Universidade do Minho
Escola de Engenharia

Mariana Neves Ramalho Ferreira Alão

Deep and Transfer Learning Approaches for Glioblastoma Patient Survival Prediction from Pre-treatment MRI

Dissertação de Mestrado
Mestrado Integrado em Engenharia Biomédica
Ramo de Eletrónica Médica

Trabalho efetuado sob a orientação de
Professor Doutor Carlos Alberto Batista Silva
Professor Doutor Mauricio Reyes

dezembro de 2019

Direitos de Autor e Condições de Utilização do Trabalho por Terceiros

Acknowledgement

I want to express my deepest appreciation to my supervisors, Prof. Mauricio Reyes and Prof. Carlos Silva, for the opportunity to work in a topic that fascinates me, for the guidance during this work and for allowing me to grow and learn about scientific research.

I would also like to extend my sincere thanks to Yannick for sharing the passion and experience on this topic and for all the advice, help and practical suggestions provided.

Especially helpful during this time was my family, mainly my mum and dad. Thank you for the constant encouragement and for supporting me with love and understanding.

To the new friends made through these five years, thank you for sharing the good moments, for all the laughs and the memories that will stay forever with me. I am also very grateful for mutual support when the work was too much. The old friends should also be remembered, for the continuous support and strong friendship. Especially, I gratefully acknowledge the patience, support and kindness that André gave me.

Statement of Integrity

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Resumo

Título: Abordagens de *Deep* e *Transfer Learning* para a previsão do tempo de vida em pacientes com Glioblastoma utilizando imagens de RM pré-tratamento

O Glioblastoma Multiforme é um tumor cerebral nocivo com uma sobrevida mediana de apenas catorze semanas. Prever o tempo de sobrevida a partir de imagens de ressonância magnética é intrinsecamente difícil, não apenas devido às características morfológicas, mas também devido ao impacto dos tratamentos. Durante os últimos anos, o *deep learning* começou a despertar o interesse da comunidade científica devido às excelentes performances obtidas. Na área da imagem médica, as redes neuronais convolucionais estão a obter melhores performances que os médicos em diversas tarefas.

Esta dissertação tem como objetivo desenvolver um modelo de regressão de *deep learning* para prever o tempo de sobrevida através de ressonâncias magnéticas pré-tratamento de pacientes com glioblastoma. Para tal, três diferentes abordagens foram utilizadas: a transferência de conhecimento de uma rede neuronal convolucional 2D e de outra 3D e ainda treinar do zero uma rede neuronal convolucional 3D. Os efeitos de introduzir na rede diferentes modalidade de ressonância magnética com diferentes pré-processamentos, tal como a região de interesse considerada ou a normalização padrão efetuada, foram estudados. Após uma primeira seleção dos modelos com melhores métricas baseada em dois critérios distintos foi efetuado o aumento artificial dos dados.

Durante este trabalho, as diferentes abordagens foram analisadas individualmente e posteriormente comparadas entre si e com o estado da arte. Embora não seja possível concluir sobre modalidade de ressonância magnética e o tipo de normalização padrão que beneficiam a previsão do tempo de sobrevida, e preferência por um contexto visual mais abrangente foi evidente. Os mapas de importância relativos às redes neuronais 3D mostram que as regiões consideradas mais importantes pela rede coincidem com as que o tumor é mais frequente. A rede que revelou um maior potencial para prever o tempo de sobrevida é a rede neuronal convolutional 3D treinada do zero que utiliza a sequência FLAIR conjugada com a segmentação do tumor como entrada da rede, a região de interesse é toda a imagem e a normalização padrão é efetuada a toda a sequência. Por fim, este modelo não foi capaz de ultrapassar o estado da arte.

Palavras-chave: Glioblastoma Multiforme, redes neuronais convolucionais, Ressonância Magnética pré-tratamento, tempo de sobrevida, *transfer learning*

Abstract

Title: Deep and Transfer Learning Approaches for Glioblastoma Patient Survival Prediction from Pre-treatment MRI

Glioblastoma Multiforme (GBM) is a harmful brain tumor with a median overall survival (OS) of only fourteen months. Predicting the OS from pre-treatment Magnetic Resonance Imaging (MRI) is intrinsically tricky not only due to the morphologic characteristics of the tumor but also due to the impact of the treatment. In the last years, deep learning began to arouse the interest of the scientific community due to the excellent performances achieved. In the medical imaging field, the convolutional neural networks (CNN) are achieving better performances than the clinicians in several tasks.

This dissertation aims to develop a deep learning model for regression to directly predict the OS using pre-treatment MRI scans from patients with glioblastoma. For this, three different approaches were used: the transference of knowledge from a 2D and a 3D CNN and training from scratch on a 3D CNN. The effects of inputting different MRI modalities with different preprocessing techniques, such as the region of interest (ROI) and z-score normalizations, were studied. After a first selection of the top-ranked models based on two distinct criteria, data augmentation was performed.

Throughout this work, the approaches were analyzed individually and then compared with each other and with a state-of-the-art approach. Although it was not possible to conclude on the MRI modalities and z-score normalization that benefits the most the OS prediction, a preference for a broader context while training the model was evident. The importance maps from the 3D networks developed showed that the regions considered most important by the network overlap with the ones where the tumor is more frequent. The network with more potential for the OS prediction is the 3D CNN trained from scratch that used FLAIR and tumor segmentation as input, ROI as all image and z-score normalization to all the image. In the end, this model was not able to outperform the state of the art.

Key words: CNN, GBM, Glioblastoma overall survival, pre-treatment MRI, transfer learning

Table of Contents

List of Figures

List of Tables

Acronyms

BBB blood-brain barrier.

BraTS Brain Tumor Segmentation.

CNN Convolutional Neural Network.

CNN-S Slow Convolutional Neural Network.

CT Computer Tomography.

FLAIR T2-weighted fluid-attenuated inversion recovery.

GBM Glioblastoma Multiforme.

GPU Graphics Processing Unit.

ICC Intraclass correlation coefficient.

ILSVRC ImageNet Large Scale Visual Recognition Challenge.

IXI Information eXtraction from Images.

KPS Karnofsky Performance Status.

LASSO Least Absolute Shrinkage and Selection Operator.

LRN Local Response Normalization.

LS Long Survivers.

medianSE Median Square Error.

MRI Magnetic Resonance Imaging.

MS Medium Survivers.

MSE Mean Square Error.

NIfTI Neuroimaging Informatics Technology Initiative.

OCCC Overall Concondance Correlation Coeficient.

OS Overall Survival.

ReLU Rectified Linear Unit.

RISE Randomized Input Sampling for Explanation of Black-box Models.

ROI Region of Interest.

SpearmanR Spearman's Correlation Coeficient.

SS Short Survivers.

stdSE Standard-deviation Square Error.

T1 T1-weighted.

T1ce T1-weighted contrast enhanced.

T2 T2-weighted.

WHO World Health Organization.

Chapter 1

Introduction

In this first chapter, the motivation of the work is presented. Then, the objectives of this dissertation are defined, and, to conclude, a summary of the following chapters is provided.

1.1 Motivation

Glioblastoma, also named Glioblastoma Multiforme (GBM), is a harmful brain tumor in humans that arises from the glial cells and represents more than 60% of all brain tumors. This type of high-grade glioma is classified as grade IV on the World Health Organization (WHO) grading scale and it is considered a deadly disease. Due to the inter- and intra-patients' tumor heterogeneity and impact of treatments such as chemo- and radiotherapy and tumor resection, the prediction of the survival time is a challenging task. Despite how the survival can be improved with these treatments, owing to the fast growth of the tumor, short survival time after diagnosis and its incidence, it is fundamental to find pre-treatment flags that give early signs of the prognostic of the disease [1, 2].

Consequently, much investigation has been developed in the last years in order to find those pre-treatment prognostic signals. They are being searched in Magnetic Resonance Images (MRI), genetic maps or based on clinical factors such as patients' age or Karnofsky Performance Status (KPS) factor. Discovering these factors can help the clinicians to plan the following treatment steps by selecting the ones that would benefit more the patient, having a crucial impact on their survival time and quality of life [1–3].

1.2 Objectives

Several studies using pre-treatment MRI scans have been developed in order to predict the OS of patients with glioblastoma. The present work aims to develop a deep learning model to predict the survival time of patients with glioblastoma directly. With this model, it is expected that patterns beyond human knowledge for predicting the OS using pre-treatment MRI scans are discovered. Moreover, the work must lead to a well-understood interpretable neural network. During its development, the application of data augmentation and transfer learning techniques occurs as an attempt to improve the patterns found.

For the development of the work, the combination of knowledge from the medical and computational field is indispensable. In particular, deep learning knowledge for medical image analysis is required, as well as brain anatomy knowledge for a better interpretation of the results.

1.3 Structure of the Thesis

This dissertation has six chapters, besides this one. Chapter 2 starts with a clinical overview of high-grade gliomas, where information about its emergence, morphology, diagnosis and treatment is presented. In the same chapter, machine, deep and transfer learning concepts essential to the development of this work are introduced. The following chapter gives the current state of the art of predicting the survival time of patients with glioblastoma. All the details on the used datasets and methods implemented are presented in Chapter 4 and all the results obtained are presented and discussed in Chapter 5. Here, a comparison with state of the art is also made. Next, the conclusions of the work are shown and future outlooks are suggested in Chapter 6. Finally, there is an attachment section.

Chapter 2

Theoretical Foundations

As soon as it was possible to load images into a computer, the development of systems to automatically analyze the images started to occur. Between the 1970s and 1990s, systems for analyzing the images were designed by humans: low-level pixel processing and mathematical modeling was used to encode clinicians' decisions as rules. At the end of the 1990s, systems that perform the image analysis started to be trained using computers, beeing them that learn the rules for the analysis by themself. For this purpose, machine and deep learning methods have been used since then [4, 5].

This chapter starts with a clinical overview of the glioblastoma, the brain tumor in the scope of this dissertation. Responses to the question of what it is and its appearance, how it emerges and the current treatment are provided. After that, computational concepts used to perform the image analysis related to artificial intelligence are reviewed, in particular machine and deep learning. Some details on a successful type of deep learning models, the convolutional neural networks, are provided. An introduction to transfer learning is also provided. Finally, a brief discussion on the importance of the interpretation of the developed models is done.

2.1 Clinical Context

GBM is a brain cancer that almost certainly leads to neurological demise and dead, with a median survival of fifteen months after diagnostic [2]. It is the most vascularized and mortal type of primary tumor that appears in the central nervous system. This type of tumor arises from glial cells and their name derives from the word *glial*, meaning glue in greek. These nerve cells are not directly part of the synaptic interactions and electrical signaling. However, they have supportive functions that help define synaptic contacts and maintaining the signaling abilities of the neurons. GBM is a highly locally invasive tumor, invading other brain tissues, but it rarely spreads to other organs [6].

World Health Organization (WHO) has a classification and grading scale for brain tumor classification, which comprises four grades. The grading is dependent on the growth rate of the tumor and its cell differentiation. GBM is classified as grade IV, the highest rate [1, 7].

Gliomas can appear as primary or secondary tumors. The first ones encompass 90% of the diagnosed patients which means that their tumor grows through multistep tumorigenesis from healthy glial cells within three months. Glioblastoma is the most common type of primary brain tumor and comprises 80% of them. The secondary gliomas cover the remaining ten percent of the patients and arise from lower-grade tumors, which takes about four to five years.

2.1.1 Epidemiology

Epidemiological data show that this brain tumor represents 60% among all the brain tumors and affects 0.002% of adults in most European countries. These studies also show evidence that its incidence is reported to be higher on men than women, the peak age where they appear is 55 to 60 years and its incidence is reported to be higher in Asians, Latinos and Caucasians [1, 2]. An incrementing number of cases have been diagnosed during the last 20 years, which may be more due to the improvement in the diagnostic of this type of tumors rather than a higher incidence of GBM [8]. Due to the low median survival time of fifteen months already mentioned and its five-year survival rate is less than 3%, the evidence of the mortality of this tumor is shown and it is considered a public health issue [2].

2.1.2 Causes and risk factors

Not much information is known about the etiology of GBM. It is believed to be a spontaneous tumor and its development is associated with a deregulation of checkpoint G1/S of a cell cycle and genetic perturbation on glial cells [1]. Despite this, some factors have been considered to influence the appearance of this type of gliomas. Due to the high incidence reported in older patients and men, age and gender can be referred as risk factors. Genetics are also a risk factor since some cases of relatives have been reported. Glioblastoma is also reported to arise a course of some genetic diseases, like tuberous sclerosis [1, 2].

2.1.3 Morphological features

When macroscopically observing the shape of this type of tumor, it is possible to realize its large size and irregular shape. As the word multiforme in the designation of this type of gliomas suggests, when observing the tumor microscopically, a big difference in appearance between its regions is noted [6]. The necrotic zones are yellow and soft and, generally, the others are white and firm. Moreover, some regions show marked cystic degeneration and hemorrhage. As a consequence of the last ones, red and brown colors are also exhibited [2, 9].

Moreover, this lesion is located on the supratentorial region 95% of the cases, more frequently on the frontal and temporal lobes [2, 10]. A study developed by Larjavaara et al. [11] demonstrate that the area with the densest occurrence was the anterior subcortical brain. A higher frequency of tumors on the right hemisphere has also been reported [11]. Even though the development of glioblastoma occurs in the trans-barrier space of the blood-brain barrier (BBB), those zones are intact [8].

2.1.4 Diagnosis and treatment

The primary diagnostic method of this disease is the MRI scan, once it allows high soft-tissue contrast enabling a good visualization of the complexity and heterogeneity of the tumor. With T1-weighted (T1) and T2-weighted (T2) scans, the hypointense and hyperintense lesions can be better observed, respectively. When using gadolinium as contrast, this tumor shows a central area of necrosis surrounded by white matter edema. Its use also help to detect changes in the integrity of the BBB [2, 12]. When the MRI cannot be applied to the patient, a Computed Tomography, CT, is used [2].

Currently, the standard care starts with surgical resection and is followed by chemo- and radiotherapy [1, 13, 14]. The surgical resection aims to perform a maximal safe resection of the enhancing part of the tumor. That is, to remove as much as possible from the enhancing part of the tumor, always keeping in mind which structures are affected and the changes to the neurological functions that may result from the resection. However, some tumor cells invade healthy surrounding tissues, making the task of removing everything even harder. In order to selectively kill the remaining tumor cells, prevent its regrowth and alleviating any symptoms, chemo- and radiotherapy are performed. The ideal dose of the last one is between 50 and 60 Gray and the most used drug for chemotherapy in patients with GBM is temozolomide [1]. There are some other treatment modalities such as RNA interference [15], immunotherapy [16] and hormone treatment [17].

It is important to mention that the treatments applied are heavily dependent on the prognostic factors of a patient like clinical characteristics (Karnofsky Performance Status (KPS) and patient age), tumor size, location, degree of its malignancy and molecular characteristics (genetic profile and proliferation activity) [1, 14, 18]. For example, if a patient has a low KPS and an advanced age, the surgery is most likely not to occur [13]. Also, the use of the surgery, chemo- and radiotherapy can increase the survival rate in young people [1].

2.2 Computational Background

2.2.1 Machine Learning

Over the last years, the scientific community has been observing a boosting on the application and investigation of technologies that take advantage of machine learning models, not only on the computer vision field but also on language modeling and robotics. Step-by-step, we are incorporating these technologies into our daily lives. The sudden interest in these technologies is motivated by the increased access to data, powerful processors for parallel computations, like Graphics Processing Unit (GPU), tweaks on algorithms and the development of user-friendly frameworks for creating these models [19, 20].

The machine learning goal is to solve problems by learning useful representations of the data through the creation of mathematical models. Formally, Mitchell [21] defined machine learning as:

> A computer program is said to learn from experience E with respect to some class of task T and performance measure P, if its performance at tasks in T, as measured by P,

improves with experience E.

In this context, the model has to learn generalized patterns to be able to make predictions for unseen data. It is important to mention that the created model is task-specific [20].

Depending on how the mathematical model uses the data for learning the patterns, machine learning can be categorized into three branches:

- Supervised Learning: On this class of transfer learning, the model receives the data and the corresponding answers, named targets, and learn to map one to the other. This is the most frequent type of machine learning.
- Unsupervised Learning: Here, the model receives only the data and try to find inherent and interesting patterns and relationships without having any information on the targets, that is, without any human supervision. With this branch, it is possible to discover hidden correlations present in the data, which may be a key for a better understanding of the data.
- Reinforcement Learning: On this branch of transfer learning, an agent, the central entity, receives information about its environment and learn to choose an action that will maximize its reward, by trial and error. A practical application of this branch is the model that Google DeepMind trained to play Atari games [20, 22, 23].

In supervised machine learning, the learning units can be sorted into different groups, depending on the target task. The most common categories are classification and regression. If the targets are categorical in nature, that is, each one belongs to only one discrete class, we are in the presence of a classification task. Specifically, the model will have to produce a function with the domain: $f : \mathbb{R}^n \to \{1, ..., k\}$. An example of this task is to recognize a numeric digit from scanned handwritten images. On the other hand, we are facing a regression task if the targets are continuous in nature so, the model function should be $f : \mathbb{R}^n \to \mathbb{R}$. These last targets can also be categorized, that is, organized into groups for the propose of visualization. Predicting the temperature for the next days is an example where this task can be applied [22, 24].

One of the most famous models in supervised machine learning is explained in the next section.

2.2.1.1 Artificial neural networks

In 1958, artificial neural networks were introduced by Rosenblatt [25] and their structure is inspired by the ability of the human brain to learn complicated patterns in data by changing the strengths of synaptic connections between neurons [26]. An example of a multilayer perceptron, also called feedforward neural network, is illustrated in Figure 1.

