts and how the model is "inverted" to obtain a segmentation.

4.1.1 Segmentation prior

The segmentation prior encodes spatial information about the labels $\mathbf{l}$. The prior is defined as a probabilistic atlas, implemented as a deformable tetrahedral mesh (Van Leemput, 2009; Puonti et al., 2016). Let $\mathbf{x}$ denote the node positions of the mesh with prior distribution

$$p(\mathbf{x}) \propto \exp\left[-\kappa \sum_{m=1}^{M} U_m(\mathbf{x}, \mathbf{x}_{ref})\right].$$

where κ controls the stiffness of the mesh and U_m is the cost contribution of the m-th tetrahedron as the mesh is deformed from its reference position $\mathbf{x}_{ref}$ (Ashburner et al., 2000).

Given a node position $\mathbf{x}$, an assumption of conditional independence between labels across voxels is made, allowing the prior to factorize:

$$p(\mathbf{l}|\mathbf{x}) = \prod_{i=1}^{I} p(l_i|\mathbf{x}).$$

For a particular label k, its probability at voxel i is the defined as:

$$p(l_i = k|\mathbf{x}) = \sum_{m=1}^{M} \alpha_m^k \psi_m^i(\mathbf{x}),$$

where α_m^k denotes the probability of observing label k at vertex m, and $\psi_m^i(\mathbf{x})$ is an interpolation function attached to the m^{th} vertex and evaluated at the i^{th} voxel.

Finally, the full segmentation prior is defined as:

$$p(\mathbf{l}) = \int_{\mathbf{x}} p(\mathbf{l}|\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

In Van Leemput (2009), an algorithm is described for learning the topology of the mesh from training data.

The atlas used in SAMSEG was learned from 20 manual segmentations randomly chosen from a pool of 28 healthy subjects and 11 subjects with questionable or probable Alzheimer's disease with ages ranging from under 30 years old to over 60 years old (Puonti et al., 2016).

4.1.2 Likelihood function

In SAMSEG, a multivariate normal (MVN) distribution is associated with each label. An assumption is made that the bias field imaging artifact can be modeled as a multiplicative and spatially smooth effect. However, modelling the log-transformed image intensities is preferred for computational reasons and therefore bias field is considered an additive effect in the model (Wells et al.,

1996; Van Leemput et al., 1999). To model the bias field, a linear combination of spatially smooth basis functions is used.

Let $\boldsymbol{\theta}$ denote the set of parameters of the likelihood function; the means and variances of the MVNs and the bias field parameters. The likelihood function is defined as:

$$p(\mathbf{D}|\mathbf{l},\boldsymbol{\theta}) = \prod_{i=1}^{I} p(\mathbf{d}_i|l_i,\boldsymbol{\theta}),$$

$$p(\mathbf{d}_i|l_i=k,\boldsymbol{\theta}) = \mathcal{N}(\mathbf{d}_i|\boldsymbol{\mu}_k + \mathbf{C}\boldsymbol{\phi}_i, \boldsymbol{\Sigma}_k),$$

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_N^T \end{pmatrix}, \quad \mathbf{c}_n = \begin{pmatrix} c_{n,1} \\ \vdots \\ c_{n,P} \end{pmatrix}, \quad \boldsymbol{\phi}_i = \begin{pmatrix} \phi_1^i \\ \vdots \\ \phi_P^i \end{pmatrix}.$$

where P is the number of basis functions of the bias field model, ϕ_p^i is the p-th basis function evaluated at voxel i, and $\mathbf{c}_n$ are the bias field coefficients of the n^{th} MRI contrast.

To regularize the parameters of the covariance matrices, their prior distribution is chosen to be a inverse-Wishart distribution:

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} IW(\boldsymbol{\Sigma}_k|\gamma I, \gamma - N - 1),$$

where γ is a hyperparameter, chosen to be very small.

Finally, the likelihood function is defined as:

$$p(\mathbf{D}|\mathbf{l}) = \int_{\boldsymbol{\theta}} p(\mathbf{D}|\mathbf{l},\boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

4.1.3 Segmentation

Now that we have defined the segmentation prior and likelihood function, the posterior distribution is defined as:

$$p(\mathbf{l}|\mathbf{D}) \propto \underbrace{\int_{\mathbf{x}} p(\mathbf{l}|\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x}}_{\text{Segmentation prior}} \underbrace{\int_{\boldsymbol{\theta}} p(\mathbf{D}|\mathbf{l},\boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}_{\text{Likelihood function}}$$

The segmentation posterior is intractable, requiring some simplifications to be made. Towards this end, the posterior distribution can be written as:

$$p(\mathbf{l}|\mathbf{D}) = \frac{\int_{\mathbf{x}} \int_{\boldsymbol{\theta}} p(\mathbf{D}, \mathbf{l}, \mathbf{x}, \boldsymbol{\theta})}{p(\mathbf{D})} = \int_{\mathbf{x}} \int_{\boldsymbol{\theta}} p(\mathbf{l}, \mathbf{x}, \boldsymbol{\theta}|\mathbf{D}) = \int_{\mathbf{x}} \int_{\boldsymbol{\theta}} p(\mathbf{l}|\mathbf{D}, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{D}),$$

and the following approximation is made:

$$p(\mathbf{l}|\mathbf{D}) \approx p(\mathbf{l}|\mathbf{D}, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}), \tag{4.1}$$

where $\hat{\boldsymbol{\theta}}, \hat{\mathbf{x}}$ are point estimates of the model parameters, obtained by solving the following optimization problem:

$$\{\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\} = \arg\max_{\mathbf{x}, \boldsymbol{\theta}} p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{D}). \tag{4.2}$$

Therefore, to obtain segmentations, we first solve the optimization problem of Eq. (4.2), and then maximize Eq. (4.1).

What follows is a brief description of these two steps, further detailed in Puonti et al. (2016).

Computing point estimates

A generalized EM (GEM) algorithm, derived in Van Leemput et al. (1999), is used for obtaining point estimates of the model parameters. In an iterative scheme: first, the position of the mesh nodes $\mathbf{x}$ is estimated using the L-BFGS algorithm followed by an EM algorithm to estimate $\boldsymbol{\theta}$. Estimating $\boldsymbol{\theta}$ is achieved by iteratively constructing a lower bound to the objective function by computing the soft label assignments

$$w_{i,k} = \frac{\mathcal{N}(\mathbf{d}_i|\boldsymbol{\mu}_k + \mathbf{C}\boldsymbol{\phi}_i, \boldsymbol{\Sigma}_k) p(l_i = k|\mathbf{x})}{\sum_{k'=1}^{K} \mathcal{N}(\mathbf{d}_i|\boldsymbol{\mu}_{k'} + \mathbf{C}\boldsymbol{\phi}_i, \boldsymbol{\Sigma}_{k'}) p(l_i = k'|\mathbf{x})}, \quad 0 \leq w_{i,k} \leq 1 \tag{4.3}$$

and improving the lower bound by updating the likelihood parameters $\boldsymbol{\theta}$ given the current soft assignments $w_{i,k}$. Once this process of estimating $\boldsymbol{\theta}$ converges, the mesh nodes are estimated again with the current estimates of $\boldsymbol{\theta}$ and so forth. Once this process converges, the current estimates of the model parameters are taken as the desired point estimates $\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}$.

Computing final segmentations

Having obtained the point estimates of model parameters, the final segmentation is obtained by

$$\hat{\mathbf{l}} = \arg\max_{\mathbf{l}} p(\mathbf{l}|\mathbf{D}, \hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}),$$

where each voxel is assigned to the label with the highest probability:

$$\hat{l}_i = \arg\max_{k} \hat{w}_{i,k},$$

where $\hat{w}_{i,k}$ (Eq. (4.3)) is evaluated at the estimated parameters $\{\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\}$.

4.2 SAMSEG-Tumor

SAMSEG-Tumor extends the SAMSEG model (cf. Fig 4.1) to include tumors. The main goal of the SAMSEG-Tumor project is to bring the work of Agn et al. (2019) into a publicly and easily accessible state, by implementing it in python and integrating it into the open-source FreeSurfer toolbox. To include brain tumors in the model, we introduce a tumor vector $\mathbf{z} = (z_1, ..., z_I)$, denoting the tumor segmentation, where $z_i \in \{0, 1, ..., K_T\}$, is the assignment of voxel i to one of K_T different tumor classes ($z_i = 0$ when voxel i doesn't contain tumor). We use $K_T = 3$ tumor classes, representing edema, contrast-enhancing core and non-enhancing core. Furthermore, we introduce new model parameters $\mathbf{h}$ and $\boldsymbol{\theta}_z$ to model the shape and appearance of tumor, respectively. The goal of SAMSEG-Tumor is to compute the joint segmentation labels $\{\mathbf{l}, \mathbf{z}\}$ given the data $\mathbf{D}$.

For the purpose of segmenting scans with brain tumors, the SAMSEG model is combined with a spatial regularization model of tumor shape using generative neural networks (Agn et al., 2019). Although in its original formulation, Agn et al. (2019) used restricted Boltzmann machines (Lee et al., 2011) for this purpose, our current implementation has variational autoencoders Kingma and Welling (2013) (described in Section 3.1) since these have a deeper structure and can therefore better represent lesion shape (Cerri et al., 2021). Fig. 4.2 shows the graphical model of SAMSEG-Tumor.

What follows is a description of the segmentation prior and likelihood function of the new extended model; a description of how the segmentation is computed; a description of the implementation; discussion of the performance on benchmark data; and finally discussion about the model and future work (Section 4.3).

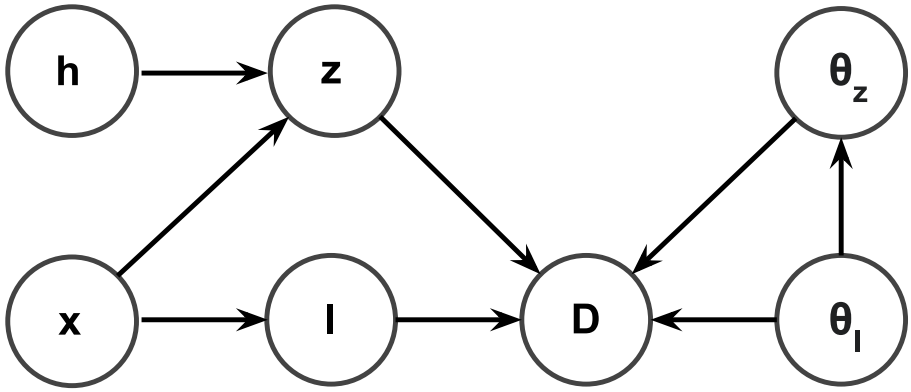

Fig. 4.2: Graphical representation of SAMSEG-Tumor. The additional variables shown here compared to the original SAMSEG model (cf. Fig. 4.1), are the tumor intensity model parameters $\boldsymbol{\theta}_\mathbf{z}$, the tumor labels $\mathbf{z}$ and the tumor shape variables $\mathbf{h}$. $\boldsymbol{\theta}_\mathbf{l}$ are the parameters of the intensity model for healthy brain tissue.

4.2.1 Segmentation prior

The joint segmentation prior over $\{\mathbf{l}, \mathbf{z}\}$, has the following form:

$$p(\mathbf{l}, \mathbf{z}|\mathbf{h}, \mathbf{x}) = p(\mathbf{l}|\mathbf{x})p(\mathbf{z}|\mathbf{h}, \mathbf{x}).$$

Here, $p(\mathbf{l}|\mathbf{x})$ is the probabilistic atlas model described in Section 4.1.1, and $p(\mathbf{z}|\mathbf{h}, \mathbf{x})$ is modeled with a variational autoencoder.

4.2.2 Likelihood function

For the likelihood function, we model the image intensity $\mathbf{D}_i$ as being drawn from a MVN distribution associated with the segmentation $\{l_i, z_i\}$ at voxel i. Let $\boldsymbol{\theta}_\mathbf{l} = \{\boldsymbol{\mu}_\mathbf{l}^{(k)}, \boldsymbol{\Sigma}_\mathbf{l}^{(k)}\}$ denote the mean and variance parameters of the MVN for the anatomical structures and $\boldsymbol{\theta}_\mathbf{z} = \{\boldsymbol{\mu}_\mathbf{z}^{(k)}, \boldsymbol{\Sigma}_\mathbf{z}^{(k)}\}$ for the tumor classes. The likelihood of the parameters is defined as $p(\boldsymbol{D}|\mathbf{l}, \mathbf{z}, \boldsymbol{\theta}_\mathbf{l}, \boldsymbol{\theta}_\mathbf{z}) = \prod_{i=1}^{I} p(\mathbf{d}_i|l_i, z_i, \boldsymbol{\theta}_\mathbf{l}, \boldsymbol{\theta}_\mathbf{z})$, where

$$p(\mathbf{d}_i|l_i, z_i, \boldsymbol{\theta}_\mathbf{l}, \boldsymbol{\theta}_\mathbf{z}) = \begin{cases} \mathcal{N}\big(\mathbf{d}_i|\boldsymbol{\mu}_\mathbf{z}^{(z_i)} + \boldsymbol{C}^T\boldsymbol{\phi}_i, \boldsymbol{\Sigma}_\mathbf{z}^{(z_i)}\big) & \text{if } z_i \neq 0 \\ \mathcal{N}\big(\mathbf{d}_i|\boldsymbol{\mu}_\mathbf{l}^{(l_i)} + \boldsymbol{C}^T\boldsymbol{\phi}_i, \boldsymbol{\Sigma}_\mathbf{l}^{(l_i)}\big) & \text{else} \end{cases} \tag{4.4}$$

Constraints are imposed on $\boldsymbol{\theta}_\mathbf{z}$ that limit the range of the tumor means relative to means of white matter and gray matter. These constraints and their effect on the optimization are described in (Agn et al., 2019), but we add yet another constraint on the mean value of the enhancing core in T1w-c, limiting it to higher values than the mean of edema voxels.

4.2.3 Segmentation

To compute the posterior segmentation we follow the same procedure as we described for the SAMSEG model in Section 4.1.3: First, we compute point estimates $\{\hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\boldsymbol{\theta}}_\mathbf{z}, \hat{\mathbf{x}}\}$ of the model parameters and then we estimate the final segmentation $p(\mathbf{l}, \mathbf{z}|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\boldsymbol{\theta}}_\mathbf{z}, \hat{\mathbf{x}})$, as in Agn et al. (2019) and Cerri et al. (2021).

Computing point estimates

To find point estimates of the model parameters, we use a simplified model that doesn't contain the VAE shape variables $\mathbf{h}$. However, instead of completely ignoring the lesion shape during this step as is described in Agn et al. (2019) and Cerri et al. (2021), we include spatial regularization of tumor shape in the form of Markov random field (MRF). The MRF penalizes configurations of the segmentation where tumor voxels are surrounded by many non-tumor voxels, encouraging tumor voxels to be spatially clustered. The strength of the MRF regularization is a hyperparameter that we fixed to a value that we found empirically. The optimization problem we aim to solve in this step is:

$$\{\hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\boldsymbol{\theta}}_\mathbf{z}, \hat{\mathbf{x}}\} = \underset{\{\boldsymbol{\theta}_\mathbf{l}, \boldsymbol{\theta}_\mathbf{z}, \mathbf{x}\}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}_\mathbf{l}, \boldsymbol{\theta}_\mathbf{z}, \mathbf{x}|\mathbf{D}). \tag{4.5}$$

This optimization problem is similar to the one of Eq. (4.2), where only few alterations of the GEM algorithm are needed to estimate the parameters.

Computing final segmentation

After we obtain point estimates $\{\hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}}\}$, we infer the final segmentation using the following factorization:

$$p(\mathbf{l}, \mathbf{z}|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}}) = p(\mathbf{z}|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}}) p(\mathbf{l}|\mathbf{z}, \mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}})$$

This suggests a two-step procedure where we first estimate $\mathbf{z}$ from $p(\mathbf{z}|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}})$ (see step 1 below), and then use that estimate to obtain $\mathbf{l}$ from $p(\mathbf{l}|\mathbf{z}, \mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}})$ (step 2):

Step 1: Evaluating $p(\mathbf{z}|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}})$ involves marginalizing over both $\mathbf{h}$ and $\boldsymbol{\theta}_\mathbf{z}$, which we approximate by drawing S Monte Carlo samples $\{\mathbf{h}^{(s)}, \boldsymbol{\theta}_\mathbf{z}^{(s)}\}_{s=1}^{S}$ from $p(\mathbf{h}, \boldsymbol{\theta}_z|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}})$:

$$\begin{aligned} p(\mathbf{z}|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}}) &= \int_{\mathbf{h}, \boldsymbol{\theta}_z} p(\mathbf{z}|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}}, \mathbf{h}, \boldsymbol{\theta}_\mathbf{z}) p(\mathbf{h}, \boldsymbol{\theta}_\mathbf{z}|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}}) \, \mathrm{d}\mathbf{h}, \boldsymbol{\theta}_\mathbf{z} \\ &\simeq \frac{1}{S} \sum_{s=1}^{S} p(\mathbf{z}|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}}, \mathbf{h}^{(s)}, \boldsymbol{\theta}_\mathbf{z}^{(s)}). \end{aligned} \tag{4.6}$$

From the samples of the tumor posterior, we obtain a "hard" segmentation $\hat{\mathbf{z}}$ by assigning the i-th voxel to the most likely tumor label:

$$\hat{z}_i = \underset{k \in \{0, \ldots, K_T\}}{\operatorname{argmax}} \; p(z_i = k|\mathbf{D}, \hat{\boldsymbol{\theta}}_\mathbf{l}, \hat{\mathbf{x}})$$

Step 2: Voxels that are not assigned to tumor in the previous step (i.e., $\hat{z}_i = 0$) are then assigned to the most likely healthy tissue label according to the conditional posterior probability $p(l_i = k|\hat{z}_i = 0, \mathbf{d}_i, \hat{\boldsymbol{\theta}}, \hat{\mathbf{x}})$, which simply involves computing

$$\hat{l}_i = \arg\max_k \hat{w}_{i,k},$$

where $\hat{w}_{i,k}$ is defined as in Eq. (4.3).

4.2.4 Implementation details

Our implementation is based on the existing code of SAMSEG, which is written in python and C++ and is part of the FreeSurfer (Fischl, 2012) neuroanatomical toolbox. Furthermore, our implementation is similar to the already existing SAMSEG-Lesion method (Cerri et al., 2021), also available through FreeSurfer.

The VAE that we use to regularize tumor shape is the same model we described in Section 3.2. The VAE was trained on the manual segmentations from the BraTS2020 (Bakas et al., 2018) training dataset, which consists of 369 subjects with grade II-IV gliomas. The dataset, as we mentioned in Section 3.2, has been standardized such that all subjects are co-registered to template with 1mm^3 resolution.

4.2.5 Validation

At this stage, we have only validated the tumor segmentation accuracy of our implementation on the BraTS2020 data. The validation was done by comparing the output of the automatic segmentation to the respective ground truth segmentation, provided in the public dataset. The mean and median dice scores, computed from the whole set of 369 subjects, are shown in Table 4.1 for the three tumor components and the whole tumor (the union of the three tumor components). The whole-tumor score is substantially higher than the individual components, indicating that while the method can accurately label voxels as tumor, the sub-classification into the three tumor components needs improvement.

An example of the segmentation output is shown in Fig. 5.5 (E) for a pre-operative scan that has been skull-stripped and has 4 available input modalities (Fig. 5.5 (A-D)). The tumor components are shown in blue (non enhancing core), yellow (enhancing core) and green (edema). The method easily handles variability in the inputs, such as available modalities and type pre-processing. To demonstrate that, in Fig. 5.4 (D), we show the automatic segmentation of a post-operative scan that was not skull-stripped and has 3 available modalities (Fig. 5.4 (A-C)).

Table 4.1: Dice scores, computed by comparing the output of the SAMSEG-Tumor method to the manual segmentations in the BraTS20 dataset.

	non-enhancing core	enhancing core	edema	whole tumor
Dice mean	0.38	0.54	0.46	0.80
Dice median	0.33	0.66	0.46	0.86

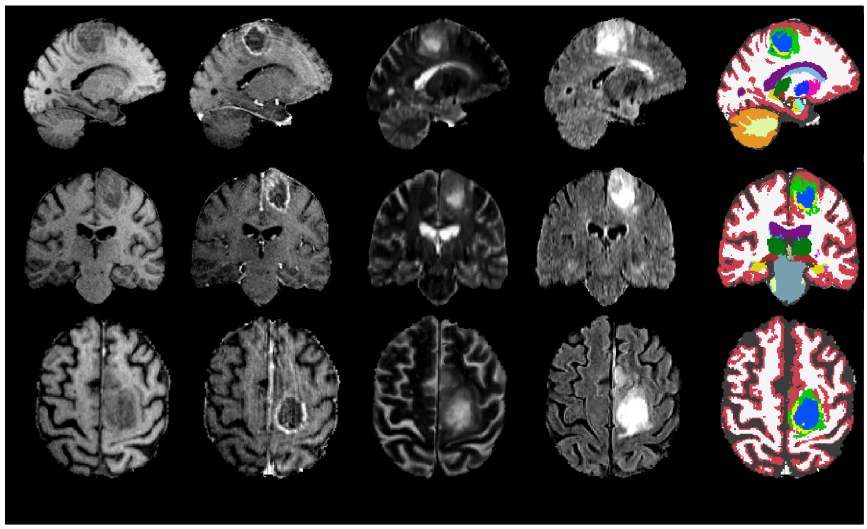

Fig. 4.3: Example of resulting segmentation using SAMSEG-Tumor. From top to bottom: sagittal, coronal and axial view. The columns show (A) T1w, (B) T1w-c, (C) T2w, (D) FLAIR, (E) Automatic segmentation output. Although the method segments right- and left sided healthy brain structures separately, for visualization we merged them into one color.

