REPÚBLICA PORTUGUESA
POCH
UNIÃO EUROPEIA
Fundo Social Europeu
FCT Fundação para a Ciência e a Tecnologia

Automatic Segmentation and Classification of Brain Tumors based on Multisequence MRI Images with Deep Learning Methods
Sérgio Rafael Mano Pereira
UMinho | 2018

Universidade do Minho
Escola de Engenharia

Sérgio Rafael Mano Pereira

Automatic Segmentation and Classification of Brain Tumors based on Multisequence MRI Images with Deep Learning Methods

Programa de Doutoramento em Informática (MAP-i) das Universidades do Minho, de Aveiro e do Porto

Universidade do Minho
universidade de aveiro
U.PORTO

julho de 2018

Universidade do Minho
Escola de Engenharia

Sérgio Rafael Mano Pereira

Automatic Segmentation and Classification of Brain Tumors based on Multisequence MRI Images with Deep Learning Methods

Programa de Doutoramento em Informática (MAP-i)
das Universidades do Minho, de Aveiro e do Porto

Universidade do Minho

universidade de aveiro

U.PORTO

Trabalho realizado sob a orientação do
Professor Doutor Carlos Alberto Batista Silva
e do
Professor Doutor Victor Manuel Rodrigues Alves

julho de 2018

STATEMENT OF INTEGRITY

I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.
I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho

University of Minho, July 30th, 2018

Full Name: Sérgio Rafael Mano Pereira

Signature: Sérgio Rafael Mano Pereira

Acknowledgments

I would like to express my gratitude to my supervisors. I am very thankful to Professor Carlos A. Silva for the opportunity for conducting this work. I also thank the motivation, dedication, guidance, knowledge sharing, and for pushing me beyond what I thought I would be able to do. All the discussions and brainstorms that certainly shaped this thesis, as well. I also thank Professor Victor Alves for the opportunity, the feedback, and all the support. For all this, I am very grateful, and I wish them all the best.

I also thank Professor Higino Correia for integrating me in his research team after my Biomedical Engineering studies. Without it, I would not be around and have the chance to do this thesis work.

I thank Professor Mauricio Reyes for welcoming me in his Medical Image Analysis group at the Institute for Surgical Technologies & Biomechanics of the University of Bern. I also thank Dr. Raphael Meier, with whom I collaborated. They were always very open, provided insightful discussion, and shared their knowledge with me. This research stay was a memorable experience that certainly contributed for my professional and personal growth. Also, the friends I made there were crucial for making my stay a very pleasant one. Carlos, Raphael, and Vimal, thanks a lot!

A special thanks goes to Adriano Pinto for his friendship and all the constructive discussions and brainstorms. Also, it was nice to work with Américo during his Master's thesis. Some of the insights and discussions ended up in parts of this thesis. I wish all the best for both of them; I am pretty sure they will have many successes.

Several factors contribute for the success of a PhD. Arguably, a good work environment and a lab's weekly football match are some of them. I was very lucky for being part of a lab of cool people, whom I call friends.

A huge thanks goes to my dear friends for their friendship and support. Hanging out with them made wonders for refreshing the soul and mind. A very special thanks goes to Sieun for all the support, companionship, and cheering up.

Finally, my parents and brother. Their unconditional love, support, and encouragement was a constant, not only during this thesis, but all my life. This thesis is also theirs.

This work was supported by a scholarship awarded to Sérgio Rafael Mano Pereira by Fundação para a Ciência e Tecnologia (FCT), Portugal, with scholarship reference PD/BD/105803/2014.

Abstract

Automatic Segmentation and Classification of Brain Tumors based on Multisequence MRI Images with Deep Learning Methods

Gliomas are the most common primary brain tumors. Unfortunately, these neoplasms hold the worst prognosis among all brain tumors, as well. They can be broadly categorized as low or high grade gliomas. Magnetic Resonance Imaging is the standard imaging technique for their assessment. Using it, physicians can extract measurements that are crucial for treatment planning and follow-up. Notwithstanding, manual segmentation is time-demanding and prone to variability. Also, tumor grading by biopsy is very important, but it is invasive, and prone to sampling error. Therefore, automatic approaches for both segmentation and grading are needed. However, these tasks are quite challenging due the the heterogeneity of gliomas, as well as the variability among Magnetic Resonance Imaging scans. This makes it difficult to model brain tumors from prior knowledge.

Machine Learning algorithms can learn how to perform a task directly from the data. Some of these algorithms may be categorized as Representation Learning if they can learn features directly from the data. Among these methods, Deep Learning is a group of Representation Learning algorithms that learn multiple levels of representations.

In the past years, Deep Learning-based methods have shown remarkable performances. Hence, the aim of this work was to investigate Deep Learning methods, and use them for the automatic segmentation and grade classification of brain tumors in multisequence structural Magnetic Resonance Imaging. Additionally, an often cited setback of these complex models is their "black box" behavior. Thus, in this work we also studied interpretability of Machine Learning algorithms applied to medical imaging. Therefore, we built our work on: brain tumor image analysis in Magnetic Resonance Imaging, Machine Learning with focus on Representation Learning, and its interpretability.

We investigated Convolutional Neural Networks for the task of segmentation. As a starting point, we studied a classification Convolutional Neural Network. We were able to show its effectiveness, as well as the importance of careful pre-processing. However, afterwards we adopted a more efficient Fully Convolutional Network approach. In this setting, we proposed a hierarchical approach for dealing with class imbalance. Finally, the relationships among channels of feature maps were studied. We proposed and showed the benefits of recombination and recalibration of feature maps in the context of Fully Convolutional Networks for semantic segmentation.

Automatic glioma grading from structural Magnetic Resonance Imaging images is challenging due to their large heterogeneity. Additionally, a tumor mass must be graded as a whole. Therefore, we propose 3D Convolutional Neural Networks for automatic glioma grading. Since Convolutional Neural Networks learn features directly from the data, it allows one to bypass the need for a very accurate segmentation that is often seen in radiomics-based approaches, which use hand-crafted features.

"Black box" systems may pose trusting issues when deployed in critical domains, such as the medical field. This is due to professionals not being able to explain certain predictions. Therefore, interpretability of machine learning systems is a crucial field of research, given the high performances currently achieved with these systems. We first investigated this topic in a Restricted Boltzmann Machine and Random Forest classifier system in the context of segmentation. We proposed methodologies for both global and local interpretability. The former is targeted at understanding if the system learned the relevant relations in the data, while the latter is focused on explaining individual predictions. We were able to confirm if the system learned correct patterns, but we also found a bias in the database. Later, we employ interpretability methodologies to inspect the 3D Convolutional Neural Network for glioma grading. With it, we were able to catch and correct an issue during pre-processing. Hence, we provide tools and study cases that show how interpretability not only helps in increasing trust, but it may also be useful during the development cycle.

Finally, all methodologies developed in this work were validated in publicly available databases. This ensures a fair comparison with the state of the art. Additionally, it enables future work to be directly compared with us.

Resumo

Segmentação e Classificação Automática de Tumores Cerebrais Baseado em Imagem por Ressonância Magnética Multi-sequência com Métodos de Deep Learning

Os gliomas são os tumores cerebrais primários mais frequentes. Infelizmente, estas neoplasias são também as que que têm os priores prognósticos entre os tumores cerebrais. Estes podem ser categorizados em gliomas de baixo ou de alto grau. A Imagem por Ressonância Magnética é a técnica imagiológica padrão para avaliar tumores cerebrais. Desta forma, os médicos podem extrair medições que são da maior importância para o planeamento do tratamento e para monitorização. Não obstante, a segmentação manual das imagens é um processo demorado e suscetível a variabilidade. A classificação dos gliomas quanto ao seu grau através de biópsia é também muito importante, mas é um processo invasivo, e suscetível a erros de amostragem. Assim sendo, são necessárias abordagens automáticas para ambas as tarefas. Contudo, são problemas bastante complexos devido à heterogeneidade dos gliomas, mas também devido à variabilidade das imagens de Ressonância Magnética. Isto faz com que seja difícil modelar os tumores cerebrais a partir de conhecimento *a priori*.

Os algoritmos de Aprendizagem Automática conseguem aprender a executar uma determinada tarefa diretamente a partir dos dados. Alguns destes algoritmos podem ser categorizados como Aprendizagem de Características se forem capazes de aprender características diretamente a partir dos dados. Entre estes métodos, Aprendizagem Profunda é um grupo de algoritmos de Aprendizagem de Características que aprendem múltiplos níveis de características.

Nos últimos anos, métodos baseados em Aprendizagem Profunda têm mostrado desempenhos notáveis. Assim, um dos objetivos deste trabalho foi a investigação de Aprendizagem Profunda no contexto de segmentação. Um segundo objetivo foi explorar Aprendizagem Profunda para a classificação automática dos graus dos gliomas a partir de Imagem por Ressonância Magnética estrutural. Finalmente, um problema que é muitas vezes apontado a estes modelos complexos é a sua natureza de "caixa preta". Assim sendo, neste trabalho também foi investigada a interpretabilidade de sistemas de Aprendizagem Automática. Portanto, há três grandes temas sobre os quais nós construímos o nosso trabalho: análise de imagem de tumores cerebrais em Imagem por Ressonância Magnética, Aprendizagem Automática com foco em Aprendizagem de Características, e interpretabilidade.

Foram investigadas Redes Neuronais Convolucionais para a tarefa de segmentação. Como ponto de partida, nós estudamos Redes Neuronais Convolucionais para classificação. Assim, conseguimos mostrar

a sua eficácia, tal como a importância de um pré-processamento cuidadoso. Contudo, posteriormente, nós adotamos uma abordagem mais eficiente denominada por Redes Totalmente Convolucionais. Com esta nova rede, nós propusemos uma abordagem hierárquica que nos permitiu lidar melhor com o desbalanceamento de classes. Por fim, as relações entre os canais dos mapas de características foram estudadas. Nós propusemos e mostrámos as vantagens da recombinação e recalibração dos mapas de características no contexto de Redes Totalmente Convolucionais para segmentação semântica.

A classificação automática do grau dos gliomas a partir de Imagem por Ressonância Magnética estrutural é complexa devido à sua grande heterogeneidade. Adicionalmente, uma massa tumoral deve ser classificada como um todo. Assim sendo, nós propusemos Redes Neuronais Convolucionais 3D para a classificação automática do grau dos gliomas. Uma vez que as Redes Neuronais Convolucionais aprendem características diretamente a partir dos dados, é possível evitar a necessidade de uma segmentação muito precisa, tal como é comumente observado nas abordagens mais tradicionais baseadas em *radiomics*.

Os sistemas "caixa preta" podem colocar problemas relacionados com confiança quando são colocados em domínios críticos, como é o caso do campo da medicina. Isto deve-se aos profissionais não serem capazes de explicar certas predições. Assim sendo, a interpretabilidade de sistemas de Aprendizagem Automática é uma área de investigação crucial, dado os elevados desempenhos atualmente atingidos com estes sistemas. Nós investigamos este tópico inicialmente com um sistema constituído por uma *Restricted Boltzmman Machine* e um classificador *Random Forest* no contexto de segmentação. Nós propusemos metodologias para interpretabilidade global e local. A primeira está direcionada para se perceber se o sistema aprendeu relações relevantes nos dados, enquanto a última foca-se mais em explicar predições individuais. Nós conseguimos confirmar se o modelo aprendia padrões corretos, mas também conseguimos encontrar um enviesamento na base de dados. Posteriormente, também aplicamos metodologias de interpretabilidade para inspecionar a Rede Neuronal Convolucional 3D para classificação do grau dos gliomas. Assim, conseguimos identificar e corrigir um problema durante o pré-processamento. Desta forma, nós fornecemos ferramentas e casos de estudo que mostram como a interpretabilidade é útil não só para aumentar a confiança, mas também durante o ciclo de desenvolvimento.

Finalmente, todas as metodologias desenvolvidas neste trabalho foram validadas em bases de dados publicamente disponíveis. Isto garante uma comparação justa com o estado da arte. Adicionalmente, isto permite que trabalhos futuros possam ser diretamente comparados com os nossos métodos.

Contents

List of Figures

List of Tables

Acronyms

Adam Adaptive Moment Estimation.

BRATS Brain Tumor Segmentation Challenge.

CADx Computer-aided Diagnosis.

CD Contrastive Divergence.

CNN Convolutional Neural Network.

DT Decision Tree.

FCN Fully Convolutional Network.

FLAIR T2-weighted Fluid-Attenuated Inversion Recovery.

GradCAM Gradient-weighted Class Activation Mapping.

HGG High Grade Glioma.

LGG Low Grade Glioma.

MDI Mean Decrease Impurity.

MI Mutual Information.

MRF Markov Random Field.

MRI Magnetic Resonance Imaging.

RANO Response Assessment in Neuro-Oncology Criteria.

RBM Restricted Boltzmann Machine.

RECIST Response Evaluation Criteria in Solid Tumors.

ReLU Rectified Linear Unit.

RF Random Forest.

ROI Region of Interest.

RR Recombination and Recalibration.

SE Squeeze-and-Excitation.

SGD Stochastic Gradient Descent.

T1 T1-weighted.

T1c Post-contrast T1-weighted.

T2 T2-weighted.

WHO World Health Organization.

Chapter 1

Introduction

The main goals of this work are the study of Deep Learning, and automatic brain tumor segmentation and grade classification from multisequence structural Magnetic Resonance Imaging (MRI). Given its importance in the critical medical domain, we also investigate the interpretability of Machine Learning-based methods. Therefore, we build our work on: brain tumor image analysis in Magnetic Resonance Imaging, Machine Learning with focus on Representation Learning, and Interpretability of Machine Learning-based systems. This chapter introduces these topics in a general way, and how they are connected in the context of the work here described. In Section 1.2, the main objectives driving this thesis are outlined. In line with it, in Section 1.3, we describe the main contributions, and we list the academic outputs – publications and results in international segmentation challenges. Finally, the chapter closes with a description of the structure of the remaining document and a general overview of the overall work.

Contents

1.1 Context and Motivation

Every year, several types of cancer affect millions of people worldwide. This is a huge problem, since it impairs the quality of life of those patients and their families. In 2012, 8.2 million people succumbed from causes related to cancer, while 14.1 million new cancer cases were diagnosed (Ferlay et al., 2015). From the 27 studied cancers, brain tumors are not the most common, being placed in the 16^{th} position, which

accounts for 1.8% of the new cancer diagnosis. Nevertheless, they can be associated with high mortality by being the 11th most lethal neoplasms, accounting for 2.3% of cancer-related deaths. The prevalence of brain tumors is higher on developed countries. However, it is likely due to the better diagnostic tools in those countries (Ferlay et al., 2015).

The World Health Organization (WHO) broadly divides brain tumors into four grades (Louis et al., 2007, 2016). The higher the grade, the more malignant and proliferative is the tumor. Patients with brain tumors of grades I and II have an average life expectancy of more than 5 years. Unfortunately, life expectancy quickly drops as the grade increases, being 2-3 years in grade III, and only 14 months in the case of grade IV (Louis et al., 2007, 2016; Van Meir et al., 2010). Among the types of brain tumors, gliomas are the most common and malignant ones.

Magnetic Resonance Imaging stands as the standard imaging technology for assessing the structure of brain tumors in clinical practice. This is due to its good contrast for soft tissues, and the availability of multi-sequence acquisitions that are conspicuous for the various tumor compartments (Mabray et al., 2015; Suetens, 2017; Bauer et al., 2013). Therefore, MRI is used for diagnosis, evaluation, treatment planning, and follow-up. To that end, tumor segmentation is a crucial step. However, it is problematic. Manually done, it is very time consuming since it can take up to one hour per patient, and highly prone to intra- and inter-rater variability (Menze et al., 2015; Meier et al., 2016). Hence, there is a need for semi- or fully-automatic computerized methods.

Visualization of brain tumors through imaging techniques is crucial. But, the identification of the tumor grade plays a very significant role, as well. Namely, it is important for deciding the treatment strategy. The gold standard for tumor grading consists in biopsy followed by histological studies, when possible. Despite that, the procedure is time-demanding and invasive. Moreover, gliomas are very heterogeneous, so, the samples collected during biopsy may not be optimal for characterizing the tumor. Therefore, the procedure is also prone to sampling error (Zacharaki et al., 2009). Thus, predicting the tumor grade from standard MRI images may help physicians during the diagnosis process. A system based on such prediction would avoid sampling error because it analyzes the whole brain, while being not invasive. Moreover, in the case that biopsy is compulsory, it could expedite the treatment planning, as there is a waiting time from diagnosis, biopsy, histological studies, and the definitive report.

The brain tumor segmentation task was under-studied, despite its need. For this reason, the Brain Tumor Segmentation Challenge (BRATS) was created in 2012 to boost research on the topic. After the beginning of the challenge, a multitude of proposed methods approached brain tumor segmentation following different strategies (Bauer et al., 2013; Menze et al., 2015). However, after 2013/2014 it became apparent that Machine Learning-based methods were the most promising ones, with a special highlight for supervised-learning. Regarding glioma grading, although a few approaches can be found (Zacharaki et al., 2009; Khawaldeh et al., 2017), it is still a very under-studied field. Nevertheless, Machine Learning seems a possible approach for tackling this task, as well.

Machine Learning algorithms learn directly from the data. They can be broadly divided into supervised and unsupervised. In the former, during learning there is a target driving training to perform some task. In unsupervised algorithms, there is no target variable, so the algorithms learn a distribution of the data. In any case, the algorithms require some representation of the data as input. When these representations are

designed by feature engineering they are called handcrafted features. These representations are effective, but its design may be a long journey that requires expert domain knowledge (LeCun et al., 2015). In another direction, Representation Learning approaches learn the features directly from the data. Deep Learning is a group of Representation Learning algorithms that learn a hierarchy of concepts, in the form of multiple levels of representations (Goodfellow et al., 2016). Deep Learning approaches take the form of Deep Artificial Neural Networks, where several layers of learning blocks as stacked. In this case, where we have several layers, more complex and discriminative features can be learned directly from the data, with much less domain knowledge (Bengio et al., 2013; LeCun et al., 2015). Several Deep Learning architectures gathered a lot of attention in the recent years. Arguably, Convolutional Neural Networks (CNN) reached the spotlight after the outstanding work of Krizhevsky et al. (2012). Note, however, that Deep Learning algorithms are computational and data demanding algorithms.

In glioma segmentation, we can see a common element among all of the most successful handcrafted feature-based approaches – they all require dozens, or hundreds, of features, with some of them having some randomness during their computation (Zikic et al., 2012; Pinto et al., 2015a, 2018b; Tustison et al., 2015; Meier et al., 2014a). This is due to the high heterogeneity in the appearance, location, and shape of gliomas. In this scenario, Deep Learning methods may be able to learn discriminative and data-driven features, provided that enough data is available. Despite this, one still needs to understand how reliable Deep Learning is in segmentation. Also, a main issue is which architecture better learns the segmentation-needed patterns. Furthermore, MRI brain tumor segmentation data is highly imbalanced, with most voxels in a MRI image belonging to normal tissue. CNNs optimize features that are discriminative for a given task. However, such task may have several classes to distinguish, and features may have different importance for each class. For addressing this issue, one may try to suppress the irrelevant features for the class being predicted (Hu et al., 2018). Although this was shown for classification problems, it is not obvious in segmentation tasks.

The fact that CNNs can effectively process raw images make them a potential candidate for automatic glioma grading from structural MRI. Khawaldeh et al. (2017) employed CNNs, but its 2D single-sequence method does not really follows the consensus in brain tumor imaging, neither it reflects the definitions by the WHO (Louis et al., 2007, 2016; Ellingson et al., 2015). Moreover, low and high grade gliomas share some characteristics. Therefore, it is not clear if Machine Learning-based methods can learn how to discriminate the grades.

Despite the remarkable performances of Deep Learning, and Machine Learning in general, a major issue may raise some resistance in its acceptance and deployment in clinical practice – their lack of explainability. Usually, better performances come at the expense of using larger and more complex models. Of course, this leads to models that are harder to interpret. In critical domains, like the medical field, physicians may be reluctant in accepting support from a system whose predictions are not explainable (Wang and Summers, 2012). Hence, it is imperative to develop reliable systems, while, at the same time, investing efforts in making algorithms more transparent, or develop the tools to inspect their inner workings. Not only this is crucial for increasing trust, but also to understand the failures of the algorithms during development.

So, in recent years, Representation Learning have been bringing a boost and novelties to Machine

Learning-based approaches. As these methods are pervasively being adopted, it is needed to investigate their potential in Medical Image Analysis. If explored and understood, they may lead to breakthrough performances in this domain. In the case of this work, both brain tumor segmentation and grading are open and challenging problems. Therefore, we propose and evaluate methods for automatic brain tumor segmentation and grading. Most of our approaches are based on CNNs. Finally, the search for top performing methods should not suppress the important aspect of interpretability. So, we additionally investigate the interpretability of Machine Learning-based methods.

1.2 Objectives of the thesis

In this work, we aim at exploring Deep and Representation Learning methodologies, and conceive reliable solutions to the relevant tasks of glioma segmentation and classification from on multisequence MRI images.

The *first objective* is to investigate Representation Learning for semantic segmentation, and successfully apply them to brain tumor segmentation. Although Deep Learning, especially CNNs, show impressive potential in object recognition tasks, it is not so obvious for segmentation. Segmentation pose some different challenges: it needs to be detailed, and it deals with high class imbalance. Furthermore, is some locations, some feature detectors are more important for some classes than others.

The *second objective* is to explore, propose, and evaluate an automatic glioma grading method from conventional structural MRI, using Representation Learning techniques. This must be a 3D approach, since a glioma must be assessed globally.

Finally, the *third objective* is to investigate interpretability of opaque Machine Learning methods, such as Deep and Representation Learning-based systems. This objective is transversal to the previous two. We are interested in providing insights about the decisions in both segmentation and classification. This insight is crucial for increasing trust about the models. Moreover, it can help driving the decision-making and provide hints during the development of the systems.