The basic units that form the above Figure are the neurons, which are organized in layers. These units are responsible for the computational process. The first layer, the input layer, receives the data where the information that is going to be learned from and the last one, the output layer, produces the output. If the task is a regression, the output layer is composed of one neuron and, if the network learned a classification

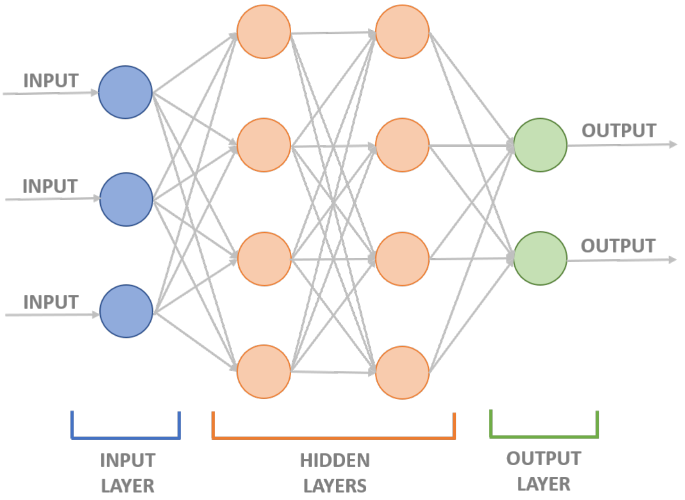


Figure 1: Multilayer perceptron.

task, this layer would have two or more neurons, depending on the number of possible classes. Between these layers, the hidden layers learn are present to connect them and transform the data.

The training process takes place as follows: first, a small group, called batch, of training points is fed into the model and produces an output that is compared to the desired output using a cost function or loss function, which measures quantitatively how far from the desired target the network is. Then, the backward propagation occurs. As the name suggests, the loss values are propagated backwards through the network and, with the help of an optimization algorithm, called gradient descend, the new weights are calculated using a chain rule to compute the contributions that each parameter had in the loss value. Finally, the training process is repeated all over again, until the patterns identified by the network make reasonable predictions for the training data.

After training the model, as above mentioned, the patterns created must be robust for new data. So, while the training occurs, another dataset is used, called validation dataset. This dataset is used to see how accurate the model predictions are for new data and, consequently, make adjustments on the hyperparameters of the model in order to improve its generalization capacity. If several tunings are made considering the performance on the validation dataset, somehow, information about the validation dataset will incorporate the model. To this respect, is used another dataset, called test dataset, to evaluate the model in a never-seen-before dataset and no tuning should be performed using this dataset [19, 20, 22].

2.2.2 Deep Learning

Deep Learning is a sub-field of machine learning, that is currently getting a lot of attention in the computer science field, due to the performances achieved. These models have already exceeded human performances on some tasks. Conversely, to what happens with machine learning, where the features have to be extracted manually, with deep learning, the features are directly extracted from the images by the computer using the model, which means that with one model, it is possible to both extract the features and do the classification or regression task [20]. With this representation-learning method, the most crucial step in machine learning, the selection of the features to model, is eliminated and, in general, more representative features are extracted [19, 22].

Currently, deep learning models are achieving results that outperform human performance [26]. An example of a type of deep neural network where this is happening are the Convolutional Neural Networks (CNN). In the next section, an overview of the building blocks of this feedforward network is present.

2.2.2.1 Convolutional Neural Networks

CNN are a type of deep learning models used in computer vision applications, especially in the medical imaging field. The main goal of this multilayered neural network is to identify shape and textured patterns with a high degree of invariance to translation, scaling and rotation in multiple arrays that can be 1D, for signals and sequences, 2D, for images, or 3D, for video and volumetric images [19, 23]. Figure 2 presents a typical structure CNN.

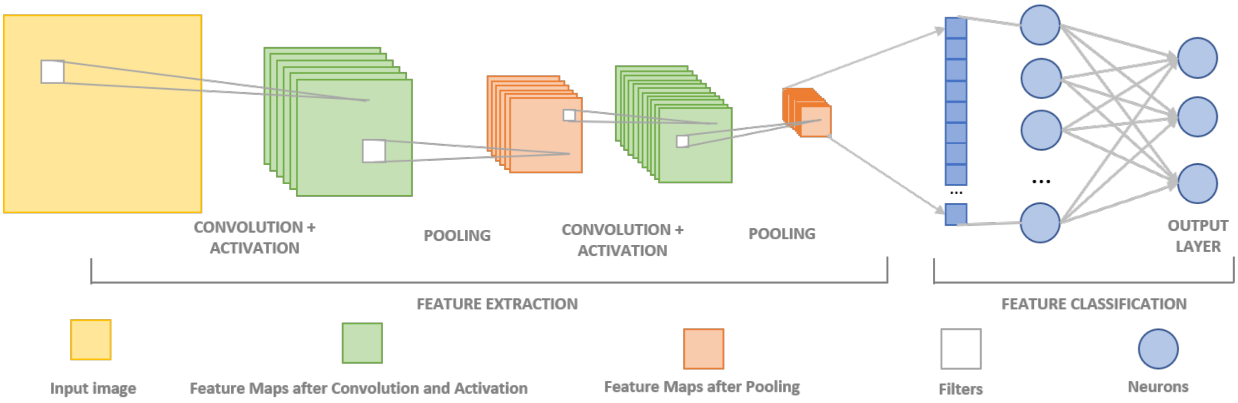


Figure 2: A typical structure of a CNN

On the feature learning stage of this network, at least two types of layers are present: the convolution layer and the pooling layer. Conversely to what happens on a dense connected layer, where global patterns are learned, on a convolutional layer, local patterns are learned instead. As the name suggests, the mathematical operation convolution occurs on this layer. In deep learning applications, the convolution can be applied at more than one axis at a time. This operation is defined by $S(i,j) = (I * K)(i,j)$, where I is the input, a multidimensional array of data, K is the kernel or filter, a multidimensional array of parameters that are updated during the training process and S the output, denoted as feature map. Since the kernel has typical dimensions between three and eleven, several convolutions have to be performed for covering all the input. The number of pixels that the filters move after each convolution is defined as stride. This, plus the size of the filters and the padding added, that consists of adding the appropriate number of columns and rows to the input feature map to preserve its size, leads to the capture of different areas of the image and, consequently, a different field of view. Its importance should not be discarded to ensure that no information from the image is left out. Since the filters used are the same for every convolution in the same layer, a weight sharing exists and, at each layer, is present a translational equivariance. Figure 3 represents an example of a 2D convolution.

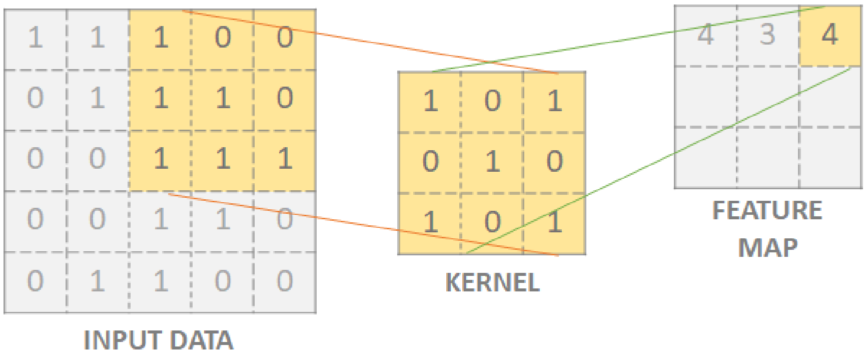


Figure 3: A 2D convolution operation with a stride of 1 and padding 0. Addapted from [27].

Further, the pooling layer is present. Its main purpose is to reduce the sensitivity of the feature maps output to shifts and rotations and perform a downsampling effect. For this, a small grid region, typically two per two, is considered and a single number is produced from each region. The number is calculated using an average or a maximum function. Figure 4 shows an example of a max-pooling operation.

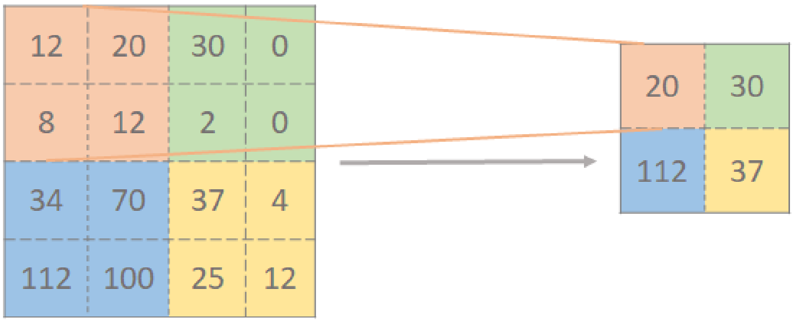


Figure 4: A two by two 2D max-pooling operation. Addapted from [23].

Between the abovementioned layers, an activation layer is stacked. Non-linear neurons compose it and, as the name suggests, endow them with a non-linearity function that will allow the network to approximate these types of functions. Some examples of these functions are the Logistic Sigmoid, Hyperbolic Tangent, Rectified Linear Unit (ReLU) and Softmax [20, 23].

The block of layers previously mentioned is repeated various times, depending on the depth desired for the network. On the first convolutional layer, usually small patterns of edges with particular orientations and locations are learned. Then, larger patterns from the previous layers are learned and so on. Like this, a CNN can learn more complex and abstract patterns that are crucial to identify the learned object [19, 22].

Finally, on the classification stage, fully connected layers are used to make the prediction. On the last layer, an activation is needed to constrain the network output. The activation function used depends on what the network prediction is: if we are facing a binary classification problem, the activation function used on the last layer should be the Sigmoid and if the problem is a regression, no activation function is needed.

2.2.2.2 Overvitting, Underfitting and Regularization techniques

While training the model, the optimizer will try to get the best performance on the training data. However, the best performance does not mean, necessarily, a better generalization, that is, a good performance on unseen data. While the training phase takes place, in some cases, it is possible to observe an increment in the difference between training and validation error. When this occurs, a case of overfitting is present. Moreover, when the training error stops decreasing, it is considered a case of underfitting, meaning that the model cannot make a useful generalization of that problem. These problems are also present in machine learning models.

Currently, there are strategies to avoid overfitting and reduce generalization errors. These strategies are called regularization techniques. Goodfellow et al. [24] defined regularization as:

> Any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error.

These techniques make slight modifications to the learning algorithm such that the model generalizes better and, consequently, improves the model's performance [24]. Currently, there are many regularization techniques and the ones used during this work are briefly described below.

Data augmentation: The best way to get a better-generalized model is to add more data but, in practice, the existing data is limited. So, this method generates more training data from the existing training samples. Some techniques for this augmentation are geometric transformations, such as translations, rotations and scaling, or noise injection. These procedures are carried out to expose the model to more aspects of the data and, consequently, extract more information from them, leading to a better generalization of the model.

Dropout: This regularization method modifies the network itself by randomly dropping out, which is turning off, a defined percentage of neurons in the hidden layers during the training process. The dropout rate corresponds to the percentage of neurons that are dropped out at a layer. Since different neurons are affected at each epoch, the network trains a different configuration. Thus, this procedure introduces noise to the output layer where it is applied and the network is forced to learn more robust features instead of relying on the prediction capability of a small subset of neurons.

Batch normalization: Before entering the model for training, a good practice of preprocessing is to normalize the pixel values of the input image. Based on this idea, after the activation layer, the batch is normalized by subtracting the batch mean and dividing by the standard deviation of the batch. This helps the gradient propagation, being mainly a benefit in deeper networks.

Local Response Normalization (LRN): This normalization scheme was designed by Krizhevsky et al. [28] to be applied with ReLU. The lateral inhibition of a neuron inspires this response normalization.

The neurons create competition for bigger activities among the output values when using different kernels and the activations made on those neurons are more pronounced when compared with the neighborhood.

Early stopping: While the network is being trained and the training loss is decreasing, the validation loss may stop improving, as shown in Figure 5, meaning that the generalization ability of the model starts do decrease. The point right before the validation loss starts to increase is where the model has the best generalization and so, the parameters from this model should be kept and used instead of the ones from the last epoch. Due to its simplicity and effectiveness is one of the most popular methods of generalization [22–24, 29, 30].

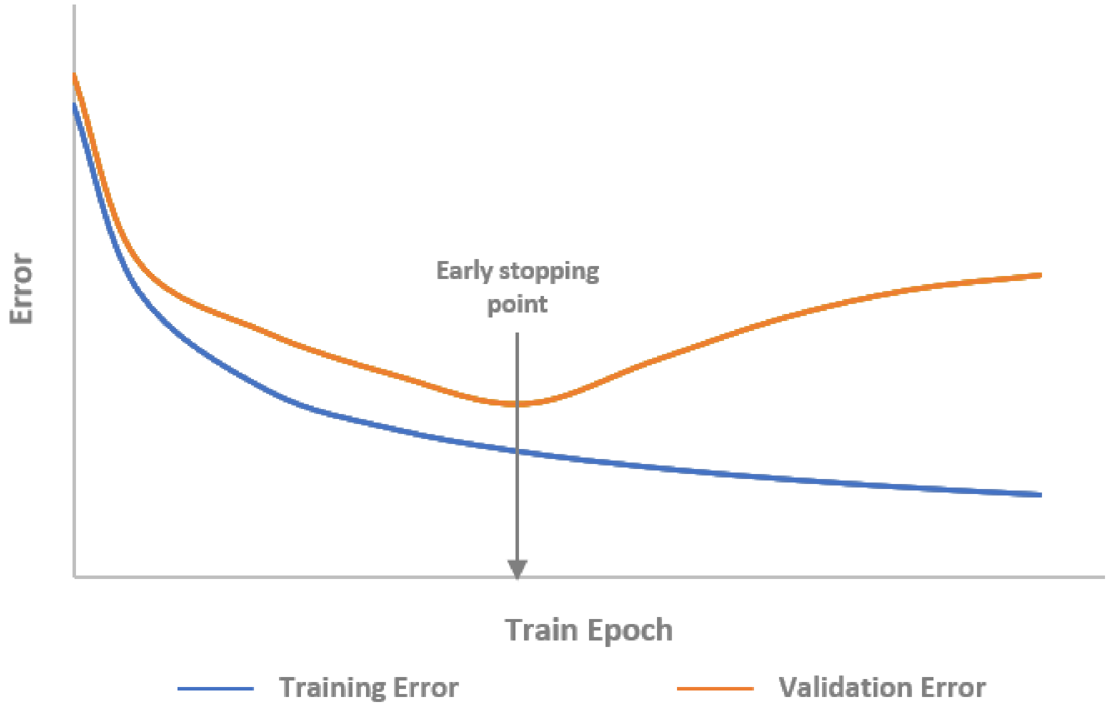


Figure 5: Early stopping: training and validation curves. Adapted from [23].

It is important to note that CNN already has an inbuilt method to reduce overfitting called weight-sharing. It consists of the same weights being used in different layers in the network. The fact that the convolution layer has connections between local regions also contributes to intrinsic help to avoid overfitting.

2.2.3 Transfer Learning

Transfer learning, in contrast with machine learning, that learns every task from source, reuses knowledge from other related tasks. This is based on the native capability of humans to apply relevant acquired knowledge from previous experiences into an unseen task. Pan and Yang [31] consider a domain $\mathcal{D}$ (denoted as $\mathcal{D} = \{\mathcal{X}, P(X)\}$) with two components: a feature space, $\mathcal{X}$, and a marginal probability distribution, P(X), where X = {x_1, ..., x_n } $\in \mathcal{X}$. They also denote a task as $\mathcal{T}$ = {$\mathcal{Y}$, $f(\cdot)$} and is composed by a label space $\mathcal{Y}$ and an objective predictive function $f(\cdot)$. Therefore, they state that this knowledge transference aims to help the improvement of the target predictive funcion $f_{\mathcal{T}}(\cdot)$ in the target domain $\mathcal{D}_{\mathcal{T}}$ using the knowledge present in the domain source $\mathcal{D}_{\mathcal{S}}$ and target source $\mathcal{T}_{\mathcal{S}}$, taking into account that $\mathcal{D}_{\mathcal{S}} \neq \mathcal{D}_{\mathcal{T}}$ and $\mathcal{T}_{\mathcal{S}} \neq \mathcal{T}_{\mathcal{T}}$.

As for humans is easier to learn to play the electric organ if they already know how to play the piano, it is believed that, with transfer learning, the knowledge transfer will help to improve the learning on the target task. The improve in learning can occur in three ways. First, the performance of the baseline model can be upgraded. That is, before the learning process occurs, the performance of the model is already better when compared with the one that is going to be trained from scratch. This is translated in Figure 6 as an higher start. Secondly, a decrease in the overall time needed to learn the target task can happen too. As seen in Figure 6, a higher slope appears on the model that benefited from the transfer learning. Finally, a better final performance can also occur and it is manifested as a higher asymptote in Figure 6.