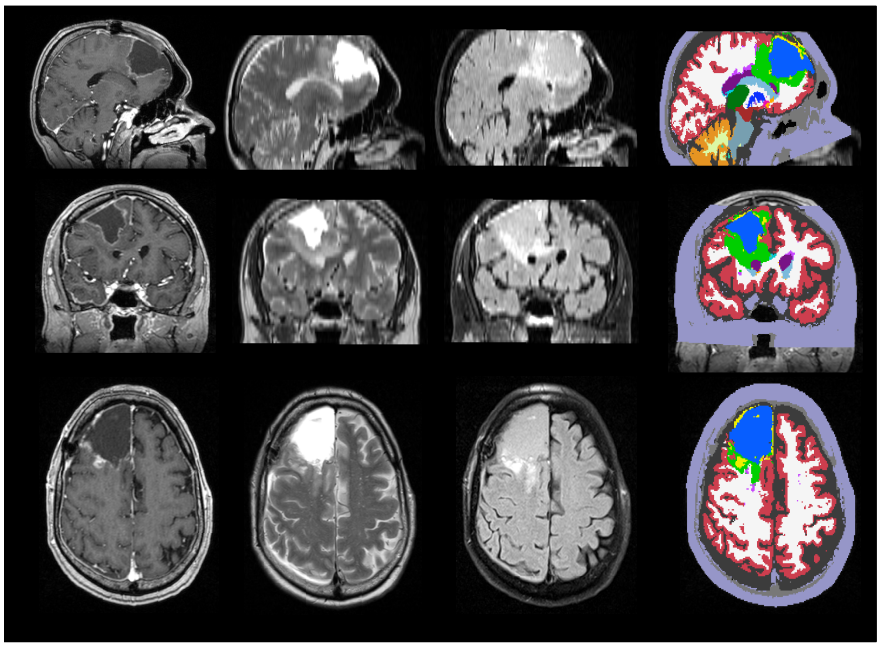

Fig. 4.4: Example of resulting segmentation using SAMSEG-Tumor. From top to bottom: sagittal, coronal and axial view. The columns show (A) T1w-c, (B) T2w, (C) FLAIR, (D) Automatic segmentation output. The tumor components are shown in blue (non enhancing core), yellow (enhancing core) and green (edema). Although the method segments right- and left sided healthy brain structures separately, for visualization we merged them into one color.

4.3 Discussion

The method we described simultaneously segments the whole-brain and tumor. The generative modeling approach allows the method to generalize across scanning platforms and the specific types of imaging sequences used. The main improvements we made on the work in Agn et al. (2019) was to simplify the model and use a variational autoencoder to regularize tumor shape instead of restricted Boltzmann machines. We have only validated the SAMSEG-Tumor method on one dataset, and only for the tumor segmentation accuracy. Further validation of segmenting both tumor and healthy structures on different

datasets, ideally both pre- and post-operative, is an avenue for future work.

In our implementation, we found it crucial to put additional constraints on the appearance of tumor with respect to the estimated white and gray matter parameters, compared to the implementation in Agn et al. (2019). Because these constraints are decided on based on training data, this poses a potential trade-off between generalizability and the benefit from these constraints.

We added a resection cavity label to handle post-operative data. The resection cavity appears in the image as a large connected area with rather uniform intensity, usually similar to that of CSF. We therefore fixed the resection cavity parameters to the CSF parameters and regularized the resection cavity shape with a MRF in the GEM part of the method.

There are several unresolved issues with the method that future work needs to address. We used a MRF to regularize tumor shape, but the strength of regularization should vary based on voxel spacing of the input image. We currently only use 1mm^3 resolution data, but to generalize to other resolutions the regularization strength needs to adapt. The method currently uses constraints on tumor appearance in T1w-c and FLAIR, but if the user is missing a FLAIR scan, the constraints should perhaps be based on T2 instead, if available. To integrate the VAE, in each sampling step, in our implementation, we currently have to transform the data into the space that the VAE was trained on and then back to the subject's space. It could be better to implement the method such that transforming between the two spaces is not required between sampling steps.

Chapter 5

Survival prediction using robust and interpretable features

Glioblastoma is a malignant primary brain tumor with very poor prognosis. As we mentioned in our discussion of brain tumors in Section 2.2, median overall survival (OS) of glioblastoma is lower than 15 months, and only about 10% of patients survive longer than 5 years, despite standard treatment being aggressive. The topic of this chapter is glioblastoma survival prediction, which has clinical applications such as guiding treatment and stratification of patients for clinical trials.

Glioblastoma is an incredibly complex and heterogeneous disease, and has many sub-types (Lauko et al., 2021). Although understanding of the disease has improved drastically over recent years, survival has hardly improved and treatment of glioblastoma follows mostly one standard course, where the tumor is resected to the highest possible extent, followed by radio and chemotherapy. Researching glioblastoma from the aspect of survival prediction can give valuable insight into the disease, and may help advance treatment methods to improve survival. In this chapter we will describe the research of paper B, where we propose robust and interpretable image features to predict overall- and progression-free survival (PFS). The chapter is organized as follows:

- We begin with an overview of imaging features, that have previously been proposed for glioblastoma survival prediction.
- We describe the proposed method of paper B, where we develop a survival prediction method with novel, interpretable imaging features.
- Finally, we discuss the advantages and limitations of our proposed methods and future work.

5.1 Biomarkers for brain tumors

This section gives a brief overview of biomarkers, or "features", that have been studied in relation to glioblastoma survival.

5.1.1 Conventional clinical features

Response to treatment and OS following standard therapy has been shown to correlate with various patient-specific features. Age, performance status (PS), expression of O^6-methylguanine-DNA methyltransferase (MGMT) are reported in many studies as prognostic features (Poulsen et al., 2017; Michaelsen et al., 2013; Hegi et al., 2005; Gorlia et al., 2008).

5.1.2 Tumor location

The prognostic effect of tumor location has previously been studied. The presence of midline shift in individuals with good PS has been shown to be a significant indicator of poor survival (Gamburg et al., 2000). Significant differences between frontal, temporal, occipital, and parietal tumor locations have not been found, but central location or multi-focal (i.e. present in more than one lobe) tumors have been associated with worse prognosis (Gorlia et al., 2008). Occurrence in the left, rather than the right, cerebral hemisphere having prognostic value is disputed (Yersal, 2017; Abou Jaoude et al., 2019).

5.1.3 Tumor size

Size of post-operative enhancing tumor and pre-operative necrosis have been shown to negatively impact OS and PFS (Iliadis et al., 2012; Michaelsen et al.,

2013). Studies on the relevance of FET PET Wester et al. (1999) in radiotherapy planning have shown the biological tumor volume, derived from post-operative FET PET images, to be a significant prognostic factor of both OS and PFS while residual tumor in MR images is not. (Suchorska et al., 2015; Pirotte et al., 2009; Poulsen et al., 2017)

5.1.4 Advanced imaging features

Tumor size and location are perhaps the simplest and most intuitive of imaging features. However, their prognostic value is very limited, and as we discussed in Section 2.3.2, more advanced imaging features can be extracted with computational methods.

Recently, radiomics (Lambin et al., 2012) features for glioblastoma survival prediction have been proposed (Isensee et al., 2017; Weninger et al., 2019; Agravat and Raval, 2019; Sun et al., 2019; Baid et al., 2018, 2020; Shboul et al., 2017; Ingrisch et al., 2017) which usually count hundreds, even thousands, of features extracted from MR images. Radiomics has been successful in various medical imaging applications. They main advantages of radiomics for glioblastoma survival prediction is the correlation of the textural features with tumor sub-type, which has implications for survival (Fathi Kazerooni et al., 2020). Radiomics has its drawbacks, being hard to interpret and has difficulties generalizing across scanning platforms and pulse sequences (Traverso et al., 2018; Zwanenburg et al., 2020; Welch et al., 2019; Gillies et al., 2016). Furthermore, most studies on radiomics have been done with pre-operative images only. Since radiomics features are only extracted from within the tumor region of images, they may not generalize well to post-operative images.

5.2 New robust and interpretable biomarkers for glioblastoma

In this section we will give an overview of the research of paper B, where we address the main issues we mentioned with radiomic features. We introduce new imaging features, automatically obtained from MR images. The features are obtained by comparing shapes of automatically segmented structures in the patient's brain an average healthy structure. Difference in the shape between the average and segmented structure is measured with the 95% Hausdorff distance. The features do not depend on raw data from within the tumor region, thus being applicable to post-operative images, which have been much less studied

in the context of survival prediction. We build machine learning models for survival prediction based on these features and show that they carry prognostic value in terms of overall- and progression-free survival, and show substantial improvement over models that only consider conventional non-imaging features. Our experiments involve both pre- and post-operative data.

5.2.1 Proposed method

The method we propose for survival prediction consists of the three steps illustrated in figure 5.1. In the first step, we segment the images using SAMSEG-Tumor, described in Section 4.2. In the second step, we compute features by comparing each segmented structure to a model healthy structure using the 95% Hausdorff distance. In the third step, we select features automatically and feed them to the survival prediction model.

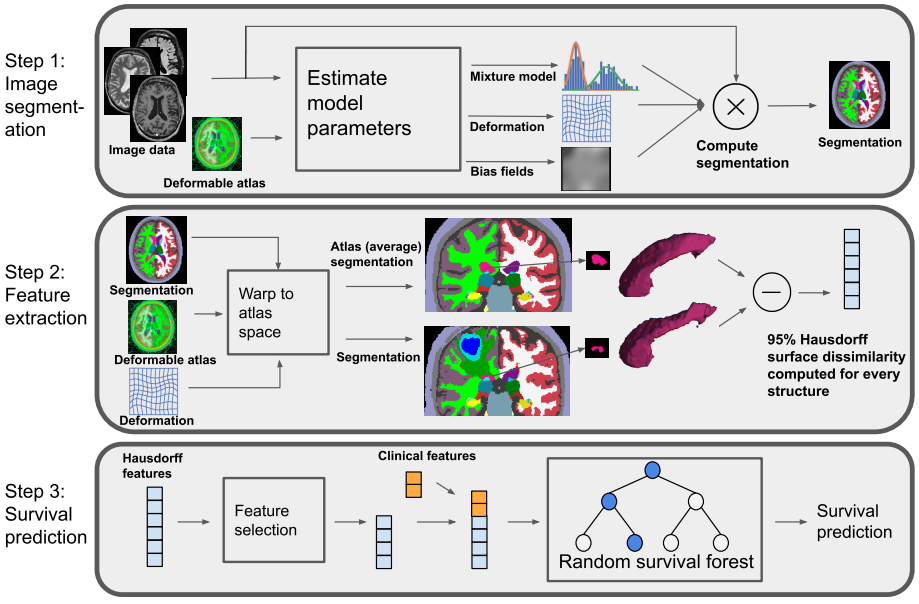

Fig. 5.1: From MR images to survival prediction in three steps: segmentation, feature extraction and survival prediction.

Feature extraction

Once segmentations are available, the goal is to extract features that can accurately measure the effect the tumor has on the shapes of the various neuroanatomical structures, compared to their appearance in healthy individuals. In order to compare segmentations, we compute the features in atlas space by warping the automatic segmentations back onto the average-shaped head model. This is achieved by applying the deformation fields that were computed as part of the segmentation procedure (see Fig. 5.1). The result is a warped, subject-specific segmentation which can be compared to the "average" head segmentation that is obtained by assigning voxels to the structure with the highest probability in the atlas. This "average" head segmentation will be referred to as the *atlas segmentation*. For healthy subjects, the subject-specific warped segmentations should be similar to the atlas segmentation in non-cortical structures after warping into the atlas space, while for brain tumor patients the difference will be much larger.

To measure difference of two segmentations, we compute a robust version of the Hausdorff distance for each of 26 structures. These structures are: Accumbens area, amygdala, brain stem, caudate, cerebellum cortex, cerebral cortex, hippocampus, lateral ventricle, optic chiasm, pallidum, putamen, thalamus, ventral diencephalon, 3rd- and 4th-ventricles. For all the aforementioned structures, except the brain stem and 3rd- and 4th-ventricles, we look at the left- and right-sided structures separately.

The Hausdorff distance computes the distance between the outer borders of a pair of segmentation masks and its robust version is an often-used metric to quantify the performance of automatic segmentation methods with respect to manual "ground truth" delineations performed by human experts (Menze et al., 2014). Let A and B denote the outer border of the segmentation masks of a particular brain structure, obtained from the atlas and warped segmentation, respectively. The Hausdorff distance computes, for all voxels on the border A, the shortest Euclidean distance to voxels on the border B, and vice versa, and returns the maximum value over all the computed distances. Because the maximum distance is highly sensitive to outliers, the robust version instead returns the 95th percentile of the distances (Huttenlocher et al., 1993) (illustrated in Fig. 5.2). The robust version is often called the 95% Hausdorff distance but for short, will be referred to as Hd95 throughout the rest of the thesis.

In cases where no voxel is assigned to a structure when obtaining the automatic segmentation, the Hd95 is not defined. In such cases, we instead use a single voxel located at the center of mass of the corresponding atlas segmentation.

For an example of how Hd95 captures the deformation of brain structures, figure 5.3 shows two subjects with glioblastoma (Fig. 5.3 (B-C)) and the corresponding atlas segmentation (Fig. 5.3 (A)) for comparison. The tumor in figure 5.3 (B) has a clear effect on the shape of the left hippocampus, putamen and pallidum, with an estimated Hd95 of 21.7, 28.5 and 49.3 [mm], respectively. While also showing a clear deformation of the left hippocampus, the left pallidum and putamen in figure 5.3 (C) seems largely unaffected, with Hd95 of 24.6, 2.5 and 2.2 [mm], respectively.

The proposed Hd95 features contain some information about where the tumor is located in the brain and its size, both of which have been studied before and shown to carry prognostic value. To verify that any prognostic value of our features is not solely based on tumor size and location, we also evaluate the performance of our survival prediction models when they are trained directly on the estimated tumor size and the whole tumor's center-of-mass (CoM) coordinates. The contrast-enhancing tumor volume (CEV) is the tumor size definition most widely used clinically, but we will also consider the volume of each tumor component (TCV), including resection cavity in case of post-operative images.

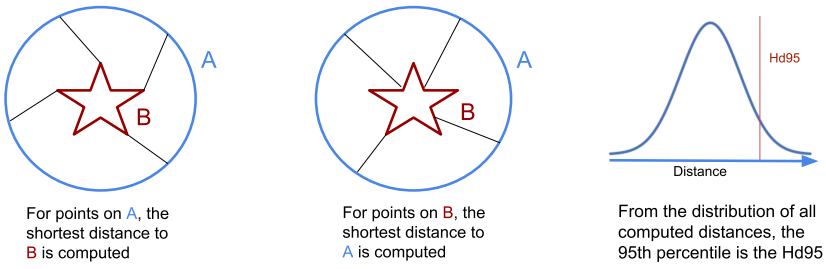

Fig. 5.2: An illustration of how the Hausdorff 95% distance is computed between two example shapes.

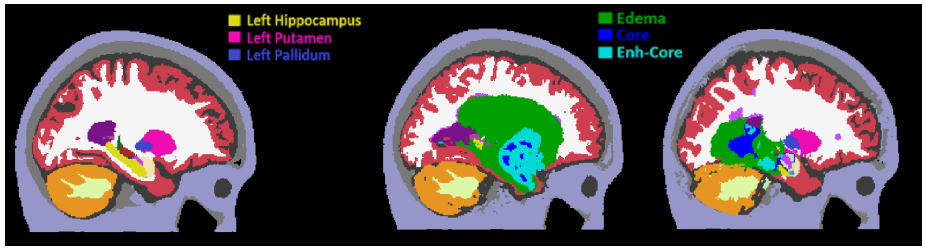

Fig. 5.3: The atlas segmentation reflecting average anatomy (A) and automatic brain segmentations of two subjects with glioblastoma (B-C). The subject in (B) has highly deformed left hippocampus, pallidum and putamen, which is reflected in high Hd95 values for these structures. While the hippocampus in (C) is also deformed, the pallidum and putamen are largely unaffected.

Survival prediction

Survival predictions models were trained following a standard machine learning workflow. The workflow consists of feature selection to remove uninformative features, and subsequent fitting of a survival prediction model to the remaining features. For feature selection, we used the univariate Cox proportional hazards (Cox PH) model (Cox, 1972), considering one feature at a time and selecting the ones whose coefficient was significantly nonzero. We used two sided P-values and considered $P < 0.05$ statistically significant.

A random survival forest (RSF) Ishwaran et al. (2008) was used as the prediction model. RSF extends the random forest model Breiman (2001) to handle right-censored data, i.e. subjects who had not died at the end of the study. Even though the right-censored subjects didn't die, knowing that they survived at least until their recorded time still contributes to fitting the RSF parameters. The RSF is an ensemble of trees whose leaf nodes estimate the subject's survival function from training data seen by the node. The survival prediction for a subject is taken as the expected survival of the average survival function across all terminal nodes the subject visits.

Due to the small number of subjects in our datasets, we did not optimize over the RSF hyperparameters but left them at the default setting in the survival analysis software: 100 trees, no maximum depth, 6 subjects minimum to split a node and minimum 3 subjects in leaf nodes. Models were trained via K-fold cross-validation where K was chosen such that in each fold, 5 subjects were left

out while the model was trained on the remaining N-5 subjects (K= $N/5$); the model was then used to predict survival of the 5 left-out subjects. We repeated this procedure 100 times for more accurate estimation of model performance.

5.2.2 Data

The data we base our results on are two different datasets, one pre-operative and the other post-operative.

Copenhagen dataset (post-operative): The dataset contains MR scans of 146 histologically verified glioblastoma patients. Each patient received radiation therapy with concomitant and adjuvant temozolomide (see Poulsen et al. (2017) for details about the treatment). OS and PFS were recorded in months for all subjects with 14 and 6 censored subjects (i.e. still alive/non-progression at the end of the study), respectively. MR scans were acquired for radiation planning 2-3 weeks post-operatively. The acquired MR modalities included 3D T1 (MPRAGE) post-administration of gadolinium contrast (T1c), T2 and FLAIR (Fig. 5.4 (A-C)), using a 1.5T Siemens Espree scanner. The T1c scans were acquired using a voxel size of $0.5 \times 0.5 \times 1.0$ mm^3 (matrix size $384 \times 512 \times 176$); the FLAIR scans with a voxel size of $0.45 \times 0.45 \times 3.3$ mm^3 (matrix size $448 \times 512 \times 40$); and the T2 scans using a voxel size of $0.3 \times 0.3 \times 3.3$ mm^3 (matrix size $672 \times 768 \times 39$). As the only form of pre-processing, intra-subject registration and resampling to 1mm^3 resolution was performed using FLIRT Jenkinson et al. (2002). Three out of the 146 subjects were excluded as their post-operative MR data was unavailable. Out of the remaining 143 subjects, 11 were missing FLAIR scans and 3 were missing T2. However, our segmentation algorithm is robust with respect to missing modalities, allowing all 143 subjects to be included in the study.

BraTS20 dataset (post-operative): The Multi-modal Brain Tumor Segmentation Challenge 2020 (BraTS20) released a publicly available set of 235 high grade glioma subjects with known OS. This dataset contains both glioblastoma and anaplastic astrocytoma Menze et al. (2014), although more detailed information on the subjects' sub-classification is not provided. For each subject, information on their age and OS is provided, but PFS or other clinical features are not available. None of the 235 subjects are censored. The MR scans originate from multiple clinics and were acquired on different scanners, with magnetic field strengths of 1.5T and 3T. For each subject, the dataset contains a T1 pre- and post-administration of gadolinium contrast, a T2 and a T2 FLAIR scan (Fig. 5.5 (A-D)). In a pre-processing step, the images were aligned to a brain template, interpolated to 1mm^3 isotropic resolution and skull-stripped by the challenge organizers Menze et al. (2014); Bakas et al. (2018).

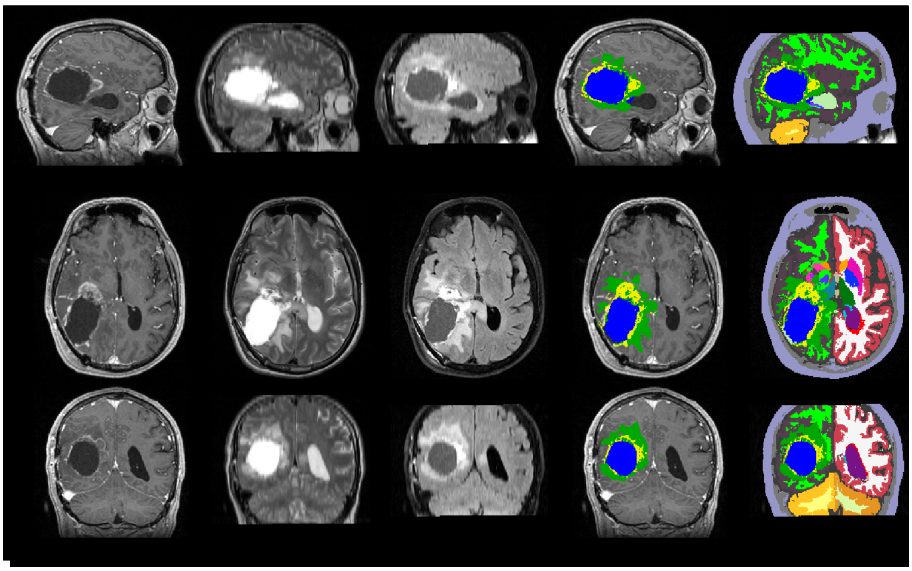

Fig. 5.4: A sample from the Copenhagen (post-operative) dataset. From top to bottom: sagittal, axial and coronal view. The columns show: (A) T1c, (B) T2, (C) FLAIR, (D-E) the automatic segmentation output, (D) the tumor components only and (E) the full segmentation output. The tumor components in (D-E) are edema (green), enhancing core (yellow) and non-enhancing core (blue). Resection cavity is shown in light green color in the sagittal view of (E).

Despite differences in available MR contrasts and in pre-processing compared to the Copenhagen dataset, our segmentation method did not need adjustment to handle the BraTS20 data (see in Fig. 5.5 (E-G), the manual and automatic tumor segmentation along with the whole-brain segmentation).