Ultimately, we hope to make advances not only in computer science and machine learning areas of knowledge, but, also, somehow improve the quality of life of patients and contribute to the acceptance of Artificial Intelligence in clinical practice.

1.3 Contributions from this thesis

The contributions of this thesis work are summarized as follows:

1. We have investigated and proposed brain tumor segmentation approaches based on CNNs (Pereira et al., 2015, 2016a, 2017, 2018b). These contributions can be further enumerated as:
   (a) Inspired by object recognition CNNs, we first propose classification CNNs for brain tumor segmentation. We studied the use of small 3×3 kernels, and cross glioma grade data augmentation (Pereira et al., 2015, 2016a).

(b) Using the classification CNN, we proposed to use histogram standardization as a data pre-processing before CNNs. We demonstrated its benefits experimentally. This showed, contrarily to the belief that CNNs could handle the raw data, that careful pre-processing leads to better performance (Pereira et al., 2016a).

(c) We developed and proposed a hierarchical brain tumor segmentation approach based on Fully Convolutional Networks (FCN). We first segmented the whole tumor. Then, a second network segmented the inner tissues. This tackles the problem of class imbalance and misclassified isolated clusters (Pereira et al., 2017).

(d) We proposed adaptive feature recombination and recalibration for semantic segmentation using FCNs. Not all feature maps are equally important for all pixels. Hence, we proposed to recalibrate them accordingly (Pereira et al., 2018b).

2. We investigated interpretability of Machine Learning systems. The object of study was a system comprising a Restricted Boltzmann Machine for feature learning, and a Random Forest classifier. The problems under interpretation were related with segmentation (Pereira et al., 2018c). These contributions can be further enumerated as:

   (a) We explored a strategy for global interpretability by inspecting which parts of the data contributed the most for important features (Pereira et al., 2018c).

   (b) We interpreted image segmentations locally by assessing the spatial relevance of the features distributed in the image space (Pereira et al., 2018c).

   (c) We proposed a Mutual Information-driven feature selection scheme to deal with large feature vectors that impair interpretation (Pereira et al., 2018c).

3. We developed and propose 3D CNNs for automatic glioma grading. We studied two settings for glioma grading: 1) from the whole image, and 2) from an automatically defined region of interest (Pereira et al., 2018a).

   (a) We employed interpretability methods to inspect the predictions. In this way, we not only increase trust, but we also show how it enables us to detect and correct problems during development (Pereira et al., 2018a).

1.3.1 Publications

During the course of the doctoral studies, the author of this work authored the following manuscripts that are directly related with this thesis, and are part of this document.

Peer-reviewed Journal articles

- Pereira, S., Meier, R., McKinley, R., Wiest, R., Alves, V., Silva, C. A., & Reyes, M. *Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation.* Medical image analysis, 44, 228-244, 2018.

- Pereira, S., Pinto, A., Alves, V., & Silva, C. A. *Brain tumor segmentation using convolutional neural networks in MRI images.* IEEE transactions on medical imaging, 35(5), 1240-1251, 2016 (top 1%

cited paper in Clinical Medicine in early 2018 by Web of Knowledge).

Peer-reviewed Conference and Workshop articles

- Pereira, S., Alves, V., & Silva, C. A. *Adaptive feature recombination and recalibration for semantic segmentation: application to brain tumor segmentation in MRI*. Medical Image Computing and Computer Assisted Intervention (MICCAI), 2018.
- Pereira, S., Meier, R., Alves, V., Reyes, M., & Silva, C. A. *Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment*. Workshop on Interpretability of Machine Intelligence in Medical Image Computing, Lecture Notes in Computer Science, 2018.
- Pereira, S., Oliveira, A., Alves, V., & Silva, C. A. *On hierarchical brain tumor segmentation in MRI using fully convolutional neural networks: A preliminary study*. IEEE 5th Portuguese Meeting on Bioengineering (ENBENG), 2017.
- Pereira, S., Pinto, A., Alves, V., & Silva, C. A. *Deep convolutional neural networks for the segmentation of gliomas in multi-sequence MRI*. International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Lecture Notes in Computer Science, Springer, 2015.

The author of this thesis also authored, or co-authored, the following publications in parallel with his doctoral studies.

Peer-reviewed Journal articles

- Oliveira, A., Pereira, S., & Silva, C. A. *Retinal Vessel Segmentation based on Fully Convolutional Neural Networks*. Expert Systems with Applications, 2018.
- Pinto, A., Pereira, S., Rasteiro, D., & Silva, C. A. *Hierarchical Brain Tumour Segmentation using Extremely Randomized Trees*. Pattern Recognition, 2018.
- Oliveira, J., Pereira, S., Gonçalves, L., Ferreira, M., & Silva, C. A. *Multi-surface segmentation of OCT images with AMD using sparse high order potentials*. Biomedical optics express, 8(1), 281-297, 2017.
- Pereira, S., Pinto, A., Oliveira, J., Mendrik, A. M., Correia, J. H., & Silva, C. A. *Automatic brain tissue segmentation in MR images using random forests and conditional random fields*. Journal of neuroscience methods, 270, 111-123, 2016.

Peer-reviewed Conference articles

- Pinto, A., Pereira, S., Meier, R., Alves, V., Wiest, R., Silva, C. A, & Reyes, M. *Enhancing clinical MRI Perfusion maps with data-driven maps of complementary nature for lesion outcome prediction*. Medical Image Computing and Computer Assisted Intervention (MICCAI), 2018.

- Oliveira, A., Pereira, S., & Silva, C. A. *Augmenting data when training a CNN for retinal vessel segmentation: How to warp?*. IEEE 5th Portuguese Meeting on Bioengineering (ENBENG), 2017.
- Mestre, J., Pereira, S., Silva, C. A., & Rasteiro, D. M. L. D. *Modelling brain tissues intensities using dirichlet process.* IEEE 5th Portuguese Meeting on Bioengineering (ENBENG), 2017.
- Cavadas, B., Branco, P., & Pereira, S. *Crime Prediction Using Regression and Resources Optimization*. Portuguese Conference on Artificial Intelligence, Lecture Notes in Computer Science. Springer, 2015.
- Pinto, A., Pereira, S., Correia, H., Oliveira, J., Rasteiro, D. M., & Silva, C. A. *Brain tumour segmentation based on extremely randomized forest with high-level features*. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015.
- Oliveira, J., Pereira, S., Gonçalves, L., Ferreira, M., & Silva, C. A. *Sparse high order potentials for extending multi-surface segmentation of OCT images with drusen*. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015.
- Pinto, A., Pereira, S., Dinis, H., Rasteiro, D. M., & Silva, C. A. *Random decision forests for automatic brain tumor segmentation on multi-modal MRI images*. IEEE 4th Portuguese Meeting on Bioengineering (ENBENG), 2015.
- Pereira, S., Mariz, J. A., Sousa, N., Correia, J. H., & Silva, C. A. *A Fully Automatic Tool for Counting Virchow-Robin Spaces in Magnetic Resonance Imaging for Lacunar Stroke Study*. BIOIMAGING, 2015.

1.3.2 Participation in international challenges/competitions

Brain Tumor Segmentation Challenge (BraTS), Munich, 2015

Our method was **ranked 2nd among the 13 participating teams** in the overall ranking.

This was an international challenge integrated in MICCAI conference, where the objective was to segment a set of MRI images of patients with brain tumor. The participants were provided with 274 MRI acquisitions with manual segmentation for training their algorithms. Afterwards, a Challenge set without manual segmentation was provided for the participants to segment, and the evaluation was performed by the organizers.

Pereira et al. (2015) and Pereira et al. (2016a) were based on the system we used for this challenge. The methods and results can be found in Section 4.1.

1.3.3 Organization of Events

Workshop on Interpretability of Machine Intelligence in Medical Image Computing (iMIMIC), Granada, 2018

This workshop was held on September 16th as a satellite event of the Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2018, in Granada, Spain. The author of this thesis was a co-organizer of the workshop.

The aim of the workshop was to introduce the challenges and opportunities related to the topic of interpretability of Machine Learning systems in the context of medical image computing and computer-assisted intervention.

1.4 Structure of the Thesis and General Overview

This work is organized in seven chapters. We first present a couple of chapters regarding basic medical and technical background related with the work in this document. Then, each chapter delves into one of the objectives defined in Section 1.2. The remainder of this section closes this introductory chapter, and provides a description of the remaining document and general overview of the work.

In Chapter 2 we start by providing a medical perspective on gliomas, and their appearance on MRI images. Then, we delve into the problems of automatic segmentation and grading. It is described why it is needed to automate these tasks, and how it can positively impact the clinical practice. We also expose the main challenges regarding each task, and the main lines of research.

We proceed by providing a technical overview in Chapter 3. In this chapter, we present the Machine Learning-related essentials of the methods developed in this work. The chapter closes with why interpretability is needed.

In Chapter 4 we delve into the proposed methods for brain tumor segmentation using CNNs. We start with our work using classification CNNs. This represents a thorough exploration of the potential of this algorithm for segmentation. We investigate its capabilities, use of small kernels, problem-dependent data augmentation, and careful pre-processing. Then, in Section 4.2, we follow the most recent advancements and develop a segmentation approach based on FCNs. We study hierarchical segmentation as a means to reduce the class imbalance problem and false positive detections. The hierarchical approach serves as basis for the last Section of the chapter, where we investigate the relationships among feature maps for semantic segmentation in FCNs. We propose to adaptively recombine and recalibrate feature maps.

In Chapter 5 we investigate the interpretability of a Machine Learning-based system in the context of brain lesion segmentation. In this chapter, we also explore unsupervised feature learning with a Restricted Boltzmann Machine. We identify and build upon two interpretability principles: global and local interpretability. In the former, we propose interpretability for reasoning about the system as a whole, and how it learned. Local interpretability is related with explaining the predictions on a sample basis. We also propose a joint Mutual Information-based feature selection approach.

The insights obtained from CNNs for segmentation and from interpretability methodologies converge

in Chapter 6, where we explore and propose 3D CNNs for automatic glioma grading. We investigate and propose two grading settings: from the whole brain, and from a previously identified region of interest. The region of interest is defined through segmentation, similarly with Sections 4.2 and 4.3. We further employ interpretability methodologies to inspect and explain the predictions. This is especially important in a problem where data is scarce. In fact, we show how interpretability helps in identifying patterns that were erroneously learned and how it helped in devising strategies for correcting it.

Finally, in Chapter 7, we close this document by drawing the main conclusions and perspectives on opened lines of research.

Chapter 2

Brain tumors — a medical perspective

In this chapter, the reader will be presented with a medical perspective on brain tumors, with focus on gliomas. These neoplasms are the most common and aggressive among primary brain tumors. Hence, gliomas gather a big research interest. We will also discuss the sub-structure of gliomas and relate it with their appearance in MRI. Image analysis of gliomas is the main application of the methodologies developed in this thesis, specifically glioma segmentation and grade prediction from MRI images. So, the main challenges and opportunities regarding these problems will be presented.

Contents

2.1 A general overview

Brain tumors can be benign or malign; in the latter, the tumors may be called as cancer. If their origin is in the brain itself, the tumors are designated as primary. Otherwise, if they result from tumor cells moving from other organs, they are named metastatic brain tumors (Jones and Hreib, 2012). Furthermore, according to the WHO (Louis et al., 2007, 2016), it is possible to grade brain tumors into four grades, depending on their aggressiveness. Tumor malignancy increases with the grade. Moreover, the four grades can be subdivided into two larger classes: low and high grade. Grades I and II are considered low grade, being less aggressive and proliferative than the highest grades. However, they can progress into grade III and IV, which are malignant tumors (high grade). Tumors in this grade are more proliferative

and grow in a faster pace than the low grade ones (Jones and Hreib, 2012). Life expectancy is higher for patients with low grade tumors. In fact, it can be more than 5 years for patients with brain tumors of grades I and II. Nevertheless, life expectancy drops to 2-3 years when patients face a grade III tumor. Tumors graded as IV are commonly glioblastomas. These neoplasms are especially deadly, with most patients not surviving more than 14 months after diagnosis, on average, even if the patient is under treatment (Louis et al., 2007, 2016; Van Meir et al., 2010). In general, the most common treatments include surgery, radiotherapy, chemotherapy, or a combination of them (Tabatabai et al., 2010).

Brain tumors can have different origins. Meningiomas (WHO grade I – III) originate from the meningeal layers, and account to up to 25% of the primary brain tumors. Although their grade can be up to III, the majority (90%) of these tumors are benign. Schwannomas are composed by Schwann cells and represent 5 – 10% of the primary brain tumors. These neoplasms are mostly benign, as well (Drevelegas, 2005; Larjavaara et al., 2008). Gliomas (WHO grade II – IV) arise from the glial cells of the brain, being the most common primary brain tumors. Moreover, these tumors account for around 80% of the malignant brain tumors. Given these facts, more attention is being given to gliomas.

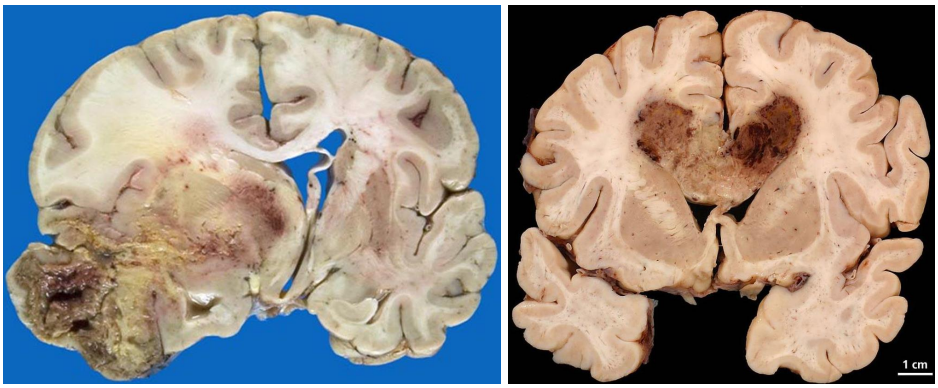
2.2 Gliomas — clinical foundations and treatment

Gliomas are brain tumors originated from the glial cells, such as astrocytes, or oligodendrocytes. The glial cells are around five times more abundant in the brain than the neurons, and they essentially play a supporting role for the latter. Hence, their functions include physical support, electric insulation among neurons, transportation of oxygen and nutrients, and defense against pathogenic entities. Unlike the neurons, the glial cells have some capacity of replication through mitosis, which explains why they can originate gliomas (Fox, 2006).

It was observed that cells from gliomas may have similarities with the glial cells from where they are originated, which lead the tumors to be designated as astrocytomas, oligodendrogliomas, or oligoastrocytomas (mix of abnormal astrocytes and oligodendrocytes). The factors associated with gliomas are the age, the gender (men have two times more chances than women), and radiation exposure. Still, the development of brain cancer is a rather random event, with small connections with family occurrences (Jones and Hreib, 2012). Regarding spatial distribution, the frontal lobe of the brain has the highest frequency of gliomas (40%), followed by the temporal (29 %) and the parietal (14%) lobes, with higher incidence in the right hemisphere than the left (Larjavaara et al., 2007).

As the other brain tumors, gliomas can be graded according to their malignancy into grade II – IV, following the WHO classification (Louis et al., 2016). Gliomas can be further broadly graded as low grade gliomas (LGG) if they are WHO grade II, and high grade gliomas (HGG) for WHO grade III and IV. In fact, the determination of the brain tumor grade is imperative for treatment planning and for prognosis. LGGs are mostly constituted by highly dense, but almost normal, cells. These gliomas are mainly astrocytomas and oligodendrogliomas with a slow growth rate, being mostly benign. In fact, patients can survive for several years with these tumors. However, these patients need close monitoring, since LGGs eventually evolve into higher grades and become fatal. Anaplastic gliomas (WHO grade III) are in-between LGGs

and the highest grade gliomas. While these tumors proliferate at a faster rate than LGGs, they still do not present necrotic tissues and the aggressiveness of the grade IV gliomas. Finally, glioblastomas (WHO grade IV) are the most common, and aggressive gliomas. These tumors are always malignant, almost certainly leading to death, as the median survival time of these patients is only 12-18 months (Jones and Hreib, 2012). These cancers may affect one hemisphere of the brain (Fig. 2.1(a)), or progress through the corpus callosum into the contralateral hemisphere. The latter are denominated as butterfly glioblastomas (Fig. 2.1(b)), being the most aggressive of all (Dziurzynski et al., 2012). Glioblastomas can be primary, or secondary. The former appears for the first time in patients without signs of a previous tumor, usually older patients, while the latter evolves from a low grade astrocytoma in younger patients (Louis et al., 2016; Ohgaki and Kleihues, 2013). Regularly, these neoplasms include regions of necrosis, hemorrhages, and fleshy tumor. Additionally, they are accompanied by perilesional edema. Due to their heterogeneous appearance, glioblastomas are also known as glioblastomas multiformes. Usually, there is an associated mass effect that compresses the surrounding tissues; in Fig. 2.1(a) it is observed a severe mass effect, with strong deformation of the ventricles (Jones and Hreib, 2012).



(a) (b)

Figure 2.1: Examples of glioblastomas affecting a) one hemisphere of the brain, or b) both (butterfly glioblastoma). Cases courtesy of Prof. Frank Gaillard, Radiopaedia.org, rID: 27812 and rID: 27877.

Treatment approaches in gliomas depend on the tumor grade, and location. Location is important, since the removal of the tumor mass may affect cognitive functions, such as language and memory, leading to worse quality of life. In glioblastomas, the current treatment standard is complete tumor surgery resection[1]. This is followed by radiation and chemo-therapy. In LGGs, surgery may be used when it will not impact the cognitive performance of the patient. Sometimes, it may be beneficial to do a partial resection to release the pressure caused by mass effect. Radiation therapy may be prescribed as well, especially if the patient is older, the tumor is big, it is a butterfly glioma, or if it is an astrocytoma (oligodendrogliomas are less aggressive) (Jones and Hreib, 2012). Also, due to the indolent and slower progression of LGGs, a "watch and wait" approach may be preferred to avoid impacting the quality of life of the patient, especially in younger people (Grier and Batchelor, 2006).

[1] If not possible, ressection may be done until the post-surgery cognitive functions impact is acceptable

2.3 Imaging in gliomas

The gold standard for glioma diagnosis and grade determination is biopsy and histology (Bauer et al., 2013). Notwithstanding, brain tumor assessment from imaging data is crucial for rapid diagnosis, follow-up, and intervention planning.

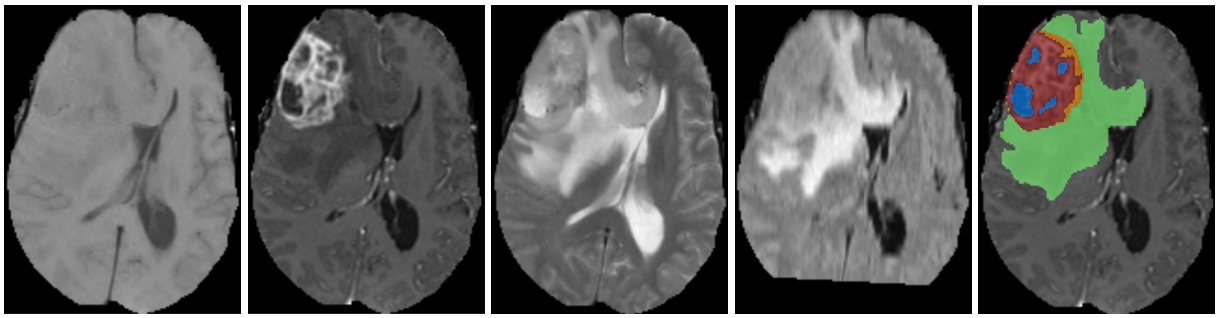
Computed Tomography may be the first imaging technique employed in a patient suffering from brain tumor. This is due to its readily availability, and fast scanning time. It may show the tumor's mass effect and calcification, but it lacks contrast for soft tissues, which is the case of the brain. In fact, in the case of non-enhancing tumors, such as many LGGs, physicians may fail to detect the lesion in Computed Tomography images (Mabray et al., 2015; Suetens, 2017; DeAngelis, 2001). Positron Emission Tomography plays an important role in assessing the evolution of the neoplasm and response to treatment, because advanced tumors have a higher intake of the radioactive tracer than the normal tissues. However, this technique has a low spatial resolution (around 4 mm) and does not provide detailed structural information (Mabray et al., 2015; Suetens, 2017). Therefore, Positron Emission Tomography is useful to assess the physiology of the tumor, but not its structure and sub-structure. So, it is not suited for measurements and volumetric analysis. Magnetic Resonance Imaging has good spatial resolution and very good contrast in soft tissues. Hence, it allows us to observe the tumor compartments, such as necrosis, but also the extent of edema (Mabray et al., 2015; Suetens, 2017). In fact, MRI have become the imaging standard for the assessment of brain tumors (DeAngelis, 2001; Bauer et al., 2013; Wen et al., 2010). Therefore, the work in this thesis is focused on gliomas image analysis in the structural MRI sequences recommended by the consensus for standard brain tumor imaging (Ellingson et al., 2015). More specifically, we tackle the problems of automatic glioma segmentation and grading. Perfusion MRI acquisitions are considered a plus, since they may provide information on tumor physiology, such as clues of the grade. However, this kind of acquisition is not part of the consensus, neither it is readily acquired in clinical practice (Ellingson et al., 2015; Essig et al., 2013). Hence, it is also out of the scope of this work.

Structural MRI allows us to examine the anatomy and pathology of the brain. So, in a prior evaluation of the tumor, it allows one to determine its location for surgery or biopsy planning, and evaluate the mass effect on the normal brain tissues (Mabray et al., 2015). This is possible because MRI images are three-dimensional. Moreover, with structural MRI, it is possible to acquire a variety of sequences that result in images with different tissue contrasts. In fact, this is not only desirable, but mandatory for glioma image analysis, since different MRI sequences are conspicuous for different tissues and tumor structures.