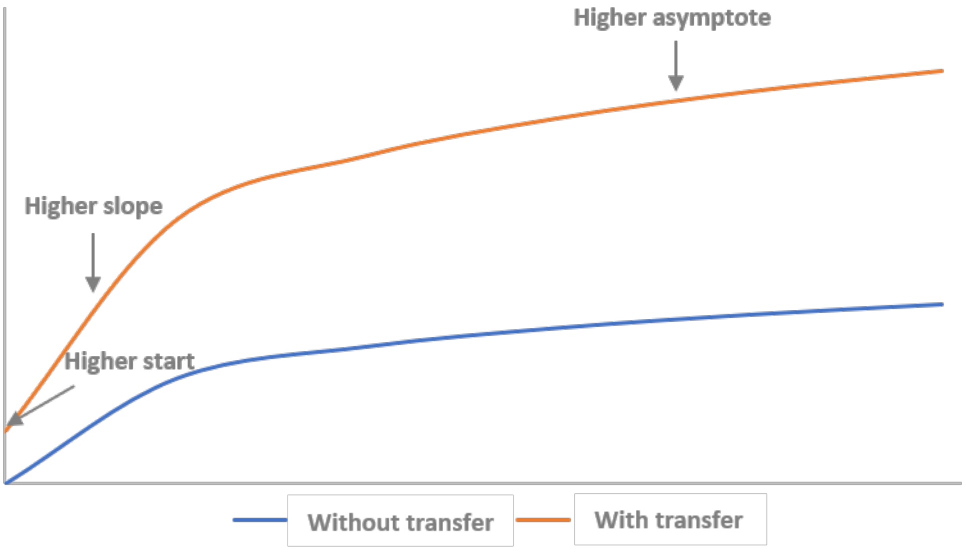


Figure 6: Ways in which transfer learning might help improve learning. Adapted from [32].

Considering a target task, the efficiency of the learning transfer process is highly related to the model used to extract knowledge from: not only the source task but its target too. So, it is imperative that, during the transfer process, three questions are taken into account:

- What to transfer: On this question, considered the most important one, it is imperative to identify the portion of knowledge that is source-specific and the part that the source and target share and, consequently, help the improvement of the learning.
- When to transfer: It is essential to be aware of which situations the transference should be done in order not to deteriorate the performance of the model. When this happens, a negative transference has occurred. It happens when the source task is not sufficiently related or the transfer method cannot translate the relationship between the source and target tasks very well.

- How to transfer: After identifying what and when to transfer, an investigation on the ways of transferring this knowledge has to be done: learning algorithms have to be developed or changed for the transference of the learning between the domain and tasks occur.

Going back to the question of what to transfer, four approaches can be applied. The first one is instance transfer, where the knowledge from the source domain is reused to the target task. The aim of the second approach, called feature-representation transfer, is to identify a good feature representation that can be used from the source to target domains by minimizing the domain divergence and reduce error rates. The third method, parameter transfer, assume that the source and target tasks share some parameters or prior distribution of hyperparameters of the model. Lastly, the approach relational-knowledge transfer deals with data that is not independent and identically distributed, which means that some data points have a relationship with other data points.

Focusing on the previous how to transfer question, there are three categories that transfer learning methods can be categorized into:

- Inductive transfer: This transference method make use of the inductive biases of the source domain to help improve the target task. Here, the source and target tasks have to be different but related whereas source and target domains can be the same or not. This type can also be split into two categories: multitask learning, where the data in the source domain is labeled, and self-taught learning, where it is not labeled.

- Unsupervised transfer: This technique is similar to the previous one, but focuses on solving unsupervised learning tasks. Hence, labeled data is not present in both domains.

- Transductive transfer: On this method, the source and target tasks are the same, but the source and target domains are different. In this situation, the source domain has a lot of labeled data whereas the target domain does not have any labels.

A way of transferring knowledge is using a pretrained network. A pretrained network is a saved network that was already trained on a large dataset. When this dataset is large and general enough, the network has learned a robust hierarchy of features. It can act as a generic model for the visual world and the learned features are cross-cutting to various problems. An example of these networks are AlexNet [28], Inception [33] and ResNet [34] and they were trained using the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [35], which contains a total of about 1,4 million diversified images and each belongs to one of the 1000 categories.

When reusing these networks for learning transfer, that usage can occur in two ways: or use it as a feature extractor or fine-tuning the network. On the feature extractor, the representations learned by a pretrained network are used to extract interesting features from a new dataset. The features extracted run through a new classifier that is trained from scratch. On the fine-tuning approach, as its name suggests, the features learned by the pretrained network are retrained to make them more relevant to the new problem [22, 23].

Transfer learning has been widely and successfully used in tasks involving text data, speech/audio and computer vision. Related with the last application, several applications have been developed in the health field such as classification and location of thorax diseases [36] and early detection of Alzheimer's Disease [37].

2.2.4 Interpretability

A growth of investigation on deep learning networks, especially in the medical field, has occurred during the last years and some authors have already developed models that reached better performances than clinicians [38].

Since human lives are on the line, due to the capacities of those algorithms to detect early diseases or help the recovery of the patients, clinicians are concerned about trusting those deep learning models. One of the main reasons is because they do not understand its underlying decision-making process, despite the good results obtained by those models. On their perception, deep learning models are ”black-boxes” [39]. The clinician's fears are legitimate since deep learning models rely on complicated interconnected hierarchical representations of the relationship between the input and output variables. So, it is imperative to produce descriptions that are simple enough for a clinician to understand how a network comes to a decision, how certain it is and if it can be trusted [20, 39, 40].

Only when having models explained from a human-centered perspective and designed to think like the doctors, the clinicians will be able to trust them [38]. A way for making the ”black-box” models more transparent that is, for understanding its decision-making process is the generation of importance maps that give cues on how important each pixel is, helping to find the ones that the network ”focus” more [40, 41].

The usage of methods for understanding the importance of each pixel can also help to improve the performance of the deep learning networks: understand why their predictions were not accurate or what decisions lead to a classification outlier [40].

2.3 Summary

With this chapter, some theoretical concepts essential to this work were described. It was possible to understand that glioblastoma is a deadly spontaneous brain tumor that arises from glial cells with a median survival time of fifteen months. This highly locally invasive tumor arises from primary tumor 90% of the cases and their preferred location are the frontal and temporal lobe on the right hemisphere. Currently, the diagnosis is made through the observation of MRI scans and the current standard of care involves tumor resection together with chemo- and radiotherapy.

Deep learning models have shown an emerging potential. In the computer vision field, CNN are able to extract features from an image, learn patterns and make predictions using unseen data. This network uses the convolution operation for feature extraction and it is endowed with an inbuilt capacity for reducing overfitting, due to the existing weight-sharing. Overfitting occurs when the model learns during the training patterns that are irrelevant on new data. Some other techniques such as data augmentation, dropout or early stopping were described. In the same context, transfer learning reuses knowledge from other related tasks, leading the model already start the training with a better initialization. Because these models are ”black-boxes” for the clinicians, they struggle to trust their results. Thus, there is a need for constructing interpretable models.

Chapter 3

State of the Art

Over the years, MRI scans stopped to be seen only as images but also as information sources: where details that are not provided by clinical reports can be found. For extracting these informations, radiomics and deep learning methods have been used [42]. In the last years, the usage of machine learning methods has been increasing exponentially. This has been happening due to the increasing numbers of new datasets available, advances in computing power and new-found machine learning algorithms. The prevalence of competitions/challenges has also helped, naturally, to an increment on the investigation [43].

Currently, there are several applications on the medical field that used MRI scans combined with deep learning or radiomics. One such example is the segmentation of kidneys [44] and distinction between rejected and non-rejected renal transplants [45]. In the field of breast cancer, models for predicting the cancer subtype [46] and treatment failure [47] are already developed. In the cardiovascular field, an algorithm for calculating the coronary calcium has been developed [48]. Additionally, the segmentation of the prostate [49], its cancer occurrence [50] and a preview on the biochemical recurrence [51] has also been studied. In the case of the brain, several studies related to the recognition of different stages of Alzheimer [52], skull stripping the brain [53], tumor segmentation[54], prediction of the age [55] and and of the OS of patients with glioblastoma [56]. In the next sections, the two last applications are further described since they are going to be in the scope of this work.

3.1 Estimating subjects' age

It has been found that the human brain changes during life and have some specific patterns when it comes to healthy aging [57]. These changes in the brain are owing to progressive and regressive neural processes such as cell growth and myelination and cell death and atrophy, respectively. By learning these changes from healthy brain images with different chronological ages, deep learning models have been able to grasp patterns and make predictions of chronological age with a high level of accuracy.

For example, Cole et al. [55] used a 3D CNN with five blocks of convolutional layers with a fully connected layer in the end to predict the chronological age, using as input images normalized brain volumes maps of grey and white matter. Using the volume maps of grey matter as input, the network achieved a mean absolute error of only 4.16 years and a determination coefficient between chronological age and the brain predicted age of 0.92.

Currently, these models have been used to identify pathologic brain developments such as Alzheimer, schizophrenia and epilepsy. This is possible because the predicted age is greater than the chronological age. Moreover, the same age models have also been able to identify factors that prevent the brain from aging, such as meditation, increased levels of education and physical exercise [55, 57].

3.2 Prediction of overall survival of patients with glioblastoma

Finding imaging biomarkers for predicting the OS of patients with GBM using pre-operative MRI scans has been an emerging task in the last few years. Due to the inter- and intra-patients tumor heterogeneity, the impact of the tumor resection during the operation and the response to the chemo- and radiotherapy, predicting survival is a puzzling task.

In 2017, the prediction of OS started to be part of the International Brain Tumor Segmentation (BraTS) Challenge, which brought an increase in the investigation on this task. Moreover, with this introduction, a fair comparison can be made between different approaches, since they use all the same dataset. The use of a multi-institutional dataset is also a plus [56].

The growing body of literature has already investigated diverse approaches for making this prediction. A typical workflow comprises a feature extraction stage, followed by feature selection when needed and then, the features are modeled. Finally, an evaluation of the overall performance is made. In the following sections, each stage of this workflow is explained.

3.2.1 Feature extraction

Firstly, the extraction of high dimension feature data to quantitatively describe attributes from a region of interest (ROI) occurs [42]. In the literature, the type of features extracted from scans used for the OS prediction can be split into three categories:

- Classical features: They are used by numerous authors and embrace features such as first-order statistic, shape-based, volumetric, positional and texture like neighborhood gray-tone difference, gray level co-occurrence matrix and dependence matrix features [18, 58–60].

- Deep features: These features extracted using a neural network are also common among literature and CNN are the network architecture of choice. The models can either be trained from scratch [61, 62] or trained using a pretrained network as an initializer [63].

- Combination of classical and deep features: The combination of these two types of features is present in literature as well [61, 63, 64].

These features are extracted from MRI sequences, such as T1, T1-weighted contrast enhancement (T1ce), T2, T2-weighted fluid-attenuated inversion recovery (FLAIR) or functional MRI. Regions consistently reported to be used as ROI are the whole brain or the whole tumor or even specific regions from the last one. These regions can be necrosis, edema, enhancing, non-enhancing or tumor core (a combination of necrosis and edema) and they can be obtained through the segmentation of the tumors present in the brain. Such segmentations can be performed manually, by doctors, or by automatic methods [54].

The image features aforementioned are frequently combined with non-image features such as age, sex, resection status or the KPS [63, 65–67].

Before the feature extraction, some authors apply specific preprocessing on the scans. Regardless of the type of features that are extracted, an image resampling and registration between scans is usually performed [60]. If the features extracted depend on the intensity levels of the sequences, the N4 bias field algorithm or intensity normalization is also carried out [42, 63]. Moreover, the application of filters such as wavelets or laplacian of gaussian is also used by state-of-the-art methods [59, 61].

3.2.2 Feature selection

Before modeling the extracted features, when the initial amount of features is not reduced, a feature selection is performed due to the curse of dimensionality phenomenon. This phenomenon happens when there are more features than training samples, which helps to avoid overfitting during the training of the model [24]. In this context, non-informative and highly correlated features are removed. Firstly, some authors reduce the features based on their robustness against different raters, using intraclass correlation coefficient (ICC) or overall concordance correlation coefficient (OCCC) [63, 68, 69]. Moreover, the concordance index is used to keep only features with prognostic value and the correlation coefficient to remove highly correlated features [58, 63, 68].

Following this process, usually the features are also ranked to define the most significant ones, using decision trees, least absolute shrinkage and selection operator (LASSO) Cox regression or Spearman’s correlation coefficient [63, 70, 71] and only these are used in the model.

3.2.3 Modeling of the extracted features

Finally, the modeling of the features can be done by a constructed fixed-parameter signature [63] or a predictive model. For the last one, the most used approaches are linear regression models [72], random forest [70], support vector machine algorithm [65] and fully connected layers [73]. The models can be trained as regression [58, 66] or 2-class or 3-class classification models [60, 61] and the class boundaries vary widely from article to article.

3.2.4 BraTS Challenge

As previously mentioned, the BraTS Challenge in 2017 started to include the task of predicting the OS of patients with GBM. Here, the participants are required to preview the OS using features that they consider appropriate and analyze them using a regression machine learning method. For performing these predictions, the T1, T1ce, T2 and FLAIR sequences are available as well as the age of each patient.

On Table 1 are presented the validation metrics from the top-ranked teams for the survival prediction task of BraTS challenge in 2017 and 2018, since not all the authors mention all their test metrics.

When looking at the features used to predict the survival time by these authors, the usage of texture, volumetric and features related to tumor location stands out. Moreover, all of them use the non-image feature age, except Shboul et al. [58], the number one ranked in 2017. Sun et al. [72] investigated the performance of modeling age with the volume in voxels of enhancing tumor, edema and necrosis individually and all together using a linear regression. It was also tested the combination of age with the distance between the centroid of the brain to the centroid of the tumor. They found that all the other features, when combined with age, deteriorate the performance of age by itself.

Table 1: Validation metrics from the top-ranked teams for the survival prediction task on BraTS challenge in 2017 and 2018. The metrics that are being evaluated are the Mean Square Error (MSE), Spearman's Correlation Coeficient (SpearmanR), Median Square Error (medianSE) and Standard-deviation Square Error (stdSE). For calculating the accuracy, the predictions are divided into 3 classes, depending if the survival is greater than ten months or fifteen or if is between these two values.

		Metrics					
Year	**Ranking**	**Accuracy**	**MSE**	**SpearmanR**	**medianSE**	**stdSE**	**Paper**
2017	1	0.636	2.18×10^5	0.590	3.91×10^4	5.53×10^5	[58]
	2	0.424	2.46×10^5	0.050	5.63×10^4	5.41×10^5	[73]
	3	0.424	1.97×10^5	0.284	5.80×10^4	4.93×10^5	[74]
2018	1	0.321	1.04×10^5	0.247	7.84×10^4	1.12×10^5	[66]
	2(tie)	0.607	9.79×10^4	0.501	1.84×10^4	1.81×10^5	[75]
	2(tie)	0.393	1.06×10^5	0.217	4.75×10^4	1.46×10^5	[70]
	3(tie)	0.583	1.31×10^5	0.446	2.34×10^4	3.15×10^5	[71]
	3(tie)	0.500	9.78×10^4	0.267	4.61×10^4	1.40×10^5	[72]

3.2.5 Overview of performance evaluation

Depending on whether the prediction is made using regression or classification, different metrics for evaluating the model are used. For regression, the MSE and SpearmanR are the most used. Moreover, these are also part of the metrics evaluated by BraTS. With classification, the most frequent metric is accuracy. A confusion matrix is also used by some authors, like Chato and Latifi [61]. Other authors, make a statistical analysis using a Kaplan-Meier plot and the log-rank test [59].

Mainly due to the different class boundaries of the classification models, it became tough to make a fair comparison between the different approaches that exist. Even comparing regression models is not easy, since not all the authors report the same test metrics. This is also verified with the articles that got the best performances on BraTS Challenge.

A low consistency between the performance on training and testing datasets is reported by some authors, mainly when deep features are used [61, 65]. The reason given by them is usually related to overfitting while training.

In short, the majority works show that age, the position of the tumor in the brain and texture information play an essential role in the prediction of OS in patients with GBM. Even in the first studies carried out in 1993 by Simpson et al. [67] identified the age and the position of the tumor in the brain as important features for predicting OS. However, some recent studies that rely on deep features have shown interesting results [61, 63–65].

3.3 Summary

Two applications of deep learning models on MRI images were presented during this chapter. One of them, estimating subjects' age, is based on the fact that the human brain changes as we get older. Cole et al. [55] suggests an approach that uses a CNN using volume maps as input with a mean absolute error of 4.16 years. The prediction of OS of patients with glioblastoma was also reviewed. The image features generally used are classical and deep features and they are frequently combined with non-image features. To avoid the curse of dimensionality, a feature selection is performed. After modeling those features and comparing several state-of-the-art approaches, the features with a better predictive capacity are the position of tumor, texture information and age.

Chapter 4

Data and Methods

Despite still being in its early stages, medical imaging has attracted more attention in the last years due to the excellent performances achieved and the vital component in healthcare applications.

After an introduction to the theoretical concepts, in this chapter, the datasets used for the development of all the work and some demographic information on them are stated in the first section. The developed strategy for directly predicting the OS involves the training of deep models from scratch and the reuse of knowledge from other networks. In the next section, details on this strategy and its implementation procedures are thoroughly described.

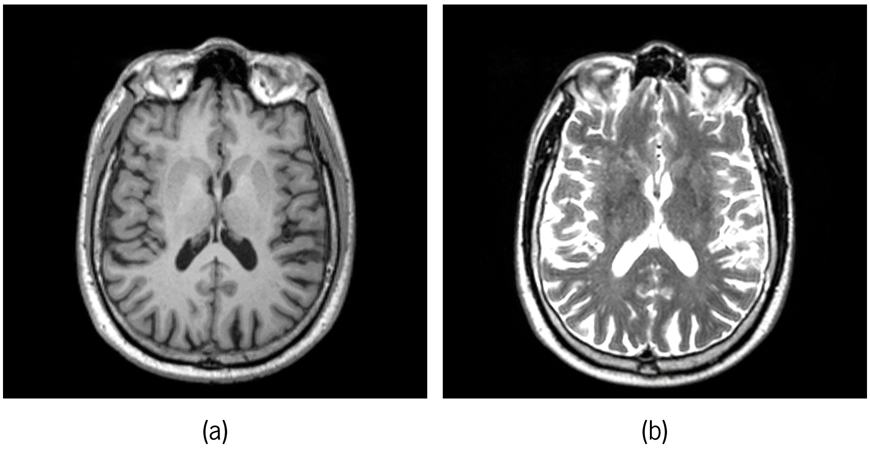
4.1 Data

During this work, two datasets where used: Information eXtraction from Images (IXI) database and BraTS 2018 Training dataset. This last dataset used to predict the OS and the IXI dataset is used to train a model from where knowledge is going to be extracted in order to be reused when predicting the OS.