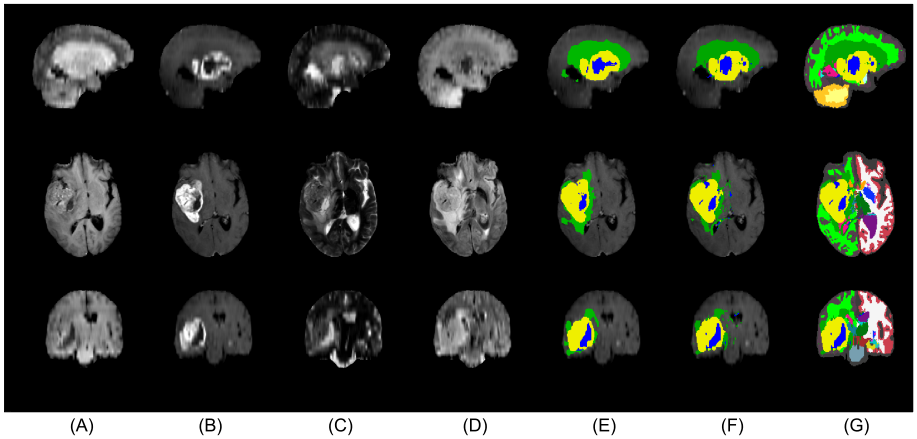

Fig. 5.5: A sample from the BraTS20 (pre-operative) dataset. From top to bottom: sagittal, axial and coronal view. The columns show (A) T1, (B) T1c, (C) T2, (D) FLAIR, (E) manual segmentation of tumor, (F-G) show the automatic segmentation output, (F) shows the tumor components only but (G) shows the full segmentation output.

5.2.3 Experiments and results

We will only briefly discuss the results described in paper B, where we present several different aspects of the proposed prediction method. First, we look at which Hd95 features were automatically selected for inclusion in the survival models. We then make a comparison between models trained on different feature sets, and we test the proposed method's ability to stratify patients into high- and low-risk groups based on their predictions. Finally, we evaluate the discriminative power of individual features for predicting short and long survival.

Feature selection

On the Copenhagen dataest, our feature selection resulted in 10 retained Hd95 features for OS prediction, and 4 for PFS, while 4 were retained for OS prediction

on the BraTS20 dataset.

Feature	OS (Copenhagen)	PFS (Copenhagen)	OS (BraTS20)
Amygdala	✓(L)		✓(L)
3rd-Ventricle	✓		
Hippocampus			✓(L)
Lateral ventricle	✓(L)		
Pallidum	✓(L,R)	✓(L)	✓(L)
Putamen	✓(L)	✓(L)	✓(L)
Thalamus	✓(L,R)	✓(L)	
Ventral diencephalon	✓(L,R)	✓(L)	

Table 5.1: Brain structures whose Hd95 feature was selected by the feature selection method are marked with a check mark, accompanied by L and R denoting left and right sided structures.

Subject-level prediction performance

To evaluate the prognostic value of the Hd95 features, in this section we investigate the performance of RSF prediction models trained on different sets of input features. In particular, we are interested in the comparison of models trained with the clinical features alone; the Hd95 features alone; and the combination of both. In addition, we compare with models that use tumor size (either TCV or CEV) and center-of-mass (CoM) as input features and models that only use age. Note that feature selection was only performed on the Hd95 features as the clinical, size and location features have all been previously shown to carry prognostic value.

To quantify the performance of a model, its predictions were compared with the ground truth survival times using Harrell's concordance index (C-index) (Harrell et al., 1982). The C-index computes the probability that for a pair of randomly selected subjects, their predicted survival is correctly ordered with respect to their true survival times. A C-index value of 1 means perfect prediction performance while 0.5 is the expected result of blindly guessing.

Copenhagen dataset: The best model for OS was achieved by combining the proposed features with the previously known prognostic clinical features. Further addition of CoM, TCV and CEV did not provide significant improvement. Individually, the clinical, size and location features all showed lower performance than the Hd95 features for OS prediction, and when combined they achieved only 0.624 C-index, compared to the 0.670 C-index when Hd95 was included.

The Hd95 features, thus seem to bring prognostic value that is not contained in simple size and location based features. For PFS, the best model was achieved by combination of Hd95, clinical, CoM and CEV, achieving a C-index of 0.637. Individually, the CoM was the best predictor of PFS and combining it with clinical and size features achieved a C-index of 0.622. The benefit of including the Hd95 features is clear for PFS, but is considerably lower than for OS. Age alone is not a reliable predictor for OS or PFS. Note that because we do not pass the clinical, size or location features through feature selection, including them can hurt model performance due to the additional dimensionality of data that has very little prognostic value.

BraTS20 dataset: The best OS prediction model was obtained with a combination of Hd95, CEV and age. This is in line with the results we obtained for OS prediction on the Copenhagen data, where the best model was one combining Hd95 with other features. The results for size and location features are similar between the datasets; neither are good OS predictors individually. However, individually, the age was the best feature, achieving a C-index of 0.581, which is substantially higher than in the Copenhagen dataset where age alone only achieved 0.509. Although the performance of the proposed Hd95 features and CEV individually was quite low (0.571 and 0.534, respectively), combining them both with the age achieved a C-index of 0.631. This best performance was still considerably lower than the best model obtained for the Copenhagen dataset, indicating that the Hd95 features may be more relevant for post-operative data than pre-operative. While the best model for OS prediction on the Copenhagen dataset was one combining Hd95 with clinical features, that combination only achieved a C-index of 0.612 on the BraTS20 dataset. In both cases, this is an improvement over considering either of the two individually, and it's important to note that the age is the only clinical feature provided in the BraTS20 dataset. Addition of MGMT and performance status information could improve the performance and possibly outperform the model using Hd95, CEV and age.

Risk group stratification

The RSF models, trained on the combination of clinical and selected Hd95 features, were chosen to stratify the two datasets into low- and high risk groups. Visualization of the resulting groups is shown with Kaplan-Meier survival curves (Kaplan and Meier, 1958) in Fig. 5.6. The results show that these survival models can stratify patients into significantly different survival groups for both OS and PFS.

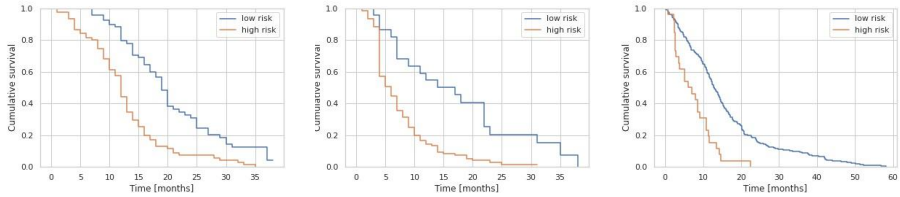

Fig. 5.6: Kaplan-Meier curves, showing the cumulative survival (fraction of the population alive/without progression at a given time), for (A) OS and (B) PFS in the Copenhagen dataset and (C) for OS in the BraTS20 dataset stratified by cross-validated predictions of the RSF models.

Prognostic potential of individual features

The Hd95 features we propose have a clear biological interpretation: higher values reflect more severe deformation in the corresponding brain structures. To test the intuition that highly deformed individual structures are associated with poor outcomes, we concentrated on subjects with very high deformations and tested to what degree their survival differs from that of the remaining subjects. Specifically, for each of the 26 brain structures for which we computed Hd95 features, we split the subjects into two groups according to whether or not they are in the highest 10% range of feature values. We then computed 1. the percentage of short survivors (below the median survival of the cohort) among the subjects in the highest 10% range, and 2. the log-rank test between the two groups.

The results of this experiment are in listed in Table 5.2 for structures where the log-rank P-value was significant. The results show that for several brain structures, high Hd95 value is a strong predictor of short survival.

Hd95 features	OS (Copenhagen) % short	OS (Copenhagen) P	PFS (Copenhagen) % short	PFS (Copenhagen) P	OS (BraTS20) % short	OS (BraTS20) P
Left lateral ventricle	92	1×10^{-3}	71	4×10^{-2}	-	-
Left putamen	85	2×10^{-3}	71	2×10^{-3}	71	2×10^{-3}
Left pallidum	83	4×10^{-3}	71	7×10^{-3}	67	5×10^{-3}
Left thalamus	82	7×10^{-3}	91	2×10^{-2}	65	2×10^{-2}
Left ventral diencephalon	77	6×10^{-3}	69	2×10^{-2}	81	2×10^{-4}
4th ventricle	58	4×10^{-2}	-	-	-	-
Left amygdala	-	-	-	-	79	1×10^{-2}
Left hippocampus	-	-	-	-	75	7×10^{-4}

Table 5.2: Percentage of short survivors among the subjects in the highest 10% range of individual Hd95 feature values. The table also shows the P-value of a log-rank test between the survival times of subjects within and outside the highest 10% range. Brain structures where the log-rank P-value > 0.05 are omitted.

5.3 Discussion

In this chapter, we gave an overview of the imaging biomarkers used for predicting survival of glioblastoma. We described the proposed features in paper B, the methods applied in that paper and our results. Our main goal was to introduce imaging features that are interpretable and can be computed regardless of the available MR modalities, scanning equipment or preprocessing. The proposed features can be interpreted as measuring the deformation of the brain anatomy due to glioblastoma and are computed by comparing the whole-brain segmentation to an atlas segmentation based on healthy subjects. To achieve robustness to missing MR modalities, scanning equipment or preprocessing, the segmentation method used was chosen to be a generative model that has these properties. On two different datasets – one post-operative and one pre-operative, we measured deformation due to tumor using the proposed Hd95 metric for dozens of brain structures. On our Copenhagen (post-operative) dataset, we looked at the relation of these deformation features to both OS and PFS, and OS in the BraTS20 (pre-operative) dataset. We showed through our experiments that the proposed features carry prognostic information and can improve survival models that use conventional clinical features (age, MGMT and performance status). Group analysis based on the output of models showed that they could clearly stratify the datasets into low- and high-risk groups with significantly different survival characteristics. Furthermore, individual feature predictiveness was explored, indicating that for some brain structures, very high deformation is a

reliable indicator of short survival.

While radiomics studies focus on patterns within the tumor region, in this study we have focused on the rest of the brain and ignored the tumor region itself entirely. Using such an approach, we demonstrated that considering out-of-region deformation features together with conventional clinical prognostic factors significantly improves survival models. A recent study Bae et al. (2018) showed how 18 radiomic features could similarly improve RSF model accuracy when combined with clinical features. Future work may therefore involve combining both within-tumor radiomic features and our Hd95 features to further improve model accuracy.

Chapter 6

MGMT prediction for glioblastoma

In this chapter we describe the research related to paper C, where we predict MGMT methylation of glioblastoma. We developed the method for the RSNA-BraTS 2021 MGMT prediction challenge (Baid et al., 2021), which provided data and evaluation platform.

6.1 MGMT prediction of glioblastoma

Expression of O^6-methylguanine-DNA-methyltransferase (MGMT) in glioblastoma is of clinical importance as it has implications of the patient's overall survival (Michaelsen et al., 2013; Gorlia et al., 2008). The prognostic information of MGMT is believed to be due to resistance of tumors with unmethylated MGMT promoter to Temozolomide (Hegi et al., 2005; Kitange et al., 2009), a drug used in standard therapy (Stupp et al., 2009). Inference of the MGMT status in the clinic is done by histological analysis, as currently available non-invasive techniques are still too unreliable.

The RSNA-BraTS 2021 challenge (Baid et al., 2021) contains two tasks: tumor segmentation and MGMT methylation prediction from pre-operative magnetic

resonance (MR) images. The challenge organizers released a large dataset with the goal of facilitating comparison between methods and advancing state-of-the-art methods in these domains. In paper C, we focus on the prediction task only.

As we have discussed in this thesis, radiomics has gained much interest for prediction tasks related to brain tumors and has been successfully applied to MGMT methylation prediction (Xi et al., 2018; Li et al., 2018). We propose a method for inference of the MGMT methylation that combines the use of radiomics with shape features learned by the variational autoencoder (VAE) we described in Section 3.2. VAE, implemented with deep neural networks, can learn high level features that are specific to the data structure it is trained on. By training the VAE on tumor segmentations, we may be able to extract complex tumor shape features that radiomics does not include. Combining hand-crafted features with a learned latent representation of medical images for classification has been previously studied (Cui et al., 2019), showing improved model classification performance.

6.2 Data

For every subject, the available modalities are T1 weighted, post-contrast T1 weighted (Gadolinium), T2 weighted and T2-FLAIR (Fig. 6.1 (A-D)). A detailed description of the data and pre-processing applied to it by the challenge organizers is given in Baid et al. (2021). The segmentation task dataset was registered to a standard template and provided as NIFTI files, while the classification data are not co-registered and are provided as DICOM files. For the prediction task, the training data consists of 585 subjects while the validation data consists of 87 subjects.

6.3 Proposed method

In this section, we give a brief overview of the proposed method, which is described in greater detail in paper C.

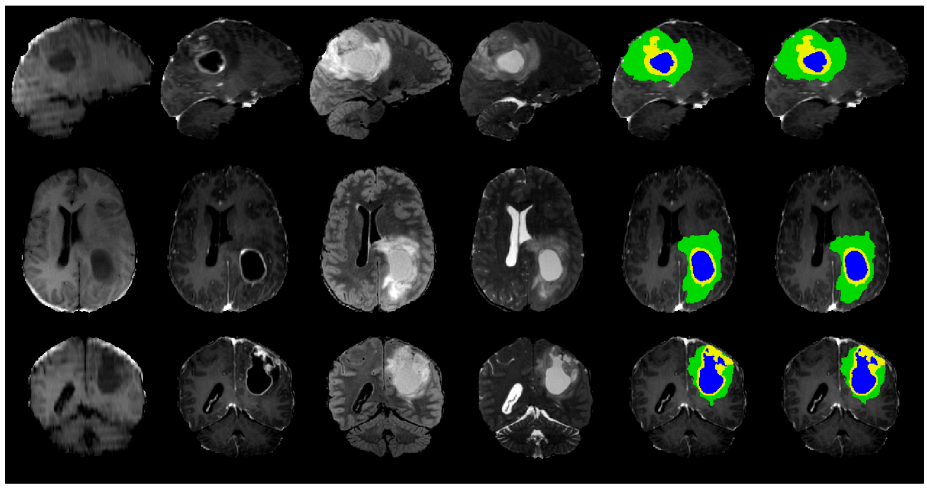

Fig. 6.1: MR images of a brain tumor patient. From top to bottom: sagittal, axial and coronal view. The columns show (A) T1w, (B) T1w-ce, (C) T2w-FLAIR, (D) T2w, (E) manual tumor segmentation, (F) automatic tumor segmentation.

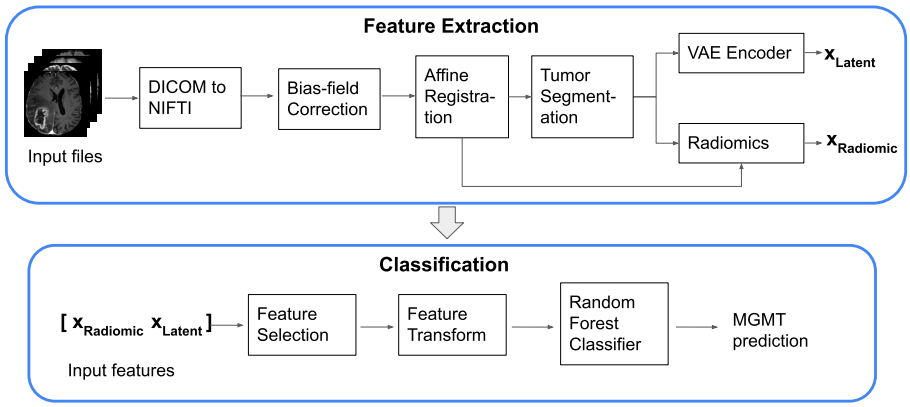

Fig. 6.2: Overview of our method. The figure shows the main components involved in going from input images to MGMT methylation prediction. The figure is borrowed from paper C.

Latent shape features

We obtain latent shape features from the VAE model described in Section 3.2. We train the VAE on 1251 segmentations from the segmentation training dataset. To extract features from a given tumor segmentation, it is passed through the encoder network and its output is taken as the latent features. We set the number of latent features to 64.

Radiomics

We extract radiomic features from three automatically segmented tumor regions and from each provided modality, resulting in a total of 1172 extracted radiomic features. The radiomic features comprise seven categories: first-order statistics, shape descriptors, gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), gray level dependece matrix (GLDM), and neighboring gray tone difference matrix (NGTDM). We use the PyRadiomics (Van Griethuysen et al., 2017) python implementation of radiomics for the feature extraction. The three tumor regions we consider are the whole tumor, enhancing core and non-enhancing core. The whole tumor is the union of all the three tumor components that are segmented.

Classification

We use a random forest (Breiman, 2001) to obtain predictions of MGMT methylation status, given the input features we extracted. The model is trained on 585 subjects via K-fold cross validation, with K chosen such that in each fold, 5 subjects are held out while the remaining subjects are used to train a model ($K = 117$ in our case). In each fold, the model is trained on 580 subjects and predictions on the 5 held-out subjects are obtained. Once predictions are obtained for all subjects, a performance score is calculated. The performance score we use is the area under the receiver operating characteristic curve (AUC). Using grid search, we tune two hyperparameters of the RF; the number of samples to split a node and maximum depth of trees. At test time, given an unseen subject, the 117 models are all used to predict the MGMT methylation status, each predicting either 0 or 1 for the unmethylated or methylated group, respectively. The average of the predictions is interpreted as the probability of belonging to the methylated group.

6.4 Experiments and results

In this section, we briefly describe the results obtained.

Feature selection

Feature selection is performed to reduce the number of uninformative variables. The list of selected radiomic features is given in Table 6.1. We observe selected radiomic features from 6 out of 7 categories mentioned in 6.3, from 3 out of 4 modalities and from all 3 tumor regions.

Table 6.1: List of selected radiomic features.

Category	Feature name	Modality	Region
Shape	Maximum 3D Diameter	-	Enh-core
First order	Interquartile Range	T1-ce	Core
First order	Mean Absolute Deviation	T1-ce	Core
First order	Mean	T1-ce	Core
First order	Median	T1-ce	Core
First order	Median	T1-ce	Whole
First order	Variance	T1-ce	Core
First order	10Percentile	FLAIR	Core
GLRLM	Graylevel non-uniformity normalized	FLAIR	Whole
GLRLM	Graylevel variance	FLAIR	Whole
GLSZM	Small area emphasis	FLAIR	Whole
GLSZM	Small area high graylevel emphasis	FLAIR	Whole
GLSZM	Small area low graylevel emphasis	FLAIR	Whole
NGTDM	Busyness	FLAIR	Whole
GLSZM	Small area high graylevel emphasis	T2	Whole

Classification

To test the benefit of using the latent shape features in the model along with the radiomic features, we train the RF on both feature sets separately and together and measure the AUC score. For a more accurate performance measure on the training dataset, we run our cross validation 10 times (each time the dataset is shuffled) and in Table 6.2, we report the mean AUC score across the 10 iterations. The true labels of the validation dataset are unknown to us, but by submitting our predictions to the challenge platform, we obtain a validation

Table 6.2: Classification performance measured by AUC. For three feature sets, the table shows AUC score for both cross-validated training set predictions and predictions on the validation set.

Features	Training	Validation
Radiomics + Latent	0.603	0.598
Radiomics	0.582	0.632
Latent	0.568	0.488

AUC score reported in Table 6.2. We observe a substantial disagreement between the training and validation scores: the training results show improvement with the combination of feature sets, while the validation scores indicate that using radiomics alone is preferred and that the latent shape features have very low predictive value.

6.5 Discussion

We have described the research of paper C, where we propose a method for MGMT methylation prediction that combines the use of radiomics with high level shape features learned by a variational autoencoder. The method was submitted to the challenge and obtained a validation score (AUC) of 0.598.

As we discuss in Section 6.1, radiomic features have already been shown to be applicable to this prediction task while tumor shape has not been proven to predict MGMT methylation. Therefore, to test whether the feature set combination we propose performs better than simply using the radiomic features alone, we experiment with training the classifier on them separately. On our training data, we observe a performance benefit of using the shape features (cf. Table 6.2), but this is not reproduced on the validation set where the radiomic features alone achieve a score of 0.632 but the latent features only 0.488. This may be due to overfitting of our feature transform and hyperparameter selection to the training data, or high uncertainty stemming from the small number of samples in the validation dataset. At this stage, we have only been able to run our method on the validation data, but we hope to gain more insight by submitting our method to the testing phase of the challenge, which will contain a substantially larger number of subjects.

Chapter 7

Conclusions and future work

In this thesis, we developed robust imaging biomarkers for brain tumors and applied them to prediction tasks for glioblastoma.

In Chapter 3, we described a tumor shape model implemented as a variational autoencoder and used it directly to predict survival in a semi-supervised setting (Paper A). The semi-supervised approach generally enables the use of unlabeled data to improve prediction accuracy, which can be very useful in medical applications where data is limited. However, we based the prediction on shape alone, which doesn't seem to have a strong relationship to survival. A more promising approach might be to model the tumor region intensities directly, using a similar method.

In Chapter 4, we then described the SAMSEG-Tumor model which simultaneously segments tumor and the whole-brain using a generative modeling approach. The benefit of a generative segmentation modeling approach is that the model generalizes across scanning platforms and the types of imaging sequences used. At this stage, we have only validated the SAMSEG-Tumor method on one dataset, and only for the tumor segmentation accuracy. Further validation of both healthy structure segmentation and tumor segmentation is needed on different datasets, ideally both pre- and post-operative.

In Chapter 5, we described the research related to paper B, where we devel-

oped interpretable and robust imaging biomarkers based on segmentations of MR images using SAMSEG-Tumor, and used them for survival prediction of glioblastoma patients. The proposed features can be interpreted as measuring the deformation of the brain anatomy due to glioblastoma and are computed by comparing the whole-brain segmentation to an atlas segmentation based on healthy subjects. The method was tested on two different datasets – one post-operative and one pre-operative – showing improved performance for OS and PFS compared to using only conventional non-imaging features, size and location. Potential further improvements in survival prediction may come from combining the proposed features with ones that do depend on the image intensities. Further improvements may even come from the use of PET data, which is available for the Copenhagen dataset (Poulsen et al., 2017).