According to Ellingson et al. (2015), the accepted consensus for brain tumor imaging include the following MRI sequences: T1-weighted (T1), 3D isotropic 1 mm post-contrast T1-weighted (T1c), T2-weighted (T2), and T2-weighted Fluid-Attenuated Inversion Recovery (FLAIR). T1 has very good contrast for normal brain tissues, such as gray and white matter. Gliomas may appear as abnormal regions. However, after the injection of the gadolinium contrast, it is possible to observe the enhancing regions of the tumor. These regions are a feature of HGGs, although it may appear in some LGGs. The enhancing regions result from a disruption of the blood-brain barrier that leads to accumulation of blood. This region corresponds to active tumor regions, often with a ring-like shape. The pathological death of cells and tissues is designated as necrosis (Fox, 2006). In glioblastomas it is often observed the necrotic tissues as a hypointense region

inside the enhancing tumor in the T1c sequence. The T1 and T1c MRI sequences must be acquired with the same parameters to allow their subtraction (T1c - T1), which facilitates the observation of the enhancing regions. In the case of non-enhancing and very diffuse gliomas, usually LGGs, the T1 and T1c sequences may be inefficient for assessing the tumor. Additionally, those sequences are not optimal for observing the edema regions. However, the whole tumor region, including edema, is hyperintense in both T2 and FLAIR. Hence, these sequences are important for assessing the non-enhancing regions of the tumor, which includes non-enhancing active tumor and edema. T2 is the preferred sequence for these regions. But, in FLAIR the image intensities of the cavities filled with cerebrospinal fluid are suppressed, which makes it useful for examining tumors located near the ventricles and the sulci. Structural MRI allows one to observe the proliferation of the glioma in the surrounding tissues to a certain extent in the form of edema. However, it may not provide the real tumor extension, as gliomas may be infiltrating a larger region at a cellular level (DeAngelis, 2001; Wen et al., 2010; Ellingson et al., 2015; Bauer et al., 2013; Menze et al., 2015).

In Fig. 2.2 it is depicted an example of a brain with a glioblastoma. In the T1 sequence (Fig. 2.2(a)) it is possible to observe the clear mass effect, especially through the deformation of the ventricles. The enhancing tumor is conspicuous in the T1c sequence (Fig. 2.2(b)), with the necrotic regions inside it. The whole tumor is hyperintense in both T2 (Fig. 2.2(c)) and FLAIR (Fig. 2.2(d)) sequences, while in the latter the cerebrospinal fluid response is suppressed, which is crucial to contrast the tumor from the ventricles. It is also possible to observe a narrow region of non-enhancing active tumor close to the enhancing tumor in the manual segmentation (Fig. 2.2(e)). In the case of Fig. 2.2, the non-enhancing active tumor may be perceived as a lower intensity region in-between the enhancing tumor and the edema in the T2 sequence. This region is difficult to identify, as it does not enhance in the T1c sequence, and shows the need for multi-sequence acquisitions. Indeed, during manual segmentation, the non-enhancing active region is identified as the remaining tumor part, after defining the edema, necrosis, and enhancing tumor (Menze et al., 2015).



(a) (b) (c) (d) (e)

Figure 2.2: MRI acquisition of a patient with a glioblastoma. a) T1 sequence. b) T1c. c) T2. d) FLAIR. e) Manual segmentation; colors identify different tumor regions: blue – necrosis, red – enhancing tumor, orange – non-enhancing tumor, and green – edema.

2.4 Glioma image analysis – challenges and opportunities

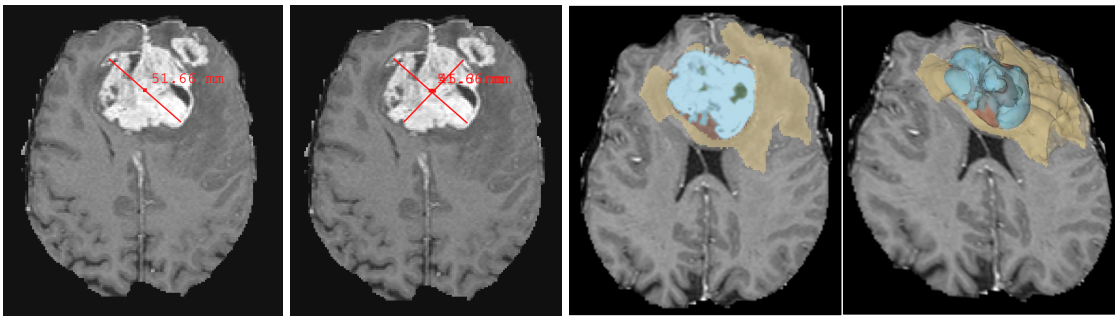
Brain tumor imaging offers the possibility to non-invasively assess the brain. Moreover, images are a rich source of data that allows physicians to extract measurements, locate the tumors and their extension, thus enabling a better treatment and surgery planning. For these purposes, segmentation is a required step. Additionally, images may contain information about the pathophysiology of the tumor. The study of such features is in the domain of radiomics.

In this subsection it is presented some open challenges and opportunities for developing Machine Learning-based approaches to support physicians' work, and decision making. The focus will be on brain tumor segmentation and radiomics.

2.4.1 Image segmentation

Brain tumor measurements, such as the tumor size, are important for treatment planing, as well as monitoring of tumor evolution and response to treatment. Historically, rough estimates of the tumor size have been used. The Response Evaluation Criteria in Solid Tumors (RECIST) stipulates the use of unidimensional tumor measures for estimating the overall tumor burden. To that end, experts must measure the largest tumor axis in the axial plane (Therasse et al., 2000; Eisenhauer et al., 2009), as observed in Fig. 2.3(a). However, in the case of HGGs, the bidimensional Macdonald Criteria (Macdonald et al., 1990) remains as one of the most widely used measurement procedures. In this criteria, physicians should measure the two largest orthogonal axis of the enhancing tumor in the axial plane (Fig. 2.3(b)). One big drawback of these criteria is that it only considers the enhancing tumor. According to it, complete response to therapy is obtained when no enhancing tumor is measurable, whereas progression happens if there is an increase of at least 25% of the enhancing tumor measures. This immediately demonstrates two more drawbacks: 1) it is not suited for non-enhancing tumors, such as many LGGs, and 2) the enhancing tumor amount may be affected by non-tumor related causes, such as treatment (Wen et al., 2010). To account on these issues, the Response Assessment in Neuro-Oncology Criteria (RANO) establishes that measurements should still be performed on T1c sequence, but tumor progression or remission must take into account the non-enhancing regions observed in the T2 and FLAIR sequences. Uni- and bidimensional measurements have the advantage of being fast to calculate. Notwithstanding, the intra- and inter-rater variabilities are large because of difficulties in finding the best slice, largest axis, or the irregular shape of the tumor. Moreover, these measurements assume symmetric tumor progression and remission, and are sensitive to the positioning of the patient's head across acquisitions (Wen et al., 2010). Both RECIST and RANO criteria establish a minimum tumor size for being measurable, which may make them unsuitable for post resection surgery (Wen et al., 2010; Therasse et al., 2000; Eisenhauer et al., 2009).

Volumetric measurements may be more robust to some of the limitations of the uni- and bidimensional measures (Wen et al., 2010; Ellingson et al., 2014). For instance, it does not require the identification of the largest axis, it measures asymmetrical tumor responses, it is more robust to different head positioning and inter-slice thickness, and it provides more realistic measures of the tumor size. Studies found that volumetric measurements are a more accurate indicator of tumor burden and post-surgery residual tumor



(a) (b) (c)

Figure 2.3: Example of a post-contrast T1 sequence of a glioblastoma with a) unidimensional, and b) bidimensional measurements. In c) it is shown the same patient with a tumor manual segmentation (left) and its volumetric rendering; colors mean: yellow – edema, green – necrosis, red – non-enhancing tumor, and blue – enhancing tumor.

than uni- and bidimensional measurements (Ellingson et al., 2014; Dempsey et al., 2005; Chow et al., 2014). So, although uni- and bidimensional measurements are widely used, it is expected that volumetric measurements become more and more interesting for clinicians. Additionally, volumetric measures open the opportunity to assess the several sub-compartments of the gliomas. The division of HGGs into meaningful sub-regions was found relevant for treatment response and survival estimation (Meier et al., 2016; Iliadis et al., 2012; Pope et al., 2005). Such sub-regions are: necrosis, edema, non-enhancing tumor, and enhancing tumor (Meier et al., 2016; Iliadis et al., 2012; Pope et al., 2005). In Fig. 2.3(c) it is possible to observe the volume of the sub-compartments of the tumor. It is obvious the gross approximation that a uni- or bidimensional measure represents for such a heterogeneous HGG.

A compulsory step for volumetric delineation of the tumors is segmentation. Image segmentation consists in detecting and partitioning an image into its regions or objects. If at the time of segmentation each region is classified, i.e., if each segment has a semantic meaning, it is designated as semantic segmentation.

Manual tumor segmentation requires an experienced radiologist for manually delineating the tumor regions. However, it is time demanding, as it may take until one hour, and it is highly prone to intra- and inter-rater variability (Porz et al., 2016; Menze et al., 2015). Therefore, semi-automatic procedures are being used in clinical practice. Although these methods significantly accelerate the annotation procedure, they do not solve all of the problems of manual segmentation, since they still require a human user, and, because of that, they are still prone to intra- and inter-rater variability (Porz et al., 2016; Bauer et al., 2013). Additionally, semi-automatic methods cannot be integrated in fully automatic pipelines that run on servers before assessment by physicians. Thus, fully automatic, but reliable, segmentation methods are needed.

2.4.1.1 Challenges of automatic segmentation

Although fully automatic segmentation approaches are desirable, it is a complex task. Gliomas are highly heterogeneous in appearance, size, and shape. Moreover, these neoplasms can appear in diverse

locations of the brain, and cause none to severe mass effect. Different MRI sequences are needed to identify the several sub-regions of the tumor, which may make brain tumor segmentation computationally demanding (Menze et al., 2015).

Besides the tumor characteristics, MRI poses some challenges for segmentation, as well. The bias field is an artifact that appears due to imperfections in image acquisitions. It translates as intensity inhomogeneity, such that a smooth intensity variation occurs in the image. For this reason, the same tissue may have different image intensity in different locations (Vovk et al., 2007). Also, the image intensities in MRI do not have a fixed meaning, as they are not calibrated. In fact, image intensities may vary even if two images are acquired from the same patient using the same scanner, but at different moments. This problem may be more severe across several acquisition sites, and equipments (Nyúl et al., 2000). Finally, acquisition protocols and parameters may vary across clinical centers, resulting in different image contrasts (Ellingson et al., 2015).

2.4.1.2 Main trends in automatic brain tumor segmentation

In the literature, one can find two main approaches for automatic brain tumor segmentation: unsupervised generative probabilistic methods, and supervised discriminative methods (Menze et al., 2015).

In unsupervised generative probabilistic approaches, strong priors about the tissues spatial distribution and appearance are enforced. These models should describe the underlying data, such that it represents the joint distribution $p\left(x, y\right)$ of both observed (x) and target variables (y). Still, it can be transformed into a conditional distribution $p\left(x|y\right)$ to classify the data. Generative methods, in principle, can generalize better to unseen data because the model is fit to the unseen data itself (e.g. using an Expectation-Maximization algorithm), offering some robustness to acquisition parameter differences when faced with small datasets (Menze et al., 2015; Murphy, 2012). The models often assume image intensities to be generated by Gaussian Mixture Models (Menze et al., 2010; Gooya et al., 2012; Kwon et al., 2014b). Furthermore, generative approaches might also explicitly model other kind of prior knowledge besides image intensity distributions. Such prior knowledge may be in the form of smoothness constraints on the neighborhood of the pixel through Markov Random Fields (Menze et al., 2010; Kwon et al., 2014b). Some models also take into account the spatial location of normal and tumor tissues with probabilistic atlases (Menze et al., 2010; Gooya et al., 2012; Kwon et al., 2014b), or the tumor mass effect with tumor growth models (Gooya et al., 2012; Kwon et al., 2014b). However, it is difficult to translate prior knowledge into a formal and tractable model. A way of modeling the tissues spatial distribution is through probabilistic atlases. Still, atlases require registration and adaptation to account for the tumor abnormalities. Growth models can represent the tumor mass effect, but their simulation may be complex. Hence, probabilistic generative models may be approximations that do not represent the real phenomenon of the data and require complex and computationally demanding methods.

Supervised discriminative approaches learn to predict the target variable (y) directly from the observable variable (x) as a conditional distribution $p\left(x|y\right)$. Hence, these approaches do not attempt to learn the phenomenon generating the underlying data. Usually, when the prior assumptions in generative methods are incorrect, discriminative approaches achieve superior results because these methods do not need

to model the distribution of the data (Murphy, 2012).

Discriminative methods are inherently supervised, as they need annotated training data, i.e., for each training instance, we need the class it belongs to. Hence, discriminative approaches may not generalize well if the training set is small, biased, or very different from the testing conditions (Murphy, 2012). However, supervised discriminative approaches have achieved the state-of-the-art results in glioma segmentation. In fact, in all editions (2012 – 2017) of the Brain Tumor Segmentation Challenge (BRATS) (Menze et al., 2015) the top performing approaches were based on supervised discriminative methods. Hence, in this thesis work, supervised discriminative approaches were mostly followed. Until 2014, handcrafted features followed by a discriminative classifier was the main approach regarding supervised discriminative methods for brain tumor segmentation. Given the complexity of the problem, a large spectrum of features was studied, such as context (Meier et al., 2014b; Zikic et al., 2012; Pinto et al., 2015a), first-order and fractals-based texture (Meier et al., 2013, 2014b; Reza and Iftekharuddin, 2014; Pinto et al., 2015a; Islam et al., 2013), gradients (Meier et al., 2013, 2014b), brain symmetry (Meier et al., 2013, 2014b; Tustison et al., 2015), and physical properties (Tustison et al., 2015). All of these methods used the Random Forest (RF) classifier (Breiman, 2001). Instead of handcrafted features, Representation Learning approaches were pursued in this thesis work. These methods learn the optimal set of features directly from the data (LeCun et al., 2015). In fact, CNNs have achieved state-of-the-art results in brain tumor segmentation. Apart from the works described in this thesis (Pereira et al., 2016a, 2017, 2018b), other authors have concurrently or posteriorly followed CNN approaches (Lyksborg et al., 2015; Havaei et al., 2017; Kamnitsas et al., 2017b; Zhao et al., 2018). Lyksborg et al. (2015) used an ensemble of 2D CNNs for identifying the whole tumor, followed by intra-tumor segmentation. Havaei et al. (2017) changed fully-connected layers by convolutional layers. Furthermore, a cascade of CNNs was employed, such that the second network could learn contextual information from the output of the first. Additionally, a complex two-stage training procedure was proposed. Kamnitsas et al. (2017b) proposed a fully convolutional neural network without max-pooling layers, where a sub-volume of voxels was segmented at once in one forward pass. This approach is much more efficient than the previous ones, where each voxel requires a forward pass of the patches through the network. Finally, Zhao et al. (2018) proposed a fully convolutional neural network followed by a Conditional Random Field formulated as a Recurrent Neural Network. Conditional Random Fields are able to regularize the segmentation by taking into account both the unary predictions, and the appearance of a neighborhood. By formulating the Conditional Random Field as a Recurrent Neural Network, it was possible to couple the CNN and the Conditional Random Field as a single network.

2.4.2 Computer-aided Diagnosis and Radiomics

A crucial step during glioma assessment relies upon tumor grading (Zacharaki et al., 2009). The identification of the grade is important for treatment planning. For instance, in the case of a LGG, it may be more suited to not perform any intervention besides monitoring. For grading, the gold standard is biopsy followed by a histology study. However, the former is a time consuming and invasive procedure, and sometimes impossible. Moreover, biopsy is prone to sampling error, because it retrieves just a few samples of the tumor in different time points, which may not represent the whole tumor. This is due to

the high heterogeneity in tumor's location and evolution over time. In fact, only a low percentage of biopsy samples contains adequate tumor content and genetic material (Hu et al., 2015).

Computer-aided Diagnosis (CADx) systems help physicians in decision making. Besides image segmentation, image-based CADx systems may provide predictions about the clinical status and survival estimation of the patient (Yang et al., 2015), or the characterization of the disease (Zacharaki et al., 2009). These automatic systems can take the big amount of imaging data as input, and provide the physician with some information for decision support. Hence, an image-based CADx system for tumor grade prediction from volumetric MRI data could be useful in clinical practice. Such a system may assess the macroscopic structure of the whole tumor and surroundings in a noninvasive way. Thus, it has the potential of bypassing the sampling problem of biopsy, and expedite treatment planning, since it is much faster than biopsy and histology studies (Zacharaki et al., 2009). It would also be useful during treatment monitoring to assess the tumor progression.

Digital imaging encodes information regarding more complex phenomenons than just image intensities. Indeed, thorough image analysis may provide better prognostic capabilities, or predict the effectiveness of treatments, than simple uni- and bidimensional measures (Kumar et al., 2012). Radiomics deal with identifying features from medical images related to the pathophysiology of the medical condition of the patients. Usually, a large pool of features is extracted (200+). In this way, faster and higher level decision support systems may be developed. Radiomics features may be the image intensity, texture, tumor shape and size, or image transforms, such wavelet transforms. These features may be related with some variable of interest, such as survival, cancer phenotype, or treatment response. Usually, radiomics workflows include tumor segmentation, feature extraction, feature selection, and prediction model construction (Lambin et al., 2012, 2017; Bakas et al., 2017; Kotrotsou et al., 2016). There are evidences that radiomics can, indeed, assess the tumor's subtypes, and predict survival time (Yang et al., 2015; Bakas et al., 2017; Kotrotsou et al., 2016).

So, CADx and radiomics may provide tools for monitoring the development and progression of the disease, predict survival, response to treatment, as well as characterize the tumors themselves (Lambin et al., 2012; Aerts et al., 2014; Lambin et al., 2017; Kotrotsou et al., 2016; Yang et al., 2015). Although radiomics alone may not substitute biopsy and histological studies as gold standard, it reflects the pathophysiology of glioblastomas (Kotrotsou et al., 2016).

2.4.2.1 Challenges of automatic tumor grading from MRI

Automatic glioma grading from structural MRI has practical utility in clinical practice. However, it is a complex task. Regarding gliomas, these heterogeneous neoplasms have some characteristics more associated with HGGs or LGGs. For instance, LGGs exhibit less mass effect and are more diffuse than HGGs. Also, the enhancing rim is a characteristic more often associated with HGGs, as well as being larger, more heterogeneous, and presenting necrotic regions. Nevertheless, those characteristics are interchangeable and may occur in both grades. This high variability makes it hard to develop a CADx system. However, we note that the presence of enhancing rim in LGGs is a sign that it needs close monitoring, and, probably, intervention (Grier and Batchelor, 2006; Steed et al., 2018; Van Meir et al.,

2010).

Some studies suggest that perfusion MRI-derived maps, such as the relative cerebral blood volume map, offer better potential to distinguish HGGs and LGGs than conventional structural MRI sequences, e.g. T1, T1c, T2, and FLAIR sequences (Zacharaki et al., 2009; Law et al., 2003). But, perfusion MRI sequences are consensually considered a plus regarding MRI acquisitions for brain tumor assessment, whereas the referred conventional structural MRI sequences are crucial (Ellingson et al., 2015). In fact, perfusion MRI is being studied in academic contexts, but it is not widely used in clinical practice (Essig et al., 2013). Thus, although complex, tumor grading from conventional structural MRI has more potential for being used.

Radiomics studies suggest that it is possible to extract pathophysiology-related information from imaging data. First, the image must be segmented in order to define regions of interest (ROIs), such as tumor, or normal tissue. This segmentation step is usually done manually, or in a (semi-)automatic way (Parmar et al., 2014; Velazquez et al., 2015). Manual segmentation makes it unpractical to employ radiomics in large databases. Semi-automatic approaches are much faster, but it does not bypass the inter- and intra-observer variability problem, since a user needs to provide some input. Automatic approaches solve both problems. Still, radiomics needs segmentation to be robust in order to consistently identify well the sub-compartments of the tumors in all subjects in order to maintain the conclusions of radiomics studies across subjects (Kumar et al., 2012). Hence, it would be desirable to bypass the segmentation stage, or, at least, decrease the need on highly accurate sub-compartment segmentation. After segmentation, radiomics procedures extract a large number of features, which are then subjected to feature selection to reduce the risk of overfitting (Kumar et al., 2012). This suggests that general features, such as tumor shape, texture, or histogram, are extracted, without any certainty about its correlation with the target variable. These challenges (segmentation, and features extraction and selection) pose an opportunity for Representation Learning approaches, where features are extracted directly from the data for the desired task.

2.4.2.2 Main trends in automatic brain tumor grading

Glioma grading from structural MRI is a relatively unexplored field. First approaches studied glioma grading using handcrafted features followed by a Machine Learning classifier, such as Support Vector Machines, *k*-Nearest Neighbors, Decision Trees, or Naïve Bayes (Zacharaki et al., 2009, 2011). Considered features included image intensities, volumetric data, and Gabor textures. These features were extracted from manually defined ROIs, which represents a disadvantage of the work. The authors used the relative cerebral blood volume perfusion map together with the conventional structural MRI images. However, for tumor grading, from the top eight most important features only the first was computed from that perfusion map, being the others from structural MRI (Zacharaki et al., 2009). This suggests that perfusion MRI may, indeed, help during tumor grade classification, but structural MRI encodes relevant information, as well. More recently, Khawaldeh et al. (2017) used a CNN for automatic tumor grading without the need of defining ROIs. However, the method comes with a major drawback. It processes individual 2D slices, instead of the whole 3D MRI study. So, radiologists need to select the relevant slices, which makes it prone

to some inter-rater variability. Moreover, it processes several slices per study, which makes it possible to classify slices of the same tumor with different grades.