4.1.1 IXI dataset

Consisting on about 600 healthy subjects, this dataset contains scans using different acquisition protocols as T1,T2 and proton density weighted images, magnetic resonance angiography images and diffusion-weighted images acquired in fifteen directions. All the data were acquired from three different hospitals in London, using 1.5T and 3T scanners [76].

Among all the available subjects, only five hundred and twelve subjects were used. The other subjects where rejected because they did not have both T1 and T2 sequences (these are the ones used in this work) or the subject's age was not available. The data was given as Neuroimaging Informatics Technology Initiative (NIfTI) files and had no preprocessing applied. An example of the T1 and T2 scans are present in Figure 7 and some demographic information on the training and validation groups are presented on Table 2 and the respective distribution of the age of the subjects for those groups is shown in Figure 8.



(a) (b)

Figure 7: Axial slice from the scans of a subject from IXI dataset: (a) T1 sequence, (b) T2 sequence.

Table 2: Demographic information of the training and validation groups on the IXI dataset

	Training group	**Validation group**
Nr. of Patients	447 (87%)	67 (13%)
Age in years		
Minimum	19,98	20,91
Maximum	86,32	79,41
Mean	48,60	48,85
Standard Deviation	16,63	15,66

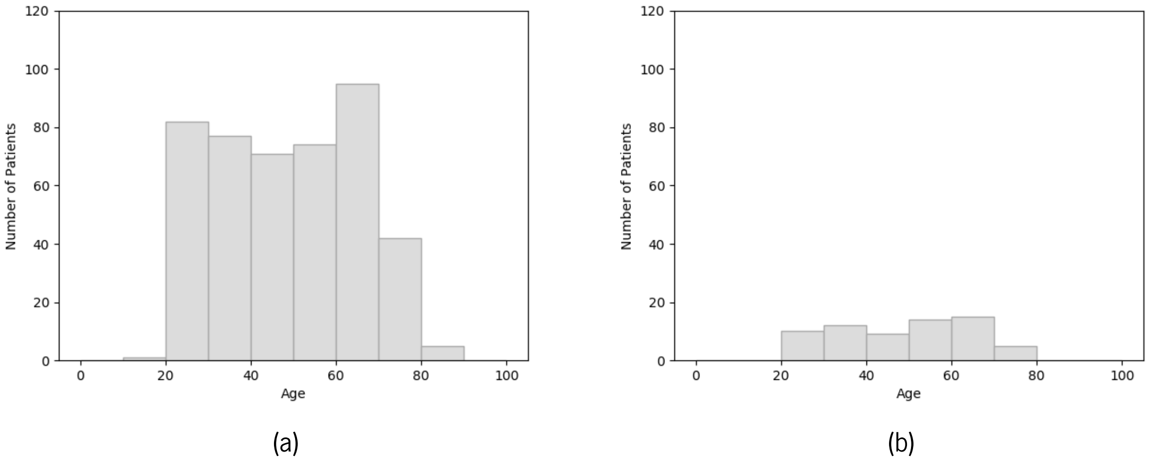


Figure 8: Distribution of subject's age on (a) training and (b) validation group of the IXI dataset.

4.1.2 BraTS 2018 dataset

This dataset comprises 3T multimodal pre-treatment MRI scans from patients with a confirmed diagnosis of glioblastoma and lower grade glioma [54, 56, 77]. The scans include T1, T1ce, T2 and FLAIR sequences. Following this information, the OS of the patients is reported and it consists of the time from the diagnosis date until the patient dies. The age in years of each patient and the resection status were also made available. It can be classified as GTR or STR, when a total or partial resection was performed, respectively. When there is no information on that, an indication with NA (Not Available) is present. Accompanying the scans, a manual segmentation exists too. This segmentation was performed by one to four experienced neuro-radiologists, following the annotation protocol described in [54]. Three zones of the tumor have been segmented: GD-enhancing tumor, peritumoral edema and necrosis together with non-enhancing tumor core. Additionally, all the data is available as NIfTI files and have already some preprocessing applied. The scans are co-registered to the same anatomical template, interpolated to the same resolution and skull-stripped. An axial slice from each sequence, as well as the respective tumor segmentation, are shown in Figure 9.

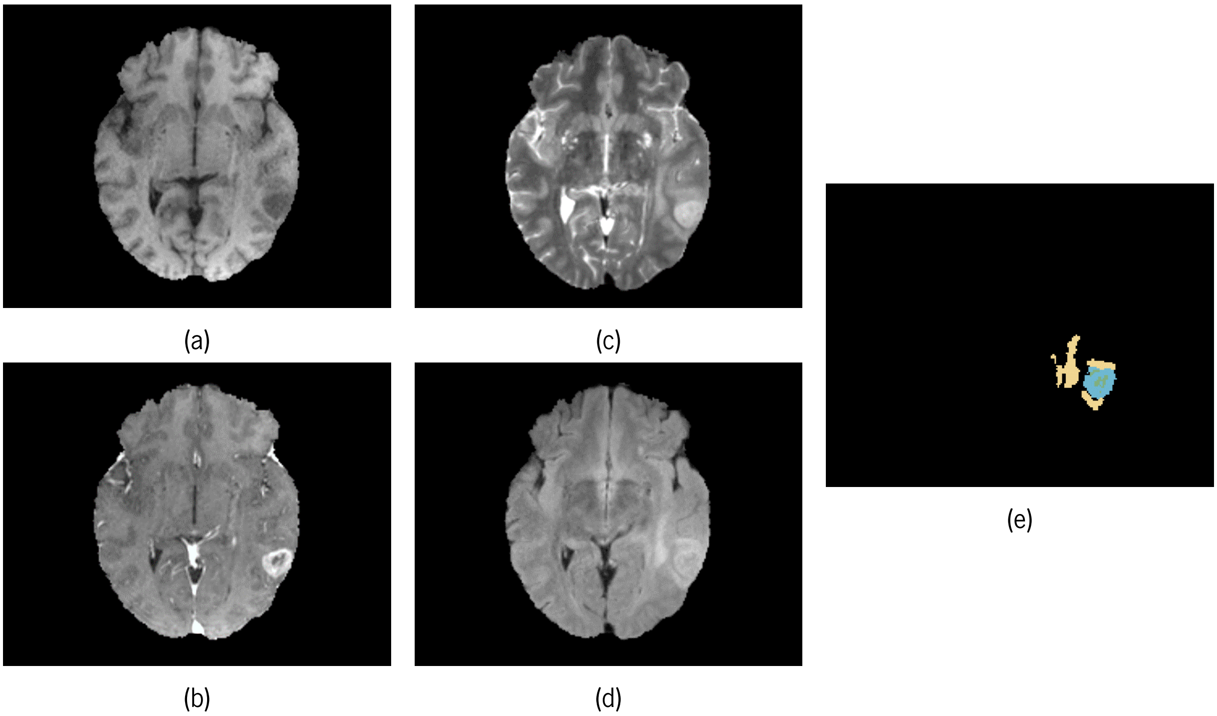


Figure 9: Axial slice from the scans (a) T1, (b) T1ce, (c) T2, (d) FLAIR and (e) tumor segmentation of a patient from BraTS dataset. On tumor segmentation the green color corresponds to necrosis and non-enhancing are, the yellow to the edema and blue to the enhancing contrast.

This dataset encompasses 163 subjects. It will be divided into three groups: training, validation and testing. The demographic information on each group is presented in Table 3.

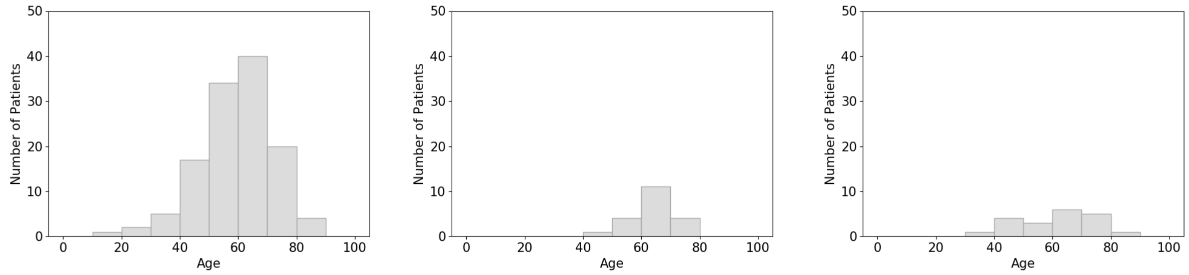
The distribution on the age and survival on each group of the dataset can be observed in Figures 10 and 11, respectively.

4.2 Methods

After a description of the datasets used through this work, on the following subsections, there is an explanation of the whole pipeline of the work. Briefly, the preprocessing applied to each dataset is explained. This includes the preparation of the data for the next step, that is the feature extraction and modeling. Here, the three main deep learning approaches used for the prediction of the OS are

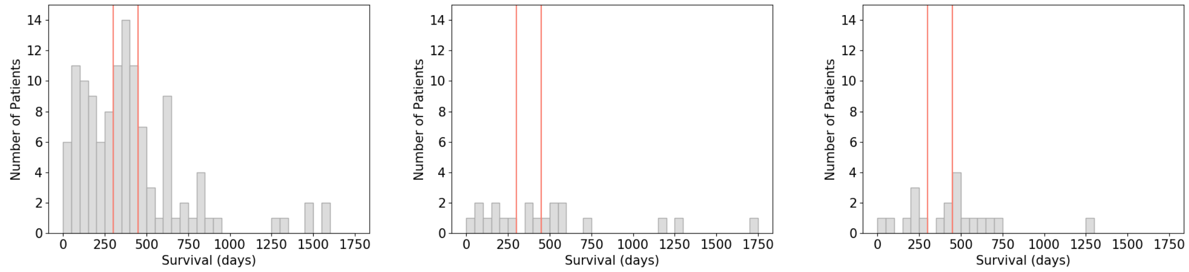
Table 3: Demographic information of the training, validation and test groups on the BraTS 2018 training dataset

	Training group	**Validation group**	**Testing group**
Nr. of Patients	123 (75,46%)	20 (12,27%)	20 (12,27%)
Age in years			
Minimum	18,97	47,97	36,85
Maximum	85,761	78,74	81,21
Mean	59,59	63,84	61,36
Standard Deviation	12,44	7,88	12,29
Survival in days			
Minimum	5	23	32
Maximum	1767	1731	1283
Mean	409,09	493,45	437,75
Standard Deviation	343,27	430,51	272,12



(a) (b) (c)

Figure 10: Distribution of patient's age on (a) training, (b) validation and (c) testing group on BraTS 18 Training Dataset. The vertical orange lines indicates the 300 and 450 days class boundaries.



(a) (b) (c)

Figure 11: Distribution of patient's survival on (a) training, (b) validation and (c) testing group on BraTS 18 Training Dataset. The two vertical orange lines point out the 300 and 450 days.

described: two networks received knowledge obtained from other tasks and another network was trained from scratch. For each approach, different inputs and preprocessing techniques were considered. After training and validating the networks, the combination of input scans and preprocessing that obtained a better performance are selected, accordingly to the two criteria defined. Then, the top-ranked models were retrained using data artificially augmented and a thorough analysis of the results is performed.

The whole pipeline was implemented using the programing language *Python* and, for the implementation of the deep network, the *Pytorch* framework, version 1.0 was used. All the calculations, including the training of the networks, were performed on UBELIX (`http://www.id.unibe.ch/hpc`), the HPC cluster at the University of Bern, which has three types of GPUs, that were used for the training of the deep learning networks.

4.2.1 Image Pre-processing

4.2.1.1 IXI dataset

As aforementioned, the scans from this dataset did not have any preprocessing applied. In the first place, a skull stripping of the sequences needs to be performed and, for that, the brain imaging software *FreeSurfer* was used. Using the first processing stage of the *FreeSurfer* cortical reconstruction process, a skull-stripped T1 sequence with a size of $255 * 255 * 255$ was obtained. In Figure 12, an axial slice before and after removing the skull is presented. Since this software is only able to skullstrip T1 scans, a mask is produced using the skull stripped scan by thresholding the sequence: if the pixel value is greater than one, it will take value one in the mask, otherwise it takes the value zero. Moreover, since the skull stripped image has a different shape from the original images and, in order to T1 and T2 sequences have the same preprocessing, after aligning the sequences with each other and the mask, the last one was applied to the image obtained after the *Motion Correction and Conform* step of the first processing stage of the *FreeSurfer* cortical reconstruction process. For the alignment of the scans and remove the skull, *SimpleITK*, an image analysis library, was used.

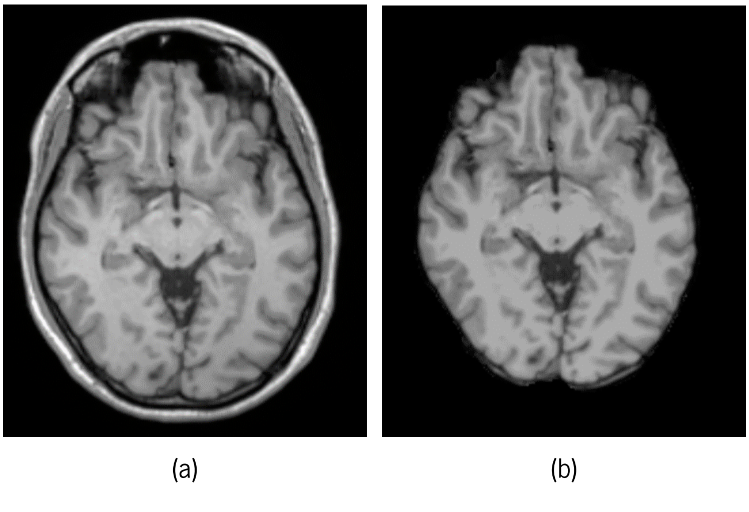

(a) (b)

Figure 12: Axial slice from a scan from IXI dataset (a) before and (b) after being skull-striped.

After removing the skull, a bias field correction was applied and, using the mask previously generated, the brain was cropped from the scan and resized to $240 * 240 * 155$, using a BSpline interpolator. This size was used so that the images have the same size as the ones from the BraTS dataset. To perform these last steps, *SimpleITK* was also used.

4.2.1.2 BraTS 2018 dataset

Additionally to the preprocess that already comes by default, the N4 bias field algorithm was computed to all the scans on the BraTS dataset. It was made to remove unwanted non-uniform low-frequency intensity from the image, using *SimpleITK*. The further processing depends on the neural network used and the desired studies to perform with each sequence. An overview of the tested preprocessing methods can be found in Table 4.

Table 4: Tested preprocessing methods and the variables considered on each method

Preprocessing methods	**Variables considered**
Intensity Rescale	0 - 255 (8-bits) 0 - 4095 (12-bits)
Z-score Normalization	Whole Image Only to the brain
ROI delimitation for extracting the features	Original Image Whole Brain Whole Tumor

While the 8-bits normalization range was chosen due to the restrictions of the network, the other one was chosen based on the image: the maximum intensity of the scans was bellow and near the maximum value of 4095. The intensity rescale allows the direct comparison between the intensity values of the scans.

A z-score normalization is also applied to the scan. The normalization can be performed considering all the pixels of the image (whole image) or only the pixels that are part of the brain. In the second case, the background pixels continue to be zero unlike in the other case. This can be an issue when the training is performed with the segmentation, where the background is zero.

Figure 13 presents the different ROI for extracting the features. Moreover, it is important to mention that, when considering the ROI the whole brain or the whole tumor, since the number of background pixels is reduced, the type of z-score normalization will not have a big effect. The z-score normalization only to the brain is performed.

It is important to keep in mind that, from all the preprocessing abovementioned, only the ROI delimitation is applied to the tumor segmentations.

4.2.2 Features extraction and modeling

In order to perform the critical step of this work, the feature extraction and the respective modeling, deep neural networks for directly predicting the OS are going to be used, more specifically CNN for regression architectures. With two of those networks, a transfer learning process occurs: the networks were already trained on other tasks. The first one was trained for predicting the chronological age of a subject and the other one acquired knowledge that allows distinguishing between several objects, such as cars, boats, cats, televisions and flowers. These networks are going to be finetuned using the BraTS dataset with the aim that their filters are good initializers for finding features for the prediction of OS. The other

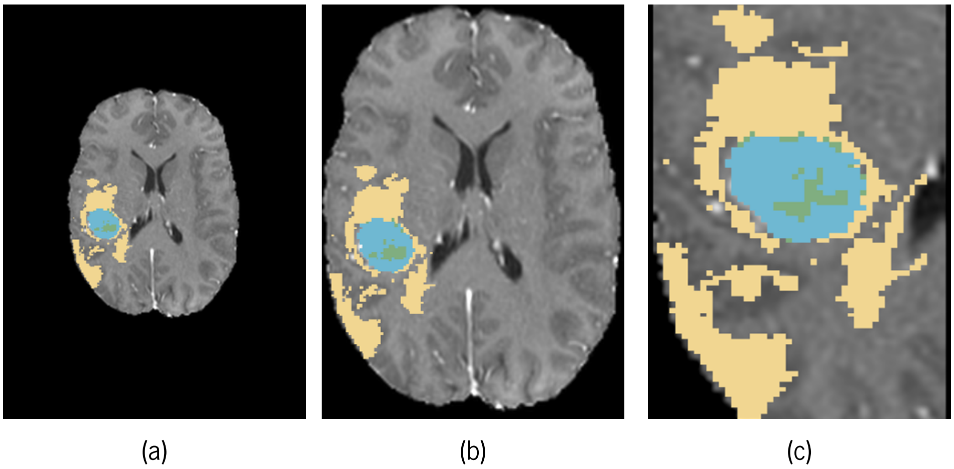


Figure 13: T1 scan and respetive tumor segmentation showing different ROI delimitations used for extracting the features: (a) Original Image, (b) Whole Brain, (c) Whole Tumor.

network is trained from scratch and has the same architecture as one of the pretrained ones. In the upcoming sections, the procedures to perform the training or finetuning of the CNN architectures are going to be explained.