In Chapter 6, we described our work in paper C, where we predict MGMT promoter status of glioblastoma patients from MR images using a combination of shape features and radiomics features. The results using radiomics features showed significant accuracy but the addition of our proposed shape features did not improve the model performance.

Chapter 8

Paper A

Semi-supervised Variational Autoencoder for Survival Prediction

Sveinn Pálsson[1(✉)], Stefano Cerri[1], Andrea Dittadi[2], and Koen Van Leemput[1,3]

[1] Department of Health Technology, Technical University of Denmark, Lyngby, Denmark
svpa@dtu.dk
[2] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark
[3] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, USA

Abstract. In this paper we propose a semi-supervised variational autoencoder for classification of overall survival groups from tumor segmentation masks. The model can use the output of any tumor segmentation algorithm, removing all assumptions on the scanning platform and the specific type of pulse sequences used, thereby increasing its generalization properties. Due to its semi-supervised nature, the method can learn to classify survival time by using a relatively small number of labeled subjects. We validate our model on the publicly available dataset from the Multimodal Brain Tumor Segmentation Challenge (BraTS) 2019.

Keywords: Survival time · Deep generative models · Semi-supervised VAE

1 Introduction

Brain tumor prognosis involves forecasting the future disease progression in a patient, which is of high potential value for planning the most appropriate treatment. Glioma is the most common primary brain tumor and patients suffering from its most aggressive form, glioblastoma, have generally very poor prognosis. Glioblastoma patients have a median overall survival (OS) of less than 15 months, and a 5-year OS rate of only 10% even when they receive treatment [1]. Automatic prediction of overall survival of glioblastoma patients is an important but unsolved problem, with no established method available in clinical practice.

The last few years have seen an increased interest in brain tumor survival time prediction from magnetic resonance (MR) images, often using discriminative

S. Pálsson, S. Cerri and A. Dittadi Contributed equally.

A. Crimi and S. Bakas (Eds.): BrainLes 2019, LNCS 11993, pp. 124–134, 2020.
https://doi.org/10.1007/978-3-030-46643-5_12

methods that directly encode the relationship between image intensities and prediction labels [2]. However, due to the flexibility of MR imaging, such methods do not generalize well to images acquired at different centers and with different scanners, limiting their potential applicability in clinical settings. Furthermore, being supervised methods, they require "labeled" training data where for each training subject both imaging data and ultimate survival time are available. Although public imaging databases with survival information have started to be collected [3–6], the requirement of such labeled data fundamentally limits the number of subjects available for training, severely restricting the prediction performance attainable with current methods.

In this paper, we explore whether the aforementioned issues with supervised intensity-based methods can be ameliorated by using a semi-supervised approach instead, using only segmentation masks as input. In particular, we adapt a semi-supervised variational autoencoder model [7] to predict overall survival from a small amount of labeled training subjects, augmented with *unlabeled* subjects in which only imaging data is available. The method only takes segmentation masks as input, thereby removing all assumptions on the image modalities and scanners used.

The Multimodal Brain Tumor Segmentation Challenge (BraTS) [3] has been held every year since 2012, and focuses on the task of segmenting three different brain tumors structures ("enhancing tumor", "tumor core" and "whole tumor") and "background" from multimodal MR images. Since 2017, BraTS has also included the task of OS prediction. In this paper we focus on the latter, classifying the scans into three prognosis groups: **long-survivors** (>15 months), **short-survivors** (<10 months), and **mid-survivors** (between 10 and 15 months), all relative to the time of diagnosis.

2 Model

We begin by formally describing the problem we aim to solve. The available training data consists of a set of N_l labeled pairs $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_{N_l}, y_{N_l})\}$, possibly augmented with a set of N_u *unlabeled* data points $\{\boldsymbol{x}_{N_l+1}, ..., \boldsymbol{x}_{N_l+N_u}\}$, where $\boldsymbol{x}_i \in \{1, ..., M_x\}^D$ is the i-th subject's image data in the form of a segmentation map with D voxels, and the target variable $y_i \in \{1, ..., M_y\}$ denotes the survival group the subject belongs to. In our case we have the segmentation of $M_x = 4$ different tumor structures as input to the model, and $M_y = 3$ different survival groups. For convenience, we will omit the index i when possible in the remainder.

We assume that the data is generated by a random process, illustrated in Fig. 1, that involves some latent variables $\boldsymbol{z} \in \mathcal{R}^L$, assumed to be independent of y, where $L \ll D$. These latent variables encode high-level tumor shape and location features shared across survival groups. Specifically, we assume a generative model of the form

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|y, \boldsymbol{z})p(\boldsymbol{z})p(y), \tag{1}$$

where $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \mathbf{I})$ is a zero-mean isotropic multivariate Gaussian, $p(y) \propto 1$ is a flat categorical prior distribution over y, and $p_{\boldsymbol{\theta}}(\boldsymbol{x}|y, \boldsymbol{z})$ is a conditional distribution parameterized by $\boldsymbol{\theta}$.

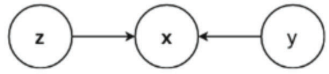

Fig. 1. Probabilistic graphical model of the generative process.

Our task is to find the maximum likelihood parameters, i.e., the parameter values $\boldsymbol{\theta}$ that maximize the probability of the training data under the model. This is equivalent to maximizing

$$\sum_{i=1}^{N_l} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i, y_i) + \sum_{i=N_l+1}^{N_l+N_u} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \tag{2}$$

with respect to $\boldsymbol{\theta}$, where

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = \int_{\boldsymbol{z}} p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z}) d\boldsymbol{z} \tag{3}$$

and

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{y} p_{\boldsymbol{\theta}}(\boldsymbol{x}, y). \tag{4}$$

Once suitable parameter values are found, the survival group of a new subject with image data $\boldsymbol{x}$ can be predicted by assessing $p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}, y)/p_{\boldsymbol{\theta}}(\boldsymbol{x})$.

2.1 Semi-supervised Variational Autoencoder

Maximizing Eq. (2) for $\boldsymbol{\theta}$ directly is not feasible due to intractability of the integral over the latent variables in Eq. (3). We therefore use an Expectation-Maximization (EM) [8] algorithm to exploit the fact that the optimization would be easier if the latent variables were known. The algorithm iteratively constructs and maximizes a lower bound to Eq. (2) in a process that involves "filling in" the missing latent variables using their posterior distribution. Since this posterior distribution is intractable, we follow [7] and approximate $p_{\boldsymbol{\theta}}(\boldsymbol{z}, y|\boldsymbol{x})$ using a specific functional form $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)$ with parameters $\boldsymbol{\phi}$:

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}, y|\boldsymbol{x}) = q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y) q_{\boldsymbol{\phi}}(y|\boldsymbol{x}),$$

where $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)$ is a multivariate Gaussian distribution with diagonal covariance matrix, and $q_{\boldsymbol{\phi}}(y|\boldsymbol{x})$ is a categorical distribution. This approximation can be used

to obtain a lower bound to Eq. (2) as follows. The probability of each *labeled* data point (first term in Eq. (2)) can be rewritten as:

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, y)] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log\left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, y)}\right]\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log\left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)}\frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, y)}\right]\right] \\
&= \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log\left[\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)}\right]\right]}_{=\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x},y)} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log\left[\frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, y)}\right]\right]}_{=D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},y)||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x},y))}
\end{aligned}$$

where D_{KL} denotes the Kullback-Leibler (KL) divergence. Since the KL divergence is always non-negative, we have that

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, y) \geq \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}, y). \tag{5}$$

Using a similar derivation, the probability of each *unlabeled* data point can be bounded as follows:

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) &\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(y,\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|y, \boldsymbol{x})} - \log q_{\boldsymbol{\phi}}(y|\boldsymbol{x})\right] \\
&= \sum_{y} q_{\boldsymbol{\phi}}(y|\boldsymbol{x})(\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}, y)) + \mathcal{H}(q_{\boldsymbol{\phi}}(y|\boldsymbol{x})) = \mathcal{U}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}),
\end{aligned} \tag{6}$$

where $\mathcal{H}(\cdot)$ denotes the entropy of a probability distribution.

By combining (5) and (6), a lower bound to Eq. (2) is finally obtained as:

$$\mathcal{J}_{\boldsymbol{\theta},\boldsymbol{\phi}} = \sum_{i=1}^{N_l} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}_i, y_i) + \sum_{i=N_l+1}^{N_l+N_u} \mathcal{U}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}_i), \tag{7}$$

which we optimize with respect to both the variational parameters $\boldsymbol{\phi}$ and the generative parameters $\boldsymbol{\theta}$. We use stochastic gradient ascent for the optimization, approximating gradients of the expectations in (7) as described in [9]. Implementation details are discussed in Sect. 4.

From a information theory point of view, the latent unobserved variables $\boldsymbol{z}$ can be interpreted as a code. Therefore, we can refer to the distributions $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, y)$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}|y, \boldsymbol{z})$ as a probabilistic *encoder* and *decoder*, respectively [9]. The label predictive distribution $q_{\boldsymbol{\phi}}(y|\boldsymbol{x})$ has the form of a discriminative *classifier*, and can be used as an approximation to $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ for classifying new cases after training.

2.2 Model Modifications

Here we describe a few model modifications for making the parameter learning process faster and less prone to overfitting.

Classification Objective. Note that in the objective function (7), the label predictive approximation $q_\phi(y|\boldsymbol{x})$ only appears in the bound for unlabeled data. To let $q_\phi(y|\boldsymbol{x})$ also learn from labeled data, we follow [7] and add a weak classification loss, resulting in the modified objective

$$\mathcal{J}^{\alpha}_{\boldsymbol{\theta},\phi} = \mathcal{J}_{\boldsymbol{\theta},\phi} + \alpha \sum_{i=1}^{N_l} \log q_\phi(y_i|\boldsymbol{x}_i) \tag{8}$$

where α controls the relative weight between generative and purely discriminative learning.

Gumbel-Softmax. One of the issues of training a semi-supervised VAE is that the marginalization over $q_\phi(y|\boldsymbol{x})$ in Eq. (6) can be computationally expensive. This marginalization can be avoided by using Gumbel-Softmax [10,11], a continuous distribution on the probability simplex that approximates a categorical sample and can be smoothly annealed (through a temperature parameter) to the categorical distribution. Gumbel-Softmax is reparameterizable so that the gradient of the loss function can be propagated back through the sampling step $y \sim q_\phi(y|\boldsymbol{x})$ for single-sample gradient estimation.

Regularization. The lower bound for labeled data can be rewritten as

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\theta},\phi}(\boldsymbol{x}, y) &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, y, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x}, y)}\right] \\ &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},y)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, y)\right] + \log p(y) - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}, y)||p(\boldsymbol{z}))\end{aligned}$$

where $\log p(y)$ is a constant, the first term can be interpreted as expected negative reconstruction error, and the last term is the negative KL divergence from the prior to the approximate posterior. Similarly, we can express the bound for unlabeled data as follows:

$$\mathcal{U}_{\boldsymbol{\theta},\phi}(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z},y|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}, y)\right] - D_{KL}(q_\phi(\boldsymbol{z}, y|\boldsymbol{x})||p(\boldsymbol{z}, y))$$

In both cases, the KL divergence acts as a regularization term that encourages the approximate posterior to be close to the prior, thereby constraining the amount of information encoded in the latent variables. The overall lower bound (7) thus trades off reconstruction error with this regularization term. When training a VAE, we can control such trade-off in order to favor more accurate reconstructions or more constrained latent space, by simply multiplying the KL term by a factor $\beta > 0$ as proposed in [12]. Similarly, we found it beneficial in practice to scale the entropy of $q_\phi(y|\mathbf{x})$ in Eq. (6) by a factor $\gamma > 1$. Intuitively, the entropy term acts as a regularizer in the classifier by encouraging $q_\phi(y|\mathbf{x})$ to have high entropy: the amplification of this term helps to further reduce overfitting in the classifier.

3 Data

The BraTS 2019 challenge is composed of a training, a validation and a test set. The training set is composed of 335 delineated tumor images, in which 210 images have survival labels. The validation set is composed of 125 non-delineated images without survival labels, in which only 29 images with resection status of GTR (i.e., Gross Total Resection) are part of the online evaluation platform (CBICA's Image Processing Portal). Finally, the test set will be made available to the challenge participants during a limited time window, and the results will be part of the BraTS 2019 workshop.

In all our experiments we performed 3-fold cross-validation by randomly splitting the BraTS 2019 training set with survival labels into a "private" training (75%) and validation set (25%) in each fold, in order to have an alternative to the online evaluation platform. This help us having a more informative indication of the model performance, since the online evaluation platform includes just 29 cases (vs. 53 cases in our private validation sets). With this set-up, which we call **S0** in the remainder, we effectively trained the model on a training set of $\mathbf{N_l = 157}$ and $\mathbf{N_u = 125}$ for each of the three cross-validation folds. These models were subsequently tested on their corresponding private validation sets of 53 subjects, as well as on the standard BraTS 2019 validation set of 29 subjects.

In order to evaluate just how much the proposed method is able to learn from *unlabeled* data (i.e., subjects with tumor delineations but no survival time information), we used three open-source methods [13–15] to automatically segment both the entire BraTS 2019 training and validation sets in order to have many more unlabeled training subjects available. We further augmented these unlabeled data sets by flipping the images in the coronal plane. With this new set-up, which we call **S1**, we then trained the model on an "augmented" private training set of $\mathbf{N_l = 157}$ and $\mathbf{N_u = 2268}$ for each of the three cross-validation folds. Ideally, dramatically increasing the set of unlabeled data points this way should help the model learn to better encode tumor representations, thereby increasing classification accuracy.

4 Implementation

We implemented the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x}, y)$, the decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z}, y)$ and the classifier $q_\phi(y|\boldsymbol{x})$ all as deep convolutional networks using PyTorch [16]. The segmentation volumes provided in the BraTS challenge have size $240 \times 240 \times 155$, but since large parts of the volume are always zero, we cropped the volumes to $146 \times 188 \times 128$ without losing any tumor voxels. We further reduced the volume by a factor of 2 in all dimensions, resulting in a shape of $73 \times 94 \times 64$, roughly a 95% overall reduction in input image size. This leads to faster training and larger batches fitting in memory, while losing minimal information.

We optimized the model end-to-end with Adam optimizer [17], using a batch size of 32, learning rate $2 \cdot 10^{-5}$, latent space size 32, $\alpha = 10^{-5} \cdot D \approx 4.4$ with D the data dimensionality (number of voxels), β from 0 to $6 \cdot 10^3$ in $3 \cdot 10^4$ steps,

$\gamma = 50$, and exponentially annealing the Gumbel-Softmax sampling temperature from 1.0 to 0.2 in $5 \cdot 10^4$ steps. Hyperparameters were found by grid search, although not fine-tuned because of the computational cost. The total number of parameters in the model is around 2.7×10^6.

4.1 Network Architecture

The three networks consist of 3D convolutional layers, with the exception of a few fully connected layers in the classifier. There are nonlinearities (Scaled Exponential Linear Units, [18]) and dropout [19] after each layer, except when noted. What follows is a high-level description of the network architecture, represented in diagrams in Fig. 2. For more details, the code is available at https://github.com/sveinnpalsson/semivaebrats.

The inference network consists of a convolutional layer (B1_e) with large kernel size and stride (7 and 4, respectively), followed by two residual blocks [20] (B2_e and B3_e). The input to each block is processed in parallel in two branches, one consisting of two convolutional layers, the other of average pooling followed by a linear transformation (without nonlinearities). The results of the two branches are added together. The output of the first layer is also fed into the classifier network, which outputs the class scores (these will be used to compute the classification loss for labeled data). A categorical sample from $q_\phi(y|\mathbf{x})$ is drawn using the Gumbel-Softmax reparameterization given the class scores, and is embedded by a fully connected layer into a real vector space. Such embedding is then concatenated to the output of the two encoder blocks, so that the means and variances of the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x}, y)$, that are computed

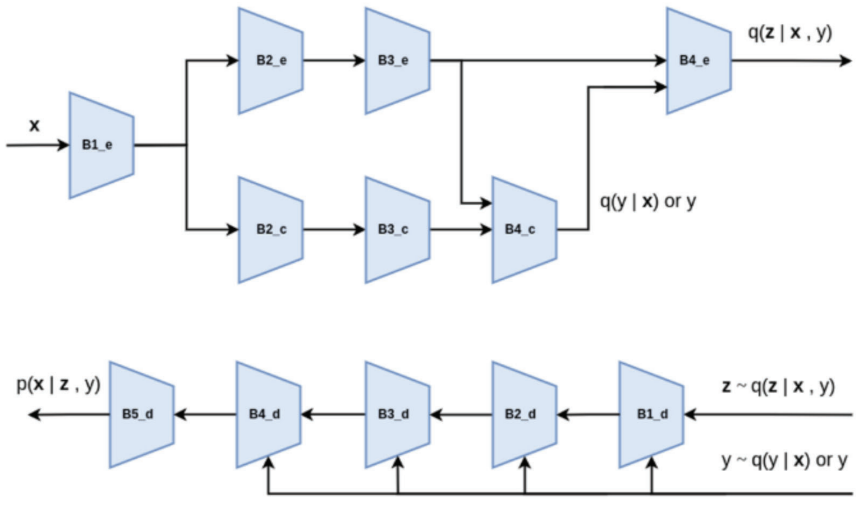

Fig. 2. Networks architectures: encoder, decoder and classifier architectures.

by a final convolutional layer, are conditioned on the sampled label. The classifier consists of two residual blocks similar to the ones in the encoder (B2_c and B3_c), followed by two fully connected layers (B4_c).

The decoder network consists of two convolutional layers (B1_d and B2_d), two residual blocks similar to those in the encoder (B3_d and B4_d), and a final convolution followed by a sigmoid nonlinearity (B5_d). In the decoder, most convolutions are replaced by transposed convolutions (for upsampling), and pooling in the residual connections is replaced by nearest neighbour interpolation. The input to the decoder network is a latent vector $\mathbf{z}$ sampled from the approximate posterior. The embedding of y, computed as in the final stage of the inference network, is also concatenated to the input of each layer (except the ones in the middle of a block) to express the conditioning of the likelihood function on the label. Here, the label is either the ground truth (for labeled examples) or a sample from the inferred posterior (for unlabeled examples).

5 Results

5.1 Conditional Generation

We visually tested whether the decoder $p_{\theta}(\boldsymbol{x}|y,\boldsymbol{z})$ is able to generate tumor-like images after training, and whether it can disentangle the classes. For this purpose we sampled $\boldsymbol{z}$ from $\mathcal{N}(\boldsymbol{z}|\mathbf{0},\mathbf{I})$ and varied y between the three classes, namely, short survivor, mid survivor and long survivor. Figure 3 shows the three shapes generated accordingly by one of the models trained in set-up **S0**. From the images we can see that the generated tumor for the short survivor class has an irregular shape with jagged edges while the long survivor generated tumor has a more compact shape with rounded edges.

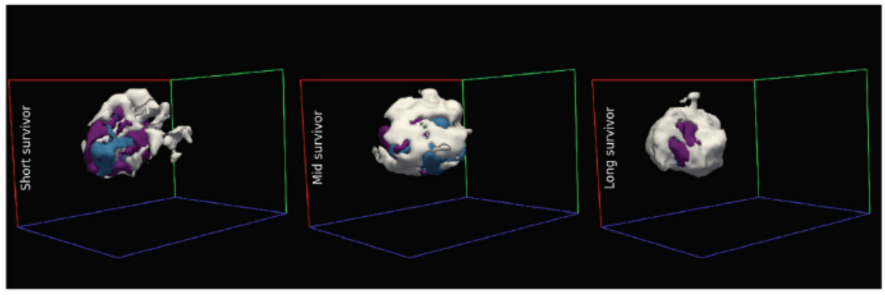

Fig. 3. Generated tumor from $p_{\theta}(\boldsymbol{x}|y,\boldsymbol{z})$ where we sampled $\mathbf{z}$ from $\mathcal{N}(\mathbf{z}|\mathbf{0},\mathbf{I})$ and we varied y between short survivor, mid survivor and long survivor.

5.2 Quantitative Evaluation

All the classification accuracies are reported with binomial confidence interval with normal approximation [21], defined as

$$a \pm z^* \sqrt{\frac{a(1-a)}{n}}$$

where a is the classification accuracy, $z^* = 1.96$ is the critical value with confidence level at 95% and n is the number of subjects. In Table 1 we show the classification accuracy of the proposed method on the "private" validation set of 53 subjects for each of the three cross-validation folds, both for the set-up with fewer (**S0**) and more (**S1**) unlabeled training subjects. The corresponding results based on the online evaluation platform (29 validation subjects) are summarized in Table 2, where we submitted the majority vote for survival group prediction across the three models trained in the cross-validation folds. The online evaluation platform takes the estimated number of days as input and returns the accuracy along with mean- and median squared error and Spearman's rank correlation coefficient. To make these predictions we input the average survival from each class. Our scores on the challenge leaderboard for set-up **S0** are as follows: 37.9% accuracy, 111214.828 mean squared error, 51076.0 median squared error and a correlation of 0.36. When testing the models we found that they are insensitive to the segmentation method used to produce the input.

Table 1. Classification accuracies [%] for both set-ups on the "private" validation set for each of the three cross-validation folds.

Set-up	Fold 1	Fold 2	Fold 3	Avg
S0	42.18 ± 13.30	35.90 ± 12.91	39.53 ± 13.16	39.20 ± 7.59
S1	47.55 ± 13.45	41.13 ± 13.40	42.91 ± 13.32	43.86 ± 7.71

Table 2. Classification accuracies [%] for both set-ups on the BraTS 2019 online evaluation platform.

Set-up	Majority voting
S0	37.90 ± 17.57
S1	31.00 ± 16.83

The results show that in none of the experiments our model achieved a significant improvement over always predicting the largest class, which constitutes around 40% of the labeled cases.

6 Discussion and Conclusions

In this paper we evaluated the potential of a semi-supervised deep generative model for classifying brain tumor patients into three overall survival groups, based only on tumor segmentation masks. The main potential advantages of this approach are (1) its in-built invariance to MR intensity variations when different scanners and protocols are used, enabling wide applicability across clinics; and (2) its ability to learn from unlabeled data, which is much more widely available than fully-labeled data.