2.5 Summary

Gliomas are the most common primary brain tumors. Still, patients with this cancer yield the worst prognosis among brain tumors. Gliomas may be categorized as LGGs and HGGs depending on their degree of malignancy. Magnetic Resonance Imaging stands as the gold standard imaging technique for the assessment of these tumors. However, their structure and appearance is highly heterogeneous, making glioma image analysis a challenging task. Among the open problems we can identify automatic and reliable glioma segmentation and grading. Due to the complexity of the task, Machine Learning-based approaches appear as the most promising techniques for both tasks, as long as enough data is available. Both automatic glioma segmentation and glioma grading are of clinical relevance and may provide physicians with a faster and more accurate assessment, as well as better treatment planning, and monitoring tools.

Chapter 3

Machine Learning

In this chapter, the reader will be introduced to the technical background related with the methods developed in the scope of this thesis work. Machine Learning algorithms serve as the basis for all the conceived methods. Therefore, a broad overview on Machine Learning is presented. In this section, we distinguish supervised and unsupervised learning, since we use algorithms of both categories.

Deep Learning and Representation Learning are also introduced. Additionally, we introduce its most relevant components: hidden layers, activation functions, and gradient-based learning. These topics are the basis of most of the methods conceived in this thesis work. Two Representation Learning methods are employed: Convolutional Neural Networks, and Restricted Boltzmann Machines. We use Convolutional Neural Networks for classification in the glioma segmentation and grading approaches. Restricted Boltzmann Machines are used together with a Random Forest classifier to form the Machine Learning system that is studied in the context of Machine Learning interpretability in Chapter 5. Restricted Boltzmann Machines learn features in an unsupervised way. Thus, a supervised Random Forest classifier was used to learn a classification task on top of those features. Finally, the chapter closes with a discussion on the need for interpretability of Machine Learning methods.

Contents

3.1 General concepts

Machine Learning can be broadly defined as the set of methods that learn directly from data. Generally speaking, it consists in a model learning how to do a certain task. The process of learning is usually called training, while the execution of the task is designated as inference, or test. During training, it is used some kind of performance metric to evaluate the model. Additionally, during training the parameters, also known as weights in some models, of the chosen model are optimized and adapted to better perform the task. The parameters controlling the model that are not adapted during training are called hyperparameters, for example, the number of hidden layers in a CNN, or the number of trees in a RF (Murphy, 2012; Goodfellow et al., 2016).

The data consists of a set $\mathcal{D}$ of samples, or examples, that represent the observed variables. The observed variables are represented as features, which are quantitative measures characterizing the data. Thus, a sample i is often represented as a features vector of m features $\boldsymbol{x^i} = [x_j : j = 1, \cdots, m]$, with $\boldsymbol{x^i} \in \mathbb{R}^m$. So, the complete observed data with all the collected n samples may be grouped in a design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$. However, in some cases, if there is structural meaning, it may be advantageous to keep the structure of the samples. For instance, $\boldsymbol{x^i}$ may represent an image, an audio signal, or a video. In this case, we may refer to it as feature maps (Murphy, 2012; Goodfellow et al., 2016).

3.1.1 Types of learning

Although it may be difficult to establish a clear separation regarding the learning mode, we can broadly divide Machine Learning methods into supervised learning and unsupervised learning. Another category of methods is called reinforcement learning, but these approaches are out of the scope of this thesis.

3.1.1.1 Supervised learning

In this context, we are usually dealing with predicting some target variable y that follows some function f of the data. So, during training, the model learns a function $\hat{f}$ that maps the observed variables $\boldsymbol{x}$ to

the predicted target variable $\hat{y}$, such that $\hat{y} = \hat{f}(\boldsymbol{x})$. The target variable can be a vector of targets $\boldsymbol{y}$, but, for the sake of simplicity, let us assume scalar predictions. In order to learn such mapping, during training, the model is presented with annotated data, i.e., for each sample it is known the corresponding target value, such that $\mathcal{D} = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^n$. Of course, at test time, the model receives the observable variables only, and infers a prediction $\hat{y}$ (Murphy, 2012; Goodfellow et al., 2016).

The need for providing y at training time may be one drawback of supervised learning, since such labeling usually requires human work, and sometimes it is not humanly possible. Still, if enough training data is available, these approaches often obtain very competitive and state-of-the-art results.

3.1.1.2 Unsupervised learning

Contrasting with supervised learning, in unsupervised learning the algorithms only experience the observable data without any target variable, or supervision. Therefore, the data consists only of samples with the corresponding features $\mathcal{D} = \{(\boldsymbol{x}^i)\}_{i=1}^n$. For this reason, the algorithms need to unveil interesting patterns in the data. In other words, these algorithms learn a distribution of the data, such that we have models of the form $p(\boldsymbol{x}|\mathcal{D}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the parameters of the model. For the sake of simplicity, let us consider $p(\boldsymbol{x})$, instead (Murphy, 2012; Goodfellow et al., 2016).

In the context of this thesis, we will emphasize unsupervised representation learning. In this case, the model needs to learn a representation of the data that keeps as much information as possible, while being subjected to some kind of constraint. There are two main reasons for the constraints: 1) to prevent the model from copying the input data into its latent variables, and 2) to learn a simpler representation than $\boldsymbol{x}$ by discarding irrelevant and redundant information. Examples of constraints include, but are not limited to, enforcing lower dimensional representations, sparse representations, or independent representations (Murphy, 2012; Goodfellow et al., 2016).

Examples of unsupervised learning algorithms include Principal Component Analysis (Hotelling, 1933), Restricted Boltzmann Machines (Smolensky, 1986; Hinton, 2002), or Autoencoders (Hinton and Zemel, 1994).

3.1.2 Classification task

Several supervised learning tasks exist, however, considering the scope of this work, classification will be emphasized. Classification deals with predicting a class from the observed data. Usually, each class is defined as an integer number, so the models need to learn a function mapping the observed data to one of the k classes, $f : \mathbb{R}^m \rightarrow \{1, \cdots, k\}$, such that $\hat{y} \in \{1, \cdots, k\}$. If $k = 2$, it is called a binary problem, otherwise it is designated as multi-class classification (Murphy, 2012; Goodfellow et al., 2016).

Instead of predicting a class directly, some models infer a probabilistic prediction. This soft classification is often not only necessary with the model at hand, but also desirable. For instance, it may be a measure of certainty about a prediction. In this scenario, a probabilistic distribution $p(y|\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta})$ is estimated, where $\boldsymbol{\theta}$ represents the parameters of the model. Hence, the probability is conditioned on the observable data $\boldsymbol{x}$, the training set $\mathcal{D}$, and the model with parameters $\boldsymbol{\theta}$. For the sake of simplicity, we may omit $\mathcal{D}$ and $\boldsymbol{\theta}$, such that the notation becomes $p(y|\boldsymbol{x})$. Having the probabilistic prediction of each

class, the final hard classification can be easily achieved as the maximum a posteriori estimate (Murphy, 2012; Goodfellow et al., 2016) as

$$\hat{y} = \hat{f}(\boldsymbol{x}) = \underset{c \in \{1,\cdots,k\}}{\text{argmax}} \; p(y = c|\boldsymbol{x}) \,. \tag{3.1}$$

Examples of supervised learning algorithms used in classification problems include Convolutional Neural Networks (LeCun et al., 1989, 1998), Random Forests (Breiman, 2001), or Support Vector Machines (Cortes and Vapnik, 1995).

3.2 Deep Learning and representation learning

As mentioned before, the observable data is represented by quantitative features. Such features may be directly measured, such as image intensities, or properties of some objects like their height. Hence, the success of a Machine Learning model is tightly related with domain knowledge of the data and its representation. So, data scientists generally employ many efforts on developing data pre-processing and transformation to achieve better and more discriminative features. Such features are called handcrafted features, and the process of their creation is designated as feature engineering. This step is required by methods that cannot extract features directly from the data (e.g. RF). The data scientist, therefore, uses domain and prior knowledge to develop discriminative features. Although this process is effective, especially in small datasets, it has some disadvantages: it requires expert knowledge on the problem, the features are often problem-dependent, and it is labor intensive. Additionally, in some domains, such as image analysis, usually the raw image cannot be directly fed to the Machine Learning algorithm and needs to be transformed into another representation. Such transformation may discard the inherent structure of the data (LeCun et al., 2015; Bengio et al., 2013).

Contrasting with feature engineering, Representation Learning refers to algorithms that learn to extract representations, or features, directly from the data. Methods with several layers/levels of non-linear Representation Learning modules are designated as Deep Learning models. In principle, each layer takes the previous representations and learns how to extract higher order and more complex features, before feeding them to the next layer. Hence, Deep Learning can be broadly defined as a field of Machine Learning dealing with algorithms that learn a hierarchy of concepts directly from the data. In principle, any method fitting in this definition can be categorized as a Deep Learning algorithm (LeCun et al., 2015; Bengio et al., 2013; Goodfellow et al., 2016). Still, in the context of this thesis, when referring to Deep Learning, we will be referring to Deep Artificial Neural Networks. Regarding the learning process, Deep Learning algorithms may be supervised, unsupervised, or a mix of the two, i.e., semi-supervised.

In handcrafted features, the development is mainly focused on designing features. In the case of Deep Artificial Neural Networks the focus moves to architecture designing. An architecture is the arrangement of layers, their hyperparameters and relationships. Generally speaking, an Artificial Neural Network has input and output layers, while the layers in-between are called hidden layers. Fig. 3.1 shows a simple feedforward neural network with one hidden layer. This kind of architecture is also known as Multi-layer Perceptron. If more hidden layers exist, it is said to be a deep feedforward neural network. The layers

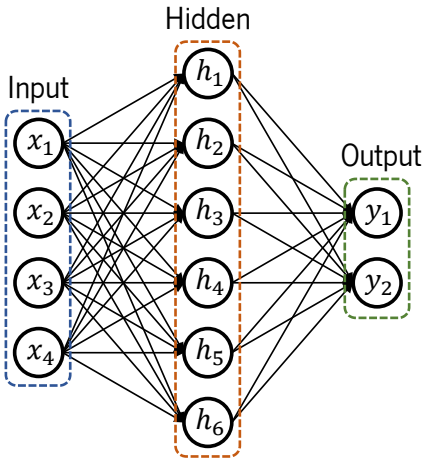


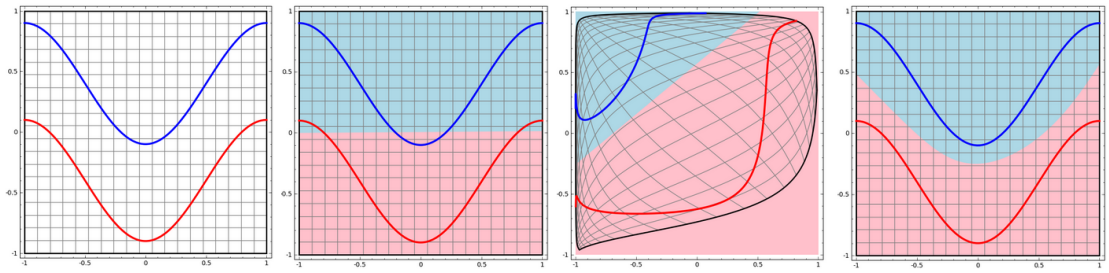
Figure 3.1: Example of an Artificial Neural Network with one input layer with four units, a hidden layer with six units, and an output layer with two units. All layers are fully-connected.

are said to be fully-connected because a unit in one layer is connected to all units in the previous and the next layers. These networks are designated as feedforward neural networks because the information flows from the input layer, through the hidden layers, until the output layer. If feedback exists, where outputs are fed back to the network, it would be called Recurrent Neural Networks, but these are out of the scope of this thesis (LeCun et al., 2015; Goodfellow et al., 2016).

3.2.1 Hidden layers

These layers are a requirement for learning more complex and abstract features. This process happens by combining the signal of the previous layer, and non-linearly transforming it. The non-linear transformation is achieved through a non-linear activation function that is applied to the output of each unit in the hidden layers. Taking into account a classification problem, having complex, high level, discriminative features is essential to transform the data into a representation that makes the classes more easily separable. In other words, it enables the Artificial Neural Network to learn a non-linear separation of the data. Fig. 3.2 depicts how this is enabled by the hidden layer. Fig. 3.2(a) shows the input data, which is not linearly separable by a two-layer model (no hidden layer), as observed in Fig. 3.2(b). However, with a hidden layer and activation function the data is transformed into a representation that enables a linear separation of the data (Fig. 3.2(c)). Hence, the network can learn a non-linear separation function (LeCun et al., 2015; Goodfellow et al., 2016).

As a Machine Learning algorithm, Artificial Neural Networks try to learn a function f of the observable variable $\boldsymbol{x}$. In a feedforward neural network, such function is the result of the chain of functions corresponding to each layer (f^l). So, if the network has three layers (excluding the input layer), we can write $f(\boldsymbol{x}) = f^3(f^2(f^1(\boldsymbol{x})))$. In Artificial Neural Networks, the output of a given layer usually results from an affine transformation of the input vector followed by an element-wise non-linear activation function ϕ. To that end, each layer indexed by l (0 corresponds to the input layer) has a weight matrix $\boldsymbol{W_l} \in \mathbb{R}^{n_{h_{l-1}} \times n_{h_l}}$ defining the connections between the $n_{h_{l-1}}$ units in the previous layer and the n_{h_l} units in the current layer l. It has, additionally, a bias vector $\boldsymbol{b_l} \in \mathbb{R}^{n_{h_l} \times 1}$. Hence, the output of a given



(a) (b) (c) (d)

Figure 3.2: Effect of having hidden layers in an Artificial Neural Network. The classes of the dataset a) are hard to be separated by a shallow network of just two layers (input and output) b). In c) it is represented the data transformed by the hidden layer, while d) shows the new separation line. Reproduced with permission from (Olah, 2018).

layer $\boldsymbol{h}_l$ is given by

$$\boldsymbol{z}_l = \boldsymbol{W}_l^T \boldsymbol{h}_{l-1} + \boldsymbol{b}_l \tag{3.2}$$

$$\boldsymbol{h_l} = \phi\left(\boldsymbol{z}_l\right), \tag{3.3}$$

where $\boldsymbol{z}_l$ is known as pre-activation (Goodfellow et al., 2016). Taking the example of the network in Fig. 3.1, we would have $\boldsymbol{W_1} \in \mathbb{R}^{4\times6}$, $\boldsymbol{b_1} \in \mathbb{R}^{6\times1}$, $\boldsymbol{h_1} \in \mathbb{R}^{6\times1}$, and $\boldsymbol{h}_0 = \boldsymbol{x}$.

3.2.2 Activation functions

Among these functions, one can use the sigmoid function (equation 3.4), the hyperbolic tangent function (equation 3.5), the rectifier linear unit (ReLU) function (equation 3.6), and its variants such as the leaky ReLU (equation 3.7).

$$\phi\left(\boldsymbol{z}_l\right) = \frac{1}{1+e^{-\boldsymbol{z}_l}}. \tag{3.4}$$

$$\phi\left(\boldsymbol{z}_l\right) = \frac{e^{\boldsymbol{z}_l} - e^{-\boldsymbol{z}_l}}{e^{\boldsymbol{z}_l} + e^{-\boldsymbol{z}_l}}. \tag{3.5}$$

$$\phi\left(\boldsymbol{z}_l\right) = \max\left(0, \boldsymbol{z}_l\right). \tag{3.6}$$

$$\phi\left(\boldsymbol{z}_l\right) = \max\left(0, \boldsymbol{z}_l\right) + \alpha \min\left(0, \boldsymbol{z}_l\right), \tag{3.7}$$

where α is the leakiness parameter. Still, many other activation functions exist, although these are arguably among the most used ones. Fig. 3.3 shows the output of the activation functions for a range of pre-activation values. It is possible to observe that sigmoid and hyperbolic tangent have limited output ranges of $]0, 1[$ and $]-1, 1[$, respectively. Therefore, it is said that they saturate, while ReLU and variants

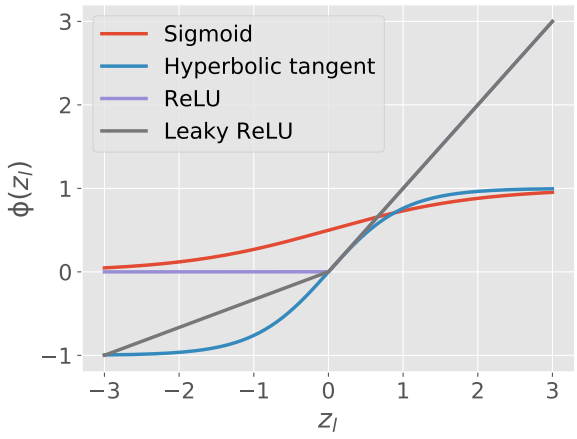


Figure 3.3: Activation functions.

do not.

In the case of classification, it is desirable to have an output probabilistic prediction. If only one output unit exists, such as in binary problems, the output of the sigmoid function may provide a probabilistic interpretation. However, in the case of multi-class problems there exists one output unit for each class. Thus, we need the softmax function, which normalizes the input across the output of all output units (Goodfellow et al., 2016). Hence, if j indexes each output unit, the probabilistic prediction of the i-th unit is given by

$$softmax\left(z^i\right) = \frac{e^{z^i}}{\sum_j e^{z^j}} \tag{3.8}$$

such that

$$\sum_i softmax\left(z^i\right) = 1. \tag{3.9}$$

3.2.3 Gradient-based learning

As mentioned before, a feedforward neural network is nothing more than a composition of functions, where each layer represents a function. So, if we define a differentiable loss function, or cost function, and all layers enforce differentiable operations, it is possible to drive learning by gradient descent. This is possible through the minimization of the loss function. To that end, the adjustable weights of the network are adapted in the direction of the negative of the gradient at a given point (the data) (LeCun et al., 2015; Goodfellow et al., 2016). We can intuitively perceive the principles of such approach from Fig. 3.4, where it is represented a simple quadratic function and tangent functions on three points. The gradient at a given point is the same as the slope of a tangent line in that point. We can easily see that the minimum stands in the direction of the negative of the slope of the tangents of each of those points. We can also observe that when the minimum is reached the gradient is 0.

In Artificial Neural Networks, the loss function is highly non-convex, with many local minima. The above-mentioned approach converges to a minimum, without guaranteeing that it is a global one. Indeed,

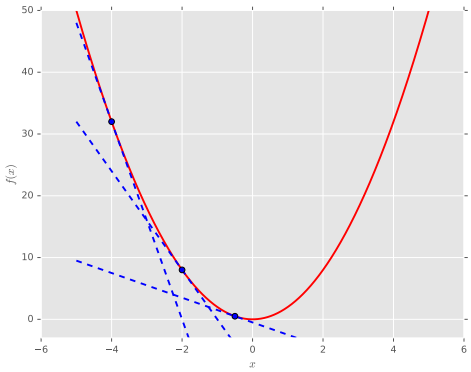


Figure 3.4: Quadratic function (solid red line) with tangent lines (dashed blue lines) on points $x = -4$, $x = -2$, and $x = -0.5$.

most likely it will converge to a local minimum. Nevertheless, as networks get deeper, arriving at a poor local minimum becomes less problematic, as many local minima are similar. In fact, the global minimum is more prone to overfitting (Choromanska et al., 2015).

3.2.3.1 Chain rule and Back-propagation

In a feedforward neural network, we feed in some data $\boldsymbol{x}$ and obtain an output $\hat{y}$, which is the predicted variable. This is called forward pass. During training, we can, then, compute the loss function $J(\boldsymbol{\theta})$, which yields a scalar measure of the prediction error. For gradient learning, we need to compute the gradient of the loss function in relation to the parameters of each layer $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. Back-propagation is an algorithm that allows one to compute such gradients (Rumelhart et al., 1986; Goodfellow et al., 2016).

Let us consider a function that is a composition of three functions f_1, f_2, and f_3, in such a way that $x = f_1(w)$, $y = f_2(x)$, and $z = f_3(y)$, or $z = f_3(f_2(f_1(w)))$. The chain rule of calculus allows us to compute the gradient of such composition of functions, provided that the gradient of each of the composing functions is known, as

$$\frac{\partial z}{\partial w} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}\frac{\partial x}{\partial w}. \tag{3.10}$$

So, the back-propagation algorithm is simply an algorithm that computes the chain rule in an efficient way (Goodfellow et al., 2016). First, given input data and a model, the forward pass is computed, where all the pre-activations and activations are calculated. Finally, the output is fed to the loss function, which provides a cost value. The forward pass of a generic feedforward neural network is depicted in Algorithm 1; ψ_l represents the pre-activation function, for instance, the affine transformation of equation 3.2.

In the backward pass (Algorithm 2) it is necessary to compute the gradient of the loss function in relation to the parameters of each layer, in order to provide the update direction and value. Therefore, first it is computed the gradient of the loss function in relation to the predicted variables $\hat{y}$; note that the target variables y are required to compute the loss, but we do not calculate the gradient in relation to them. Then, going from the last to the first layers, the gradient of the loss function in relation to each part of the network is recursively calculated. In this way, it is possible to achieve the gradient of the loss in relation to each layer's parameters, and propagate the gradient to the previous layer (Goodfellow et al.,

Algorithm 1 Forward pass in a generic feedforward neural network (Goodfellow et al., 2016). $L\left(\hat{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{\theta}\right)$ is a loss function that depends on the predicted $\hat{\boldsymbol{y}}$ and target $\boldsymbol{y}$ variables, and the parameters $\boldsymbol{\theta}$ of the network. ψ_l represents the pre-activation function in layer l using the parameters of the network, and ϕ_l is the activation function in layer l.