For training and validating these networks, the BraTS training and validation datasets, characterized in Section 4.1.2, were used. The optimizer used was the SGD and the loss criteria was the MSE. The networks were trained for 400 epochs and the early stopping technique was applied: the epoch chosen as best is the one immediately before the performance of the model in the validation dataset deteriorates.

Lately, the construction of a linear regression is also described. This regression is going to be used as the state-of-the-art comparison.

4.2.2.1 Transfer Learning networks

4.2.2.1.1 Reusing features from age prediction

As seen in Section 3.2.5, the non-image feature age seems to have an essential role in the prediction of the OS of patients with GBM. Section 3.1 refers that deep learning, in particular CNN models, are already being used for the prediction of the chronological age using MRI scans. So, the first approach is based on the fact that the brain characteristics that help to predict age may be a good start for finding other features related to survival prediction. The pretrained model is going to be based on the approach proposed by Cole et al. [55]. A schematic representation of the network is in Figure 14. It is composed of five blocks of convolutional layers and a fully connected layer for predicting age. Each block is composed of a 3D Convolution layer, with a stride of 1, a ReLU, another 3D Convolutional layer with a stride of 1, a 3D batch norm, a ReLU and, finally, a max-pooling layer. The number of feature channels doubles at the end of each block and the first block has eight feature channels.

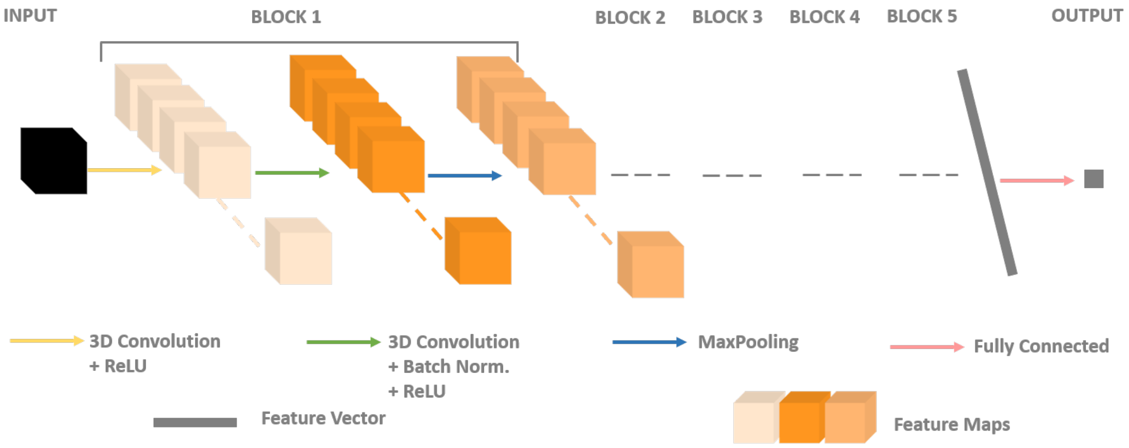


Figure 14: Overview of the deep learning model used for prediction of OS using as a start-point features related with chronological age prediction. Addapted from [55]

Although they made their pretrained model available, since it was trained in another programing language, it was not possible to reuse it and, consequently, it was necessary to replicate the training made by these authors. The same deep neural network architecture was trained with some differences in the preprocessing (the images were not normalized brain volume maps, but merely the scans) and on the inputs themselves. So, more specifically, three different types of inputs were trained, which resulted in three different pretrained models. Two of these models have only one input channel with the dimensions $240 * 240 * 155$ and their inputs are the T1 or T2 scans. The third model has four input channels, being them twice the T1 sequence followed by twice the T2 sequence. The dataset used for this pretraining was the IXI dataset, which represents only a quarter of the data used by the authors of the paper. Details on the training and the metrics obtained can be found in the attachments (Section 7.1).

After obtaining the pretrained models, it was time to finetune them. The data used for it was the training group of the BraTS training dataset. Depending on which sequence or ensemble of sequences was used while pretraining, the scan used for finetuning vary. These specifications are explained in Table 5. Moreover, using each one of the five inputs for the finetuning, different variations of the preprocessing were tested. Firstly, two types of ROI were used: the original image and the whole brain. Relatively to the variation of the z-score normalization, both variants were considering when the original image was used as ROI. The intensity rescale used the range of 0 - 4095.

Table 5: Correspondance between the inputs used for the model pretraining and the ones using for fine-tuning

Pretrain Input	**Fine-tuning Input**
[T1, T1, T2, T2]	[T1, T1ce, T2, FLAIR]
[T1]	[T1] [T1ce]
[T2]	[T2] [FLAIR]

The learning rate was finetuned. This was made for each of the variations on the preprocessing.

Mostly, the learning-rate with better performance was 1×10^{-6}, except for finetuning using FLAIR with the whole brain as ROI that was 5×10^{-7} and when the input was the four sequences: when the ROI was the original image, regardless of the z-score normalization used, the learning rate was 5×10^{-6}, and when the ROI was the whole brain, the learning rate selected was 7×10^{-8}.

4.2.2.1.2 Reusing features from CNN-S

An article written by Lao et al. [63] inspires the other transfer learning approach. They extract features from the pretrained Slow Convolutional Neural Network (CNN-S), developed by Chatfield et al. [78]. This network was trained using the dataset ILSVRC. Based on the Kaplan-Meier curves and the concordance index exhibited in Lao et al. [63], some features extracted using this network shown the potential to be an imaging biomarker for predicting the OS. In this context, the weights of CNN-S are considered to be a good start point for extracting relevant features for predicting the survival prediction, and this network is going to be used as a pretrained model, that will be finetuned using the BraTS training dataset. An overview of the CNN-S structure is presented in Figure 15. It is important to mention that a small modification on the network proposed by Chatfield et al. [78] was done in order to make it suitable for the regression task: it is only composed of one neuron. The initialization of this layer was random.

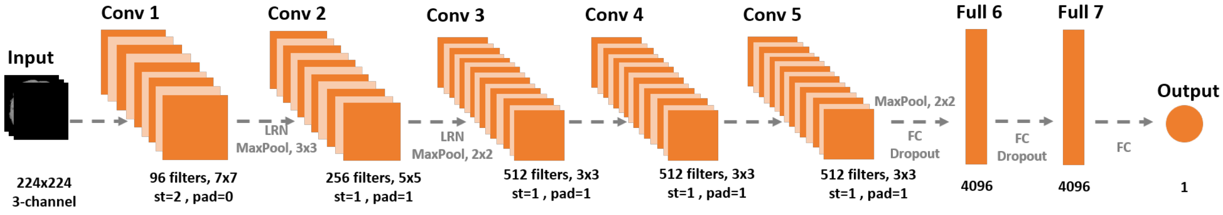


Figure 15: Modified CNN-S structure used to predict the OS. LRN, st and pad are short for Local Response Normalization, stride and padding. Addapted from [63].

The CNN-S provided by the paper authors was pretrained using the framework *Caffe*. Since the whole pipeline was already being developed using Pytorch, there was a need to convert the model to this framework. For doing that conversion, first, the Caffe model was converted to *Torch*, using `https://github.com/szagoruyko/loadcaffe.git`. Then, the conversion from *Torch* to *Pytorch* was made using `https://github.com/clcarwin/convert_torch_to_pytorch.git`.

Contrary to the network shown above, this one is pretrained on 2D images with three channels, so it is not possible to input the whole scans by once. So, as Lao et al. [63] has done, the axial slice with the biggest tumor area was chosen to be the input of the network. So, in this case, three types of input combinations were considered:

1. In one of the channels, the axial slice with the biggest tumor area is inputted in one of the three channels, while the other two are turned off. This is performed for the four scan types;
2. On each channel is inputted the axial slice with the biggest tumor area from one of the sequences. As it is only possible to input the biggest slice from three sequences at a time, one is left out. These sequences are T1 or the FLAIR;

3. For each scan, the axial slice with the biggest tumor area of each patient is inputted in the middle channel, and the slice before and after that one are inputted in the channel before and after, respectively.

Due to the characteristics of the CNN-S, especially the lack off batch normalization, all the input scans had to suffer an intensity rescale to the range of 0 to 255. Two types of ROI for extracting features from were used: with the whole brain, all three types of inputs abovementioned were used, while with the ROI delimitation being the whole tumor, only the first type of input was used. The learning-rate used was the same, regardless of the input used: 1×10^{-8}.

4.2.2.2 Training from scratch the network used for predicting age

After trying to construct and train CNN architectures from scratch and being unsuccessful, because they were overfitting too early, the same structure used in Section 4.2.2.1.1 is going to be trained from scratch. The three types of ROI are used as input and, when the ROI is the all image, the two types of normalization are tested. The intensity of the scans were rescaled to the range 0-4095. Also, the effect of adding tumor segmentation to the input was studied. To initialize the weights of the network, the initialization method proposed by He et al. [79] is used. The learning rate was tunned and the value of 1×10^{-7} was used for all the models, except when inputting the four sequences, with and without the segmentation. In that case, the learning-rate used was 5×10^{-6}.

4.2.2.3 Linear Regression using age

The chronological age of the patients with GBM seems to have a big impact when predicting their OS, as already mentioned. In this context, as a term of comparison, a linear regression between the OS and the chronological age was constructed, based on what Sun et al. [72] have done. The *scikit-learn* package was used.

4.2.3 Selecting the best performances and data augmenation

After training the deep learning models above referenced, for choosing the model with the best combination of inputs and preprocessing of each approach of training, two criteria where used. Due to the high inconsistency between the training and test results reported by several authors, a criterion used for choosing the three top-ranked models was to select the ones in which the difference between the training and validation loss was smaller. The second criterion is based only on the value fo the validation loss: the three combinations of inputs and preprocessing with the lowest loss values are chosen. From now on, consistency and best loss are going to be the used designations to identify each criterion.

The selected combinations were then trained again, but this time using data augmentation techniques. Two of the data augmentation techniques were used on this dataset: rotations and translations. A random number of rotations of 90° (between 0 and 119) or random rotations on a range of $\pm 15°$ and a random shifting between ± 15 pixels were performed. These techniques were performed on the axial or coronal plane. It is important to mention that both rotations and translations are applied only when the ROI of the input is the original image. When the ROI is the whole tumor or the whole brain, only rotation is performed, so that the scans do not get distorted. Moreover, the data augmentation is performed online and, due to the random transformations abovementioned, the input images will be different at each epoch, which is expected to improve the generalization capacity of the network.

4.2.4 Performance analysis

In order to analyze the performance of the top-ranked combinations, a detailed analysis of the same metrics evaluated by the BraTS challenge is going to be done. Those metrics are MSE, stdSE, medianSE and SpearmanR. The first three metrics measure the mean, standard deviation and median of the squared difference between the predicted and ground truth survival. The greater the euclidean distance between both values is, the greater the squared error will be. Moreover, greater euclidean distances will be more penalized than smaller ones [24]. The metric SpearmanR is a correlation coefficient that measures the correlation between two variables. This coefficient represents the two-way monotonic relationship that can exist between two variables and it can range from -1 to 1, a perfect negative or positive correlation [80]. In the case of this work, since we are interested in evaluating the relationship between the ground-truth and the predicted survival values, we are only interested in the positive values of this correlation. A rule of thumb for quantifying the strength of a correlation coefficient is present in Table 6.

Table 6: Rule of thumb for interpretating the size of a Correlation Coeficient [81]

Size of Correlation	Interpretation
0.90 to 1.00	Very high correlation
0.70 to 0.90	High correlation
0.50 to 0.70	Moderate correlation
0.30 to 0.50	Low correlation
0.00 to 0.30	Negligible correlation

In order to understand where the regression is failing the most to predict the survival, a confusion matrix is constructed. This square matrix, which has the same size as the number of classes, determines the percentage of correctly and incorrectly classified data. As the name implies, the way a class is confused with other classes can be easily noticed through this matrix [82]. However, the problem that we are considering is a regression problem, not a classification one. Therefore, in order to be able to plot the confusion matrices, the predicted OS is stratified into three different classes: a class where the predicted survival time is less than 300 days, another one where the predicted values are between 300 and 450 days and a last one where the OS predicted is greater than 450 days.

Another way used to understand the dispersion of the predicted data and to measure the agreement between two quantitative variables was through the Bland-Altman plot. This plot was introduced by Altman and Bland in 1983 to study the mean differences and to construct the limits of agreement between both variables, which are calculated using the mean and standard deviation of the variables. On this plot, the difference between the variables (A-B) is plotted against their mean ((A+B)/2) [83].

Finally, in order to understand the decision-making process of the top-ranked models, the Randomized Input Sampling for Explanation of Black-box Models (RISE) approach, developed by Petsiuk et al. [40], is going to be used. This occlusion test measures the importance of each image region by creating masks that will dim parts of the input image in random combinations, reducing their intensities down to zero. Then, these occluded images are inputted into the network, a weighted sum between the predicted value and the maps is made and the saliency map is generated. It is on it where the importance of each pixel is indicated. This approach has the advantage that it does not need to access any parameters from the network, treating it as a black-box. An implementation of this approach was made available by the authors of the paper. However, the implementation is prepared to handle only 2D images and, consequently, some adjustments were made so that it could be used with 3D imagens.

4.3 Summary

For developing the deep learning model to directly predict the OS with the help of BraTS training dataset, different models and inputs were tested. Two different CNN architectures were tested. The first one is the 2D CNN-S created by Chatfield et al. [78] and the knowledge obtained from this training is used as a weight initialization for the prediction of the OS. The second architecture used is the same proposed by Cole et al. [55] and it was trained from scratch or the knowledge for predicting a subjects' age is reused for the prediction of the OS. Aside from the application of the N4 bias field correction, different preprocessing techniques were applied to the input images: intensity rescale, z-score normalization and resizing to the different ROI. After selecting the best performance using the two criteria defined, data augmentation was performed and an analysis of the performance was done, which is presented in the following chapter.

Chapter 5

Results

Throughout this chapter, the best combinations of scans used as an input and its preprocessing for each approach are going to be discussed. Then, a comparison between the different approaches is made and the most promising ones are chosen. Finally, they are compared with a state-of-the-art method.

In order to make it easier to specify each input used and its preprocessing, a specific encoding for each approach was created and is present in Table 7.

Table 7: Encoding of each approach

Approach	**Encoding**
Reusing features from age prediction	pretrainedseq_finetunningseq_ROI_intnorm
Reusing features from CNN-S	inputR_inputG_inputB_ROI
Training from scratch	inputseq_usingseg_ROI_intnorm

On the approach that reuses features from predicting age, it is essential to identify the scans used for pretraining the network (pretrainedseq) and the ones used for finetuning (finetunningseq). When reusing features from CNN-S, it is relevant to indicate what is inputted on the red (inputR), green (inputG) and blue (inputB) channel, keeping in mind that, since this model is 2D, the slice chosen is the axial one with the biggest tumor area. On the approach that trains the model from scratch, it is important to mention the input scan used (inputseq) and if the segmentation is going to be inputted as well or not (usingseg). Relatively to the ROI used the abbreviations are AI, for all image, WB, for whole brain, and WT, for whole tumor. The identifiers for the type of z-score normalization (intnorm) were I and B if the normalization was done considering the pixels from the whole scan or only the ones that are part of the brain, respectively.

In the confusion matrices, in order to simplify their plots, the acronyms SS, MS and LS stands for short, medium and long survivors, respectively.

5.1 Reusing features from predicting age

In Section 4.2.2.1.1, the different inputs used for pretraining and finetuning the network were defined. After validating all the models with different inputs and preprocessings and, before applying different data augmentation techniques, the MSE was calculated when using the training, validation and testing datasets for the top-3 models selected using the consistency and best loss criterion. Those values are present on Table 8.

Table 8: Train, validation and testing MSE for the top-ranked combinations that reuse features from predicting age

		Training	Validation	Testing
	Input Combinations	**MSE**	**MSE**	**MSE**
Consistency	T2_FLAIR_AI_B	8.68×10^4	1.68×10^5	6.68×10^4
	T2_T2_AI_I	9.50×10^4	1.92×10^5	7.28×10^4
	4seq_4seq_AI_B	8.24×10^4	1.79×10^5	6.76×10^4
Best loss	T2_FLAIR_AI_B	8.68×10^4	1.68×10^5	6.68×10^4
	4seq_4seq_AI_B	8.24×10^4	1.79×10^5	6.76×10^4
	T1_T1ce_AI_B	1.03×10^5	1.80×10^5	5.52×10^4

Firstly, when observing the designation of the inputs, it is possible to verify that two of them are ranked top three by both criteria. When observing their ROI, the preference for the original image is evident. Moreover, the z-score normalization only to the brain seems to be helpful for the model to learn features that help to predict the OS. All the models top-ranked have this type of normalization except one, but it had the worst MSE on the test set.