We compared two different set-ups: one where fewer unlabeled subjects were available for training, and one where their number was (largely artificially) increased using automatic segmentation and data augmentation. Although the latter set-up increased classification performance in our "private" experiments, this increase did not reach statistically significant levels and was not replicated on the small BraTS 2019 validation set. We demonstrated visually that the proposed model effectively learned class-specific information, but overall failed to achieve classification accuracies significantly higher than predicting always the largest class.

The results described here are only part of a preliminary analysis. More real unlabeled data, obtained from truly different subjects pooled across treatment centers, and more clinical covariates of the patients, such as age and resection status, may be necessary to reach better classification accuracies. Future work may also involve stacking hierarchical generative models to further increase the classification performance of the model [7].

Acknowledgements. This project was funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie project TRABIT (agreement No 765148).

References

1. Poulsen, S.H., et al.: The prognostic value of fet pet at radiotherapy planning in newly diagnosed glioblastoma. Eur. J. Nucl. Med. Mol. Imaging **44**(3), 373–381 (2017)
2. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. CoRR, abs/1811.02629 (2018)
3. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATs). IEEE Trans. Med. Imaging **34**(10), 1993 2024 (2015)
4. Sotiras, A., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features (2017)
5. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection (2017)
6. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection (2017)
7. Kingma, D.P., Rezende, D.J., Mohamed, S., Welling, M.: Semi-supervised learning with deep generative models. arXiv e-prints arXiv:1406.5298, June 2014

8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc.: Ser. B (Methodol.) **39**(1), 1–22 (1977)
9. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv e-prints arXiv:1312.6114, December 2013
10. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with Gumbel-softmax. arXiv e-prints arXiv:1611.01144, November 2016
11. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: a continuous relaxation of discrete random variables. arXiv preprint arXiv:1611.00712 (2016)
12. Higgins, I., et al.: beta-VAE: Learning basic visual concepts with a constrained variational framework
13. Wang, G., Li, W., Ourselin, S., Vercauteren, T.: Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 178–190. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_16
14. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No new-net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 234–244. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_21
15. Nuechterlein, N., Mehta, S.: 3D-ESPNet with pyramidal refinement for volumetric brain tumor image segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 245–253. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_22
16. Paszke, A., et al.: Automatic differentiation in PyTorch. In: NIPS-W (2017)
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. In: Advances in Neural Information Processing Systems, pp. 971–980 (2017)
19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929 1958 (2014)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
21. Brown, L.D., Tony Cai, T., DasGupta, A.: Interval estimation for a binomial proportion. Stat. Sci. **16**(2), 101 133 (2001)

Chapter 9

Paper B

Predicting survival of glioblastoma from automatic whole-brain and tumor segmentation of MR images

Sveinn Pálsson[1,*], Stefano Cerri[1], Hans Skovgaard Poulsen[2], Thomas Urup[2], Ian Law[3], and Koen Van Leemput[1,4]

[1]Department of Health Technology, Technical University of Denmark, Denmark
[2]Department of Oncology, The Finsen Center, Rigshospitalet, Denmark
[3]Department of Clinical Physiology, Nuclear Medicine and PET, Center of Diagnostic Investigation, Rigshospitalet, Denmark
[4]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA
[*]svpa@dtu.dk

ABSTRACT

Survival prediction models can potentially be used to guide treatment of glioblastoma patients. However, currently available MR imaging biomarkers holding prognostic information are often challenging to interpret, have difficulties generalizing across data acquisitions, or are only applicable to pre-operative MR data. In this paper we aim to address these issues by introducing novel imaging features that can be automatically computed from MR images and fed into machine learning models to predict patient survival. The features we propose have a direct biological interpretation: They measure the deformation caused by the tumor on the surrounding brain structures, comparing the shape of various structures in the patient's brain to their expected shape in healthy individuals. To obtain the required segmentations, we use an automatic method that is contrast-adaptive and robust to missing modalities, making the features generalizable across scanners and imaging protocols. Since the features we propose do not depend on characteristics of the tumor region itself, they are also applicable to post-operative images, which have been much less studied in the context of survival prediction. Using experiments involving both pre- and post-operative data, we show that the proposed features carry prognostic value in terms of overall- and progression-free survival, over and above that of conventional non-imaging features.

Introduction

Glioblastoma is the most common malignant primary brain tumor in adults. Prognosis is generally very poor, with a median overall survival (OS) of less than 15 months, and a 5-year OS rate of only 10%, even when aggressively treated[1–4]. The standard treatment consists of maximal surgical resection followed by radiation therapy and chemotherapy with temozolomide[3]. Following standard therapy, OS and progression-free survival (PFS) have been shown to correlate with several patient-specific features such as age, performance status and expression of O^6-methylguanine-DNA-methyltransferase (MGMT)[4–7]. However, the prognostic value of these features is still too low to guide treatment choices in individual patients.

Magnetic resonance (MR) images of glioblastoma patients contain vast amounts of information about the disease, some of which may carry prognostic value. The literature on imaging biomarkers for glioblastoma survival prediction is currently dominated by radiomics[8], an approach in which hundreds or even thousands of features are extracted from delineated tumor regions of MR images, each quantifying some shape, texture, wavelet or histogram property. This approach has shown good performance in predicting survival in many studies[9–16], likely stemming from the correlation between the tumor's texture in MR images and its intratumoral heterogeneity and aggressiveness[17,18]. However, despite good prediction performance, radiomics suffers from three issues impeding wide-scale practical adoption:

- Lack of interpretability: Radiomic features, instead of aiming to be interpretable, are designed to be many, to maximize the chance of some having correlation to the target variable. Consequently, many radiomic features are seemingly arbitrary and hard to connect in a meaningful way to the nature of the disease. However, interpretability of features is important: If a model cannot give biologically meaningful explanations of its predictions, clinicians may not trust the model enough to factor its predictions into their decisions, even if the model is accurate[19]. Interpretable models may also uncover patterns in the data that give valuable new insight into the disease, and inspire future research.

- Difficulties generalizing: The reproducibility of studies using radiomics has been shown to be less than ideal, with results

failing to generalize well across scanners and software implementations[20–23]. Since many radiomic features depend directly on raw image intensities, they are sensitive to subtle changes in scanning equipment and image acquisition parameters. Furthermore, both textural and shape features depend on the segmentation mask that is used[24], underlining the importance of using image segmentation methods that are robust with respect to such sources of variation.

- Focus on pre-operative data: Compared to pre-operative images of glioblastoma, relatively little attention has been given to radiomics and other biomarkers in post-operative images. The reason may be that post-operatively, tumor shape and textural features are less easily detectable, as a large part of the tumor is usually removed. Nevertheless, post-operative images are collected closer to the time of disease progression and contain information about the success of operation, making them important to consider in a survival model. While the volume of tumor in post-operative images has been shown to correlate with OS[25,26], more advanced imaging biomarkers in post-operative images remain mostly unexplored.

In this paper, we propose a method that aims to address these shortcomings. Rather than focusing on in-region radiomic features of the tumor itself, we look at out-of-region features that are more straightforward to interpret and that can readily be applied both to pre- and post-operative data. For this purpose, we take advantage of a recently proposed method to robustly segment dozens of neuroanatomical structures in the presence of tumors[27]. Because this method aims to be invariant to imaging variations, it can be directly applied to data acquired at different centers with different scanners and protocols.

We demonstrate the resulting surivival prediction method on two fundamentally different datasets: one pre- and one post-operative dataset, each acquired with different scanners, MR contrasts and pre-processing workflows. Our results show that the proposed features improve the performance of survival models for both overall- and progression-free survival, compared to models based only on several previously known prognostic factors. To the best of our knowledge, this is the first time a survival model for glioblastoma has been proposed that is based on a detailed segmentation of the surrounding brain.

Methods

The method we propose for survival prediction consists of three steps, illustrated in Fig. 1. The first step is to segment the images with a contrast-adaptive *whole-brain* segmentation method, simultaneously segmenting dozens of brain structures and the tumor. In the second step, features are computed by comparing each segmented structure to its expected healthy shape using the 95% Hausdorff distance. In the third step, the extracted features are fed into a feature selector and a survival prediction model.

Image segmentation

For segmentation we use a method that we recently developed[27], in which three tumor components (edema, enhancing core and non-enhancing core) and dozens of neuroanatomical structures are automatically delineated from a patient's brain MR scan. For post-operative scans, another component is added to capture resection cavity. The method builds on a tool for whole-brain segmentation called Sequence Adaptive Multimodal SEGmentation (SAMSEG), which is distributed with the open-source software suite FreeSurfer[28]. It robustly segments head MR scans without any form of preprocessing, using an algorithm that can analyze multimodal data and adapt to variations in contrast due to differences in acquisition hardware or pulse sequences[29].

SAMSEG is centered on a probabilistic atlas that encodes the spatially varying voxel-wise prior probability of 41 different structures in an average-shaped head[30]. This atlas is augmented with a deformation model warping it to match the anatomy of individual subjects, along with models of the MR bias field and of structure-specific intensity profiles. At segmentation time, these models are fitted to the image being segmented, and then used to compute an automatic segmentation (Fig. 1).

For the purpose of segmenting scans with brain tumors, the basic SAMSEG model is further augmented with a spatial regularization model of tumor shape using generative neural networks[27]. Although in its original formulation we used convolutional restricted Boltzmann machines[31] for this purpose, our current implementation has variational autoencoders[32] since these have a deeper structure and can therefore better represent lesion shape[33].

Feature extraction

Once segmentations are available, we aim to extract features that can sensitively measure the effect the brain tumor has on the shape of the various neuroanatomical structures, compared to those seen in healthy individuals (Fig. 1 (Step 2)). To facilitate comparisons between individuals, we compute the features in atlas space, i.e., we warp the automatic segmentations back onto the average-shaped head model by applying the deformation fields that were estimated as part of the segmentation process. The resulting warped, subject-specific segmentations can then be compared to an "average" head segmentation that does not take any intensity information into account, obtained by assigning each voxel to the structure with the highest probability in the atlas. We will refer to this "average" head segmentation as *the atlas segmentation*. In healthy individuals, the subject-specific warped segmentations will be fairly close to the atlas segmentation in non-cortical structures after warping into atlas space, whereas in brain tumor patients the difference will often be much larger.

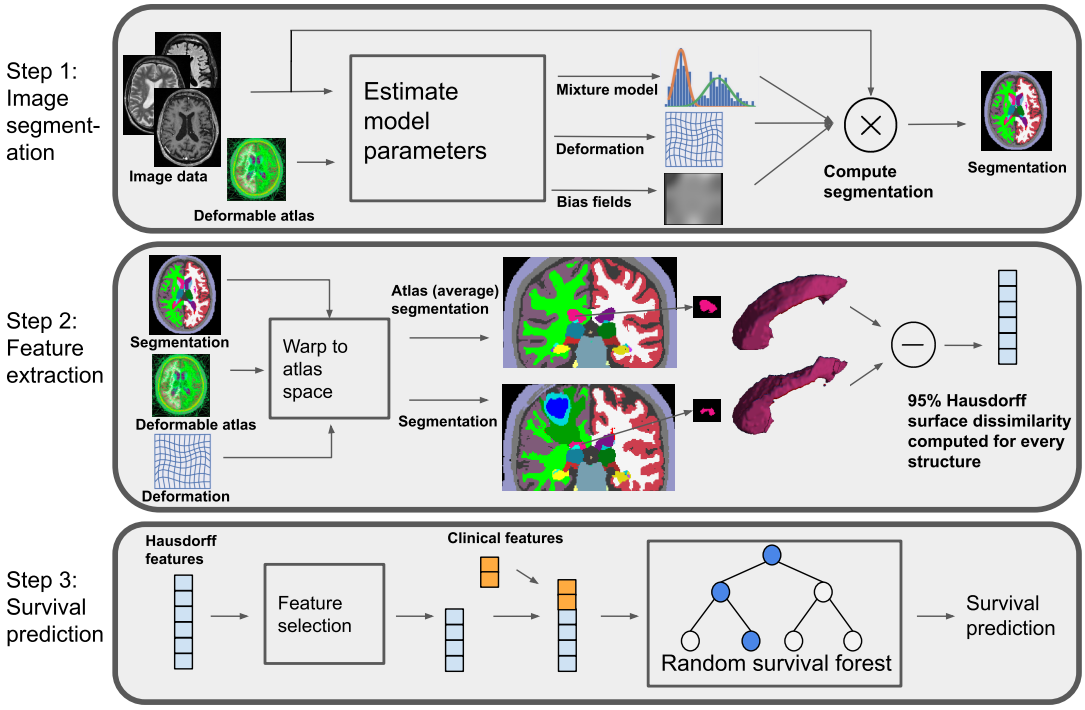

Figure 1. From MR images to survival prediction in three steps: segmentation, feature extraction and survival prediction.

In order to quantitatively compare the two segmentations, we compute a robust version of the Hausdorff distance[34] for each of 26 relevant structures: Accumbens area (L&R), amygdala (L&R), brain stem, caudate (L&R), cerebellum cortex (L&R) , cerebral cortex (L&R), hippocampus (L&R), lateral ventricle (L&R), optic chiasm, pallidum (L&R), putamen (L&R), thalamus (L&R), ventral diencephalon (L&R), 3rd- and 4th-ventricles. The Hausdorff distance measures the distance between the outer borders of a pair of segmentation masks; its robust version is an often-used metric to quantify the performance of automatic segmentation methods with respect to manual "ground truth" delineations performed by human experts[35]. Let A and B denote the outer border of the segmentation masks of a particular brain structure, obtained from the atlas and warped segmentation, respectively. The Hausdorff distance computes, for all voxels on the border A, the shortest Euclidean distance to voxels on the border B, and vice versa, and returns the maximum value over all the computed distances. Because the maximum distance is highly sensitive to outliers, the robust version instead returns the 95th percentile of the distances (Fig. 2). The robust version is often called the 95% Hausdorff distance but for short, will be referred to as Hd95 throughout the rest of the paper.

In cases where no voxel is assigned to a structure when obtaining the automatic segmentation, the Hd95 is not defined. In such cases, we instead use a single voxel located at the center of mass of the corresponding atlas segmentation.

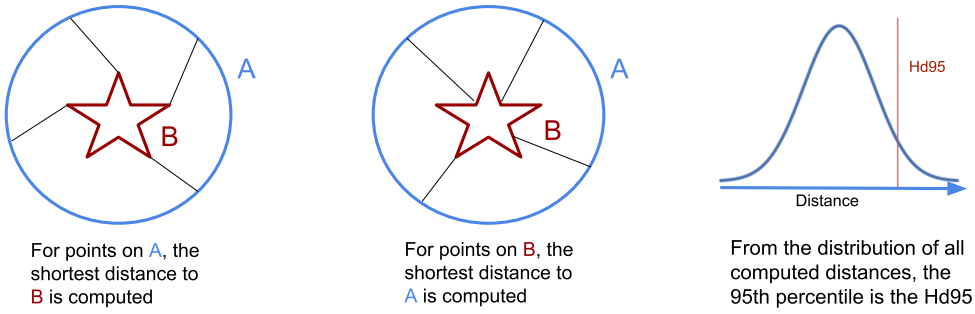

Figure 2. An illustration of how the Hausdorff 95% distance (Hd95) is computed between two example shapes.

For an example of how Hd95 captures the deformation of brain structures, Fig. 3 shows two subjects with glioblastoma (Fig. 3 (B-C)) and the corresponding atlas segmentation (Fig. 3 (A)) for comparison. The tumor in figure 3 (B) has a clear effect on the shape of the left hippocampus, putamen and pallidum, with an estimated Hd95 of 21.7, 28.5 and 49.3 [mm], respectively. While also showing a clear deformation of the left hippocampus, the left pallidum and putamen in Fig. 3 (C) seems largely unaffected, with Hd95 of 24.6, 2.5 and 2.2 [mm], respectively.

The proposed Hd95 features contain some information about where the tumor is located in the brain and its size, both of which have been studied before and shown to carry prognostic value[4,7,26,36–39]. To verify that any prognostic value of our features is not solely based on tumor size and location, in our experiments we also evaluate the performance of our survival prediction models when they are trained directly on the estimated tumor size and the center-of-mass (CoM) coordinates of the whole tumor (defined as the set of voxels assigned to any tumor component). The contrast-enhancing tumor volume (CEV) is the tumor size definition most widely used clinically, but we will also consider the volume of each tumor component (TCV), including resection cavity in case of post-operative images.

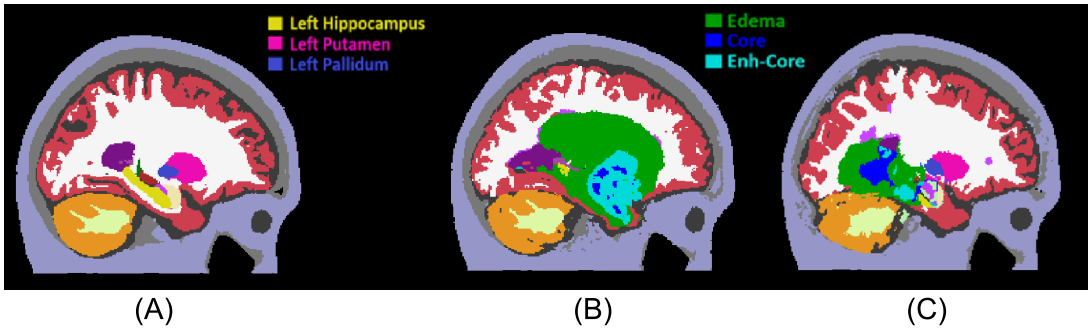

Figure 3. The atlas segmentation reflecting average anatomy (A) and automatic brain segmentations of two subjects with glioblastoma (B-C). The subject in (B) has highly deformed left hippocampus, pallidum and putamen, which is reflected in high Hd95 values for these structures. While the hippocampus in (C) is also deformed, the pallidum and putamen are largely unaffected.

Survival prediction

Survival predictions models were trained following a standard machine learning workflow. The workflow consists of feature selection to remove uninformative features, and subsequent fitting of a survival prediction model to the remaining features (see Fig. 1 (Step 3)).

For feature selection, we used the univariate Cox proportional hazards (Cox PH) model[40] (implemented in python[41]), considering one feature at a time and retaining it if its coefficient is significantly nonzero. We used two sided P-values and considered $P < 0.05$ statistically significant.

A random survival forest (RSF)[42] was used as the prediction model (implemented in python[43]). RSF extends the random

forest model[44] to handle right-censored data, i.e., subjects who had not yet died by the end of the study – knowing that these subjects survived at least until their recorded time still contributes to fitting the RSF parameters. The RSF is an ensemble of trees whose leaf nodes estimate the subject's survival function from training data seen by the node. The survival prediction for a subject is taken as the expected survival of the average survival function across all leaf nodes the subject visits. Due to the small number of subjects in our datasets, we did not optimize over the RSF hyperparameters but left them at the default setting in the survival analysis software: 100 trees, no maximum depth, 6 subjects minimum to split a node and minimum 3 subjects in leaf nodes. Models were trained via K-fold cross-validation where K was chosen such that in each fold, 5 subjects were left out while the model was trained on the remaining N-5 subjects (K= $N/5$); the model was then used to predict survival of the 5 left-out subjects. We repeated this procedure 100 times for more accurate estimation of model performance.

Experiments and Results

To demonstrate the versatility and reproducibility of the proposed method across data acquisitions, we performed experiments on two fundamentally different datasets: an in-house dataset of post-operative scans, and a publicly accessible dataset of pre-operative scans. Here we first describe these datasets, and subsequently present results for each.

Datasets

Copenhagen dataset (post-operative)

Our primary focus is on a set of post-operative scans acquired at Rigshospitalet, Copenhagen. It contains MR scans of 146 histologically verified glioblastoma patients, diagnosed in the period September 2011 - April 2014. Permission for data collection was given from the Danish Data Protection Agency (2006-41-6979). Each patient received radiation therapy with concomitant and adjuvant temozolomide (see [4] for details about treatment). OS and PFS were recorded in months for all subjects with 14 and 6 censored subjects (i.e. still alive/non-progression at the end of the study), respectively.

MR scans were acquired for radiation planning 2-3 weeks post-operatively. The acquired MR modalities included 3D T1 (MPRAGE) post-administration of gadolinium contrast (T1c), T2 and FLAIR (Fig. 4 (A-C)), using a 1.5T Siemens Espree scanner. The T1c scans were acquired using a voxel size of $0.5 \times 0.5 \times 1.0$ mm^3 (matrix size $384 \times 512 \times 176$); the FLAIR scans with a voxel size of $0.45 \times 0.45 \times 3.3$ mm^3 (matrix size $448 \times 512 \times 40$); and the T2 scans using a voxel size of $0.3 \times 0.3 \times 3.3$ mm^3 (matrix size $672 \times 768 \times 39$). As the only form of pre-processing, intra-subject registration and resampling to 1mm^3 resolution was performed using FLIRT[45]. Three out of the 146 subjects were excluded as their post-operative MR data was unavailable. Out of the remaining 143 subjects, 11 were missing FLAIR scans and 3 were missing T2. However, our segmentation algorithm is robust with respect to missing modalities, allowing all 143 subjects to be included in the study.

Additional features recorded in the clinic were the patient's age, performance status and MGMT protein status. As mentioned in the introduction, these are features that have been previously shown to have prognostic value and are thus commonly considered for radiotherapy planning. We will refer to this set of variables as the "clinical features".

BraTS20 dataset (pre-operative)

To test the reproducibility of the methods we propose, we also applied them to a fundamentally different (namely, pre-operative) dataset, obtained with other acquisition settings and preprocessing steps. The Multi-modal Brain Tumor Segmentation Challenge 2020 (BraTS20)[46–49] released a publicly available set of 235 high grade glioma subjects with overall-survival times. This dataset contains both glioblastoma and anaplastic astrocytoma[35], although more detailed information on the subjects' sub-classification is not provided. For each subject, information on their age and OS is provided, but PFS or other clinical features are not available. None of the 235 subjects are censored.