Require: $\boldsymbol{x}$, the input data
Require: $\boldsymbol{y}$, the target variable
Require: n_l, the number of layers
Require: $\boldsymbol{\theta}$, the parameters of the model
Require: L, the loss function
$\boldsymbol{h}_0 = \boldsymbol{x}$
for $l = 1, \cdots, n_l$ **do**
$\boldsymbol{z}_l = \psi_l\left(\boldsymbol{h}_{l-1}\right)$
$\boldsymbol{h}_l = \phi_l\left(\boldsymbol{z}_l\right)$
end for
$\hat{\boldsymbol{y}} = \boldsymbol{h}_l$
$J\left(\boldsymbol{\theta}\right) = L\left(\hat{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{\theta}\right)$

Algorithm 2 Backward pass for computing the gradients in a generic feedforward neural network (Goodfellow et al., 2016). The algorithm starts in the last layer and runs backwards until the first layer.

Require: $\boldsymbol{y}$, the target variable
Require: n_l, the number of layers
Require: $\boldsymbol{\theta}$, the parameters of the model
Require: L, the loss function
Compute the gradient of the loss function in relation to the output variables
$\boldsymbol{g} \leftarrow \nabla_{\hat{\boldsymbol{y}}} J = \nabla_{\hat{\boldsymbol{y}}} L\left(\hat{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{\theta}\right)$
for $l = n_l, n_l - 1, \cdots, 1$ **do**
Convert the gradient on the layer's output into the gradient on the layer's pre-activation
$\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{z}_l} J = \boldsymbol{g} \odot \nabla_{\boldsymbol{z}_l} \boldsymbol{h}_l$
Compute the gradient on the parameters of the model
$\nabla_{\boldsymbol{\theta}_l} J = \boldsymbol{g} \nabla_{\boldsymbol{\theta}} \boldsymbol{z}_l$
Propagate the gradient to the next layer's activations
$\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{h}_{l-1}} J = \boldsymbol{g} \nabla_{\boldsymbol{h}_{l-1}} \boldsymbol{z}_l$
end for

2016). Note that in Algorithm 2 it is shown the case of a single sample, but usually one processes several samples at once.

3.2.3.2 Optimization

Back-propagation yields the gradients of the loss function in relation to the parameters of the model. However, it does not represent learning. Instead, learning is achieved by optimization through the update of the parameters of the model in order to minimize the cost value. In the context of Machine Learning, optimization differs from classical settings. More specifically, in typical optimization the minimization of the cost function is the end in itself. In Machine Learning, the cost function is often used as a surrogate function that does not represent directly the loss function we may be interested in optimizing. For instance, we may want to minimize the error rate of a model, but such loss can be intractable, such as the 0-1 loss.

In this case, we may instead minimize the cross-entropy loss function, in the hope that it will lead to low error rates. Additionally, in classical problems, the minimization finishes when the lowest cost values are achieved. This is often not the case in Machine Learning. We may be periodically evaluating the desired metric, e.g. error rate, in an independent data set, and finish optimization when a satisfactory metric value is achieved, even though the loss function may be further minimized. This procedure is called early stopping (Goodfellow et al., 2016).

Stochastic Gradient Descent It is possible to compute the gradients and perform optimization while taking the whole data set. This is called batch gradient methods. Still, this is most likely computationally impractical in Deep Learning. Thus, an iterative approach where subsets of samples are uniformly drawn from the training set is used. In this case, it is designated as mini-batch stochastic gradient methods. In the case where only one sample is used in each step it is called stochastic gradient methods. However, in the Machine Learning scientific community, mini-batch approaches are called as stochastic approaches, hence, we use these terms interchangeably (Goodfellow et al., 2016).

The gradient of a mini-batch is an unbiased estimation of the true gradient over the whole dataset. Hence, in an iteration t, if we sample m training samples for each mini-batch, the estimate of the gradient is:

$$\hat{\boldsymbol{g}}_t = \frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m} L\left(\hat{\boldsymbol{y}}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{\theta}\right) \tag{3.11}$$

This interpretation of the gradient is valid only if there is no repetition of samples. However, in practice, random mini-batches are sampled until all the dataset is used. Then, we can go over the training set several times. Each pass of the training set is called an epoch. It is important to extract samples uniformly in each epoch. After estimating the gradient of the loss function in relation to the parameters of the model (equation 3.11), the parameters can be updated with the following update rule (Goodfellow et al., 2016),

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \epsilon \hat{\boldsymbol{g}}_t, \tag{3.12}$$

where ϵ is the learning rate. This procedure is called Stochastic Gradient Descent (SGD).

The learning rate is arguably one of the most important hyperparameters when training a feedforward neural network. If it is too high, learning may become unstable, and the algorithm may lose the minimum. In fact, it may render optimization impossible, and increase the cost value. If ϵ is too low, learning may be too slow and even stall. Often, it is beneficial to start with a learning rate as high as possible, and decrease it over iterations, following some scheduling (Goodfellow et al., 2016).

Momentum Although the updates of the parameters of the model may be noisy, SGD works well in practice. Nevertheless, it may be slow in regions of the loss function's surface with low curvature. Momemtum (Polyak, 1964) helps in accelerating learning in this scenario, especially if gradients are small but consistent in the same direction, or noisy (Sutskever et al., 2013; Goodfellow et al., 2016).

In momentum, a velocity variable $\boldsymbol{\nu}$ accumulates the gradients of the previous updates. In this way,

if the updates were consistent in one direction, it accelerates training in that direction. The weight given to the influence of the previous updates in a given iteration is controlled by $\mu \in [0, 1)$. The update rule becomes as follows (Sutskever et al., 2013; Goodfellow et al., 2016)

$$\boldsymbol{\nu}_{t+1} = \mu\boldsymbol{\nu}_t - \epsilon\hat{\boldsymbol{g}}_t \tag{3.13}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\nu}_{t+1}. \tag{3.14}$$

It is possible to define a schedule for μ, where it starts with low value, and is increased over time. However, it is usually more relevant to decrease the learning rate, while keeping μ constant (Goodfellow et al., 2016).

Nesterov's Accelerated Gradient (Sutskever et al., 2013) is closely related with momentum, but it behaves in a more stable way and with less oscillation, especially if μ is large. This approach tries to apply a correction to momentum, by evaluating the gradient after applying the current velocity. Hence, the update rule becomes

$$\hat{\boldsymbol{g}}_t = \frac{1}{m}\nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m} L\left(\hat{\boldsymbol{y}}^{(i)}, \boldsymbol{y}^{(i)}, \boldsymbol{\theta} + \boldsymbol{\mu}\boldsymbol{\nu}_{\boldsymbol{t}}\right) \tag{3.15}$$

$$\boldsymbol{\nu}_{t+1} = \mu\boldsymbol{\nu}_t - \epsilon\hat{\boldsymbol{g}}_t \tag{3.16}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\nu}_{t+1}. \tag{3.17}$$

Adam The learning rate value is arguably one of the most important hyper-parameters during learning. Therefore, some approaches try to automatically adapt it during the course of training. Additionally, these techniques can estimate a learning rate for each parameter of the model. One of these methods is the Adaptive Moment Estimation (Adam) (Kingma and Ba, 2015). Similarly to momentum, Adam takes into account the past iterations as the gradient (first moment) and square of the gradient (second moment) when computing the update rule. It initializes the moments as zero, therefore, there is a bias towards zero. But, the algorithm includes corrections for these biases. The Adam algorithm is depicted in Algorithm 3.

3.2.3.3 Cross-entropy loss function

Suppose we have output activation functions that allow one to interpret the output of an Artificial Neural Network as probabilistic, such as the sigmoid (equation 3.4) or softmax (equation 3.8) functions. Then, a possible loss function for classification problems is the cross-entropy loss, defined as

$$J\left(\boldsymbol{\theta}\right) = L\left(\hat{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{\theta}\right) = -\sum_{i} y_i \log \hat{y}_i, \tag{3.18}$$

where i indexes the corresponding class, $\hat{y}$ represents the probabilistic prediction, and y is the true class following one-hot encoding.

Algorithm 3 Adam optimization algorithm (Kingma and Ba, 2015).

Require: ϵ, the learning rate
Require: $\beta_1, \beta_2 \in [0, 1)$, the decay rate for moment estimates
Require: $\boldsymbol{\theta}_0$, the initial parameters of the model
Require: δ, small stability constant
  $\boldsymbol{m}_0 \leftarrow 0$ (first moment initialization)
  $\boldsymbol{v}_0 \leftarrow 0$ (second moment initialization)
  $t \leftarrow 0$ (iteration initialization)
  while Stopping criteria is not reached **do**
    $t \leftarrow t + 1$
    $\hat{\boldsymbol{g}}_t \leftarrow \nabla_{\boldsymbol{\theta}} L\left(\hat{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{\theta}_{t-1}\right)$
    $\boldsymbol{m}_t \leftarrow \beta_1 . \boldsymbol{m}_{t-1} + \left(1 - \beta_1\right) . \hat{\boldsymbol{g}}_t$
    $\boldsymbol{v}_t \leftarrow \beta_2 . \boldsymbol{v}_{t-1} + \left(1 - \beta_2\right) . \hat{\boldsymbol{g}}_t^2$
    $\hat{\boldsymbol{m}}_t \leftarrow \frac{\boldsymbol{m}_t}{1-\beta_1^t}$ (bias-corrected first moment estimate)
    $\hat{\boldsymbol{v}}_t \leftarrow \frac{\boldsymbol{v}_t}{1-\beta_2^t}$ (bias-corrected second moment estimate)
    $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \epsilon . \frac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t} + \delta}$
  end while

Mean squared error loss is also possible to use in classification, but, it leads to saturation and slower learning, due to smaller gradients. On the other hand, it is possible to observe from equation 3.18 some desirable properties of the cross-entropy loss. When a sample is correctly classified, it leads to lower gradients. Still, the gradient is larger if the probability reveals it was uncertain. Moreover, the more the model is mistaken (e.g. estimating a high probability for a wrong class), the larger the gradient is, which leads to larger weight updates in the opposite direction to try to compensate it (Goodfellow et al., 2016).

3.2.4 Factors enabling Deep Learning models

A significant part of the Artificial Neural Network models used today exist since the 1980s, and before. However, only after 2006 it was possible to train deep models. In that year, Hinton et al. (2006) showed how to train a deep model built with stacked Restricted Boltzmann Machines. Each layer was trained in an unsupervised and greedy way, by taking as input the output of the previous layer. In the end, fine-tuning was done through supervised training with backpropagation. This kind of procedure became known as pre-training.

Pre-training became a success because it provided a way for learning better parameters before the supervised stage. This points out one of the problems of Deep Learning methods – the initialization of the parameters. Without proper initialization, deep networks tended to experience unstable training because of exploding or vanishing gradients. Later, it was found how to condition the initialization of the weights (Glorot and Bengio, 2010; He et al., 2015). With proper initialization it is now possible to train deep neural networks from scratch, in an end-to-end fashion, without the tedious process of greedy layer-wise pre-training.

Activation functions played a role as well. It was observed that sigmoid and hyperbolic tangent activation functions lead to poor training. This is due to the fact that these functions can saturate, which may stall the training (Glorot and Bengio, 2010). On the other hand, rectifying non-linearities were found to make

training easier, leading to better learning and models (Glorot et al., 2011; Jarrett et al., 2009). Moreover, ReLU was shown to be helpful not only in discriminative supervised settings, but also in unsupervised generative Restricted Boltzmann Machines (Nair and Hinton, 2010).

Other factors that contributed for enabling training of deeper models include the use cross-entropy loss functions instead of mean squared error, more and efficient regularization procedures (Srivastava et al., 2014), the availability of larger datasets, and more powerful hardware. These factors, together with the use of ReLU, were explored in a CNN by Krizhevsky et al. (2012) for winning the Imagenet competition. This achievement is arguably one of the most important in showing the power of Deep Learning, leading to its adoption and development by a larger community.

3.3 Convolutional Neural Networks

Convolutional Neural Networks are a group of Artificial Neural Networks that have at least one special kind of layer called convolutional layer. In these layers, instead of matrix multiplication operations (equation 3.2), it is employed convolution operations. Hence, the weights of the layer are organized as kernels, or filters, that are convolved over the input. Note, however, that in the context of Artificial Neural Networks, the term convolution is loosely employed, and does not strictly refers to the mathematical convolution operation. In fact, convolutional layers usually employ cross-correlation operations, which are more straightforward to implement. The convolutional layer is especially designed to take advantage of structured inputs, which are organized as grid-like structures (LeCun et al., 2015; Goodfellow et al., 2016). Examples of such inputs are audio data (1D), pictures (2D), or MRI images (3D). Here, we will focus the description on 2D images, but the concepts are equivalent for 1D or 3D data.

3.3.1 Convolutional layer

Kernels are convolved over the inputs. Since kernels are usually much smaller than the inputs, the result of a convolutional layer is also organized in a grid-like structure similar to images, hence being called *feature maps*. So, the input of a convolutional layer is a stack of feature maps. In the special case of the first layer, this stack is the input image, which may have several channels, e.g. RGB. Therefore, the several feature maps in a stack may be called as channels. A convolutional layer has several filter (kernel) banks. Each filter bank consists of weights being convolved in the spatial dimensions of the input feature maps, across all channels (Fig. 3.5(a)). For instance, if there are n_c input channels, a given filter bank with kernel width k_1 and height k_2 has dimensions of $n_c \times k_1 \times k_2$ (Goodfellow et al., 2016). The set of output feature maps is $\boldsymbol{M}$. So, for each filter bank o, there is a corresponding output feature map $\boldsymbol{M}_o$ computed as

$$\boldsymbol{M}_o = b_o + \sum_{c=1}^{n_c} \boldsymbol{X}_c * \boldsymbol{W}_{o,c}, \tag{3.19}$$

where b_o is the bias term, summed element-wise, $*$ represents convolution, c indexes the input feature map, and $\boldsymbol{X}$ are the input feature maps. Therefore, it is possible to observe that a filter bank connects

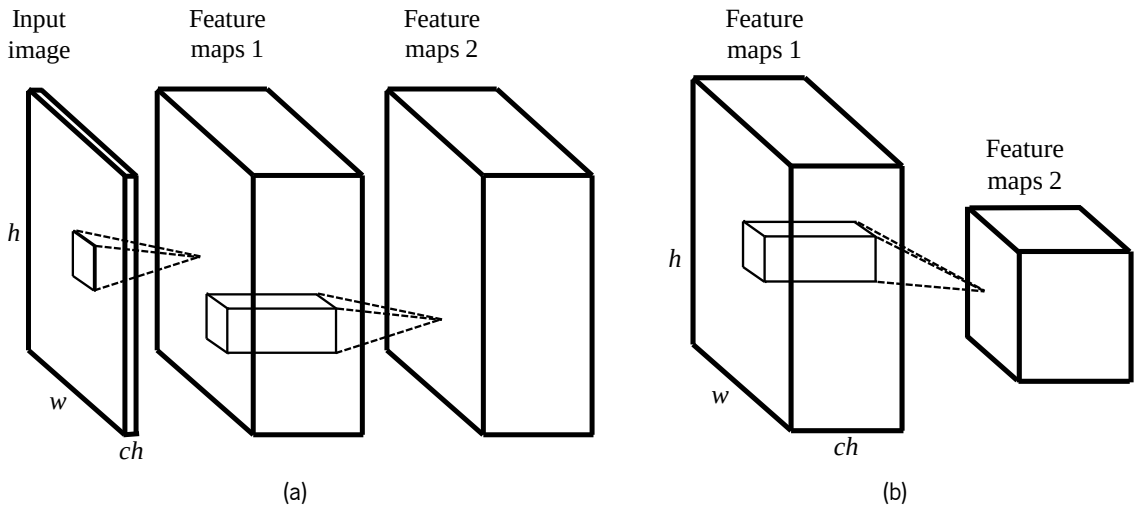


Figure 3.5: Convolutional layers. The kernel is much smaller than the input feature maps, but spreads across all channels. a) Two stacked convolutional layers, where the first one operates over the input image. b) Convolutional layer with stride of 2. Note how the resulting feature maps are half the size of the input ones.

all the input maps to a given output feature map. Now, considering a given location in the output feature maps indexed at i and j, its value given by convolution is (LeCun et al., 2015; Goodfellow et al., 2016)

$$M_{o,i,j} = \sum_{c=1}^{n_c} \sum_{m=1}^{k_1} \sum_{n=1}^{k_2} X_{c,(i-1)\times s+m,(j-1)\times s+n} W_{o,c,m,n}, \tag{3.20}$$

where $s \in \mathbb{N}^+$ is the stride. When $s = 1$, the convolution operation is applied in every location of the input. Otherwise, if $s > 1$, convolution is not computed everywhere in the input image, instead, it skips s pixels in between each location. Hence, when $s > 1$, it is obtained a downsampled output feature map (Fig. 3.5(b)), which may be sometimes desirable to decrease the computational costs at the expense of getting coarser features (Goodfellow et al., 2016). For the sake of simplicity, we considered the same stride for both width and height dimensions, but it can be different.

Finally, the way how the borders are treated impact the output feature maps. Two main modes are usually employed: *valid* and *same*. In *valid* mode, convolution is computed only in regions where the kernels are totally contained inside the input. So, considering k as the kernel spatial dimensions, the output feature maps will be smaller than the inputs by $2\frac{k-1}{2}$ units. Therefore, this may limit the number of convolutional layers. In the case of the *same* mode, the input feature maps are zero-padded in such a way that the output feature maps have the same size as the inputs. In this way, the number of convolutional layers is not limited, but artificial information may be introduced, which affects more the units closer to the borders (Goodfellow et al., 2016).

Equation 3.19 defines the pre-activation in a convolutional layer. Hence, after this operation, the feature maps go through an activation function, such as ReLU (equation 3.6). The activation is applied element-wise to each unit of the feature maps.

Advantages of convolutional layers Compared with more traditional fully-connected layers, convolutional layers yield some interesting properties: sparse connectivity, parameter sharing, and invariance to translation. In the case of fully-connected layers, all output units are connected to all input units. This is not the case of convolutional layers, because an output unit depends only on a small patch of input units, with which it is connected through the weights of the kernel. Hence, it is similar to having all weights in a fully-connected layer set to zero except for those connecting to that input patch. In this way, we can say that it yields sparse connectivity. Parameter sharing arises from the very nature of convolution. Since the same kernel is convolved all over the input, each unit in the output feature map will be computed using the same weights. So, they all share the same parameters. Sparse connectivity and parameters sharing lead to CNNs having much less parameters than Deep Artificial Neural Networks with all layers being fully-connected. Finally, parameter sharing leads to the last property that is invariance to translation. Since the same kernel is used in every position of the input, the same patterns are detected regardless of their position in the input feature maps. So, if we shift the pixels in an image, the convolution will yield the same output for the same pattern. This is a very important advantage of convolutional layers over fully-connected layers that make them better suited for image analysis because, in natural images, objects may appear in different regions of the images. Still, convolutional layers are not invariant to other transformations, such as rotation or scale (Goodfellow et al., 2016).

A unit in a feature map is connected, through the weights of the convolutional layer, only to a small region of the input data. Nevertheless, when there are several convolutional layers in series, a unit is indirectly connected to a larger region of the input image. For instance, having 3×3 kernels, the output units of the first layer are connected to a 3×3 region of the input image. However, the output units of the second layer are connected to a 3×3 region of the output of the first layer, but indirectly linked to a 5×5 region of the original input image. The total region that a unit "observes" projected in the original input image is called field of view. In this way, deeper layers of CNNs can learn more complex regions. The first layers usually capture edges. Then, in subsequent layers, the edges are aggregated into motifs, motifs into parts, and parts, finally, form objects (LeCun et al., 2015; Goodfellow et al., 2016).

In summary, convolutional layers are well suited for structured data because 1) the information in the neighborhood of a pixel is usually more relevant, since it is related. 2) Similar patterns may arise in several regions of the image. 3) Stacked convolutional layers capture higher order features. The first reason is due to local correlation, whereas the second one is because local statistics are invariant to location (LeCun et al., 2015; Goodfellow et al., 2016).

3.3.2 Pooling layer

The pooling layer is usually found in CNNs. It consists in summarizing a given region of the feature map into some statistic. So, there is a kernel defining a neighborhood to be considered, over which it computes the desired function. Pooling is usually applied with some stride ≥ 1 for reducing the computational cost of the next layers. Arguably, max-pooling (Fig. 3.6(a)), where the output is the maximum value found in a given kernel-defined region, is the most common pooling operation found in CNNs. Notwithstanding, one can use other aggregating functions, such as average (Fig. 3.6(b)), or L_2 norm (Boureau et al., 2010;

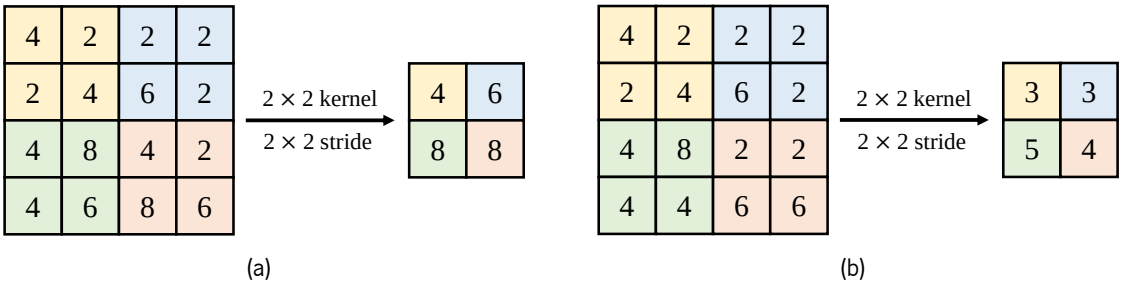


Figure 3.6: Pooling layers with kernel size of 2×2 and stride of 2. a) Max-pooling, and b) average pooling.