Table 9 reports the MSE of the top-ranked models when different data augmentation techniques were applied. From the table, we can note that the generalization capacity of the network did not improve in all the cases except one. This means that no new relevant patterns for the task were enhanced or learnt and, consequently, no benefits were obtained from the data augmentation techniques used. In this context, on Tabel 10, the MSE, medianSE, stdSE and the SpearmanR are reported for the models that were not trained using data augmentation techniques.

Table 9: Effect of data augmentation on the top ranked models that reuse features from predicting age

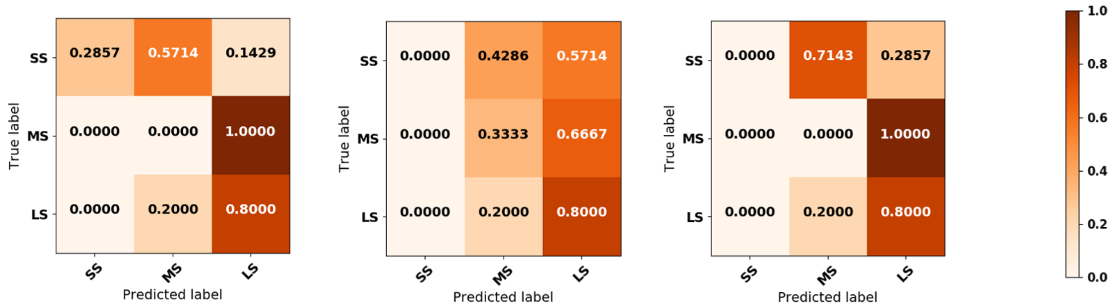
	Input Combinations	**Without DA**	**DA**	
			rotations 15°	**rotations 90°**
Consistency	T2_FLAIR_AI_B	6.68×10^4	8.47×10^4	1.12×10^5
	T2_T2_AI_I	7.28×10^4	6.90×10^4	7.43×10^4
	4seq_4seq_AI_B	6.76×10^4	7.42×10^4	1.03×10^5
Best Loss	T2_FLAIR_AI_B	6.68×10^4	8.47×10^4	1.12×10^5
	4seq_4seq_AI_B	6.76×10^4	7.42×10^4	1.03×10^5
	T1_T1ce_AI_B	5.52×10^4	7.69×10^4	1.02×10^5

Considering the MSE metric, the input combination with the lower value is the T1_T1ce_AI_B, followed by T2_FLAIR_AI_B and 4seq_4seq_AI_B. These models also have the highest SpearmanR. However, accordingly to Table 6, these models have a low correlation between the actual values of survival and the predicted ones. Among the top-3 input combinations, the T2_FLAIR_AI_B is the one where that relationship is weaker when comparing to the other two. Interestingly, the model with the normalization applied to all the scans is the one with the lowest SpearmanR, which reinforces the idea that this type of normalization is not the best approach.

Table 10: MSE, medianSE, stdSE and SpearmanR on the testing group for the top-ranked input combinations that reuse features from predicting

		Metrics			
	Input Combinations	**MSE**	**medianSE**	**stdSE**	**SpearmanR**
Consistency	T2_FLAIR_AI_B	6.68×10^4	3.12×10^4	8.36×10^4	0.396
	T2_T2_AI_I	7.28×10^4	2.30×10^4	1.52×10^5	0.187
	4seq_4seq_AI_B	6.76×10^4	2.58×10^4	1.43×10^5	0.471
Best loss	T2_FLAIR_AI_B	6.68×10^4	3.12×10^4	8.36×10^4	0.396
	4seq_4seq_AI_B	6.76×10^4	2.58×10^4	1.43×10^4	0.471
	T1_T1ce_AI_B	5.52×10^4	1.74×10^4	1.10×10^5	0.457

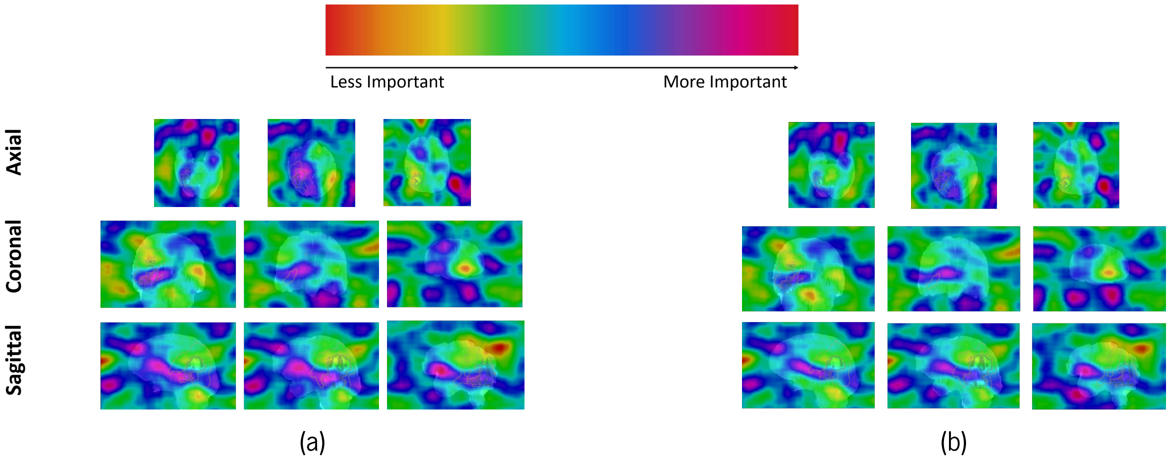
In order to have a visual interpretation of the predicted data, confusion matrices were constructed for the three best input combinations above mentioned, which are present in Figure 16. In Figure 16a, some of the survival values predicted were short survivors, meaning that the three models have found it challenging to identify the short survivors. Identifying medium survivors is also a complicated task for these networks. On the other hand, 80% of the long survivors were correctly identified as it. The problem is that, when the models fail to classify, the majority of the time, it classifies them as long survivors. The high values for the medianSE for the input combinations T2_FLAIR_AI_B and 4seq_4seq_AI_B were already an indication that this was happening.



(a) (b) (c)

Figure 16: Confusion matrices for the top-3 input combinations that reuse features from predicting age: (a) T1_T1ce_AI_B , (b) T2_FLAIR_AI_B and (c) 4seq_4seq_AI_B.

The difference between the confusion matrices of the models T1_T1ce_AI_B and 4seq_4seq_AI_B are on the prediction of the short survivals, which indicates that is where the model has mostly learned different things. However, it is important to note that the predictions for the medium and long survivors may not be the same since on these maps the predictions are stratified. The result maps from the occlusion test, for these input combinations, for the patient BraTS18_TCIA05_277_1, which has a OS of 232 days, are present in Figure 17. Unexpectedly, it was verified that both networks were giving much attention to the space between the brain and the limit of the image. This can justify the fact that the selected top-ranked input combinations only use as ROI the whole image. Moreover, the networks seem to consider foremost the edges of the brain as observed in the third coronal slice and the space between them and the tumor, like on the third image from the sagittal plane. When considering the importance given to the tumor by the network structure, it is contradictory: it can be high or low. However, it should be noted that no information on the tumor is provided to the network, in these cases. It is also interesting that the network does not give much importance to the cerebellum, zone where this kind of tumor usually is not formed. Focusing on the differences among both models, it is possible to see that they are not significant in the selected zones, but on their importance. For example, on the second axial slice, it is interesting to note that the zone where the tumor is located is considered more important for the model T1_T1ce_AI_B rather than to the 4seq_4seq_AI_B.



(a)

(b)

Figure 17: RISE maps with the contour of tumor segmentation from a patient with a short survival from the models (a) T1_T1ce_AI_B and (b) 4seq_4seq_AI_B.

5.2 Reusing features from CNN-S

This 2D network was finetuned using the different inputs described in section 4.2.2.1.2 and the top-3 input combinations selected based on both criteria previously explained are in Table 11.

Table 11: MSE of training, validation and testing without using any data augmentation (DA) technique and MSE of test dataset using DA techniques reusing features from CNN-S network

		Without DA			**With DA**
	Input Combinations	**Training**	**Validation**	**Testing**	**Testing**
Consistency	0_0_T2_WT	1.14×10^5	1.72×10^5	7.10×10^4	7.10×10^4
	FLAIR_0_0_WB	1.21×10^5	1.70×10^5	7.66×10^4	7.66×10^4
	T1_T1ce_T2_WB	1.20×10^5	1.70×10^5	9.09×10^4	9.09×10^4
Best loss	0_0_T2_WT	1.14×10^5	1.18×10^5	7.10×10^4	7.10×10^4
	big-1_T2_big+1_WB	2.30×10^4	1.46×10^5	1.21×10^5	1.21×10^5
	0_0_T1ce_WT	6.81×10^4	1.58×10^5	8.51×10^4	9.43×10^4

When taking a look at the best input combinations, the fact that the model 0_0_T2_WT is considered the best by both criteria stands out. No accordance existed about the ROI used by the top-ranked models. Concerning the input sequences used by the top-ranked input combinations, the slice with the biggest tumor area in one channel and shutting down the other ones are the most frequent type. However, the other two types are also present: inputting the slice with the biggest tumor area of the T2 scan, the slice before and the one after and inputting the axial slice with the biggest tumor area from the T1, T1ce and T2 scans together.

After this first selection, the models were retrained and the input data was augmented by rotating the scan 90° a certain random number of times in one of the planes, as previously mentioned. Observing the last column of Table 11 it is evident that the majority of the models maintained its performance with data augmentation. The knowledge transferred into the network can explain this fact: since the filters were already trained on a big dataset, they are already invariant to rotations and the benefits of data augmentation were not felt.

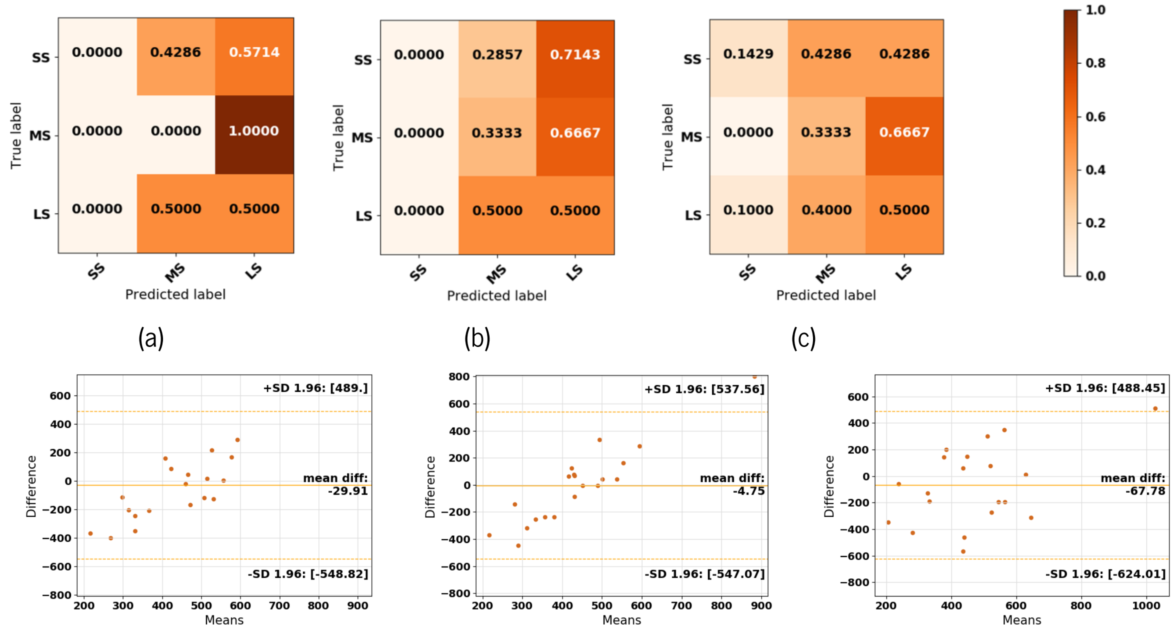
Table 12 reports the MSE, medianSE, stdSE and the SpearmanR of the top-ranked input combinations, using the test group without applying any data augmentation technique.

Table 12: MSE, medianSE, stdSE and SpearmanR metrics from the test group reusing features from CNN-S

		Metrics			
	Input Combinations	**MSE**	**medianSE**	**stdSE**	**SpearmanR**
Consistency	0_0_T2_WT	7.10×10^4	2.80×10^4	1.27×10^5	0.201
	FLAIR_0_0_WB	7.66×10^4	2.31×10^4	1.40×10^5	0.00677
	T1_T1ce_T2_WB	9.09×10^4	2.08×10^4	1.47×10^5	-0.181
Best loss	0_0_T2_WT	7.10×10^4	2.80×10^4	1.27×10^5	0.201
	big-1_T2_big+1_WB	1.21×10^5	5.12×10^4	2.02×10^5	-0.0316
	0_0_T1ce_WT	8.51×10^4	3.94×10^4	9.03×10^4	0.0053

Based on the SpearmanR value, it is possible to select three models that stand out: 0_0_T2_WT, FLAIR_0_0_WB and 0_0_T1_WT, once these are the ones with a positive value and we are interested in a direct relationship between the predicted OS and the correct one. Although they are positive, according to Table 6, the relationship between both variables is considered negligible. Nevertheless, the confusion matrices for those models, as well as the Bland-Altman plots, are present in Figure 18.

The confusion matrices on the Figures 18a and 18b show that the models 0_0_T2_WT and FLAIR_0_0_WB are only predicting survival values higher than 300 days and not predicting any subject as a short survivor. Observing the correspondent Bland-Altman plots (Figure 18d and 18e), it is possible to verify that the mean of the difference between the predicted survival and the real one is low. When observing the dispersion of the points, the difference between both values stands out and, for most of the values, is not near zero. When the mean is low, the difference is a high negative value and, when the mean is high, the difference is a high positive value. Consequently, it is possible to understand that all the predicted values are more or less the same. Since Lao et al. [63] was able to find features with a prognostic potential with the pretrained CNN-S, it was not expected this poor performance obtained by the deep model. The fact that the predicted values are more or less the same may be caused by the feature classifier of the deep network: once the second to the last layer has 4096 and the output layer has only one neuron. This difference may be too big and the network struggles to predict the OS.



(a) (b) (c)

(d) (e) (f)

Figure 18: Confusion matrices and Bland-Atman plots for the top-3 ranked input combinations from the approaches that reuses features from CNN-S network: (a) and (d) 0_0_T2_WT, (b) and (e) FLAIR_0_0_WB, (c) and (f) 0_0_T1ce_WT .

5.3 Train from scratch

In this section, the performance of training a model without receiving prior knowledge with different sequences and preprocessing inputted is evaluated. Based on the different inputs considered, mentioned in Section 4.2.2.2, the top-3 ranked input combinations based on the two criteria previously established are reported in Table 13.

Table 13: Training, validation and test MSE for the top-ranked input combinations models trained from scratch

		Training	**Validation**	**Test**
	Input Combinations	**MSE**	**MSE**	**MSE**
Consistency	4seq_wseg_WB_B	1.62×10^5	1.78×10^5	1.48×10^5
	4seq_wseg_WT_B	1.42×10^4	1.78×10^5	1.02×10^5
	FLAIR_woseg_AI_B	9.09×10^4	1.85×10^5	7.08×10^4
Best loss	FLAIR_wseg_AI_I	4.43×10^3	1.54×10^5	6.48×10^4
	T1ce_wseg_AI_B	5.82×10^3	1.66×10^5	9.58×10^4
	T1_wseg_AI_I	1.42×10^3	1.68×10^5	8.10×10^4

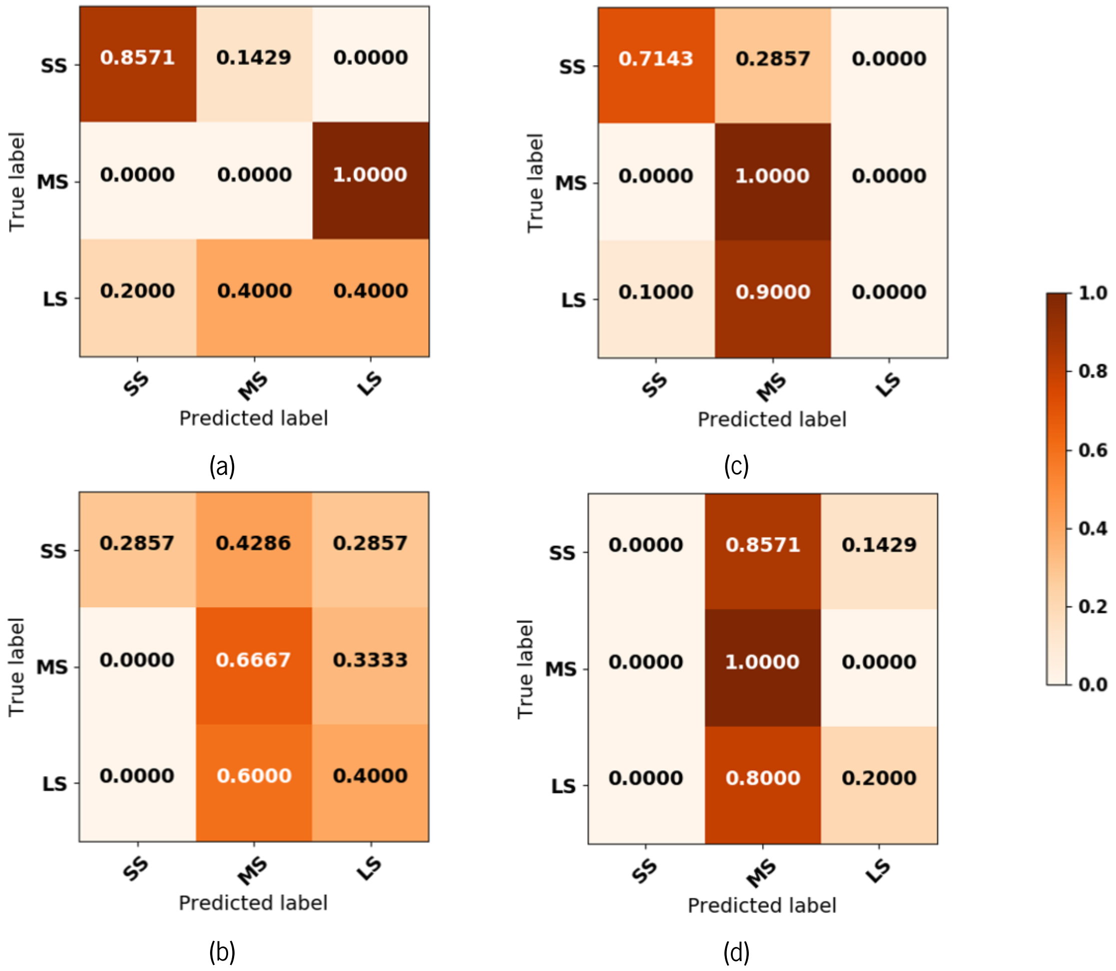
Firstly, it is possible to realize that none of the input combinations was selected as top-ranked by both criteria. Moreover, the scans used as input vary among the selected input combinations, as well as the z-score normalization method and the ROI used. After this input combinations selection, the chosen ones were retrained, but this time using data augmentation techniques, as mentioned in Section 4.2.3. It was possible to verify that the performance, based on MSE, increased or decreased with the usage of data augmentation. When the performance improved, the model considered as best was the one trained with artificial data and when the performance decreased, the one trained with the original images was considered. All the metrics for the top-ranked models with different input combinations are presented in Table 14.