The MR scans originate from multiple clinics and were acquired on different scanners, with magnetic field strengths of 1.5T and 3T. For each subject, the dataset contains a T1 pre- and post-administration of gadolinium contrast, a T2 and a T2 FLAIR scan (Fig. 5 (A-D)). In a pre-processing step, the images were aligned to a brain template, interpolated to 1mm^3 isotropic resolution and skull-stripped by the challenge organizers[35,46]. Despite differences in available MR contrasts and in pre-processing compared to the Copenhagen dataset, our segmentation method did not need adjustment to handle the BraTS20 data.

Results on the Copenhagen dataset

We present several different aspects of the proposed prediction method. First, we look at which Hd95 features were automatically selected for inclusion in the survival models. We then make a comparison between models trained on different feature sets, and we test the proposed method's ability to stratify patients into high- and low-risk groups based on their predictions. Finally, we evaluate the discriminative power of individual features for predicting short and long survival.

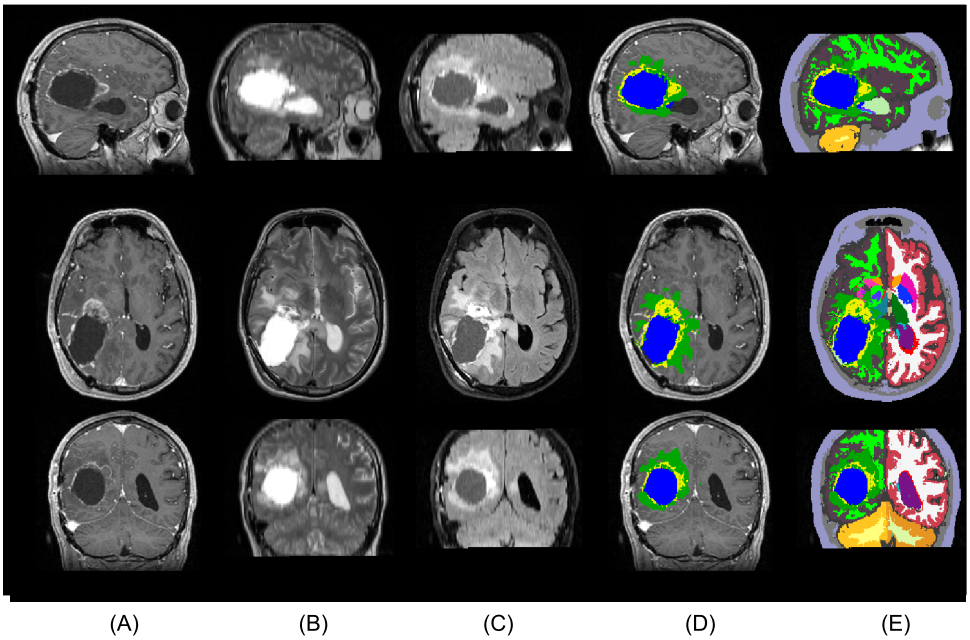

Figure 4. A sample from the Copenhagen (post-operative) dataset. From top to bottom: sagittal, axial and coronal view. The columns show (A) T1c, (B) T2, (C) FLAIR, and (D-E) the automatic segmentation output. (D) shows the tumor components only, while (E) shows the full segmentation output. The tumor components in (D-E) are edema (green), enhancing core (yellow) and non-enhancing core (blue). Resection cavity is shown in light green color in the sagittal view of (E).

Feature selection

Rather than reporting on Hd95 features selected within each fold during cross-validation, for conciseness here we present results of the Cox PH feature selection method on the entire cohort. Although this introduces information leakage between training and test sets, in our experiments we found that the selected features across folds were highly consistent, thus having minimal impact on the overall prediction performance (details provided in Appendix A). As shown in Table 1, our feature selection resulted in 10 retained Hd95 features for OS prediction, and 4 for PFS.

On the Copenhagen dataest, our feature selection resulted in 10 retained Hd95 features for OS prediction, and 4 for PFS, while 4 were retained for OS prediction on the BraTS20 dataset.

Feature	OS (Copenhagen)	PFS (Copenhagen)	OS (BraTS20)
Amygdala	✓(L)		✓(L)
3rd-Ventricle	✓		
Hippocampus			✓(L)
Lateral ventricle	✓(L)		
Pallidum	✓(L,R)	✓(L)	✓(L)
Putamen	✓(L)	✓(L)	✓(L)
Thalamus	✓(L,R)	✓(L)	
Ventral diencephalon	✓(L,R)	✓(L)	

Table 1. Brain structures whose Hd95 feature was selected by the feature selection method are marked with a check mark, accompanied by L and R denoting left and right sided structures. Shown for both the Copenhagen and BraTS20 datasets.

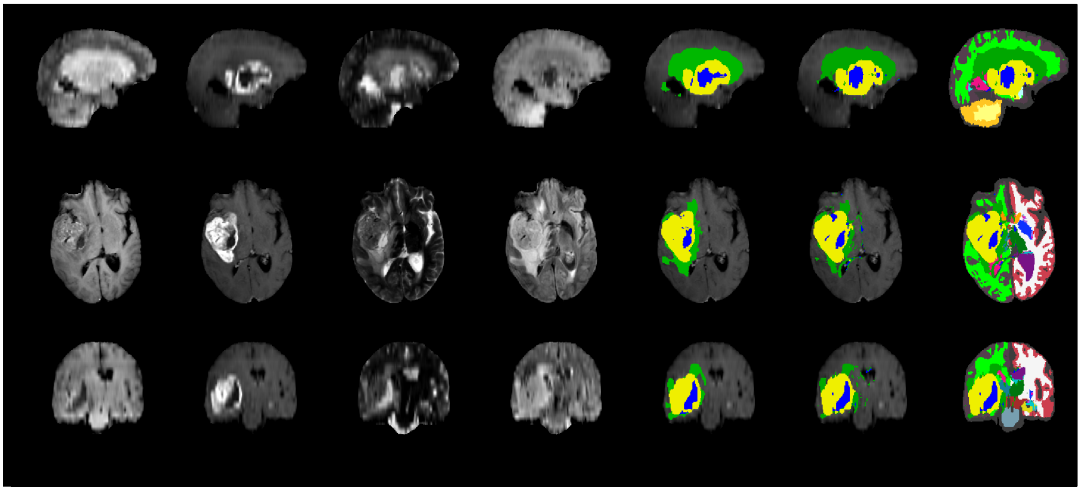

Figure 5. A sample from the BraTS20 (pre-operative) dataset. From top to bottom: sagittal, axial and coronal view. The columns show (A) T1, (B) T1c, (C) T2, (D) FLAIR, (E) manual segmentation of tumor, and (F-G) the automatic segmentation output. (F) shows the tumor components only, while (G) shows the full segmentation output. Some major differences to the Copenhagen (post-operative) dataset (see Fig. 4) can be seen in this figure.

Subject-level prediction performance

To evaluate the prognostic value of the Hd95 features, in this section we investigate the performance of RSF prediction models trained on different sets of input features. In particular, we are interested in the comparison of models trained with the clinical features alone; the Hd95 features alone; and the combination of both. In addition, we compare with models that include tumor size (either TCV or CEV) and center-of-mass (CoM) as input features, as well as with models that only use age as the clinical variable. Note that feature selection was only performed on the Hd95 features as the clinical, size and location features have all been previously shown to carry prognostic value[4,7,26,36–39].

The first two columns of Table 2 show performance of the RSF models, computed from the cross-validated predictions on the Copenhagen dataset. To quantify the performance of any given model, its predictions were compared with the ground truth survival times using Harrell's concordance index (C-index)[50]. The C-index computes the probability that for a pair of randomly selected subjects, their predicted survival is correctly ordered with respect to their true survival times. A C-index value of 1 means perfect prediction performance while 0.5 is the expected result of blindly guessing. The reported C-index is the average over the 100 repetitions of cross-validation, accompanied by the 95% confidence interval of the mean in brackets.

The best model for OS was achieved by combining the proposed features with the previously known prognostic clinical features: further addition of CoM, TCV and CEV did not provide significant improvement. Individually, the clinical, size and location features all showed lower performance than the Hd95 features for OS prediction, and when combined they achieved only 0.624 C-index, compared to the 0.670 C-index when Hd95 was also included. The Hd95 features thus seem to bring prognostic value that is not contained in simple size and location based features.

For PFS, the best model was achieved by combination of Hd95, clinical, CoM and CEV, achieving a C-index of 0.637. Individually, the CoM was the best predictor of PFS and combining it with clinical and size features achieved a C-index of 0.622. The benefit of including the Hd95 features is clear for PFS, but is considerably lower than for OS.

Features	Copenhagen		BraTS20
	OS	PFS	OS
Hd95 + Clinical + CoM + CEV	**0.670** (0.668 - 0.671)	**0.637** (0.634 - 0.641)	0.619 (0.618 - 0.620)
Hd95 + Clinical + CoM + TCV	0.657 (0.655 - 0.659)	0.629 (0.624 - 0.634)	0.612 (0.611 - 0.614)
Hd95 + Clinical + CEV	0.666 (0.665 - 0.668)	0.597 (0.595 - 0.600)	**0.631** (0.630 - 0.632)
Hd95 + Clinical	**0.669** (0.668 - 0.671)	0.614 (0.612 - 0.616)	0.612 (0.611 - 0.613)
Hd95 + CoM + CEV	0.635 (0.633 - 0.637)	0.629 (0.625 - 0.634)	0.564 (0.562 - 0.565)
Hd95 + CEV	0.643 (0.641 - 0.644)	0.580 (0.575 - 0.584)	0.594 (0.593 - 0.595)
Clinical + CoM + CEV	0.624 (0.623 - 0.625)	0.622 (0.618 - 0.626)	0.599 (0.598 - 0.600)
Clinical + CoM + TCV	0.595 (0.592 - 0.598)	0.613 (0.608 - 0.618)	0.600 (0.598 - 0.602)
Clinical + CEV	0.591 (0.589 - 0.593)	0.567 (0.563 - 0.572)	0.616 (0.614 - 0.617)
CoM + CEV	0.548 (0.545 - 0.551)	0.605 (0.599 - 0.611)	0.517 (0.516 - 0.518)
CoM + TCV	0.540 (0.537 - 0.543)	0.617 (0.612 - 0.622)	0.536 (0.533 - 0.539)
Hd95	0.644 (0.643 - 0.646)	0.552 (0.548 - 0.556)	0.571 (0.570 - 0.572)
Clinical	0.574 (0.572 - 0.576)	0.524 (0.522 - 0.527)	0.581 (0.579 - 0.583)
CoM	0.550 (0.548 - 0.551)	0.591 (0.588 - 0.594)	0.504 (0.503 - 0.506)
CEV	0.479 (0.476 - 0.482)	0.551 (0.547 - 0.554)	0.534 (0.533 - 0.535)
TCV	0.525 (0.522 - 0.528)	0.574 (0.570 - 0.577)	0.553 (0.551 - 0.555)
Age	0.509 (0.506 - 0.513)	0.519 (0.516 - 0.522)	0.581 (0.579 - 0.583)

Table 2. Prediction performance measured with the C-index for models trained on several different sets of features. The first two columns show results for OS and PFS prediction on the Copenhagen post-operative dataset, whereas the last column contains OS prediction performance on the BraTS20 pre-operative dataset. Note that the clinical features for the Copenhagen dataset include age, performance status and MGMT methylation, while the available clinical features for the BraTS dataset consist only of age. Including the proposed Hausdorff (Hd95) features in the survival model provides an improvement in prediction performance over models that only consider conventional clinical features, tumor size and location.

Risk group stratification

Here we demonstrate that the proposed survival models can be used to stratify patients into low- and high-risk groups. For this purpose, a threshold was selected by searching, among the predictions for all Copenhagen patients, for the value that best separates the dataset in terms of the recorded survival[17,51]. Separation quality was measured with the log-rank test[52], which tests the hypothesis that two groups have the same survival distribution. The prediction value yielding the lowest P-value (of the log-rank test) was chosen as the threshold separating the low- from the high-risk patients. Visualization of the resulting groups, using the RSF models trained on the combination of clinical and selected Hd95 features as prediction models, is shown with Kaplan-Meier survival curves[53] for OS and PFS in Fig. 6 (A) and (B), respectively.

We further computed the corresponding hazard ratio for the obtained splits (ratio of hazard rates between the two groups under the proportional hazards assumption[54]), using the univariate Cox proportional hazards model where the input covariate was the group membership. In addition to the hazard ratio, its 95% confidence interval and log-rank P-value were also computed. For OS, the hazard ratio was 2.65 ($1.85-3.79$), $P = 10^{-8}$ and for PFS the hazard ratio was 1.85 ($1.7-4.78$), $P = 10^{-5}$. These results show that our survival models can stratify patients into significantly different survival groups for both OS and PFS.

Prognostic potential of individual features

The Hd95 features we propose have a clear biological interpretation: higher values reflect more severe deformation in the corresponding brain structures. To test the intuition that highly deformed individual structures are associated with poor outcomes, we concentrated on subjects with very high deformations and tested to what degree their survival differs from that of the remaining subjects. Specifically, for each of the 26 brain structures for which we computed Hd95 features, we split the subjects into two groups according to whether or not they are in the highest 10% range of feature values. We then computed 1. the percentage of short survivors (below the median survival of the cohort) among the subjects in the highest 10% range, and 2. the log-rank test between the two groups.

The results of this experiment are in listed in Table 3 for structures where the log-rank P-value was significant. The results show that for several brain structures, high Hd95 value is a strong predictor of short survival. The best predictor of OS was deformation of the left lateral ventricle, where 92% of the subjects with the most deformation were short survivors. For PFS, the best predictor was the deformation of the left thalamus, with 91% of the subjects with the most deformation of that structure being short survivors.

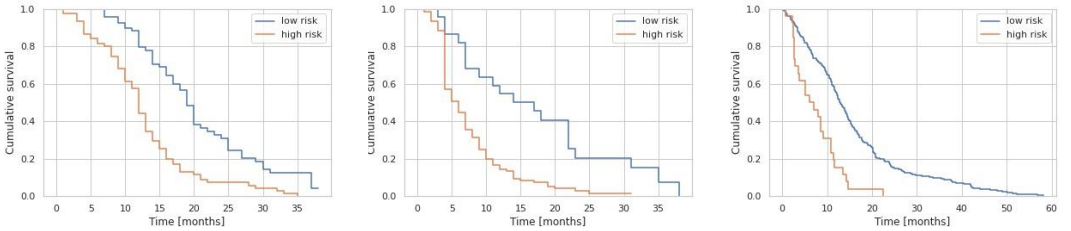

Figure 6. Kaplan-Meier curves showing the cumulative survival (fraction of the population alive/without progression at a given time), for (A) OS and (B) PFS in the Copenhagen dataset and (C) for OS in the BraTS20 dataset. The two survival groups in each figure are obtained by splitting the cohort based on the predicted survival at the threshold that best separates the cohort.

	OS (Copenhagen)		PFS (Copenhagen)		OS (BraTS20)	
Hd95 features	% short	P	% short	P	% short	P
Left lateral ventricle	92	1×10^{-3}	71	4×10^{-2}	-	-
Left putamen	85	2×10^{-3}	71	2×10^{-3}	71	2×10^{-3}
Left pallidum	83	4×10^{-3}	71	7×10^{-3}	67	5×10^{-3}
Left thalamus	82	7×10^{-3}	91	2×10^{-2}	65	2×10^{-2}
Left ventral diencephalon	77	6×10^{-3}	69	2×10^{-2}	81	2×10^{-4}
4th ventricle	58	4×10^{-2}	-	-	-	-
Left amygdala	-	-	-	-	79	1×10^{-2}
Left hippocampus	-	-	-	-	75	7×10^{-4}

Table 3. Percentage of short survivors among the subjects in the highest 10% range of individual Hd95 feature values. The table also shows the P-value of a log-rank test between the survival times of subjects within and outside the highest 10% range. Brain structures where the log-rank P-value > 0.05 are omitted.

Results on BraTS20 dataset

The same methods were applied to the BraTS20 (pre-operative) dataset, for which our goal was to predict the OS only.

Feature selection

Feature selection on the full BraTS20 cohort resulted in selection of 4 Hd95 features: left putamen, left pallidum, left hippocampus and the left amygdala (listed in Table 1 for comparison with Copenhagen dataset). Similarly to the Copenhagen dataset results, selecting features within each fold of cross-validation resulted in mostly the same features being chosen (see details in Appendix A). Note that three of the selected features on the BraTS20 data were also selected for OS prediction on the Copenhagen dataset (vs. two for PFS, cf. Table 1).

Subject-level prediction performance

Model comparison to test whether the Hd95 features contain prognostic information not included in the clinical, size or location data was done in the same manner as with the Copenhagen dataset. An important difference is that the only clinical data available here is the subject's age, while the Copenhagen data also included MGMT and performance status.

As shown in the last column of Table 2, the best OS prediction model was obtained with a combination of Hd95, CEV and age. This is largely in line with the results we obtained for OS prediction on the Copenhagen data (cf. first column of Table 2), where the best models were the ones combining Hd95 with other features. The results for size and location features are similar between the datasets: neither are good OS predictors individually. However, individually, here the age was the best feature, achieving a C-index of 0.581, which is substantially higher than in the Copenhagen dataset where age alone only achieved 0.509. Although the performance of the proposed Hd95 features and CEV individually was quite low (0.571 and 0.534, respectively), combining them both with the age achieved the best C-index of 0.631. While one of the best models for OS prediction on the Copenhagen dataset was the model combining Hd95 with clinical features, that specific combination only achieved a C-index of 0.612 on the BraTS20 dataset. Nevertheless, this is still an improvement over considering either of the

two feature sets individually. It is further worth noting that the age is the only clinical feature provided in the BraTS20 dataset – addition of MGMT and performance status information could improve the performance and possibly outperform the model using Hd95, CEV and age also here.

Risk group stratification

As in the Copenhagen dataset, the prediction model trained on the combination of clinical and selected Hd95 features was used to stratify the BRaTS20 cohort. The Kaplan-Meier curves in Fig. 6 (C) show the proportion of subjects alive at any given time point for the two resulting groups. The corresponding hazard ratio was 2.81 ($1.84 - 4.29$) and log-rank P-value 5×10^{-7}, indicating that the two resulting groups have significantly different OS.

Prognostic potential of individual features

To explore to what degree individual Hd95 features can predict short survival in the BraTS20 dataset, we repeated the experiment of exploring the percentage of short survivors among the subjects with the most deformed brain structures. As shown in Table 3, the highest 10% range of feature values is predictive of short survival for several structures. Compared to our results on the Copenhagen dataset, two new structures show high predictive power in the BraTS20 data.

Discussion

In this paper we have proposed a new set of imaging features for glioblastoma survival prediction. Our main goal was to introduce imaging features that are interpretable and that can be replicated across different MR contrasts, scanning equipment or preprocessing. The proposed Hd95 features can be interpreted as measuring the deviation from normal brain morphology due to glioblastoma, and are computed by comparing an automatic whole-brain segmentation with its expected equivalent in healthy subjects. To achieve robustness to missing MR modalities and variations in scanners or acquisition protocols, the automatic segmentations were obtained with a method that was designed to have these properties.

Using experiments on two different datasets – one post-operative and one pre-operative – we showed that the proposed features carry prognostic information and can improve survival models that use conventional clinical features such as age, MGMT and performance status. Group analysis based on the output of our models showed that they could clearly stratify the datasets into low- and high-risk groups with significantly different survival characteristics. Furthermore, individual feature predictiveness was explored, indicating that for some brain structures, very high deformation is a reliable indicator of short survival.

Through feature selection, we discovered several brain structures whose Hd95 value correlates with survival and were therefore retained for training our prediction models. Although the same set of structures was not selected in each case (OS vs. PFS and pre- vs. post-operative), two structures were selected in all three cases: the left pallidum and left putamen. Interestingly, we found that right-sided structures were overall less associated with survival. Two recent studies have explored the association of OS with left-sided glioblastoma (having higher volume in the left hemisphere than the right side) but with contrary results[38,39]. One study[38], who showed the association of left-sidedness and worse prognosis, proposed a possible explanation could be that the left hemisphere's functions may be more essential for survival.

As demonstrated in our experiments, the features proposed in this paper readily generalize across datasets: They are independent of scanner and imaging parameters, and they can be computed from both pre- or post-operative images; from data that is skull-stripped or not; and from subjects with missing modalities. We did, however, see worse prognostic performance of the Hd95 features for OS prediction on the pre-operative cohort (BraTS20) compared to the post-operative one (Copenhagen). One possible reason for this discrepancy may be that the BraTS20 dataset contains both glioblastoma and anaplastic astrocytomas, which have different survival characteristics. The fact that BraTS20 is pre-operative may play an important role as well, as the effects of surgery can not be taken into account.

The proposed Hd95 features measure how much each brain structure is deformed compared to its expected shape in the absence of pathology, and therefore they contain information about the location of the tumor, which has been shown previously to be a prognostic factor for OS[7,26,36–39]. Nevertheless, our results show that the proposed Hd95 features carry richer prognostic information for predicting OS than tumor location alone. For predicting PFS, tumor location was found to be a stronger predictor than Hd95 when considering these feature sets individually; however, substantially higher model performance was achieved with a combination of the two, together with clinical and size features. To the best of our knowledge, considerations of tumor location has not been a parameter used in the stratification of patients to treatment in clinical glioblastoma trials, although the poor prognostic feature is recognized in clinical management. Based on our results, the application of survival models exploiting advanced imaging features, such as the ones proposed here, could potentially help minimize bias in stratification in future clinical trials. High quality prognostic information could also potentially guide clinicians in adjusting the intensity of interventions, based on expected outcome and quality-of-life considerations.

While radiomics studies focus on patterns within the tumor region, in this study we have focused on the rest of the brain and ignored the tumor region itself entirely. Using such an approach, we demonstrated that considering out-of-region deformation

features together with conventional clinical prognostic factors significantly improves survival models. A recent study[17] showed how 18 radiomic features could similarly improve RSF model accuracy when combined with clinical features. Future work may therefore explore combining both within-tumor radiomic features and our Hd95 features to further improve model accuracy.