LeCun et al., 2015; Goodfellow et al., 2016). An extreme case of pooling is the global pooling, such as global average pooling (Lin et al., 2013). In this case, the whole feature map is summarized into a single value, for instance, a stack of feature maps with 10 channels of shape $10 \times 24 \times 24$ is summarized into a $10 \times 1 \times 1$ tensor. Global pooling is often used to obtain a feature vector to be fed into fully-connected layers. In general, pooling is usually employed in each channel of the feature maps independently, i.e., it summarizes only in the spatial dimensions.

Pooling layers yield some beneficial properties. They help in making CNNs invariant to small local translations. Additionally, they may keep relevant features, while discarding small and possibly meaningless details, such as noise. Pooling also increases the field of view of the network; this is especially significant if stride is > 1. Nevertheless, pooling layers do not increase the number of parameters of the network, so, in principle, they do not contribute to overfitting (Boureau et al., 2010; LeCun et al., 2015; Goodfellow et al., 2016). However, Kamnitsas et al. (2017b) argue that pooling may be not suited for tasks requiring fine grained features, such as segmentation. This is due to pooling being informative of the presence or not of a given feature, but not its precise location.

3.4 Restricted Boltzmann Machines

A Restricted Boltzmann Machine (RBM) (Smolensky, 1986) is a type of probabilistic generative Artificial Neural Network that can be used for unsupervised Representation Learning. The units of this model are arranged as a two-layered undirected bipartite graphical model. Therefore, the nodes are organized into one visible and one hidden layer, whose states are represented by the vectors $\boldsymbol{v} = [v_i : i = 1, \ldots, m]$, and $\boldsymbol{h} = [h_j : j = 1, \ldots, n]$, respectively. All nodes in one layer are connected to all nodes in the other layer with weights represented by the matrix $\boldsymbol{W} = [w_{ij}]$. There are no intra-layer connections; this property led to the name of RBM, as it represents a restricted version of the Boltzmann Machine. The structure of the typical RBM can be observed in Fig. 3.7. In contrast to feedforward neural networks, the edges are undirected, meaning that information can flow from the visible layer to the hidden layer, and vice versa (Bengio et al., 2013; Hinton, 2012).

Originally, the units in the layers of RBMs were defined as binary stochastic units, such that $\boldsymbol{v} \in \{0, 1\}^m$ and $\boldsymbol{h} \in \{0, 1\}^n$. The joint probability distribution of the visible and hidden states is given by

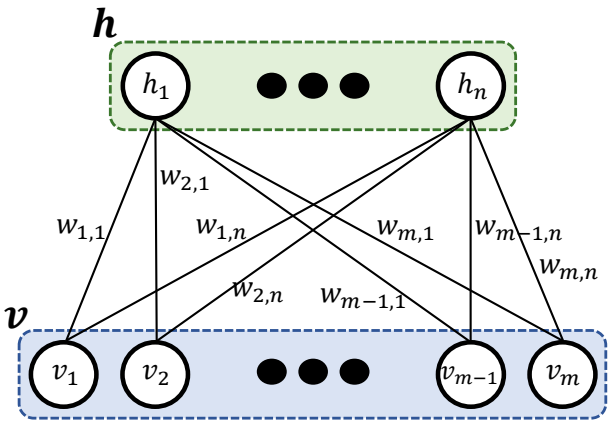


Figure 3.7: Restricted Boltzmann Machine. The weights define undirected connections between the visible layer (blue) and the hidden layer (green).

$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{e^{-E_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})}}{Z_{\boldsymbol{\theta}}} = \frac{\widetilde{p}(\boldsymbol{v}, \boldsymbol{h})}{Z_{\boldsymbol{\theta}}}, \tag{3.21}$$

where $Z_{\boldsymbol{\theta}}$ is the partition function, and $E_{\boldsymbol{\theta}}$ defines an energy over the joint configuration of the states of $\boldsymbol{v}$ and $\boldsymbol{h}$ parametrized by $\boldsymbol{\theta}$, which includes both weights and biases. The energy term is defined as

$$E_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h}) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} v_i h_j w_{ij}, \tag{3.22}$$

where a_i is the bias of the visible unit i, and b_j is the bias of the hidden unit j (Bengio et al., 2013; Hinton, 2012; Goodfellow et al., 2016).

The partition function is a sum over all possible states of $\boldsymbol{v}$ and $\boldsymbol{h}$. So, it is intractable, which means that the probability given to a visible vector computed as

$$p(\boldsymbol{v}) = \frac{\sum_{\boldsymbol{h}} e^{-E_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})}}{Z_{\boldsymbol{\theta}}} = \frac{\widetilde{p}(\boldsymbol{v})}{Z_{\boldsymbol{\theta}}} \tag{3.23}$$

is also intractable (Hinton, 2012; Goodfellow et al., 2016).

Nevertheless, given that intra-layer connections are not allowed, the conditional distribution of one layer given the other factorizes (Bengio et al., 2009, 2013) as

$$p(\boldsymbol{h}|\boldsymbol{v}) = \prod_j p(h_j|\boldsymbol{v}) \tag{3.24}$$

$$p(\boldsymbol{v}|\boldsymbol{h}) = \prod_i p(v_i|\boldsymbol{h}). \tag{3.25}$$

So, having a binary observation vector, we can get a binary sample of the hidden layer, by setting a unit to 1 with probability

$$p(h_j = 1|\boldsymbol{v}) = sigm\left(b_j + \sum_i v_i w_{ij}\right), \tag{3.26}$$

where $sigm$ is the sigmoid function (equation 3.4). Similarly, having the state of the hidden unit, it is straightforward to sample the state of the visible units as

$$p\left(v_i = 1|\boldsymbol{h}\right) = sigm\left(a_i + \sum_j h_j w_{ij}\right), \tag{3.27}$$

which is sometimes denoted as "reconstruction" (Hinton, 2012; Bengio et al., 2009, 2013).

3.4.1 Restricted Boltzmann Machines for real-valued data

Restricted Boltzmann Machines were originally designed for modeling binary data. This is done by having the sigmoid function as the activation function of the visible and hidden units, and sampling the binary state with it. Obviously, binary units are not suited for modeling real-valued data $\boldsymbol{v} \in \mathbb{R}^m$. In the case of images, voxels usually assume real-valued intensities. Still, RBMs can be adapted to deal with this kind of data, by defining an appropriate activation function. One of such models is the Gaussian-Bernoulli RBM (Hinton and Salakhutdinov, 2006). In this model, the hidden layer remains binary, but the units of the visible layer are modeled as linear units with independent Gaussian noise. Hence, the energy becomes

$$E_{\boldsymbol{\theta}}\left(\boldsymbol{v}, \boldsymbol{h}\right) = \sum_i \frac{\left(v_i - a_i\right)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_{ij} \frac{v_i}{\sigma_i} h_j w_{ij}, \tag{3.28}$$

where σ_i is the standard deviation of the Gaussian noise in visible unit i. It is possible to learn $\boldsymbol{\sigma}$, but, it is usually much simpler to just normalize each component of the data to zero mean and unit variance and get noise free reconstructions in the visible layer. In this case, $\sigma_i = 1$ in equation 3.28 (Hinton, 2012; Nair and Hinton, 2010). In this setting, having the state of the hidden layer, we can sample the state of the visible layer from a Gaussian distribution $\mathcal{N}$ (Hinton and Salakhutdinov, 2006) as

$$p\left(v_i|\boldsymbol{h}\right) = \mathcal{N}\left(a_i + \sum_j w_{ij} h_j, \sigma_i\right). \tag{3.29}$$

A downside of the Gaussian-Bernoulli RBM compared to the Binary-Binary RBM is its higher instability. This is due to the absence of a bounding value for the reconstructions. It is also possible to have linear units with independent Gaussian noise in both visible and hidden layers, although it becomes even more unstable (Hinton, 2012). In any case, the energy and hidden layer sampling function become, respectively,

$$E_{\boldsymbol{\theta}}\left(\boldsymbol{v}, \boldsymbol{h}\right) = \sum_i \frac{\left(v_i - a_i\right)^2}{2\sigma_i^2} - \sum_j \frac{\left(h_j - b_j\right)^2}{2\sigma_j^2} - \sum_{ij} \frac{v_i}{\sigma_i} \frac{h_j}{\sigma_j} w_{ij} \tag{3.30}$$

$$p\left(h_j|\boldsymbol{v}\right) = \mathcal{N}\left(b_j + \sum_j w_{ij} v_i, \sigma_i\right). \tag{3.31}$$

Instead of linear units with independent Gaussian noise, ReLU may also be used as activation function for dealing with real-valued data, or encoding real-valued features. In this case, a variant called Noisy ReLU is employed (Nair and Hinton, 2010). These units were used in the hidden layer, with linear units with

Gaussian noise in the visible layer (Nair and Hinton, 2010; van Tulder and de Bruijne, 2015). However, Noisy ReLU may be used in both visible and hidden layers (Hinton, 2012). Considering we are using these units in the hidden layer, sampling is achieved through

$$p\left(h_j|\boldsymbol{v}\right) = \max\left(0, \sum_i w_{ij}v_i + b_j + \mathcal{N}\left(0, sigm\left(\sum_i w_{ij}v_i + b_j\right)\right)\right). \tag{3.32}$$

The previous equation is used during sampling, while for feature extraction, after learning, we may use its noise-free variant. The ReLU has an interesting property: intensity equivariance. This is true for noise-free units with zero bias ($b_j = 0$). In this setting, if the input values in the visible layer are all scaled by some positive value, then, all the ReLU units with 0 output would stay off, while the positive outputs would be scaled by the same positive value (Nair and Hinton, 2010).

3.4.2 Contrastive divergence for learning

Having defined the RBM model, it is still needed to go through a learning procedure. To that end, the negative log-likelihood of the training data must be minimized (Bengio et al., 2013). The gradient is necessary for SGD, being given by

$$\nabla_{\boldsymbol{\theta}} \log p\left(\boldsymbol{v}\right) = \nabla_{\boldsymbol{\theta}} \log \widetilde{p}\left(\boldsymbol{v}\right) - \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}}, \tag{3.33}$$

where the term on the left is called positive phase and the term on the right is the negative phase (Goodfellow et al., 2016). The problem with this gradient is that the partition function is dependent on the parameters of the RBM. So, it is intractable, too. However, after some derivation (Goodfellow et al., 2016), we can see that

$$\nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{v}\ p(\boldsymbol{x})} \nabla_{\boldsymbol{\theta}} \log \widetilde{p}\left(\boldsymbol{v}\right) \tag{3.34}$$

Although the negative phase is intractable, we can approximate it by running a Markov Chain Monte Carlo. Given that no intra-layer connections are allowed in RBMs, it is possible to employ block Gibbs sampling to sample a state from the model. After the chain converges, we can extract samples that approximately represent the distribution of the model. So, starting with some given state in moment 0 as $\boldsymbol{v}^0$, we sample $\boldsymbol{h}^0 \sim p\left(\boldsymbol{h}|\boldsymbol{v}^0\right)$, followed $\boldsymbol{v}^1 \sim p\left(\boldsymbol{v}|\boldsymbol{h}^0\right)$ and $\boldsymbol{h}^1 \sim p\left(\boldsymbol{h}|\boldsymbol{v}^0\right)$, and so on until the convergence of the chain (Bengio et al., 2013; Hinton, 2012; Bengio et al., 2009; Murphy, 2012).

An obvious disadvantage of running a block Gibbs Markov Chain Monte Carlo sampling is that we need to wait until the convergence of the chain, which is time demanding for training a RBM. Hence, RBMs are more commonly trained using Contrastive Divergence (CD) (Hinton, 2002). In this case, the chain is started from a training example and runs only for a very short k Gibbs steps. CD approximates the gradient in a very rough way, but the direction of the gradient is accurate. The largest the k, the less biased the approximation is, but $k = 1$ works well in practice (Hinton, 2002; Bengio et al., 2013; Hinton, 2012; Bengio et al., 2009; Tieleman, 2008). The algorithm for CD and update rules of RBMs is presented

Algorithm 4 Training procedure for RBM with CD.

Require: $\boldsymbol{v}$, a training sample
Require: $\boldsymbol{\theta}$, the parameters of the RBM
Require: ϵ, the learning rate
Require: k, the number of Gibbs steps
$\boldsymbol{v}^0 \leftarrow \boldsymbol{v}$
$\boldsymbol{h}^0 \leftarrow \boldsymbol{h} \sim p\left(\boldsymbol{h}|\boldsymbol{v}^0\right)$
for $s = 1, \cdots, k$ **do**
$\boldsymbol{v}^s \leftarrow \boldsymbol{v} \sim p\left(\boldsymbol{v}|\boldsymbol{h}^{s-1}\right)$
$\boldsymbol{h}^s \leftarrow \boldsymbol{h} \sim p\left(\boldsymbol{h}|\boldsymbol{v}^s\right)$
end for
$\boldsymbol{W} \leftarrow \boldsymbol{W} + \epsilon\left(\boldsymbol{h}^0\boldsymbol{x}^{0^T} - \boldsymbol{h}^k\boldsymbol{x}^{k^T}\right)$
$\boldsymbol{a} \leftarrow \boldsymbol{a} + \epsilon\left(\boldsymbol{h}^0 - \boldsymbol{h}^k\right)$
$\boldsymbol{b} \leftarrow \boldsymbol{b} + \epsilon\left(\boldsymbol{v}^0 - \boldsymbol{v}^k\right)$

in Algorithm 4. For simplicity reasons, it is shown for one training example, but it can be employed for mini-batch settings, where the gradients are averaged. Also, RBMs can be trained with momentum, and weight decay.

3.5 Random Forest

Random Forests (Breiman, 2001) are ensembles of several Decision Trees (DTs), in which individual trees are decorrelated by means of randomization at training time. These algorithms can be used for several tasks, but, in this section, the classification task will be focused. Particularly, RF classifiers showed good performance in medical image segmentation (Zikic et al., 2012; Pinto et al., 2015a,b; Tustison et al., 2015; Pereira et al., 2016b).

3.5.1 Decision Trees

A DT (Breiman et al., 1984) is an algorithm that follows a "divide and conquer" strategy. Its structure is similar to a directed graph, starting in a *root* node, and followed by subsequent split nodes organized hierarchically. The root is a special split node, since it is the first. It will be considered only split nodes with binary decisions, where each one is connected to only two child nodes. Keeping the analogy with trees, terminal nodes are designated as *leaf* nodes, where the prediction is inferred (Criminisi and Shotton, 2013). A structure of a DT can be observed in Fig. 3.8.

3.5.1.1 Testing

Suppose there is a given test sample with m features $\boldsymbol{x} = \left[x_j : j = 1, \cdots, m\right]$. This sample is placed in the root node and traverses the split nodes until reaching a leaf. In each split node, a binary splitting function decides if the sample continues to the next left or right child node. The splitting function at the node i can be generally defined as

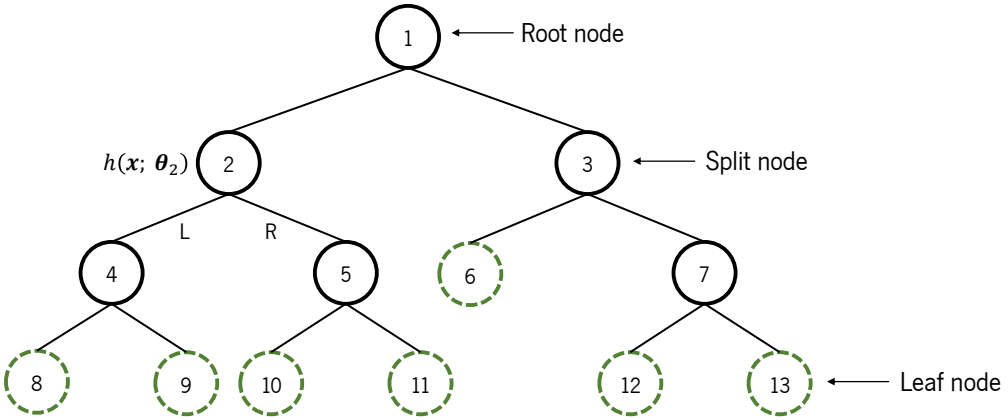


Figure 3.8: Decision Tree. Black nodes represent split nodes, with the special case of the root, and green dashed line nodes represent leaf nodes. On each split node, a split function h decides to which child (left (L) or right (R)) the data will follow.

$$h_i\left(\boldsymbol{x}, \boldsymbol{\theta}_i\right) \rightarrow \{0, 1\}, \tag{3.35}$$

where $\boldsymbol{\theta}_i$ are the parameters of the splitting function of node i, and $\{0, 1\}$ represents the binary decision of moving to the left or right child node. The splitting function is often called weak learner, as it is barely better than deciding by chance (Criminisi and Shotton, 2013).

Three main parameters govern the splitting function model: the selector function ω, the data separating function ξ, and the thresholds of the binary test τ. So, $\boldsymbol{\theta}_i = \{\omega_i, \xi_i, \tau_i\}$. The selector function ω selects which feature will be evaluated at node i. The data separation function ξ defines which function will be used at node i. The most common function is the linear axis aligned hyperplane. Finally, the threshold of the binary test τ is the decision-making value (Criminisi and Shotton, 2013). For example, considering $\omega_i\left(\boldsymbol{x}\right) = x_j$, the sample will proceed to the left child if $x_j < \tau_i$, otherwise it should continue to the right child node.

When the sample finishes its path on a leaf, a prediction is made in the form of a probability $p\left(y|\boldsymbol{x}\right)$, which can be further followed by a maximum a posteriori estimate, similarly to equation 3.1. Note that DTs can naturally handle multi-class classification problems (Criminisi and Shotton, 2013).

3.5.1.2 Training

Of course, before testing, the DT needs to be trained, which involves optimizing the parameters of the nodes, and consequently the structure of the tree. Searching for the optimal parameters and structure is infeasible. Hence, most training algorithms rely on greedy approaches by ensuring only local optimality at the node level (Criminisi and Shotton, 2013; Murphy, 2012).

Let us consider the training set $\mathcal{D} = \{(\boldsymbol{x}^u, y^u)\}_{u=1}^n$. The root node receives the full set. Then, for each feature, we optimize the threshold value that better separates the samples, according to some criteria. Finally, taking the best feature at the node, the samples of the training set are separated into two disjoint sets that will be fed to the left and right child nodes, respectively. The process repeats on the following nodes, with smaller and smaller subsets of the training set. Hence, generally speaking, node i

has access only to the subset $\mathcal{D}_i$. It then forwards the subsets $\mathcal{D}_i^L$ and $\mathcal{D}_i^R$ to the left and right nodes, respectively, in such a way that $\mathcal{D}_i = \mathcal{D}_i^L \cup \mathcal{D}_i^R$ and $\mathcal{D}_i^L \cap \mathcal{D}_i^R = \emptyset$ (Criminisi and Shotton, 2013; Murphy, 2012).

So, in each node i, we are interested in optimizing the parameters in a way that better separates the data. Thus, having an objective function I, it is defined as

$$\boldsymbol{\theta}_i = \underset{\boldsymbol{\theta}}{\text{argmax}}\, I\left(\mathcal{D}_i, \boldsymbol{\theta}\right). \tag{3.36}$$

A possible criterion for splitting the data is the degree of purity achieved in each of the children nodes. In a classification setting, children nodes are purer than the parent if they have a lower degree of mixing of classes in the samples. There are several possible criteria. However, we will consider the Information Gain. Purer child nodes lead to a gain in information, i.e., a lower entropy. The Information Gain in node i, when the data is split using feature j with threshold τ_i is defined as

$$I_j^{\tau_i}\left(\mathcal{D}_i\right) = H\left(\mathcal{D}_i\right) - \sum_{v \in \{L,R\}} \frac{|\mathcal{D}_i^v|}{|\mathcal{D}_i|} H\left(\mathcal{D}_i^v\right), \tag{3.37}$$

where $|\cdot|$ represents the cardinality of a set and H is the entropy (Criminisi and Shotton, 2013; Murphy, 2012). In classification, the frequencies of the classes may describe a discrete probabilistic distribution. So, if there are k classes in the samples at some node, the entropy can be measured as the Shannon entropy as

$$H\left(\mathcal{D}_i^v\right) = -\sum_{c=1}^{k} p\left(c\right) \log\left(p\left(c\right)\right), \tag{3.38}$$

where $p\left(c\right)$ is simply the empirical distribution, computed from the labels at the nodes as the normalized histogram of classes (Criminisi and Shotton, 2013; Murphy, 2012).

When the samples in a node belong to the same class it is a pure node, and it becomes a leaf node. Other stopping criteria exist, such as the number of samples in the node, or the depth of the tree. The samples in a leaf are stored and used as an empirical distribution for predicting $p\left(y|\boldsymbol{x}\right)$ (Criminisi and Shotton, 2013; Murphy, 2012).

3.5.1.3 Practical issues

As described, several separation functions ξ are possible. Still, we are assuming axis-aligned linear functions. Often, more complex functions lead to overfitting. Also, DTs usually evaluate the full features vector in each node for finding the optimal splitting, during training. Regarding the tree's depth, allowing a tree to grow completely often leads to overfitting. Hence, some stopping criteria must be imposed, or, alternatively, a pruning procedure should be employed (Criminisi and Shotton, 2013; Murphy, 2012).

Decision Trees have some advantages, such as naturally handling multi-class problems, being interpretable, or being efficient. Still, DTs have high variance, which makes them highly unstable for slight changes in the data, leading to poor predictive performance (Criminisi and Shotton, 2013; Murphy, 2012).

3.5.2 From Decision Trees to Random Forests

A way to decrease variance is by model ensembling. However, a naïve ensemble of DTs trained with the same training set would be wasteful, as all models would be similar and extremely correlated. Random Forests address this issue by building an ensemble of decorrelated DTs. Such decorrelation is achieved during training by injecting randomness (Criminisi and Shotton, 2013; Murphy, 2012).

Bagging consists in training models with randomly selected subsets of the training data with replacement $\mathcal{D}$‘, such that $|\mathcal{D}| = |\mathcal{D}\text{‘}|$. Therefore, each model observes a different training set, where some samples are repeated, while others are absent, resulting in less correlated trees (Criminisi and Shotton, 2013; Murphy, 2012).