Table 14: MSE, medianSE, stdSE and SpearmanR reported for the top-ranked models where the network was trained from scratch, using the test dataset. An indication if data augmentation (DA) techniques were is present

			Metrics			
	Input Combinations	**DA**	**MSE**	**medianSE**	**stdSE**	**SpearmanR**
Consistency	4seq_wseg_WB_B	yes	1.11×10^5	6.28×10^4	1.63×10^5	-0.239
	4seq_wseg_WT_B	no	1.02×10^5	5.07×10^4	1.96×10^5	-0.335
	FLAIR_woseg_AI_B	no	7.08×10^4	2.27×10^4	1.61×10^5	0.31
Best loss	FLAIR_wseg_AI_I	no	6.48×10^4	1.73×10^4	1.81×10^5	0.51
	T1ce_wseg_AI_B	yes	8.90×10^4	3.92×10^4	1.61×10^4	-0.026
	T1_wseg_AI_I	yes	7.20×10^4	2.14×10^4	1.58×10^5	0.279

Observing the MSE results, the two input combinations with the highest values use as input scans the FLAIR sequence and as ROI the whole brain and the whole tumor, while the ones with the lowest value use as input the four sequences with all the image as ROI. Observing the column with the SpearmanR results, the existence of a relationship between the predicted survival and the ground truth values for the models that use FLAIR sequence as input is evident. However, for the FLAIR_wseg_AI_I model, that relationship is considered moderate and for the model FLAIR_woseg_AI_B, the correlation between the predicted survival and the ground truth is considered low. The confusion matrices for these models, as well as the models FLAIR_wseg_AI_I and FLAIR_woseg_AI_B are present in Figure 19.

The differences in the strength of the correlation between the model FLAIR_wseg_AI_I and FLAIR_woseg_AI_B are clear when observing the confusion matrix. It is possible to observe that the input combination FLAIR_woseg_AI_B, in Figure 19d, classified almost all the predicted values as medium survivors. On the other hand, when observing the confusion matrix of the input combination FLAIR_wseg_AI_I, present in Figure 19a, it is clear that the model is predicting the short survivors mostly on their range. However, a small quantity of them is misclassified as medium survivals, which is not necessarily a problem, once the predictions are made using a regression method and a small increment can lead to a misclassification of the class. It is also interesting to note the good distinction between surviving more or less than 300 days that the input combination FLAIR_wseg_AI_I can do. Observing the Bland-Altman plot for this model, Figure 20, it is possible to see that for mean values lower than 300, the lower the mean is, the difference between the predicted survival and the ground truth is higher when compared with mean values higher than 300. The minus sign indicates that predicted values are higher than the ground truth. This reveals that the network is having trouble distinguishing the exactly survival time for short survivors and attribute to all of them a value between 200 and 300 days. For mean values greater than 350 days, it is evident that the model struggles to predict the OS value correctly. It predicts values higher and lower than the OS time.



(a)

(c)

(b)

(d)

Figure 19: Confusion matrices from the training from scratch approach: (a) FLAIR_wseg_AI_I, (b) FLAIR_wseg_AI_B, (c) FLAIR_woseg_AI_I, (d) FLAIR_woseg_AI_B models.

In order to study the impact of the usage of tumor segmentation and the type of normalization on the network training, the confusion matrices of the input combination FLAIR_wseg_AI_B and FLAIR_woseg_AI_I are shown on Figure 19b and Figure 19c. Observing the confusion matrices, it is clear that the responsibility for the good predictions related to the short survivors is an influence of the normalization to all images, once when the normalization is only made to the brain, the model identifies none or a small percentage of the short survivors. Observing the RISE maps of the same two models and also from FLAIR_woseg_AI_I, it is possible to understand that normalizing only the brain area takes the model to reject some areas outside of the brain, while with the normalization of the whole image, the region outside of the brain has medium importance. The identification with medium importance of regions outside of the brain may help the model to predict more accurately the OS for patients with a OS lower than 300 days. Moreover, comparing what Figures 21a and 21c have in common with Figure 21b is possible to see that both models give importance to some zones on the occipital lobe, what does not happen with the other model. The first axial slice represented of the FLAIR_wseg_AI_I and FLAIR_woseg_AI_I input combinations indicate that some importance is given to the tumor and surrounding tissues, which does not happen with the other model. Also, the model FLAIR_wseg_AI_B is giving a high level of importance

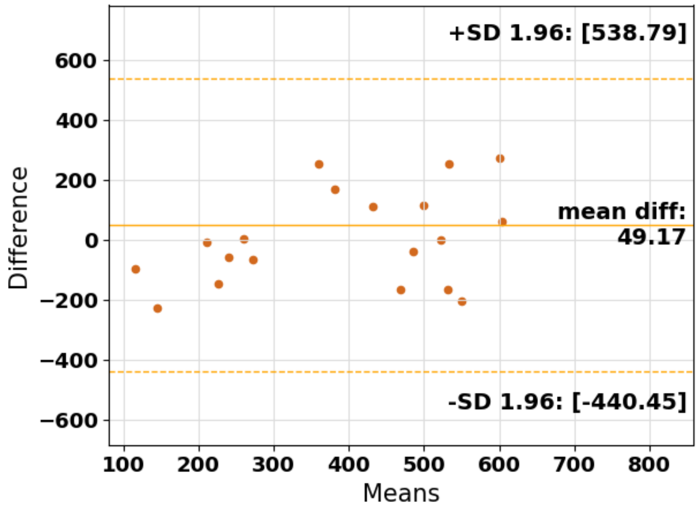


Figure 20: Bland-Altman plot from the FLAIR_wseg_AI_I model.

to the central zone of the brain, which can also harm the prediction of the short survivors. Furthermore, the importance given to the central zone of the brain seems to help to predict the survival of the long survivors. It is also important to note that these important zones seem to be a consequence of inputting the tumor segmentation during training.

5.4 Comparison between different approaches

Considering the most promising input combinations from each approach, it was found that reusing the features from pretrained CNN-S network is the approach with less potential for predicting the OS. Since it is the only one that uses 2D images as input, it is possible to affirm that the context of the scan as a whole is important. However, some problems may exist with the feature classification of the network and it may have comprised the whole performance of the model, as previously discussed.

Focusing on the other two approaches studied, the best input combinations from each approach have the following SpearmanR metrics: 0.457 when reusing features from predicting age and 0.51 when training from scratch. However, when observing the occlusion maps (Figure 17a and Figure 21a), the importance given to the zones outside of the brain are different. Moreover, the approach that reuses the features from predicting age gives importance to the whole side where the tumor is, while the other one gives importance to the center of the brain. However, there are some places on the scans that the network does not consider important on both models. For example, there is a low importance zone on the border of the brain on the right side of the axial slice of the scan. On the axial plane, it is possible to see that the zone of the tumor that is located more towards the top of the brain is also not considered important by the network. Comparing the confusion matrices of both approaches, it is evident that the model pretrained with age confuse the medium survivors with the long survivors and the sort survivors with the medium ones, while the model trained from scratch is able to predict the short survivors accurately. When comparing the metrics from these two approaches, the results are also divided: the model trained

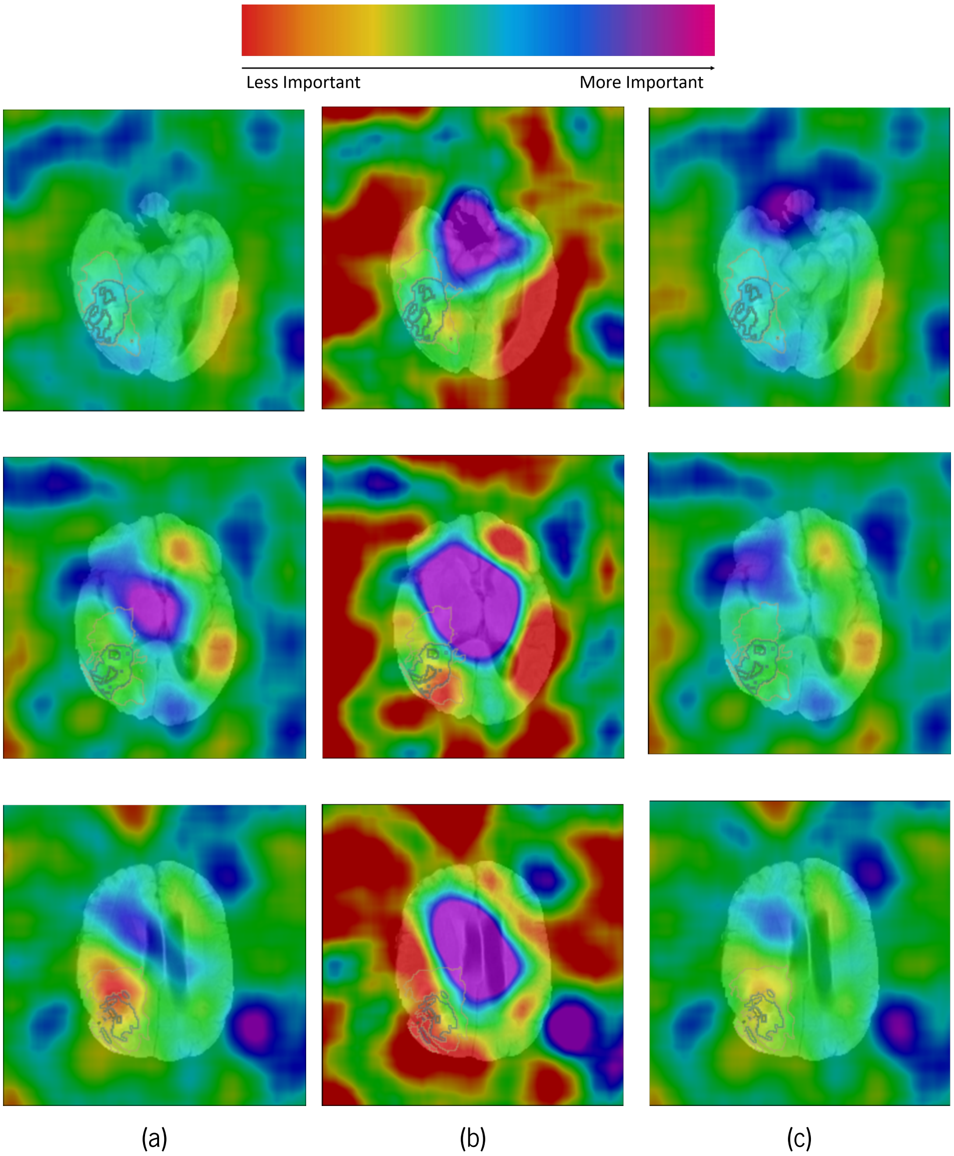


Figure 21: RISE maps with the contour of tumor segmentation from a patient with a short survival from the models with the following input combinations: (a) FLAIR_wseg_AI_I, (b) FLAIR_wseg_AI_B, (c) FLAIR_woseg_AI_I.

from scratch has higher medianSE and SpearmanR metrics, while the one that reuses the features for predicting age have a higher MSE and stdSE. Considering all the variables above mentioned makes it is hard to decide which model is better.

It is important to note that the scans used as input of the best input combination from each approach discussed in the preceding paragraph are the T1ce and FLAIR, which match with the ones that clinicians use when trying to estimate the OS. Moreover, the ROI used to input into the network was the all image in both cases. This indicates that the shape of the brain may have a word on the prediction of OS. Relatively to the different types of z-score normalization used during this work, it was hard to conclude which one was more efficient. In theory, the normalization only to the brain would be a better approach due to the impact that the background pixels would have on the calculus, however, on the training from scratch approach, the normalization of the all FLAIR scan seemed to be crucial for a better performance.

Considering that the only differences between the approach trained from scratch and the one that reuses features from predicting age are the initialization on the feature maps, as previously reported, for the best models by these approaches, it is possible to verify that the metrics obtained are close to each other. However, observing the confusion matrices is possible to report that the dispersion of the data is quite different: the approach that reuses the features from predicting age seems to have its predictions shrunk. This may be caused by the fact that the model was pretrained in a smaller range. In summary, to the knowledge of this work, transfer learning did not bring new results into the table.

Lastly, the usage of the consistency between training and validation loss as a criterion for the selection of the best models has brought no benefit into the selection of the best model of each approach, since no model that was on top-3 of this criteria was chosen as best, except when the model is also considered top-3 as the best loss criteria. However, on the approach where the model was trained from scratch, the second-best model was the third-ranked using the consistency criteria. So, maybe it is important to keep this in mind when comparing different inputs for the same network structure.

5.5 Comparison with a state-of-the-art approach

In the previous section, a comparison between the best input combinations from each approach was made and the ones with a better potential for predicting the OS of patients with GBM were discussed and chosen. Thus, in this section, their performance is going to be compared to a linear regression between the OS and the age of the patients, since age seems to have an impact on this prediction, as previously mentioned. The metrics achieved by the linear regression on the test group are listed in Table 15 and on Figure 22 is the confusion matrix of the respective outputs.

Table 15: MSE, medianSE, stdSE and SpearmanR metrics on the test group for the linear regression

MSE	medianSE	stdSE	SpearmanR
4.77×10^4	1.32×10^4	1.18×10^5	0.630

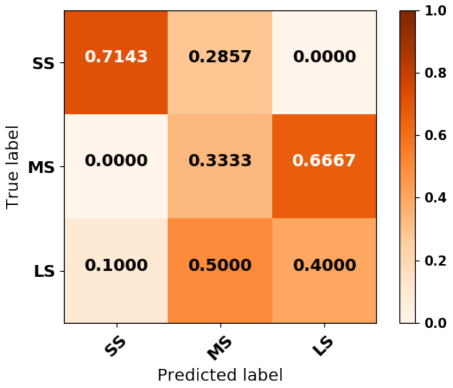


Figure 22: Confusion matrix for the test dataset for the linear regression.

When comparing the metrics from the linear regression, on Table 15, with the best input combination from the training from scratch approach and the reusing the features from age, on Table 14 and Table

10, respectively, it is clear that the age feature outperformed the other two approaches. However, when comparing the confusion matrix from the input combination FLAIR_wseg_AI_I trained from scratch (Figure 19b) with the one from the linear regression, the first one is better at predicting the short survivors. Despite the apparent improvements in the prediction of the short survivors, the linear approach is the only one that can predict the medium survivors on their range, which may be a cause of the higher SpearmanR value.

5.6 Summary

This chapter begins with the evaluation of the top-ranked input of each approach and the impact of using data augmentation, which was negative most of the cases. In addition to the MSE, medianSE, stdSE metrics, the usage of the SpearmanR, confusion matrices, Bland-Altman plots and the RISE maps helped to understand the learning process of the top-ranked input combinations. After analyzing each approach individually, the top-ranked input combinations were compared, starting from the architectures and the scans used for inputs, followed by a metrics comparison and, finally, the comparison between zones selected as important by the different approaches was made. It was not possible to conclude about the input scan and type of z-score normalization that benefits the prediction of the OS. With the RISE maps, the zones considered more important were where the tumor is most frequent. In the end, the 3D network trained from scratch showed more potential to predict the OS than the other analyzed combination of architectures and weight initializations. When comparing the performance of this model against a state-of-the-art model, a linear regression of OS against the age from the patients, the second one outperformed the developed deep learning model for the prediction of the OS of patients with glioblastoma.

Chapter 6

Conclusions and Future Outlook

During this work, a CNN was trained from scratch and two others received knowledge from previous tasks. One of them is a 2D network that was pretrained using the ILSVRC dataset and the other one is a 3D network pretrained using healthy brains from the IXI dataset. For each of the architectures, different ROI of different scans combinations with two different types of normalizations were inputted to the networks. The top-ranked combination of inputs were selected considering two criteria: the model with the lower loss or the model were the difference between the training and validation loss is lowest.