References

1. Louis, D. N. *et al.* The 2007 who classification of tumours of the central nervous system. *Acta neuropathologica* **114**, 97–109 (2007).
2. Gutman, D. A. *et al.* Mr imaging predictors of molecular profile and survival: multi-institutional study of the tcga glioblastoma data set. *Radiology* **267**, 560–569 (2013).
3. Stupp, R. *et al.* Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase iii study: 5-year analysis of the eortc-ncic trial. *The lancet oncology* **10**, 459–466 (2009).
4. Poulsen, S. H. *et al.* The prognostic value of fet pet at radiotherapy planning in newly diagnosed glioblastoma. *Eur. journal nuclear medicine molecular imaging* **44**, 373–381 (2017).
5. Michaelsen, S. R. *et al.* Clinical variables serve as prognostic factors in a model for survival from glioblastoma multiforme: an observational study of a cohort of consecutive non-selected patients from a single institution. *BMC cancer* **13**, 402 (2013).
6. Hegi, M. E. *et al.* Mgmt gene silencing and benefit from temozolomide in glioblastoma. *New Engl. J. Medicine* **352**, 997–1003 (2005).
7. Gorlia, T. *et al.* Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of eortc and ncic trial 26981-22981/ce. 3. *The lancet oncology* **9**, 29–38 (2008).
8. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. journal cancer* **48**, 441–446 (2012).
9. Booth, T. C. *et al.* Machine learning and glioma imaging biomarkers. *Clin. radiology* **75**, 20–32 (2020).
10. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M. & Maier-Hein, K. H. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In *International MICCAI Brainlesion Workshop*, 287–297 (Springer, 2017).
11. Weninger, L., Haarburger, C. & Merhof, D. Robustness of radiomics for survival prediction of brain tumor patients depending on resection status. *Front. computational neuroscience* **13**, 73 (2019).
12. Agravat, R. R. & Raval, M. S. Brain tumor segmentation and survival prediction. In *International MICCAI Brainlesion Workshop*, 338–348 (Springer, 2019).
13. Sun, L., Zhang, S., Chen, H. & Luo, L. Brain tumor segmentation and survival prediction using multimodal mri scans with deep learning. *Front. neuroscience* **13**, 810 (2019).
14. Baid, U. *et al.* Deep learning radiomics algorithm for gliomas (drag) model: a novel approach using 3d unet based deep convolutional neural network for predicting survival in gliomas. In *International MICCAI Brainlesion Workshop*, 369–379 (Springer, 2018).
15. Baid, U. *et al.* Overall survival prediction in glioblastoma with radiomic features using machine learning. *Front. computational neuroscience* **14**, 61 (2020).
16. Ingrisch, M. *et al.* Radiomic analysis reveals prognostic information in t1-weighted baseline magnetic resonance imaging in patients with glioblastoma. *Investig. radiology* **52**, 360–366 (2017).
17. Bae, S. *et al.* Radiomic mri phenotyping of glioblastoma: improving survival prediction. *Radiology* **289**, 797–806 (2018).
18. Parekh, V. S. & Jacobs, M. A. Deep learning and radiomics in precision medicine. *Expert. review precision medicine drug development* **4**, 59–72 (2019).
19. Shortliffe, E. H. & Sepúlveda, M. J. Clinical decision support in the era of artificial intelligence. *Jama* **320**, 2199–2200 (2018).
20. Traverso, A., Wee, L., Dekker, A. & Gillies, R. Repeatability and reproducibility of radiomic features: a systematic review. *Int. J. Radiat. Oncol. Biol. Phys.* **102**, 1143–1158 (2018).
21. Zwanenburg, A. *et al.* The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**, 328–338 (2020).

22. Welch, M. L. *et al.* Vulnerabilities of radiomic signature development: the need for safeguards. *Radiother. Oncol.* **130**, 2–9 (2019).

23. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).

24. Orlhac, F. *et al.* Tumor texture analysis in 18f-fdg pet: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *J. Nucl. Medicine* **55**, 414–422 (2014).

25. Ellingson, B. M. *et al.* Validation of postoperative residual contrast-enhancing tumor volume as an independent prognostic factor for overall survival in newly diagnosed glioblastoma. *Neuro-oncology* **20**, 1240–1250 (2018).

26. Awad, A.-W. *et al.* Impact of removed tumor volume and location on patient outcome in glioblastoma. *J. neuro-oncology* **135**, 161–171 (2017).

27. Agn, M. *et al.* A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning. *Med. Image Analysis* **54**, 220–237 (2019).

28. Fischl, B. Freesurfer. *Neuroimage* **62**, 774–781 (2012).

29. Puonti, O., Iglesias, J. E. & Van Leemput, K. Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling. *NeuroImage* **143**, 235–249 (2016).

30. Van Leemput, K. Encoding probabilistic brain atlases using bayesian inference. *IEEE Transactions on Med. Imaging* **28**, 822–837 (2008).

31. Lee, H., Grosse, R., Ranganath, R. & Ng, A. Y. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Commun. ACM* **54**, 95–103 (2011).

32. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

33. Cerri, S. *et al.* A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis. *NeuroImage* **225**, 117471, DOI: https://doi.org/10.1016/j.neuroimage.2020.117471 (2021).

34. Huttenlocher, D. P., Klanderman, G. A. & Rucklidge, W. J. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis machine intelligence* **15**, 850–863 (1993).

35. Menze, B. H. *et al.* The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging* **34**, 1993–2024 (2014).

36. Gorlia, T. *et al.* New prognostic factors and calculators for outcome prediction in patients with recurrent glioblastoma: a pooled analysis of eortc brain tumour group phase i and ii clinical trials. *Eur. journal cancer* **48**, 1176–1184 (2012).

37. Chaichana, K. L. *et al.* Relationship of glioblastoma multiforme to the lateral ventricles predicts survival following tumor resection. *J. neuro-oncology* **89**, 219–224 (2008).

38. Abou Jaoude, D. *et al.* Glioblastoma and increased survival with longer chemotherapy duration. *Kansas J. Medicine* **12**, 65 (2019).

39. Yersal, Ö. Clinical outcome of patients with glioblastoma multiforme: Single center experience. *J. Oncol. Sci.* **3**, 123–126 (2017).

40. Cox, D. R. Regression models and life-tables. *J. Royal Stat. Soc. Ser. B (Methodological)* **34**, 187–202 (1972).

41. Davidson-Pilon, C. lifelines: survival analysis in python. *J. Open Source Softw.* **4**, 1317, DOI: 10.21105/joss.01317 (2019).

42. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S. *et al.* Random survival forests. *The annals applied statistics* **2**, 841–860 (2008).

43. Pölsterl, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *J. Mach. Learn. Res.* **21**, 1–6 (2020).

44. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

45. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).

46. Bakas, S. *et al.* Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629* (2018).

47. Bakas, S. *et al.* Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Sci. data* **4**, 1–13 (2017).

48. Bakas, S. *et al.* Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. the cancer imaging archive. *Nat Sci Data* **4**, 170117 (2017).

49. Bakas, S. *et al.* Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *The cancer imaging archive* **286** (2017).

50. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *Jama* **247**, 2543–2546 (1982).

51. Contal, C. & O'Quigley, J. An application of changepoint methods in studying the effect of age on survival in breast cancer. *Comput. statistics & data analysis* **30**, 253–270 (1999).

52. Mantel, N. *et al.* Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* **50**, 163–170 (1966).

53. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *J. Am. statistical association* **53**, 457–481 (1958).

54. Sashegyi, A. & Ferry, D. On the interpretation of the hazard ratio and communication of survival benefit. *The oncologist* **22**, 484 (2017).

Author contributions statement

S.P., K.V.L. and I.L. conceived and planned the experiments; S.P carried out the experiments; S.P took the lead in writing the manuscript, closely collaborating with I.L and K.V.L, who supervised the project; S.C. contributed to the design and implementation of the methods; H.P. and T.U. carried out data acquisition and preparation; All authors reviewed the manuscript and provided feedback.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 765148, as well as from the National Institute Of Neurological Disorders and Stroke under project number R01NS112161.

Additional Information

Competing Interests

The authors declare no competing interests.

Appendix A

In the interest of conciseness, selection of the proposed Hd95 features was performed on the entire cohort in our experiments, i.e., outside of the cross-validation set-up. While this potentially introduces information leakage between the training and test data within each fold, here we show that the results are only minimally affected in practice. Specifically, we ran our experiments again, selecting the features *within* each fold this time, and recording the number of folds each feature was selected in. Fig. 7 shows the frequencies (proportion of the cross-validation folds) of selected features – also shown is a color indicating whether the features were selected on the entire dataset or not. As can be seen from these results, the feature selection is largely consistent across folds, and in alignment with the feature selection performed on the entire cohort.

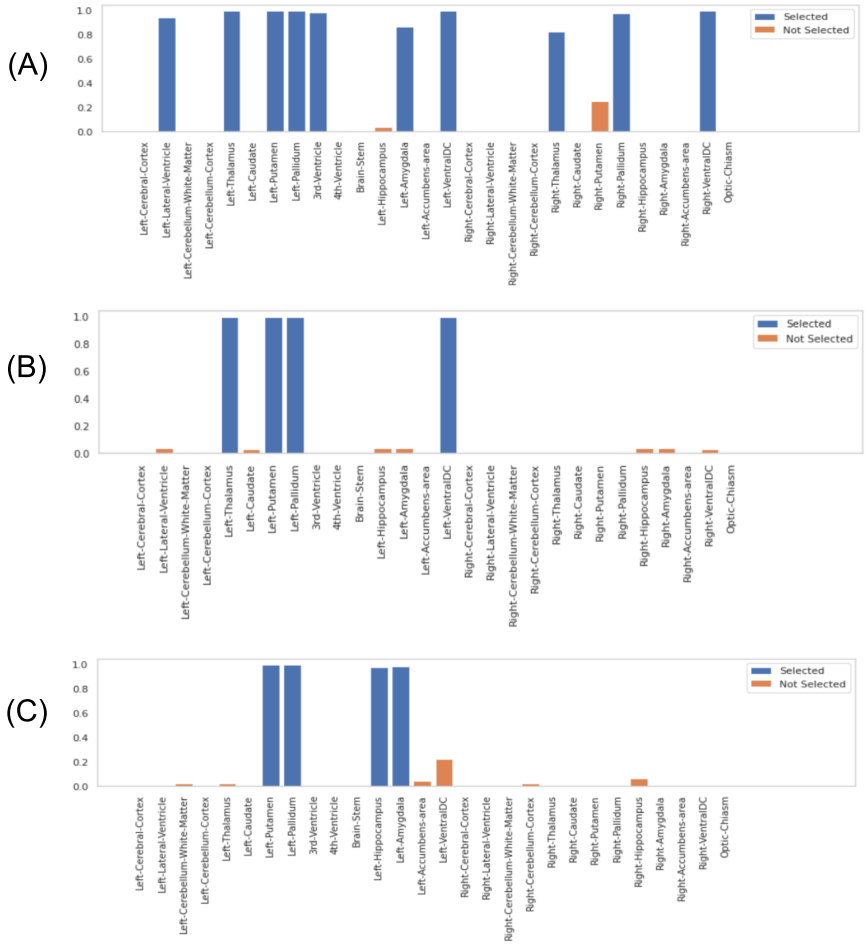

Figure 7. Frequency with which Hd95 features were selected across cross-validation folds on: (A) the Copenhagen data (OS), (B) the Copenhagen data (PFS), and (C) the BraTS20 data (OS). The colors indicate whether the features were also selected when a global feature selection was performed on the entire dataset instead.

Chapter 10

Paper C

Prediction of MGMT Methylation Status of Glioblastoma using Radiomics and Latent Space Shape Features

Sveinn Pálsson[1], Stefano Cerri[1], and Koen Van Leemput[1,2]

[1] Department of Health Technology, Technical University of Denmark, Denmark
[2] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, USA

Abstract. In this paper we propose a method for predicting the status of MGMT promoter methylation in high-grade gliomas. From the available MR images, we segment the tumor using deep convolutional neural networks and extract both radiomic features and shape features learned by a variational autoencoder. We implemented a standard machine learning workflow to obtain predictions, consisting of feature selection followed by training of a random forest classification model. We trained and evaluated our method on the RSNA-ASNR-MICCAI BraTS 2021 challenge dataset and submitted our predictions to the challenge.

Keywords: MGMT prediction · radiomics · deep learning · glioblastoma · variational autoencoder

1 Introduction

Expression of O^6-methylguanine-DNA-methyltransferase (MGMT) in glioblastoma is of clinical importance as it has implications of the patient's overall survival [1, 2]. The prognostic information of MGMT is believed to be due to resistance of tumors with unmethylated MGMT promoter to Temozolomide [3, 4], a drug used in standard therapy [5]. Inference of the MGMT status in the clinic is done by histological analysis, as currently available non-invasive techniques are still too unreliable.

The RSNA-ASNR-MICCAI BraTS 2021 challenge [6–11] contains two tasks: tumor segmentation and MGMT methylation prediction from pre-operative magnetic resonance (MR) images. The challenge organizers have released a large dataset with the goal of facilitating comparison between methods and advancing state-of-the-art methods in these domains. In this paper we focus on the prediction task only.

Radiomics [12] is a method for extracting features from MR images. The features, called "radiomic" features are a variety of statistical, shape and texture features, extracted from a target region within an MR image. Radiomics has gained much interest for prediction tasks related to brain tumors [13] and has been successfully applied to MGMT methylation prediction [14, 15].

We propose a method for inference of the MGMT methylation that combines the use of radiomics with shape features learned by a variational autoencoder (VAE) [16]. VAE, implemented with deep neural networks, can learn high level features that are specific to the data structure it is trained on. By training the VAE on tumor segmentations, we may be able to extract complex tumor shape features that radiomics does not include. Combining hand-crafted features with a learned latent representation of medical images for classification has been previously studied [17], showing improved model classification performance.

The paper is structured as follows: In Section 2 we describe our methods in detail. In Section 3, we present our results and in Section 4 we discuss our results and conclude.

2 Methods

In this section we give a detailed description of our methods (illustrated in Fig. 1). We start by describing the datasets we use. We then describe our pre-processing pipeline, consisting of DICOM to NIFTI conversion, bias-correction and registration. Next, we describe how we obtain tumor segmentations, and how we use these segmentations to compute radiomics and latent shape features.

Finally, we describe the classification model.

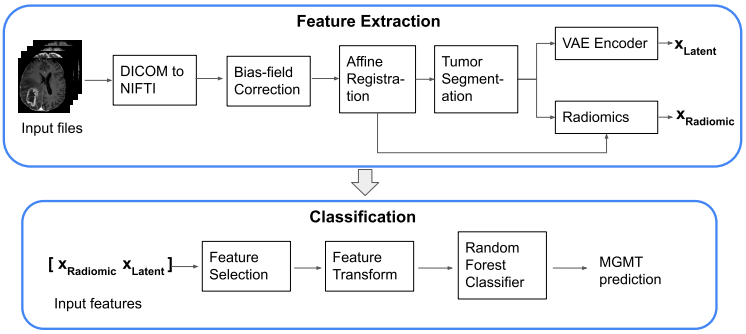

Fig. 1. Overview of our method. The figure shows the main components involved in going from input images to MGMT methylation prediction.

2.1 Data

The challenge data consists of pre-operative MR images of 2000 subjects, divided into training, validation and testing cohorts [6]. For the segmentation task, 1251 subjects are provided with ground truth labels for training segmentation models, whereas for the classification task, ground truth MGMT labels are provided for 585 of those subjects. The testing cohort is unavailable but our methods will be tested on it once it is released. Validation data for the classification task consists of image data for 87 subjects that are provided without ground truth labels but they can be used to evaluate model performance by submitting predictions to the challenge's online platform [18].

For every subject, the available modalities are T1 weighted, post-contrast T1 weighted (Gadolinium), T2 weighted and T2-FLAIR (Fig. 2 (A-D)). A detailed description of the data and pre-processing applied to it by the challenge organizers is given in [6]. The segmentation task dataset has been registered to a standard template and is provided as NIFTI files while the classification data are not co-registered and are provided as DICOM files.

2.2 Pre-processing

Our pre-processing pipeline starts with conversion of the provided DICOM files to NIFTI (implemented in python [19]). Bias field correction is then performed using N4 bias field correction implemented in SimpleITK [20]. We then register the T1 image to a template T1 image and subsequently register the other modalities to the newly registered T1 image. The template we use is the T1 image of a subject (id='00001') in the BraTS21 segmentation challenge. Affine registration was performed using the ANTs registration tool (implemented in python [21]).

2.3 Tumor segmentation

As mentioned in Section 2.1, the tumor segmentation challenge provides 1251 images for training a segmentation model. To get accurate segmentations for further analysis, we use the winning method of the BraTS 2020 challenge [22, 23], an ensemble of deep convolutional neural networks with "U-Net" [24] style architecture, which we train on the whole set of 1251 images. The model is trained to segment three different tumor components; enhancing core, non-enhancing core and edema. The resulting model achieves high segmentation performance (a representative sample is shown in Fig. 2 (F)). The resulting model is used to segment the images provided for the classification task.

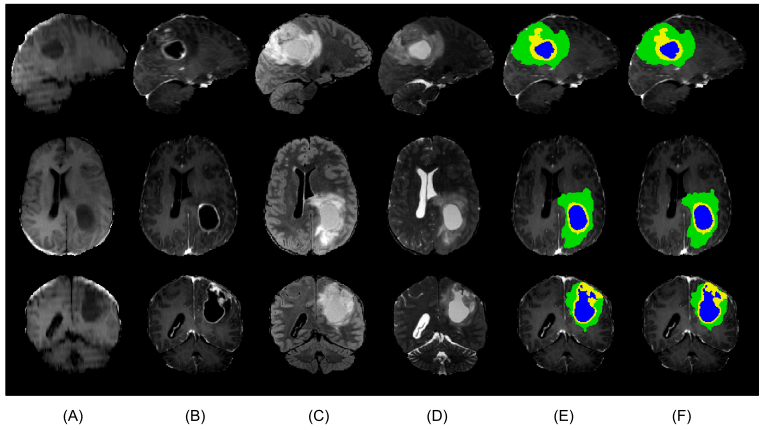

Fig. 2. The figure shows an example from the challenge dataset. From top to bottom: sagittal, axial and coronal view. Columns show (A) T1w, (B) T1c, (C) FLAIR, (D) T2w, (E) ground truth tumor segmentation, (F) automatic segmentation.

2.4 Latent shape features

We obtain our latent shape features from a variational autoencoder (VAE) [16], implemented with 3D convolutional neural networks in tensorflow [25]. The VAE model consists of two networks; a decoder, designed to generate tumor segmentations from latent variables; and encoder, to infer latent variables when given tumor segmentations. The input to the encoder network is a segmentation with size (240,240,155,4) where the last dimension is a one-hot encoding of the tumor component (or background) present at each voxel. The encoder network consists of 3 convolutional network blocks, followed by two fully connected layers. Each block consists of 2 convolutional layers followed by a max pooling layer. The decoder network has a symmetrical architecture to the encoder, where the convolutional layers are replaced with deconvolutional layers [26]. After each convolutional layer in both networks, a leaky ReLU [27] activation is applied, except at the last layer of the decoder whose output is interpreted as logits. The VAE is trained using the ADAM optimization algorithm [28].

We train the VAE on the 1251 available segmentations from the segmentation training dataset. To extract features from a given tumor segmentation, it is passed through the encoder network and its output is taken as the latent features. We set the number of latent features to 64.

2.5 Radiomics

We extract radiomic features from three automatically segmented tumor regions and from each provided modality, resulting in a total of 1172 extracted radiomic features. The radiomic features comprise seven categories: first-order statistics, shape descriptors, gray level co-occurrence matrix (GLCM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), gray level dependece matrix (GLDM), and neighboring gray tone difference matrix (NGTDM). We use the PyRadiomics [29] python implementation of radiomics for the feature extraction.

The three tumor regions we consider is the whole tumor, enhancing core and non-enhancing core. The whole tumor is the union of all the three tumor components that are segmented.

2.6 Classification

After obtaining all of our features, in the next step we perform feature preprocessing to standardize feature values and feature selection to reduce dimensionality. We then train a random forest classifier on the selected features.

For each feature, we search for a threshold value that best splits the subjects in terms of the target variable. Specifically, for each candidate threshold value, we perform a Fisher's exact test [30], testing the hypothesis that the binomial distributions (over the target variable) of the two resulting groups are the same. The feature value resulting in the lowest P-value is chosen as the threshold for that feature. The features are subsequently transformed to binary variables according to which side of the threshold they land. Features are then selected if the P-value of the best threshold is $P < P_{\text{min}}$, where P_{min} is experimentally chosen. The selected features and choice of P_{min} will be discussed further in Section 3.1.

We use a random forest [31] (implemented in python [32]) to obtain predictions of MGMT methylation status, given the input features we extracted. The model is trained on the 585 available subjects via K-fold cross validation, with K chosen such that in each fold, 5 subjects are held out while the remaining subjects are used to train a model ($K = 117$ in our case). In each fold, the model is trained on 580 subjects and predictions on the 5 held-out subjects are obtained. Once predictions are obtained for all subjects, a performance score is calculated. The performance score we use is the area under the receiver operating characteristic curve (AUC). Using grid search, we tune two hyperparameters of the model; the number of samples to split a node and maximum depth of trees. At test time, given an unseen subject, the 117 models are all used to predict the MGMT methylation status, each predicting either 0 or 1 for the unmethylated or methylated group, respectively. The average of the predictions is interpreted as the probability of belonging to the methylated group.

3 Results

3.1 Feature Selection

The number of features selected by the selection procedure described in Section 2.6 depends on our choice of $P_{\min}$, which we experimentally determine by searching a range of values and measuring model performance using the whole training cohort. We set $P_{\min} = 5 \times 10^{-4}$ which leaves 23 features remaining; 16 of which are radiomic features and 7 latent shape features. The list of selected radiomic features is given in Table 1. We observe selected radiomic features from 6 out of 7 categories mentioned in 2.5, from 3 out of 4 modalities and from all 3 tumor regions.

Table 1. List of selected radiomic features.