Further randomness is achieved during optimization of each node. DTs optimize the separating threshold by considering the best feature among all. Contrasting, in RFs it is uniformly chosen and evaluated a subset of features $\boldsymbol{x}$‘, such that $|\boldsymbol{x}\text{‘}| \leq |\boldsymbol{x}|$, in each node of each tree. Of course, the higher the number of selected features in a node, the more correlated the trees will be (Criminisi and Shotton, 2013; Murphy, 2012).

So, by means of bagging and random node optimization, we achieve decorrelated trees, such that each tree of the forest is unique. In this way, RFs are much more robust than single DTs, which results from a significant decrease in variance (Criminisi and Shotton, 2013; Murphy, 2012).

Considering a RF resulting from an ensemble of T trees, the final prediction can be obtained by averaging the probabilistic predictions $p_t\left(c|\boldsymbol{x}\right)$ of each tree, as

$$p\left(c|\boldsymbol{x}\right) = \frac{1}{T}\sum_{t=1}^{T} p_t\left(c|\boldsymbol{x}\right). \tag{3.39}$$

RFs are considered a strong classifier, whereas DTs are weak. In fact, the performance of a RF always improves with the addition of more trees, until the computational cost overcomes the benefits in accuracy. Still, with better performance of RFs comes decreased interpretability of the model (Criminisi and Shotton, 2013; Murphy, 2012).

3.5.3 Feature importance

A fortunate property of DTs and RFs is their capability to inexpensively estimate feature importances, i.e., how important is a given feature for the task. This estimate is known as Mean Decrease Impurity (MDI), and measures how much a feature contributes to decrease the impurity in a tree. The importance of a feature in a tree is simply the sum of the Information Gain over the nodes where it was used for splitting the data. Hence, the feature importance R_j of feature j in a DT is given as (Hastie et al., 2009)

$$R_j = \sum_i \mathbb{1}\left(\omega_i\left(\boldsymbol{x}\right) = x_j\right) I_j^{\mathcal{T}_i}\left(\mathcal{D}_i\right). \tag{3.40}$$

In the case of a RF, the feature importance is the average across all trees, such that (Hastie et al., 2009)

$$R_j = \frac{1}{T} \sum_{t=1}^{T} R_j^t. \tag{3.41}$$

The feature importance measured by RFs is biased. However, it is still recommended and useful, since getting an unbiased measure is quite impractical (Louppe et al., 2013).

3.5.3.1 Interpreting feature importance as Mutual Information

The MDI computed using the Information Gain as splitting criteria in the nodes is equivalent to measuring the Mutual Information between the decision in the node (left or right) and the class of the nodes (Nowozin, 2012).

3.6 The need for interpretability

In 2016, the European Union defined the General Data Protection Regulation that will take effect in 2018. This affects the Machine Learning field because it defines that individuals have the right to non-discrimination by algorithmic decisions, as well as the right to explainability (Goodman and Flaxman, 2016). Interpretability is mandatory for both aspects. On one hand, biases while acquiring data may result in models learning "discrimination" patterns. Hence, interpretability may help in inspecting the models and, ultimately, finding problems in the data. Explanation of predictions is attainable by its interpretability. Although the right to an explanation may have very specific contexts, it raises attention to the need for having explaining approaches to Machine Learning systems (Doshi-Velez et al., 2017; Wachter et al., 2017).

The recent developments in Machine Learning, and particularly in Deep Learning, show that remarkable accuracies are being achieved. Therefore, there is no doubt about the capabilities of these methods, and the potential of its economical and society impact. A common concern is related to the unprecedented degree of automation that it will enable. Such automation may lead to progress, but raises the question on how to ensure safeness (Russell et al., 2015). Interpretability of Machine Learning systems may be a way to attain some degree of safeness, by inspecting if a model learned patterns that are coherent with domain knowledge, and if its predictions are based on the right factors (Guidotti et al., 2018).

Interpretability of Machine Learning-based systems is especially important as these systems are pervasively being used in the critical medical domain (Lipton, 2016; Wang and Summers, 2012). However, there is, usually, an inverse relationship between the capacity of a model, and its interpretability (Selvaraju et al., 2017). It is certainly easier to interpret a linear model or a decision tree, than a deep neural network or (large) ensembles. Such complex models are often regarded as "black box", since it is quite difficult to understand what they have learned, and why a prediction is done. In this scenario, predictions cannot be blindly followed, as it may impact the life of patients. So, physicians may be skeptical about decision support with such Machine Learning models (Kononenko, 2001; Wang and Summers, 2012), unless the systems achieve a remarkably better performance at some task than the human experts themselves (Kononenko, 2001). There is need, thus, to build trust upon a system. Interpretability may help in

increasing such trust, by allowing one to inspect the model or its predictions.

According to Selvaraju et al. (2017), there are three main situations where interpretability is desirable and useful. 1) During development cycle and when the Machine Learning-based system performs worse than humans. In this scenario, interpretability may help in identifying when and why a model fails or succeeds, and when there are problems, e.g. biases, in the training data. In this way, it may help in devising strategies for improving the system. 2) When the Machine Learning-based system is on pair with human performance, or, it is at least ready for deployment. At this moment, interpretability is very important to increase trust in the system, by showing users what the model learned, and/or how/why a prediction was made. 3) Interpretability may help humans to learn from the model itself. For instance, by allowing the inspection of the relationships among features. This scenario is especially useful when the machine performs better than human experts.

We can broadly identify two types of interpretability: global, and local. When the model is directly interpretable, it is global interpretability. In this case, it is possible to inspect the model itself, and check how it learned. This is usually the case of simpler and smaller models, such as Decision Trees, and Linear Models. Some approaches may help in simplifying larger models by making their weights sparser, such as the case of LASSO (Tibshirani, 1996). Many times it may be useful to explain a given prediction. In this case, it is local interpretability because it explains the model only in the sub-region of the instance being predicted. While global interpretability provides a broad, yet fixed, interpretation of the model, local interpretability is more dynamic, since different samples may yield different explanations. So, approaches dealing with local interpretability may be regarded as post-hoc, since they treat the Machine Learning system as "black boxes", being concerned only with explaining predictions. In this sense, these approaches tend to be more general, as they may not be linked with a particular model (Ribeiro et al., 2016a).

Taking the above-mentioned aspects into consideration, we can think that in the medical image analysis domain interpretability is mostly important during the development of Machine Learning systems, and in increasing the trust in those systems at deployment time, especially among medical staff. One may point local and post-hoc interpretability as more important for increasing trust in the medical staff after deployment. However, global interpretability has also an important role, since it is useful before deployment for checking how the model learned.

3.7 Summary

Machine Learning methods learn to perform a given task from the data. Among these methods, Representation Learning approaches learn features for the task at hand directly from the original data, whereas the more conventional methods require a feature engineering step. Learning can be achieved through supervised or unsupervised learning.

In Deep Learning methods, current successful approaches are trained by gradient-based learning. Convolutional Neural Networks are supervised methods with recent remarkable results. The weights of the layers are organized as filters and convolved over feature maps to generate new features. Therefore, it achieves parameter efficiency, and is translational invariant. Restricted Boltzmann Machines are gen-

erative unsupervised algorithms. So, they learn the distribution of the data. These properties may make them suited for large non-annotated medical data.

Among the more conventional approaches requiring handcrafted features, Random Forests often achieve good performance. These classifiers result from the ensembling of weaker Decisions Trees trained with randomly selected samples of the training set and features in the nodes. A nice property of Random Forests is their capability to estimate features importance.

Finally, the recent remarkable performances achieved with Machine Learning often come at the cost of more complex models. This is especially critical in domains where a prediction may have severe impacts, such as the medical domain. Hence, interpretability of Machine Learning methods is necessary for increasing trust. Besides explaining predictions, it may also expedite the development of new methods, by identifying problems and their causes.

Chapter 4

Brain Tumor Segmentation using Convolutional Neural Networks

In this chapter, we investigate brain tumor segmentation in MRI images using CNN-based approaches. We start by developing a method that classifies the central voxel of a patch into some tumor class, in Section 4.1. We denote this approach as Classification CNN. Although the good results achieved, in Section 4.2 we evolved our method following two main ideas: hierarchical brain tumor segmentation, and Fully Convolutional Networks (FCN) for dense prediction. Finally, in Section 4.3, we propose a novel adaptive feature recombination and recalibration method in the context of FCNs.

Some of the sub-sections in this chapter are based on the following publications:

- Section 4.1
  - Pereira, Sérgio, et al. "Deep convolutional neural networks for the segmentation of gliomas in multi-sequence MRI." International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, 2015.
  - Pereira, Sérgio, et al. "Brain tumor segmentation using convolutional neural networks in MRI images." IEEE transactions on medical imaging 35.5 (2016): 1240-1251.
- Section 4.2
  - Pereira, Sérgio, et al. "On hierarchical brain tumor segmentation in MRI using fully convolutional neural networks: A preliminary study." Bioengineering (ENBENG), 2017 IEEE 5th Portuguese Meeting on. IEEE, 2017.
- Section 4.3
  - Pereira, Sérgio, et al. "Adaptive feature recombination and recalibration for semantic segmentation: application to brain tumor segmentation in MRI." Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2018.

Contribution The author of this thesis was responsible for conceiving and implementing the ideas and methods described in this chapter. Furthermore, he also conducted the experiments, and analyzed the resulting data, validating the methods. Finally, the author was the writer and presenter of the manuscripts listed above.

Contents

4.1 Classification Convolutional Neural Networks for Semantic Segmentation

In this section, we developed and studied Classification CNNs for brain tumor segmentation in MRI. Although CNNs were adopted and achieved remarkable results in object recognition, there were limited studies on its use for the task of brain tumor segmentation. We employed an approach that classifies the central voxel of an image patch. Further contributions are: 1) the study of small 3×3 kernels that allow deeper networks with less parameters, 2) the research of intensity normalization as pre-processing in the context of CNN, and 3) the study of data augmentation strategies for brain tumor segmentation.

This section is based on the following publications:

- Pereira, Sérgio, et al. "Deep convolutional neural networks for the segmentation of gliomas in multi-sequence MRI." International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, Cham, 2015.

- Pereira, Sérgio, et al. "Brain tumor segmentation using convolutional neural networks in MRI images." IEEE transactions on medical imaging 35.5 (2016): 1240-1251.

The methods explored here were used for participating in BRATS 2015 challenge, ranking 2nd among all participants.

4.1.1 Introduction

Gliomas are the brain tumors with the highest mortality rate and prevalence (Bauer et al., 2013). These neoplasms can be graded into LGG and HGG, with the former being less aggressive and infiltrative than the latter (Louis et al., 2007; Bauer et al., 2013). Even under treatment, patients do not survive on average more than 14 months after diagnosis (Van Meir et al., 2010). Current treatments include surgery, chemotherapy, radiotherapy, or a combination of them (Tabatabai et al., 2010). MRI is especially useful to assess gliomas in clinical practice, since it is possible to acquire MRI sequences providing complementary information (Bauer et al., 2013).

The accurate segmentation of gliomas and its intra-tumoral structures is important not only for treatment planning, but also for follow-up evaluations. However, manual segmentation is time-consuming and subjected to inter- and intra-rater errors difficult to characterize. Thus, physicians usually use rough measures for evaluation (Bauer et al., 2013). For these reasons, accurate semi-automatic or automatic methods are required (Menze et al., 2015; Bauer et al., 2013). However, it is a challenging task, since the shape, structure, and location of these abnormalities are highly variable. Additionally, the tumor mass effect change the arrangement of the surrounding normal tissues (Menze et al., 2015). Also, MRI images may present some problems, such as intensity inhomogeneity (Tustison et al., 2010), or different intensity ranges among the same sequences and acquisition scanners (Nyúl et al., 2000).

In brain tumor segmentation, we find several methods that explicitly develop a parametric or non-parametric probabilistic model for the underlying data. These models usually include a likelihood function corresponding to the observations and a prior model. Being abnormalities, tumors can be segmented as outliers of normal tissue, subjected to shape and connectivity constraints (Prastawa et al., 2004). Other approaches rely on probabilistic atlases (Menze et al., 2010; Gooya et al., 2012; Kwon et al., 2014b). In the case of brain tumors, the atlas must be estimated at segmentation time, because of the variable shape and location of the neoplasms (Menze et al., 2010; Gooya et al., 2012; Kwon et al., 2014b). Tumor growth models can be used as estimates of its mass effect, being useful to improve the atlases (Gooya et al., 2012; Kwon et al., 2014b). The neighborhood of the voxels provides useful information for achieving smoother segmentations through Markov Random Fields (MRF) (Menze et al., 2010). Zhao at al. (Menze et al., 2015) also used a MRF to segment brain tumors after a first over-segmentation of the image into supervoxels, with a histogram-based estimation of the likelihood function. As observed by Menze et al.

(2015), generative models generalize well in unseen data, but it may be difficult to explicitly translate prior knowledge into an appropriate probabilistic model.

Another class of methods learns a distribution directly from the data. Although a training stage can be a disadvantage, these methods can learn brain tumor patterns that do not follow a specific model. This kind of approaches commonly consider voxels as independent and identically distributed (Bauer et al., 2011), although context information may be introduced through the features. Because of this, some isolated voxels or small clusters may be mistakenly classified with the wrong class, sometimes in physiological and anatomically unlikely locations. To overcome this problem, some authors include information of the neighborhood by embedding the probabilistic predictions of the classifier into a Conditional Random Field (Lee et al., 2008; Bauer et al., 2011; Meier et al., 2013, 2014a). Classifiers such as Support Vector Machines (Lee et al., 2008; Bauer et al., 2011) and, more recently, Random Forests (RF) (Zikic et al., 2012; Bauer et al., 2012; Meier et al., 2013, 2014a; Reza and Iftekharuddin, 2014; Tustison et al., 2015; Geremia et al., 2013; Pinto et al., 2015a) were successfully applied in brain tumor segmentation. The RF became very used due to its natural capability in handling multi-class problems and large feature vectors. A variety of features were proposed in the literature: encoding context (Zikic et al., 2012; Meier et al., 2014a; Pinto et al., 2015a), first-order and fractals-based texture (Meier et al., 2013, 2014a; Islam et al., 2013; Reza and Iftekharuddin, 2014; Pinto et al., 2015a), gradients (Meier et al., 2014a, 2013), brain symmetry (Meier et al., 2013, 2014a; Tustison et al., 2015), and physical properties (Tustison et al., 2015). Using supervised classifiers, some authors developed other ways of applying them. Tustison et al. (2015) developed a two-stage segmentation framework based on RFs, using the output of the first classifier to improve a second stage of segmentation. Geremia et al. (2013) proposed a Spatially Adaptive RF for hierarchical segmentation, going from coarser to finer scales. Meier et al. (2014c) used a semi-supervised RF to train a subject-specific classifier for post-operative brain tumor segmentation.

Other methods known as Deep Learning deal with representation learning by automatically learning a hierarchy of increasingly complex features directly from data (Bengio et al., 2013). So, the focus is on designing architectures instead of developing hand-crafted features, which may require specialized knowledge (LeCun et al., 2015). CNNs have been used to win several object recognition (Krizhevsky et al., 2012; Dieleman et al., 2015b) and biological image segmentation (Ciresan et al., 2012) challenges. Since a CNN operates over patches using kernels, it has the advantages of taking context into account and being used with raw data. In the field of brain tumor segmentation, recent proposals also investigate the use of CNNs (Zikic et al., 2014; Urban et al., 2014; Davy et al., 2014; Havaei et al., 2017; Lyksborg et al., 2015; Rao et al., 2015; Dvorák and Menze, 2015). Zikic et al. (2014) used a shallow CNN with two convolutional layers separated by max-pooling with stride 3, followed by one fully-connected (FC) layer and a softmax layer. Urban et al. (2014) evaluated the use of 3D filters, although the majority of authors opted for 2D filters (Davy et al., 2014; Havaei et al., 2017; Lyksborg et al., 2015; Rao et al., 2015; Dvorák and Menze, 2015). 3D filters can take advantage of the 3D nature of the images, but it increases the computational load. Some proposals evaluated two-pathway networks to allow one of the branches to receive bigger patches than the other, thus having a larger context view over the image (Davy et al., 2014; Havaei et al., 2017). In addition to their two-pathway network, Havaei et al. (2017) built a cascade of two networks and performed a two-stage training, by training with balanced classes and then refining it with proportions

near the originals. Lyksborg et al. (2015) use a binary CNN to identify the complete tumor. Then, a cellular automata smooths the segmentation, before a multi-class CNN discriminates the sub-regions of tumor. Rao et al. (2015) extracted patches in each plane of each voxel and trained a CNN in each MRI sequence; the outputs of the last FC layer with softmax of each CNN are concatenated and used to train a RF classifier. Dvorák and Menze (2015) divided the brain tumor regions segmentation tasks into binary sub-tasks and proposed structured predictions using a CNN as learning method. Patches of labels are clustered into a dictionary of label patches, and the CNN must predict the membership of the input to each of the clusters.

In this paper, inspired by the groundbreaking work of Simonyan and Zisserman (2014) on deep CNNs, we investigate the potential of using deep architectures with small convolutional kernels for segmentation of gliomas in MRI images. Simonyan and Zisserman proposed the use of small 3×3 kernels to obtain deeper CNNs. With smaller kernels we can stack more convolutional layers, while having the same receptive field of bigger kernels. For instance, two 3×3 cascaded convolutional layers have the same effective receptive field of one 5×5 layer, but fewer weights (Simonyan and Zisserman, 2014). At the same time, it has the advantages of applying more non-linearities and being less prone to overfitting because small kernels have fewer weights than bigger kernels (Simonyan and Zisserman, 2014). We also investigate the use of the intensity normalization method proposed by Nyúl et al. (2000) as a pre-processing step that aims to address data heterogeneity caused by multi-site multi-scanner acquisitions of MRI images. The large spatial and structural variability in brain tumors are also an important concern that we study using two kinds of data augmentation.

The remainder of this paper is organized as follows. In Section 4.1.2, the proposed method is presented. The databases used for evaluation and the experimental setup are detailed in Section 4.1.3. Results are presented and discussed in Section 4.1.4. Finally, the main conclusions are presented in Section 4.1.5.

4.1.2 Method

Fig. 4.1 presents an overview of the proposed approach. There are three main stages: pre-processing, classification via CNN, and post-processing.

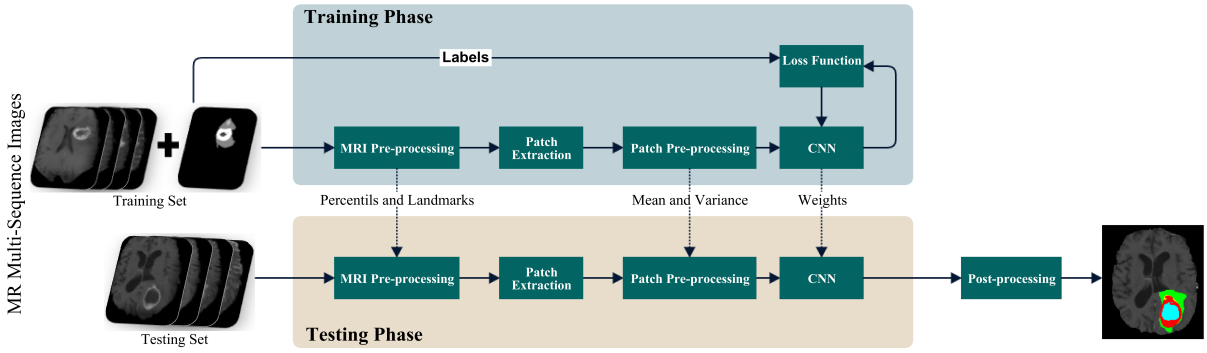


Figure 4.1: Overview of the proposed method.

4.1.2.1 Pre-processing

MRI images are altered by the bias field distortion. This makes the intensity of the same tissues to vary across the image. To correct it, we applied the N4ITK method (Tustison et al., 2010). However, this is not enough to ensure that the intensity distribution of a tissue type is in a similar intensity scale across different subjects for the same MRI sequence, which is an explicit or implicit assumption in most segmentation methods (Shah et al., 2011). In fact, it can vary even if the image of the same patient is acquired in the same scanner in different time points, or in the presence of a pathology (Nyúl et al., 2000; Nyúl and Udupa, 1999). So, to make the contrast and intensity ranges more similar across patients and acquisitions, we apply the intensity normalization method proposed by Nyúl et al. (2000) on each sequence. In this intensity normalization method, a set of intensity landmarks $I_L = \{pc_1, i_{p_{10}}, i_{p_{20}}, \cdots, i_{p_{90}}, pc_2\}$ are learned for each sequence from the training set. pc_1 and pc_2 are chosen for each MRI sequence as described by Nyúl and Udupa (1999). i_{p_l} represents the intensity at the l^{th} percentile. After training, the intensity normalization is accomplished by linearly transforming the original intensities between two landmarks into the corresponding learned landmarks. In this way, the histogram of each sequence is more similar across subjects.

After normalizing the MRI images, we compute the mean intensity value and standard deviation across all training patches extracted for each sequence. Then, we normalize the patches on each sequence to have zero mean and unit variance[1].

4.1.2.2 Convolutional Neural Network

CNNs were used to achieve some breakthrough results and win well-known contests (Krizhevsky et al., 2012; Dieleman et al., 2015b). The application of convolutional layers (LeCun et al., 1989, 1998) consists in convolving a signal or an image with kernels to obtain feature maps. So, a unit in a feature map is connected to the previous layer through the weights of the kernels. The weights of the kernels are adapted during the training phase by backpropagation, in order to enhance certain characteristics of the input. Since the kernels are shared among all units of the same feature maps, convolutional layers have fewer weights to train than dense FC layers, making CNN easier to train and less prone to overfitting. Moreover, since the same kernel is convolved over all the image, the same feature is detected independently of the location – translation invariance. By using kernels, information of the neighborhood is taken into account, which is an useful source of context information (LeCun et al., 1989, 1998, 2015). Usually, a non-linear activation function is applied on the output of each neural unit.