The findings from this study point towards the idea that 2D CNNs have less potential to predict OS of patients with GBM when compared to the other 3D CNNs models studied. These other two approaches have the same 3D CNNs architecture, but different weight initialization. The metrics for the best combination of inputs of each approach are close to each other and the RISE maps show common non-important regions that overlap with regions frequently reported by the doctors to be where the tumor is less frequent, such as the top of the brain and cerebellum. The sequences used as an input that obtained a better performance were T1ce and FLAIR, which is also in line with the doctors' findings. Relatively to the type of normalization most favorable for improving the survival prediction, no conclusions were obtained, since these two different approaches have as best input scans with the two different types of normalization. The ROI that obtained a better performance was all image. This fact, together with the RISE maps from the best combination of input, suggests that the network preferred a broader context for capturing features rather than characteristics from the tumor zone. For example, some distances inside of the brain, as well as borders of the brain, are the regions that the network considers to have the more significant potential for predicting the OS. Both approaches also seem to get some cues from the background to predict the OS. To the knowledge of this work, the learning transference did not bring any benefit into the work and the prediction of the survival is better when creating a linear regression using age and the OS. This indicates that the non-image feature age has an important paper when it comes to facilitating the pre-treatment care of patients with GBM. However, the confusion matrix of the model trained from scratch suggested that this one is better at identifying the patients with a survival smaller than 300 days.

Despite the fact that the results obtained are behind the expectations, the combination of different metrics, such as MSE, stdSE, medianSE, the usage of the SpearmanR coefficient, the confusion matrices, Bland-Altman plots and RISE saliency maps helped to interpreter the models created and, consequently, understand what was happening with the predictions of the networks and the steps that should be taken next.

Considering some limitations of the architectures used for the prediction of the OS already discussed, some future work is going to be proposed. Firstly, related to the 2D CNN, it is proposed that another hidden layer is added before the output layer, with a reasonable number of neurons that would help the network to widen its predictions. For the same propose, the non-transference of the weights relative to the feature classification on the CNN that reuses features for predicting age is suggested. Relatively to this last approach, it would be interesting to increase the data used for the pretraining of the model: use other datasets with healthy patients in order to increase the generalization of the model. The usage of the exactly pretrained model from Cole et al. [55] would also be interesting. With both approaches that use transfer learning, it would be interesting to transfer knowledge only to some layers, rejecting the classification ones, and randomly initialize the ones that do not have receive knowledge. Instead of using a deep model to directly predict the OS, a two-stage model can be developed. In the first stage, a finetuned deep network can be used for feature extraction. This network must be trained or finetuned using the BraTS dataset. Then, after performing a feature selection, a machine learning model, such as SVM or Random Forest, is used for predicting the survival time.

Once the model trained from scratch shown that was better at predicting the survival of patients with an OS less than 300 days, it would be interesting to add the non-image feature age to this prediction, due to its SpearmanR value. The addition of volumetric and distance features can also help to make a better prediction. The addition of other non-image features as KPS, genetic and molecular information can be a benefit for the prediction of the OS of patients with GBM and, consequently, choose the best pre-treatment care for those patients.

References

[1] Kaja Urba□ska, Justyna Sokołowska, Maciej Szmidt, and Paweł Sysa. Glioblastoma multiforme–an overview. *Contemporary oncology*, 18(5):307, 2014.

[2] Farina Hanif, Kanza Muzaffar, Kahkashan Perveen, Saima M Malhi, and Shabana U Simjee. Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment. *Asian Pacific journal of cancer prevention: APJCP*, 18(1):3, 2017.

[3] Nova F Smedley, Benjamin M Ellingson, Timothy F Cloughesy, and William Hsu. Longitudinal patterns in clinical and imaging measurements predict residual survival in glioblastoma patients. *Scientific reports*, 8(1):14429, 2018.

[4] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[5] Abhimanyu S Ahuja. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, 7:e7702, 2019.

[6] Eric C Holland. Glioblastoma multiforme: the terminator. *Proceedings of the National Academy of Sciences*, 97(12):6242–6244, 2000.

[7] David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.

[8] Artemiy S Silantyev, Luca Falzone, Massimo Libra, Olga I Gurina, Karina Sh Kardashova, Taxiarchis K Nikolouzakis, Alexander E Nosyrev, Christopher W Sutton, Panayiotis D Mitsias, and Aristides Tsatsakis. Current and future trends on diagnosis and prognosis of glioblastoma: From molecular biology to proteomics. *Cells*, 8(8):863, 2019.

[9] JR Simpson, J Horton, C Scott, WJ Curran, P Rubin, J Fischbach, S Isaacson, M Rotman, SO Asbell, JS Nelson, et al. Influence of location and extent of surgical resection on survival of patients with

glioblastoma multiforme: results of three consecutive radiation therapy oncology group (rtog) clinical trials. *International Journal of Radiation Oncology* Biology* Physics*, 26(2):239–244, 1993.

[10] Glioblastoma multiforme. URL `https://www.aans.org/Patients/Neurosurgical-Conditions-and-Treatments/Glioblastoma-Multiforme`.

[11] Suvi Larjavaara, Riitta Mäntylä, Tiina Salminen, Hannu Haapasalo, Jani Raitanen, Juha Jääskeläinen, and Anssi Auvinen. Incidence of gliomas by anatomic location. *Neuro-oncology*, 9(3):319–325, 2007.

[12] Prakash Ambady, Chetan Bettegowda, and Matthias Holdhoff. Emerging methods for disease monitoring in malignant gliomas. *CNS oncology*, 2(6):511–522, 2013.

[13] Taylor A Wilson, Matthias A Karajannis, and David H Harter. Glioblastoma multiforme: State of the art and future therapeutics. *Surgical neurology international*, 5, 2014.

[14] Michael Weller. Novel diagnostic and therapeutic approaches to malignant glioma. *Swiss Med Wkly*, 141(w13210):2, 2011.

[15] Dongsheng Guo, Baofeng Wang, Fuxin Han, and Ting Lei. Rna interference therapy for glioblastoma. *Expert opinion on biological therapy*, 10(6):927–936, 2010.

[16] Der-Yang Cho, Wen-Kuang Yang, Han-Chung Lee, Den-Mei Hsu, Hung-Lin Lin, Shinn-Zong Lin, Chun-Chung Chen, Horng-Jyh Harn, Chun-Lin Liu, Wen-Yuan Lee, et al. Adjuvant immunotherapy with whole-cell lysate dendritic cells vaccine for glioblastoma multiforme: a phase ii clinical trial. *World neurosurgery*, 77(5-6):736–744, 2012.

[17] Nedret Altiok, Melike Ersoz, and Meral Koyuturk. Estradiol induces jnk-dependent apoptosis in glioblastoma cells. *Oncology letters*, 2(6):1281–1285, 2011.

[18] Philipp Kickingereder, Sina Burth, Antje Wick, Michael Götz, Oliver Eidel, Heinz-Peter Schlemmer, Klaus H Maier-Hein, Wolfgang Wick, Martin Bendszus, Alexander Radbruch, et al. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*, 280(3):880–889, 2016.

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[20] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[21] Thomas M. Mitchell. *Machine learning*. McGraw-Hill, 1997.

[22] Chollet François. *Deep learning with Python*. Manning Publications Co., 2018.

[23] Dipanjan Sarkar, Raghav Bali, and Tamoghna Ghosh. *Hands-on transfer learning with Python: implement advanced deep learning and neural network models using TensorFlow and Keras*. Packt Publishing, 2018.

[24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[25] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[26] Geoffrey Hinton. Deep learning—a technology with the potential to transform health care. *Jama*, 320 (11):1101–1102, 2018.

[27] Josh Patterson and Adam Gibson. *Deep learning: a practitioners approach*. OReilly, 2017.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[29] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[30] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[31] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[32] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[36] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[37] Yiming Ding, Jae Ho Sohn, Michael G Kawczynski, Hari Trivedi, Roy Harnish, Nathaniel W Jenkins, Dmytro Lituiev, Timothy P Copeland, Mariam S Aboian, Carina Mari Aparici, et al. A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain. *Radiology*, 290 (2):456–464, 2018.

[38] Yao Xie, Ge Gao, and Xiang'Anthony' Chen. Outlining the design space of explainable intelligent systems for medical diagnosis. *arXiv preprint arXiv:1902.06019*, 2019.

[39] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Towards medical xai. *arXiv preprint arXiv:1907.07374*, 2019.

[40] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[41] Housam Khalifa Bashier Babiker and Randy Goebel. An introduction to deep visual explanation. *arXiv preprint arXiv:1711.09482*, 2017.

[42] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2015.

[43] James H Thrall, Xiang Li, Quanzheng Li, Cinthia Cruz, Synho Do, Keith Dreyer, and James Brink. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *Journal of the American College of Radiology*, 15(3):504–508, 2018.

[44] Meg F Bobo, Shunxing Bao, Yuankai Huo, Yuang Yao, Jack Virostko, Andrew J Plassard, Ilwoo Lyu, Albert Assad, Richard G Abramson, Melissa A Hilmes, et al. Fully convolutional neural networks improve abdominal organ segmentation. In *Medical Imaging 2018: Image Processing*, volume 10574, page 105742V. International Society for Optics and Photonics, 2018.

[45] Mohamed Shehata, Fahmi Khalifa, Ahmed Soliman, Mohammed Ghazal, Fatma Taher, Mohamed Abou El-Ghar, Amy C Dwyer, Georgy Gimel'farb, Robert S Keynton, and Ayman El-Baz. Computer-aided diagnostic system for early detection of acute renal transplant rejection using diffusion-weighted mri. *IEEE Transactions on Biomedical Engineering*, 66(2):539–552, 2018.

[46] Ashirbani Saha, Michael R Harowicz, Lars J Grimm, Connie E Kim, Sujata V Ghate, Ruth Walsh, and Maciej A Mazurowski. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. *British journal of cancer*, 119(4):508, 2018.

[47] HM Chan, Bas HM van der Velden, Claudette E Loo, and Kenneth GA Gilhuijs. Eigentumors for prediction of treatment failure in patients with early-stage breast cancer using dynamic contrast-enhanced mri: a feasibility study. *Physics in Medicine & Biology*, 62(16):6467, 2017.

[48] Ran Shadmi, Victoria Mazo, Orna Bregman-Amitai, and Eldad Elnekave. Fully-convolutional deep-learning based system for coronary calcium score prediction from non-contrast chest ct. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 24–28. IEEE, 2018.

[49] Ruida Cheng, Holger R Roth, Nathan S Lay, Le Lu, Baris Turkbey, William Gandler, Evan S McCreedy, Thomas J Pohida, Peter A Pinto, Peter L Choyke, et al. Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks. *Journal of Medical Imaging*, 4(4): 041302, 2017.

[50] Junichiro Ishioka, Yoh Matsuoka, Sho Uehara, Yosuke Yasuda, Toshiki Kijima, Soichiro Yoshida, Minato Yokoyama, Kazutaka Saito, Kazunori Kihara, Noboru Numao, et al. Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm. *BJU international*, 122(3):411–417, 2018.

[51] Rakesh Shiradkar, Soumya Ghose, Ivan Jambor, Pekka Taimen, Otto Ettala, Andrei S Purysko, and Anant Madabhushi. Radiomic features from pretreatment biparametric mri predict prostate cancer biochemical recurrence: preliminary findings. *Journal of Magnetic Resonance Imaging*, 48(6):1626–1636, 2018.

[52] Jyoti Islam and Yanqing Zhang. Brain mri analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain informatics*, 5(2):2, 2018.

[53] Jens Kleesiek, Gregor Urban, Alexander Hubert, Daniel Schwarz, Klaus Maier-Hein, Martin Bendszus, and Armin Biller. Deep mri brain extraction: a 3d convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, 2016.

[54] 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.

[55] James H Cole, Rudra PK Poudel, Dimosthenis Tsagkrasoulis, Matthan WA Caan, Claire Steves, Tim D Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 2017.

[56] 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.

[57] Katja Franke, Gabriel Ziegler, Stefan Klöppel, Christian Gaser, Alzheimer's Disease Neuroimaging Initiative, et al. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: exploring the influence of various parameters. *Neuroimage*, 50(3):883–892, 2010.

[58] Zeina A Shboul, Lasitha Vidyaratne, Mahbubul Alam, and Khan M Iftekharuddin. Glioblastoma and survival prediction. In *International MICCAI Brainlesion Workshop*, pages 358–368. Springer, 2017.

[59] Ahmad Chaddad, Siham Sabri, Tamim Niazi, and Bassam Abdulkarim. Prediction of survival with multi-scale radiomic analysis in glioblastoma patients. *Medical & biological engineering & computing*, 56(12):2287–2300, 2018.

[60] Parita Sanghani, Beng Ti Ang, Nicolas Kon Kam King, and Hongliang Ren. Overall survival prediction in glioblastoma multiforme patients from volumetric, shape and texture features using machine learning. *Surgical oncology*, 27(4):709–714, 2018.

[61] Lina Chato and Shahram Latifi. Machine learning and deep learning techniques to predict overall survival of brain tumor patients using mri images. In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 9–14. IEEE, 2017.

[62] Dong Nie, Han Zhang, Ehsan Adeli, Luyan Liu, and Dinggang Shen. 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 212–220. Springer, 2016.

[63] Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, 7(1):10353, 2017.

[64] Dong Nie, Junfeng Lu, Han Zhang, Ehsan Adeli, Jun Wang, Zhengda Yu, LuYan Liu, Qian Wang, Jinsong Wu, and Dinggang Shen. Multi-channel 3d deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Scientific reports*, 9(1):1103, 2019.

[65] Alexander FI Osman. Automated brain tumor segmentation on magnetic resonance images and patient's overall survival prediction using support vector machines. In *International MICCAI Brainlesion Workshop*, pages 435–449. Springer, 2017.

[66] Xue Feng, Nicholas Tustison, and Craig Meyer. Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. In *International MICCAI Brainlesion Workshop*, pages 279–288. Springer, 2018.

[67] JR Simpson, J Horton, C Scott, WJ Curran, P Rubin, J Fischbach, S Isaacson, M Rotman, SO Asbell, JS Nelson, et al. Influence of location and extent of surgical resection on survival of patients with glioblastoma multiforme: results of three consecutive radiation therapy oncology group (rtog) clinical trials. *International Journal of Radiation Oncology* Biology* Physics*, 26(2):239–244, 1993.

[68] Qihua Li, Hongmin Bai, Yinsheng Chen, Qiuchang Sun, Lei Liu, Sijie Zhou, Guoliang Wang, Chaofeng Liang, and Zhi-Cheng Li. A fully-automatic multiparametric radiomics model: towards reproducible and prognostic imaging signature for prediction of overall survival in glioblastoma multiforme. *Scientific reports*, 7(1):14331, 2017.

[69] Zhi-Cheng Li, Qihua Li, Qiuchang Sun, Ronghui Luo, and Yinsheng Chen. Identifying a radiomics imaging signature for prediction of overall survival in glioblastoma multiforme. In *2017 10th Biomedical Engineering International Conference (BMEiCON)*, pages 1–4. IEEE, 2017.

[70] Li Sun, Songtao Zhang, and Lin Luo. Tumor segmentation and survival prediction in glioma with deep learning. In *International MICCAI Brainlesion Workshop*, pages 83–93. Springer, 2018.

[71] Ujjwal Baid, Sanjay Talbar, Swapnil Rane, Sudeep Gupta, Meenakshi H Thakur, Aliasgar Moiyadi, Siddhesh Thakur, and Abhishek Mahajan. 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, 2018.

[72] Li Sun, Songtao Zhang, and Lin Luo. Tumor segmentation and survival prediction in glioma with deep learning. In *International MICCAI Brainlesion Workshop*, pages 83–93. Springer, 2018.

[73] Alain Jungo, Richard McKinley, Raphael Meier, Urspeter Knecht, Luis Vera, Julián Pérez-Beteta, David Molina-García, Víctor M Pérez-García, Roland Wiest, and Mauricio Reyes. Towards uncertainty-assisted brain tumor segmentation and survival prediction. In *International MICCAI Brainlesion Workshop*, pages 474–485. Springer, 2017.

[74] Xue Feng and Craig Meyer. Patch-based 3d u-net for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2017.

[75] Elodie Puybareau, Guillaume Tochon, Joseph Chazalon, and Jonathan Fabrizio. Segmentation of gliomas and prediction of patient overall survival: A simple and fast procedure. In *International MICCAI Brainlesion Workshop*, pages 199–209. Springer, 2018.

[76] Ixi dataset. URL `https://brain-development.org/ixi-dataset/`.

[77] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4: 170117, 2017.

[78] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[79] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[80] Mavuto M Mukaka. A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71, 2012.

[81] Dennis E. Hinkle, Stephen G. Jurs, and William Wiersma. *Applied statistics for the behavioral sciences*. W. Ross MacDonald School Resource Services Library, 2011.

[82] Anke Meyer-base. *Biomedical signal analysis - contemporary methods and applications*. Mit Press Ltd, 2010.

[83] Davide Giavarina. Understanding bland altman analysis. *Biochemia medica: Biochemia medica*, 25 (2):141–151, 2015.

Chapter 7

Attachments

7.1 Train for age prediction

Here, the MSE values for training and validation of the three models used for the age prediction are presented. While training, data augmentation was also performed and consisted of rotations between $\pm 15^\circ$ on the axial and coronal planes.

The results presented on Table 16 were obtained with a learning rate of 5×10^{-7} with constant decay of 0.8%.

Table 16: Training and Validation MSE for the different inputs when predicting age

	Training	**Validation**
Input	**MSE**	**MSE**
[T1]	12,41	65,96
[T2]	8,303	80,57
[T1,T1,T2,T2]	8,011	64,35