Category	Feature name	Modality	Region
Shape	Maximum 3D Diameter	-	Enh-core
First order	Interquartile Range	T1-ce	Core
First order	Mean Absolute Deviation	T1-ce	Core
First order	Mean	T1-ce	Core
First order	Median	T1-ce	Core
First order	Median	T1-ce	Whole
First order	Variance	T1-ce	Core
First order	10Percentile	FLAIR	Core
GLRLM	Graylevel non-uniformity normalized	FLAIR	Whole
GLRLM	Graylevel variance	FLAIR	Whole
GLSZM	Small area emphasis	FLAIR	Whole
GLSZM	Small area high graylevel emphasis	FLAIR	Whole
GLSZM	Small area low graylevel emphasis	FLAIR	Whole
NGTDM	Busyness	FLAIR	Whole
GLSZM	Small area high graylevel emphasis	T2	Whole

3.2 Classification

We find the best hyperparameters for the random forest through grid search to be 2 samples to split a node and a maximum tree depth of 4 (we leave other parameters as default). The whole training cohort is used for the hyperparameter search.

To test the benefit of using the latent features in the model along with the radiomic features, we train the model on both feature sets separately and together and measure the AUC score. For a more accurate performance measure on the training dataset, we ran our cross validation 10 times (each time the dataset is shuffled) and in Table 2, we report the mean AUC score across the 10 iterations. The true labels of the validation dataset are unknown to us, but by submitting our predictions to the challenge platform, we obtain a validation

AUC score reported in Table 2. We observe a substantial disagreement between the training and validation scores: the training results show improvement with the combination of feature sets, while the validation scores indicate that using radiomics alone is preferred and that the latent shape features have very low predictive value.

Table 2. Classification performance measured by AUC. For three feature sets, the table shows AUC score for both cross-validated training set predictions and predictions on the validation set.

Features	Training	Validation
Radiomics + Latent	0.603	0.598
Radiomics	0.582	0.632
Latent	0.568	0.488

4 Discussion

In this paper, we propose a method for MGMT methylation prediction that combines the use of radiomics with high level shape features learned by a variational autoencoder. We train a segmentation model to obtain tumor segmentations, and train a variational autoencoder on segmentations to learn high-level shape features of tumor. We use the tumor segmentation to compute radiomic features, and pass the segmentation to the encoder network of the variational autoencoder to obtain shape features from its latent space. We extracted these features from the training data provided by the RSNA-ASNR-MICCAI BraTS 2021 challenge and trained a random forest classifier. The method was submitted to the challenge and obtained a validation score (AUC) of 0.598.

As we discussed in Section 1, radiomic features have already been shown to be applicable to this prediction task while tumor shape has not been proven to predict MGMT methylation. Therefore, to test whether the feature set combination we propose performs better than simply using the radiomic features alone, we experiment with training the classifier on them separately. On our training data, we observe a performance benefit of using the shape features (cf. Table 2), but this is not reproduced on the validation set where the radiomic features alone achieve a score of 0.632 but the latent features only 0.488. This may be due to overfitting of our feature transform and hyperparameter selection to the training data or high uncertainty stemming from the small number of samples in the validation dataset. We hope to gain more insight by submitting our method to the testing phase of the challenge, which contains a substantially larger number of subjects.

5 Acknowledgements

This project was funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie project TRABIT (agreement No 765148) and NINDS (grant No R01NS112161).

References

1. Signe Regner Michaelsen, Ib Jarle Christensen, Kirsten Grunnet, Marie-Thérése Stockhausen, Helle Broholm, Michael Kosteljanetz, and Hans Skovgaard Poulsen. Clinical variables serve as prognostic factors in a model for survival from glioblastoma multiforme: an observational study of a cohort of consecutive non-selected patients from a single institution. *BMC cancer*, 13(1):402, 2013.
2. Thierry Gorlia, Martin J van den Bent, Monika E Hegi, René O Mirimanoff, Michael Weller, J Gregory Cairncross, Elizabeth Eisenhauer, Karl Belanger, Alba A Brandes, Anouk Allgeier, et al. Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of eortc and ncic trial 26981-22981/ce. 3. *The lancet oncology*, 9(1):29–38, 2008.
3. Monika E Hegi, Annie-Claire Diserens, Thierry Gorlia, Marie-France Hamou, Nicolas De Tribolet, Michael Weller, Johan M Kros, Johannes A Hainfellner, Warren Mason, Luigi Mariani, et al. Mgmt gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine*, 352(10):997–1003, 2005.
4. Gaspar J Kitange, Brett L Carlson, Mark A Schroeder, Patrick T Grogan, Jeff D Lamont, Paul A Decker, Wenting Wu, C David James, and Jann N Sarkaria. Induction of mgmt expression is associated with temozolomide resistance in glioblastoma xenografts. *Neuro-oncology*, 11(3):281–291, 2009.
5. Roger Stupp, Monika E Hegi, Warren P Mason, Martin J Van Den Bent, Martin JB Taphoorn, Robert C Janzer, Samuel K Ludwin, Anouk Allgeier, Barbara Fisher, Karl Belanger, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase iii study: 5-year analysis of the eortc-ncic trial. *The lancet oncology*, 10(5):459–466, 2009.
6. Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.
7. 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, 2014.
8. Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.
9. 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(1):1–13, 2017.
10. Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *The cancer imaging archive*, 286, 2017.
11. Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Seg-

mentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. the cancer imaging archive. *Nat Sci Data*, 4:170117, 2017.
12. 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.
13. Thomas C Booth, Matthew Williams, Aysha Luis, Jorge Cardoso, Keyoumars Ashkan, and Haris Shuaib. Machine learning and glioma imaging biomarkers. *Clinical radiology*, 75(1):20–32, 2020.
14. 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.
15. Zhi-Cheng Li, Hongmin Bai, Qiuchang Sun, Qihua Li, Lei Liu, Yan Zou, Yinsheng Chen, Chaofeng Liang, and Hairong Zheng. Multiregional radiomics features from multiparametric mri for prediction of mgmt methylation status in glioblastoma multiforme: a multicentre study. *European radiology*, 28(9):3640–3650, 2018.
16. Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
17. Sunan Cui, Yi Luo, Huan-Hsin Tseng, Randall K Ten Haken, and Issam El Naqa. Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. *Medical physics*, 46(5):2497–2511, 2019.
18. Rsna-miccai brain tumor radiogenomic classification challange. `https://www.kaggle.com/c/rsna-miccai-brain-tumor-radiogenomic-classification/`. Accessed: 2021-08-10.
19. dicom2nifti. `https://github.com/icometrix/dicom2nifti`. Accessed: 2021-08-10.
20. Richard Beare, Bradley Lowekamp, and Ziv Yaniv. Image segmentation, registration and characterization in r with simpleitk. *Journal of statistical software*, 86, 2018.
21. Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2(365):1–35, 2009.
22. Fabian Isensee and Klaus H Maier-Hein. nnu-net for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II*, volume 12658, page 118. Springer Nature, 2021.
23. Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
24. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
25. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster,

Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

26. Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
27. Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.
28. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
29. Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.
30. Ronald A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.
31. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
32. Fabian Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abou Jaoude, D., Moore, J. A., Moore, M. B., Twumasi-Ankrah, P., Ablah, E., and Moore Jr, D. F. (2019). Glioblastoma and increased survival with longer chemotherapy duration. *Kansas Journal of Medicine*, 12(3):65.

Agn, M., af Rosenschöld, P. M., Puonti, O., Lundemann, M. J., Mancini, L., Papadaki, A., Thust, S., Ashburner, J., Law, I., and Van Leemput, K. (2019). A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning. *Medical Image Analysis*, 54:220–237.

Agravat, R. R. and Raval, M. S. (2019). Brain tumor segmentation and survival prediction. In *International MICCAI Brainlesion Workshop*, pages 338–348. Springer.

Ashburner, J., Andersson, J. L., and Friston, K. J. (2000). Image registration using a symmetric prior - In three dimensions. *Human Brain Mapping*, 9(4):212–225.

Ashburner, J. and Friston, K. (1997). Multimodal image coregistration and partitioning—a unified framework. *Neuroimage*, 6(3):209–217.

Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *Neuroimage*, 26(3):839–851.

Bae, S., Choi, Y. S., Ahn, S. S., Chang, J. H., Kang, S.-G., Kim, E. H., Kim, S. H., and Lee, S.-K. (2018). Radiomic mri phenotyping of glioblastoma: improving survival prediction. *Radiology*, 289(3):797–806.

Baid, U., Ghodasara, S., Bilello, M., Mohan, S., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F. C., Pati, S., et al. (2021). The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*.

Baid, U., Rane, S. U., Talbar, S., Gupta, S., Thakur, M. H., Moiyadi, A., and Mahajan, A. (2020). Overall survival prediction in glioblastoma with radiomic features using machine learning. *Frontiers in computational neuroscience*, 14:61.

Baid, U., Talbar, S., Rane, S., Gupta, S., Thakur, M. H., Moiyadi, A., Thakur, S., and Mahajan, A. (2018). Deep learning radiomics algorithm for gliomas (drag) model: a novel approach using 3d unet based deep convolutional neural network for predicting survival in gliomas. In *International MICCAI Brainlesion Workshop*, pages 369–379. Springer.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*.

Brant-Zawadzki, M., Gillan, G. D., and Nitz, W. R. (1992). Mp rage: a three-dimensional, t1-weighted, gradient-echo sequence–initial experience in the brain. *Radiology*, 182(3):769–775.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statist. Sci.*, 16(2):101–133.

Cerri, S., Puonti, O., Meier, D. S., Wuerfel, J., Mühlau, M., Siebner, H. R., and Van Leemput, K. (2021). A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis. *NeuroImage*, 225:117471.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cui, S., Luo, Y., Tseng, H.-H., Ten Haken, R. K., and El Naqa, I. (2019). Combining handcrafted features with latent variables in machine learning for prediction of radiation-induced lung damage. *Medical physics*, 46(5):2497–2511.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dosovitskiy, A., Tobias Springenberg, J., and Brox, T. (2015). Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fathi Kazerooni, A., Bakas, S., Saligheh Rad, H., and Davatzikos, C. (2020). Imaging signatures of glioblastoma molecular characteristics: A radiogenomics review. *Journal of Magnetic Resonance Imaging*, 52(1):54–69.

Fischl, B. (2012). Freesurfer. *Neuroimage*, 62(2):774–781.

Gallego, O. (2015). Nonsurgical treatment of recurrent glioblastoma. *Current oncology*, 22(4):273–281.

Gamburg, E. S., Regine, W. F., Patchell, R. A., Strottmann, J. M., Mohiuddin, M., and Young, A. B. (2000). The prognostic significance of midline shift at presentation on survival in patients with glioblastoma multiforme. *International Journal of Radiation Oncology* Biology* Physics*, 48(5):1359–1362.

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016). Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577.

Goodenberger, M. L. and Jenkins, R. B. (2012). Genetics of adult glioma. *Cancer genetics*, 205(12):613–621.

Gorlia, T., van den Bent, M. J., Hegi, M. E., Mirimanoff, R. O., Weller, M., Cairncross, J. G., Eisenhauer, E., Belanger, K., Brandes, A. A., Allgeier, A., et al. (2008). Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of eortc and ncic trial 26981-22981/ce. 3. *The lancet oncology*, 9(1):29–38.

Gutman, D. A., Cooper, L. A., Hwang, S. N., Holder, C. A., Gao, J., Aurora, T. D., Dunn Jr, W. D., Scarpace, L., Mikkelsen, T., Jain, R., et al. (2013). Mr imaging predictors of molecular profile and survival: multi-institutional study of the tcga glioblastoma data set. *Radiology*, 267(2):560–569.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546.

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

Hegi, M. E., Diserens, A.-C., Gorlia, T., Hamou, M.-F., De Tribolet, N., Weller, M., Kros, J. M., Hainfellner, J. A., Mason, W., Mariani, L., et al. (2005). Mgmt gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine*, 352(10):997–1003.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863.

Iliadis, G., Kotoula, V., Chatzisotiriou, A., Televantou, D., Eleftheraki, A. G., Lambaki, S., Misailidou, D., Selviaridis, P., and Fountzilas, G. (2012). Volumetric and mgmt parameters in glioblastoma patients: survival analysis. *BMC cancer*, 12(1):3.

Ingrisch, M., Schneider, M. J., Nörenberg, D., de Figueiredo, G. N., Maier-Hein, K., Suchorska, B., Schüller, U., Albert, N., Brückmann, H., Reiser, M., et al. (2017). Radiomic analysis reveals prognostic information in t1-weighted baseline magnetic resonance imaging in patients with glioblastoma. *Investigative radiology*, 52(6):360–366.

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2017). Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In *International MICCAI Brainlesion Workshop*, pages 287–297. Springer.

Isensee, F. and Maier-Hein, K. H. (2021). nnu-net for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II*, volume 12658, page 118. Springer Nature.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). Random survival forests. *The annals of applied statistics*, 2(3):841–860.

Jang, E., Gu, S., and Poole, B. (2016). Categorical Reparameterization with Gumbel-Softmax. *arXiv e-prints*, page arXiv:1611.01144.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841.

Jessen, K. R. and Mirsky, R. (1980). Glial cells in the enteric nervous system contain glial fibrillary acidic protein. *Nature*, 286(5774):736–737.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. (2014). Semi-Supervised Learning with Deep Generative Models. *arXiv e-prints*, page arXiv:1406.5298.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kitange, G. J., Carlson, B. L., Schroeder, M. A., Grogan, P. T., Lamont, J. D., Decker, P. A., Wu, W., James, C. D., and Sarkaria, J. N. (2009). Induction of mgmt expression is associated with temozolomide resistance in glioblastoma xenografts. *Neuro-oncology*, 11(3):281–291.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-normalizing neural networks. In *Advances in neural information processing systems*, pages 971–980.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R. G., Granton, P., Zegers, C. M., Gillies, R., Boellard, R., Dekker, A., et al. (2012). Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446.

Lauko, A., Lo, A., Ahluwalia, M. S., and Lathia, J. D. (2021). Cancer cell heterogeneity & plasticity in glioblastoma and brain tumors. In *Seminars in Cancer Biology*. Elsevier.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2011). Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, 54(10):95–103.

Li, Z.-C., Bai, H., Sun, Q., Li, Q., Liu, L., Zou, Y., Chen, Y., Liang, C., and Zheng, H. (2018). Multiregional radiomics features from multiparametric mri for prediction of mgmt methylation status in glioblastoma multiforme: a multicentre study. *European radiology*, 28(9):3640–3650.

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., Scheithauer, B. W., and Kleihues, P. (2007). The 2007 who classification of tumours of the central nervous system. *Acta neuropathologica*, 114(2):97–109.

Lundemann, M., af Rosenschöld, P. M., Muhic, A., Larsen, V. A., Poulsen, H. S., Engelholm, S.-A., Andersen, F. L., Kjær, A., Larsson, H. B., Law, I., et al. (2019). Feasibility of multi-parametric pet and mri for prediction of tumour recurrence in patients with glioblastoma. *European journal of nuclear medicine and molecular imaging*, 46(3):603–613.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. (2014). The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024.

Michaelsen, S. R., Christensen, I. J., Grunnet, K., Stockhausen, M.-T., Broholm, H., Kosteljanetz, M., and Poulsen, H. S. (2013). Clinical variables serve as prognostic factors in a model for survival from glioblastoma multiforme: an observational study of a cohort of consecutive non-selected patients from a single institution. *BMC cancer*, 13(1):402.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Icml*.

Narang, S., Lehrer, M., Yang, D., Lee, J., and Rao, A. (2016). Radiomics in glioblastoma: current status, challenges and potential opportunities. *Translational Cancer Research*, 5(4):383–397.

Nuechterlein, N. and Mehta, S. (2018). 3d-espnet with pyramidal refinement for volumetric brain tumor image segmentation. In *International MICCAI Brainlesion Workshop*, pages 245–253. Springer.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS-W*.

Pirotte, B. J., Levivier, M., Goldman, S., Massager, N., Wikler, D., Dewitte, O., Bruneau, M., Rorive, S., David, P., and Brotchi, J. (2009). Positron emission tomography-guided volumetric resection of supratentorial high-grade gliomas: a survival analysis in 66 consecutive patients. *Neurosurgery*, 64(3):471–481.

Poulsen, S. H., Urup, T., Grunnet, K., Christensen, I. J., Larsen, V. A., Jensen, M. L., af Rosenschöld, P. M., Poulsen, H. S., and Law, I. (2017). The prognostic value of fet pet at radiotherapy planning in newly diagnosed glioblastoma. *European journal of nuclear medicine and molecular imaging*, 44(3):373–381.

Puonti, O., Iglesias, J. E., and Van Leemput, K. (2016). Fast and sequence-adaptive whole-brain segmentation using parametric Bayesian modeling. *NeuroImage*, 143:235–249.

Pérez Beteta, J., Molina, D., Ortiz-Alhambra, J., Fernández-Romero, A., Luque, B., Arregui, E., Calvo, M., Borrás, J., Meléndez, B., Rodríguez de Lope, A., Presa, R., Bayo, L., Barcia, J., Martino, J., Velásquez, C., Asenjo, B., Benavides, M., Herruzo, I., Revert, A., and Pérez-García, V. (2018). Tumor surface regularity at mr imaging predicts survival and response to surgery in patients with glioblastoma. *Radiology*, 288:218–225.

Rathore, S., Akbari, H., Doshi, J., Shukla, G., Rozycki, M., Bilello, M., Lustig, R. A., and Davatzikos, C. A. (2018). Radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma: implications for personalized radiotherapy planning. *Journal of Medical Imaging*, 5(2):021219.

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

Shaffer, R., Nichol, A. M., Vollans, E., Fong, M., Nakano, S., Moiseenko, V., Schmuland, M., Ma, R., McKenzie, M., and Otto, K. (2010). A comparison of volumetric modulated arc therapy and conventional intensity-modulated radiotherapy for frontal and temporal high-grade gliomas. *International Journal of Radiation Oncology* Biology* Physics*, 76(4):1177–1184.

Shboul, Z. A., Vidyaratne, L., Alam, M., and Iftekharuddin, K. M. (2017). Glioblastoma and survival prediction. In *International MICCAI Brainlesion Workshop*, pages 358–368. Springer.

Shortliffe, E. H. and Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *Jama*, 320(21):2199–2200.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Stupp, R., Hegi, M. E., Mason, W. P., Van Den Bent, M. J., Taphoorn, M. J., Janzer, R. C., Ludwin, S. K., Allgeier, A., Fisher, B., Belanger, K., et al. (2009). Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase iii study: 5-year analysis of the eortc-ncic trial. *The lancet oncology*, 10(5):459–466.

Suchorska, B., Jansen, N. L., Linn, J., Kretzschmar, H., Janssen, H., Eigenbrod, S., Simon, M., Pöpperl, G., Kreth, F. W., la Fougere, C., et al. (2015). Biological tumor volume in 18fet-pet before radiochemotherapy correlates with survival in gbm. *Neurology*, 84(7):710–719.

Sun, L., Zhang, S., Chen, H., and Luo, L. (2019). Brain tumor segmentation and survival prediction using multimodal mri scans with deep learning. *Frontiers in neuroscience*, 13:810.

Traverso, A., Wee, L., Dekker, A., and Gillies, R. (2018). Repeatability and reproducibility of radiomic features: a systematic review. *International Journal of Radiation Oncology* Biology* Physics*, 102(4):1143–1158.

Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J.-C., Pieper, S., and Aerts, H. J. (2017). Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107.

Van Leemput, K. (2009). Encoding probabilistic brain atlases using Bayesian inference. *IEEE Transactions on Medical Imaging*, 28(6):822–837.

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999). Automated model-based bias field correction of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):885–896.

Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., Carter, A., Casey, D. C., Charlson, F. J., Chen, A. Z., et al. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *The lancet*, 388(10053):1545–1602.

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI brainlesion workshop*, pages 178–190. Springer.

Wang, H., Naghavi, M., Allen, C., Barber, R. M., Bhutta, Z. A., Carter, A., Casey, D. C., Charlson, F. J., Chen, A. Z., Coates, M. M., et al. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015. *The lancet*, 388(10053):1459–1544.

Wang, Q., Li, Q., Mi, R., Ye, H., Zhang, H., Chen, B., Li, Y., Huang, G., and Xia, J. (2019). Radiomics nomogram building from multiparametric mri to predict grade in patients with glioma: a cohort study. *Journal of Magnetic Resonance Imaging*, 49(3):825–833.

Welch, M. L., McIntosh, C., Haibe-Kains, B., Milosevic, M. F., Wee, L., Dekker, A., Huang, S. H., Purdie, T. G., O'Sullivan, B., Aerts, H. J., et al. (2019). Vulnerabilities of radiomic signature development: the need for safeguards. *Radiotherapy and Oncology*, 130:2–9.

Wells, W. M., Grimson, W. E. L., Kikinis, R., and Jolesz, F. A. (1996). Adaptive segmentation of mri data. *IEEE transactions on medical imaging*, 15(4):429–442.

Weninger, L., Haarburger, C., and Merhof, D. (2019). Robustness of radiomics for survival prediction of brain tumor patients depending on resection status. *Frontiers in computational neuroscience*, 13:73.

Wesseling, P. and Capper, D. (2018). Who 2016 classification of gliomas. *Neuropathology and applied neurobiology*, 44(2):139–150.

Wester, H. J., Herz, M., Weber, W., Heiss, P., Senekowitsch-Schmidtke, R., Schwaiger, M., and Stöcklin, G. (1999). Synthesis and radiopharmacology of o-(2-[18f] fluoroethyl)-l-tyrosine for tumor imaging. *Journal of Nuclear Medicine*, 40(1):205–212.

Xi, Y.-b., Guo, F., Xu, Z.-l., Li, C., Wei, W., Tian, P., Liu, T.-t., Liu, L., Chen, G., Ye, J., et al. (2018). Radiomics signature: a potential biomarker for the prediction of mgmt promoter methylation in glioblastoma. *Journal of Magnetic Resonance Imaging*, 47(5):1380–1387.

Yersal, Ö. (2017). Clinical outcome of patients with glioblastoma multiforme: Single center experience. *Journal of Oncological Sciences*, 3(3):123–126.

Yip, S. S. and Aerts, H. J. (2016). Applications and limitations of radiomics. *Physics in Medicine & Biology*, 61(13):R150.

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., et al. (2020). The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338.