If we stack several convolutional layers, the extracted features become more abstract with the increasing depth. The first layers enhance features such as edges, which are aggregated in the following layers as motifs, parts, or objects (LeCun et al., 2015).

The following concepts are important in the context of CNN:

Initialization it is important to achieve convergence. We use the Xavier initialization (Glorot and Bengio, 2010). With this, the activations and the gradients are maintained in controlled levels, otherwise

[1]The mean and standard deviation computed in the training patches are used to normalize the testing patches.

back-propagated gradients could vanish or explode.

Activation Function it is responsible for non-linearly transforming the data. Rectifier linear units (ReLU), defined as

$$f(x) = max(0, x), \tag{4.1}$$

were found to achieve better results than the more classical sigmoid, or hyperbolic tangent functions, and speed up training (Jarrett et al., 2009; Krizhevsky et al., 2012). However, imposing a constant 0 can impair the gradient flowing and consequent adjustment of the weights (Maas et al., 2013). We cope with these limitations using a variant called leaky rectifier linear unit (LReLU) (Maas et al., 2013) that introduces a small slope on the negative part of the function. This function is defined as

$$f(x) = \max(0, x) + \alpha \min(0, x) \tag{4.2}$$

where α is the leakiness parameter. In the last FC layer, we use softmax.

Pooling it combines spatially nearby features in the feature maps. This combination of possibly redundant features makes the representation more compact and invariant to small image changes, such as insignificant details; it also decreases the computational load of the next stages. To join features it is more common to use max-pooling or average-pooling (LeCun et al., 2015).

Regularization it is used to reduce overfitting. We use Dropout (Srivastava et al., 2014; Hinton et al., 2012) in the FC layers. In each training step, it removes nodes from the network with probability p. In this way, it forces all nodes of the FC layers to learn better representations of the data, preventing nodes from co-adapting to each other. At test time, all nodes are used. Dropout can be seen as an ensemble of different networks and a form of bagging, since each network is trained with a portion of the training data (Srivastava et al., 2014; Hinton et al., 2012).

Data Augmentation it can be used to increase the size of training sets and reduce overfitting (Krizhevsky et al., 2012). Since the class of the patch is obtained by the central voxel, we restricted the data augmentation to rotating operations. Some authors also consider image translations (Krizhevsky et al., 2012), but for segmentation this could result in attributing a wrong class to the patch. So, we increased our data set during training by generating new patches through the rotation of the original patch. In our proposal, we used angles multiple of 90°, although another alternative will be evaluated.

Loss Function it is the function to be minimized during training. We used the Categorical Cross-entropy,

$$H = -\sum_{j \in voxels} \sum_{k \in classes} y_{j,k} \log(\hat{y}_{j,k}) \tag{4.3}$$

where $\hat{y}$ represents the probabilistic predictions (after the softmax) and y is the target.

In the next subsections, we discuss the architecture and training of our CNN.

Architecture We aim at a reliable segmentation method; however, brain tumors present large variability in intra-tumoral structures, which makes the segmentation a challenging problem. To reduce such complexity, we designed a CNN and tuned the intensity normalization transformation for each tumor grade – LGG and HGG.

The proposed architectures are presented in Tables 4.1 and 4.2, and graphically in Fig. 4.2. The architecture used for HGG is deeper than the one for LGG, because going deeper did not improve results in the latter. To go deeper, one must include more layers with weights, which may increase overfitting, given the smaller training set of LGG. This is supported by the need of setting Dropout with $p = 0.5$ in LGG, while it is $p = 0.1$ in HGG, since the database used for evaluation contained more HGG then LGG cases. Additionally, the appearance and patterns are different in HGG and LGG. Since we are doing segmentation, we need a precise sense of location. Pooling can be positive to achieve invariance and to eliminate irrelevant details, however, it can also have a negative effect by eliminating important details. We apply overlapping pooling with 3×3 receptive fields and 2×2 stride to keep more information of location. In the convolutional layers the feature maps are padded before convolution, so that the resulting feature maps could maintain the same dimensions. In the case of HGG there are 2,118,213 weights to train, while in LGG it lowers to 1,933,701 weights because it has two less convolutional layers. All sequences were used as input. LReLU is the activation function in all layers with weights, with the exception of the last that uses softmax. Dropout was used only in the FC layers.

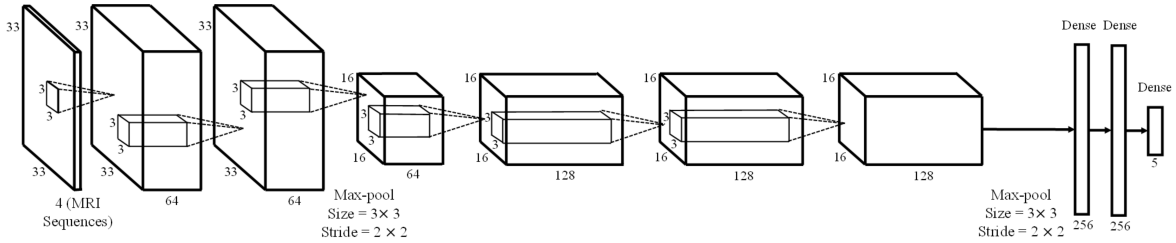
Training To train the CNN the loss function must be minimized, but it is highly non-linear. We use Stochastic Gradient Descent as an optimization algorithm, which takes steps proportionally to the negative of the gradient in the direction of local minima. Nevertheless, in regions of low curvature it can be slow. So, we also use Nesterov's Accelerated Gradient to accelerate the algorithm in those regions. The momentum ν is kept constant, while the learning rate ϵ was linearly decreased, after each epoch. We consider an

Table 4.1: Architecture of the HGG CNN. In inputs, the first dimension refers to the number of channels and the next two to the size of the patch, or feature maps. Conv. refers to convolutional layers and Max-pool. to max-pooling.

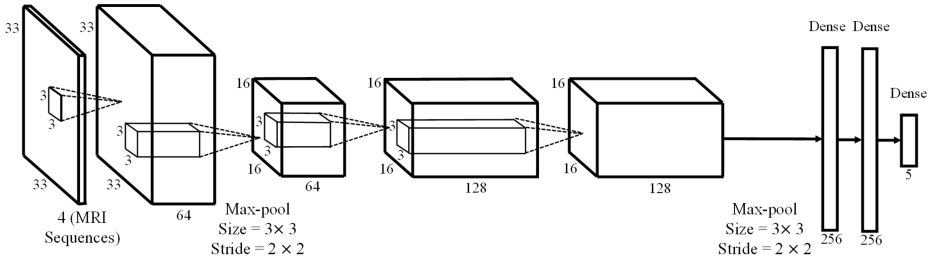
			HGG			
	Type	**Filter size**	**Stride**	**# filters**	**FC units**	**Input**
Layer 1	Conv.	3×3	1×1	64	-	4×33×33
Layer 2	Conv.	3×3	1×1	64	-	64×33×33
Layer 3	Conv.	3×3	1×1	64	-	64×33×33
Layer 4	Max-pool.	3×3	2×2	-	-	64×33×33
Layer 5	Conv.	3×3	1×1	128	-	64×16×16
Layer 6	Conv.	3×3	1×1	128	-	128×16×16
Layer 7	Conv.	3×3	1×1	128	-	128×16×16
Layer 8	Max-pool.	3×3	2×2	-	-	128×16×16
Layer 9	FC	-	-	-	256	6272
Layer 10	FC	-	-	-	256	256
Layer 11	FC	-	-	-	5	256

Table 4.2: Architecture of the LGG CNN. In inputs, the first dimension refers to the number of channels and the next two to the size of the patch, or feature maps. Conv. refers to convolutional layers and Max-pool. to max-pooling.

LGG						
	Type	**Filter size**	**Stride**	**# filters**	**FC units**	**Input**
Layer 1	Conv.	3×3	1×1	64	-	4×33×33
Layer 2	Conv.	3×3	1×1	64	-	64×33×33
Layer 3	Max-pool.	3×3	2×2	-	-	64×33×33
Layer 4	Conv.	3×3	1×1	128	-	64×16×16
Layer 5	Conv.	3×3	1×1	128	-	128×16×16
Layer 6	Max-pool.	3×3	2×2	-	-	128×16×16
Layer 7	FC	-	-	-	256	6272
Layer 8	FC	-	-	-	256	256
Layer 9	FC	-	-	-	5	256



(a)



(b)

Figure 4.2: Graphical architectures of the CNN for a) HGG and b) LGG.

epoch as a complete pass over all the training samples.

4.1.2.3 Post-processing

Some small clusters may be erroneously classified as tumor. To deal with that, we impose volumetric constraints by removing clusters in the segmentation obtained by the CNN that are smaller than a predefined threshold τ_{VOL}.

4.1.3 Experimental Setup

4.1.3.1 Database

The proposed method was validated on the BRATS 2013 and 2015 databases[2] (Menze et al., 2015; Kistler et al., 2013). For every patient in BRATS there are four MRI sequences available: T1-weighted (T1), T1 with gadolinium enhancing contrast (T1c), T2-weighted (T2) and FLAIR. The images of each subject were already aligned with the T1c and skull stripped. BRATS 2013 contains three data sets: Training, Leaderboard, and Challenge, comprising 65 MR scans from different patients — histological diagnosis: astrocytomas or oligoastrocytomas, LGG, and anaplastic astrocytomas and glioblastoma multiforme tumors, HGG. The Training set contains 20 HGG and 10 LGG, with manual segmentations available. The Leaderboard set is composed by 21 HGG and 4 LGG, while the Challenge set includes 10 HGG. Metrics for these two sets are computed through the online evaluation platform (VirtualSkeleton, 2013), given that the manual segmentations are not publicly available. In BRATS 2015, the Training set comprises 220 and 54 acquisitions of HGG and LGG, respectively. The Challenge set contains 53 cases, including both grades. In this case, the evaluation metrics were computed by the organizers of the challenge. The manual segmentation identifies four types of intra-tumoral classes: necrosis, edema, non-enhancing, and enhancing tumor. However, the evaluation is performed for the enhancing tumor, the core (necrosis + non-enhancing tumor + enhancing tumor), and the complete tumor (all classes combined).

4.1.3.2 Setup

Some of the hyperparameters of the architectures were shown in Tables 4.1 and 4.2. The remaining are depicted in Table 4.3. All hyperparameters were found using the validation set, consisting of one subject in both HGG and LGG.

We approached brain tumor segmentation as a multi-class classification problem with 5 classes (normal tissue, necrosis, edema, non-enhancing, and enhancing tumor). However, in brain tumor, the classes are imbalanced. So, we used all samples from the underrepresented classes and randomly sampled from the other. Additionally, the number of samples of necrosis and enhancing tumor is small in the LGG training set. To cope with that, we also normalized the intensities of HGG using the landmarks calculated with LGG to extract samples of those classes from HGG to use as training samples in LGG. To train the CNNs for HGG and LGG, we extracted around 450,000 and 335,000 patches, respectively. Note that, with data augmentation, we end up having roughly four times these numbers as effective training samples. Approximately 40% of these patches represent normal tissue in HGG and 50% in LGG. The learning rate was linearly decreased after each epoch during the training stage.

The CNNs were developed using Theano (Bastien et al., 2012; Bergstra et al., 2010) and Lasagne (Dieleman et al., 2015a). The trained architectures are available online[3].

[2] The data set of BRATS 2014 is not currently available.

[3] `http://www.dei.uminho.pt/pessoas/csilva/brats_cnn/`

Table 4.3: Hyperparameters of the proposed method.

Stage	Hyperparameter	Value
Initialization	bias	0.1
	weights	Xavier
Leaky ReLU	α	0.333
Dropout	p – HGG	0.1
	p – LGG	0.5
Training	epochs – HGG	20
	epochs – LGG	25
	ν	0.9
	Initial ϵ	0.003
	Final ϵ	0.00003
	Batch	128
Post-processing	τ_{VOL} – HGG	10000
	τ_{VOL} – LGG	3000

4.1.3.3 Evaluation

The evaluation of the segmentations considered three metrics: Dice Similarity Coefficient (DSC), Positive Predictive Value (PPV), and Sensitivity. The DSC (Dice, 1945) measures the overlap between the manual and the automatic segmentation. It is defined as,

$$DSC = \frac{2TP}{FP + 2TP + FN}, \tag{4.4}$$

where TP, FP, and FN are the numbers of true positive, false positive, and false negative detections, respectively. PPV is a measure of the amount of FP and TP, defined as,

$$PPV = \frac{TP}{TP + FP}. \tag{4.5}$$

Sensitivity is useful to evaluate the number of TP and FN detections, being defined as

$$Sensitivity = \frac{TP}{TP + FN}. \tag{4.6}$$

Finally, in the Challenge set of BRATS 2015 the metrics were provided by the organizers of the Challenge. The two used metrics were DSC and robust Hausdorff Distance. The Hausdorff Distance measures the distance between the surface of computed (∂P) and manual (∂T) segmentation, as

$$Haus(\partial P, \partial T) = \max\{ \sup_{p \in \partial P} \inf_{t \in \partial T} d(p,t), \sup_{t \in \partial T} \inf_{p \in \partial P} d(t,p) \} \tag{4.7}$$

In the robust version of this measure, instead of calculating the maximum distance between the surface of the computed and manual segmentation, it is taken into account the 95% quantile.

4.1.4 Experimental Results and Discussion

In this section, we analyze the effect of key components and the choice of the plane over which we extract patches on the performance of the proposed method. Also, we compare our method with the state of the art using the same database, including also methods based on Deep Learning for brain tumor segmentation. Lastly, we report our result during the participation on BRATS Challenge 2015.

4.1.4.1 Validation of Key Components

We evaluate the effect of each component on the proposed approach by studying the improvement in performance. This increment in performance is evaluated as the mean gain in the metrics (DSC, PPV and Sensitivity), which is obtained in the following way: we compute all metrics using the proposed method for the data sets; then, we remove or substitute the component under study, and compute the metrics for this alternative method. Finally, we subtract each metric for the two systems and calculate the average across the subtractions, obtaining the mean gain, μ_{gain}. The metric of each experiment is reported in Table 4.4, Fig. 4.3 and Fig. 4.4 present the boxplots in the Leaderboard and Challenge data set, respectively, and in Fig. 4.5 we exemplify the effect of the experiments in the segmentation of tumor in two patients (HGG and LGG). In the experiments, we maintained the hyperparameters presented in Table 4.3 as possible to preserve the same conditions[4]. Also, only the images in the Training data set are used in the learning phase of the intensity normalization method. All tests in this section use patches extracted from planes perpendicular to the Axial axis of the MRI image, except in subsection 4.1.4.2, where it is evaluated the choice of the best axis.

Pre-processing The effect of the pre-processing on the segmentation was evaluated by comparing with an alternative method described in (Tustison et al., 2015). We chose this method, because it is also utilized in a CNN-based brain tumor segmentation method (Havaei et al., 2017). This alternative pre-processing starts by applying a 1% winsorizing over the intensities within the brain. Then, the N4ITK is used to correct the bias field in each MRI sequence and the intensities are linearly transformed to $[0, 1]$. Finally, we normalized each sequence to have zero mean and unit variance. During the training stage of the CNN with this pre-processing for LGG, we found to be necessary to decrease the initial and final learning rate to 3×10^{-5} and 3×10^{-7}, respectively, otherwise the optimization would diverge. Observing Table 4.4, we verify that the pre-processing using the intensity normalization method by Nyúl et al. (2000) improved most of the metrics, obtaining a mean gain of 4.6% (Leaderboard: 4.2%, Challenge: 4.9%). This improvement was especially larger for LGG, indicating that the proposed pre-processing increased the detection of the complete as well as the core of the tumor, which is considered a difficult task (Menze et al., 2015). Also, comparing the drop in performance, when removing our pre-processing, and the one verified when removing any other component, we verify that this pre-processing was the key component for improving the segmentation in LGG. The result of this experiment in both grades is interesting, because we know that the features learned by the CNN are computed in local regions by a bank of band-pass

[4]The learning rate was kept constant after 25 epochs; although the validation error may fluctuate, we verified that it stabilized before 30 epochs, so, we trained that amount of epochs and selected the one with the best validation metrics.

Table 4.4: Study of key components of the proposed method. In each test, just the referred component was modified in the Proposed method. Results in bold represent metrics with p-value < 0.05 computed with the two-sided paired Wilcoxon Signed-Rank Test when comparing the results with each component of the Proposed method in each grade, or combination of grades; underlined results represent the one with the highest metric for each region in each grade, or combination of grades.

			DSC			PPV			Sensitivity		
Dataset	Method	Grade	Comp.	Core	Enh.	Comp.	Core	Enh.	Comp.	Core	Enh.
Leaderboard	**Proposed**	HGG	0.88	0.76	0.73	0.91	0.90	0.72	0.86	0.74	0.81
		LGG	0.65	0.53	0.00	0.54	0.42	0.00	0.86	0.86	0.00
		Combined	0.84	0.72	0.62	0.85	0.82	0.60	0.86	0.76	0.68
	Using pre-processing as in (Tustison et al., 2015)	HGG	0.87	0.74	**0.71**	**0.89**	**0.92**	0.73	0.86	0.69	0.75
		LGG	0.34	0.33	0.00	0.29	0.29	0.00	0.63	0.44	0.00
		Combined	0.78	0.67	**0.60**	**0.79**	0.82	0.61	0.82	0.65	0.63
	Using no training samples from HGG into LGG	HGG	0.88	0.76	0.73	0.91	0.90	0.72	0.86	0.74	0.81
		LGG	0.46	0.34	0.00	0.37	0.27	0.00	0.71	0.63	0.00
		Combined	0.81	0.69	0.62	0.82	0.80	0.60	0.84	0.72	0.68
	Using no rotations	HGG	0.87	0.77	**0.73**	**0.86**	**0.83**	**0.70**	**0.89**	**0.78**	0.83
		LGG	0.47	0.31	0.00	0.39	0.25	0.00	0.68	0.66	0.00
		Combined	**0.80**	0.69	**0.61**	**0.78**	**0.74**	**0.59**	**0.85**	0.76	0.70
	Random rotations (5.625°)	HGG	0.87	0.77	0.74	**0.92**	**0.89**	**0.76**	**0.84**	**0.76**	**0.79**
		LGG	0.62	0.49	0.00	0.49	0.38	0.00	0.91	0.87	0.00
		Combined	**0.83**	0.72	0.62	0.85	**0.81**	**0.64**	**0.85**	**0.78**	**0.66**
	Using ReLU	HGG	0.87	0.77	0.73	**0.87**	**0.88**	**0.69**	**0.89**	**0.77**	**0.86**
		LGG	0.53	0.47	0.00	0.40	0.37	0.00	0.86	0.89	0.00
		Combined	0.82	0.72	0.61	**0.79**	**0.80**	**0.58**	**0.88**	**0.79**	**0.72**
	Large/small kernels 1	HGG	**0.85**	0.74	0.72	**0.92**	**0.87**	**0.73**	**0.81**	**0.71**	**0.77**
		LGG	0.52	0.36	0.00	0.42	0.27	0.00	0.84	0.71	0.00
		Combined	**0.80**	0.68	0.60	**0.84**	0.77	**0.61**	**0.81**	**0.71**	**0.65**
	Large/small kernels 2	HGG	**0.85**	0.74	0.72	**0.92**	**0.91**	**0.78**	**0.81**	**0.71**	**0.77**
		LGG	0.52	0.34	0.00	0.42	0.26	0.00	0.85	0.71	0.00
		Combined	**0.79**	**0.67**	0.60	**0.84**	**0.81**	**0.66**	**0.81**	**0.71**	**0.64**
	Coronal patches	HGG	**0.86**	0.75	0.72	**0.88**	**0.83**	**0.74**	0.86	0.74	**0.76**
		LGG	0.59	0.44	0.00	0.46	0.34	0.00	0.92	0.86	0.00
		Combined	**0.82**	**0.70**	0.61	**0.81**	**0.75**	**0.62**	0.87	0.76	**0.64**
	Sagittal patches	HGG	**0.86**	0.75	**0.71**	**0.86**	**0.79**	**0.70**	0.87	**0.78**	**0.79**
		LGG	0.45	0.32	0.00	0.36	0.26	0.00	0.87	0.70	0.00
		Combined	**0.79**	**0.68**	**0.60**	**0.78**	**0.70**	**0.59**	0.87	0.76	**0.66**
Challenge	**Proposed**	HGG	0.88	0.83	0.77	0.88	0.87	0.74	0.89	0.83	0.81
	Using pre-processing as in (Tustison et al., 2015)	HGG	0.80	**0.78**	0.73	**0.75**	0.86	0.71	0.92	**0.74**	0.77
	Using no rotations	HGG	**0.85**	**0.79**	0.74	**0.81**	**0.78**	**0.70**	**0.91**	**0.86**	0.82
	Random rotations (5.625°)	HGG	0.88	0.82	0.76	**0.90**	**0.84**	**0.76**	**0.86**	0.84	**0.78**
	Using ReLU	HGG	**0.86**	0.81	**0.74**	**0.82**	**0.80**	**0.66**	**0.90**	**0.85**	**0.86**
	Large kernels/shallow arq. 1	HGG	0.87	**0.81**	**0.75**	**0.90**	**0.89**	0.76	**0.84**	**0.78**	**0.76**
	Large kernels/shallow arq. 2	HGG	0.87	**0.81**	**0.75**	**0.90**	0.88	**0.75**	**0.84**	**0.78**	**0.76**
	Coronal patches	HGG	**0.85**	**0.81**	**0.74**	**0.81**	**0.85**	0.75	0.90	0.79	**0.75**
	Sagittal patches	HGG	**0.84**	**0.76**	0.73	**0.78**	**0.73**	**0.67**	**0.92**	0.85	0.82

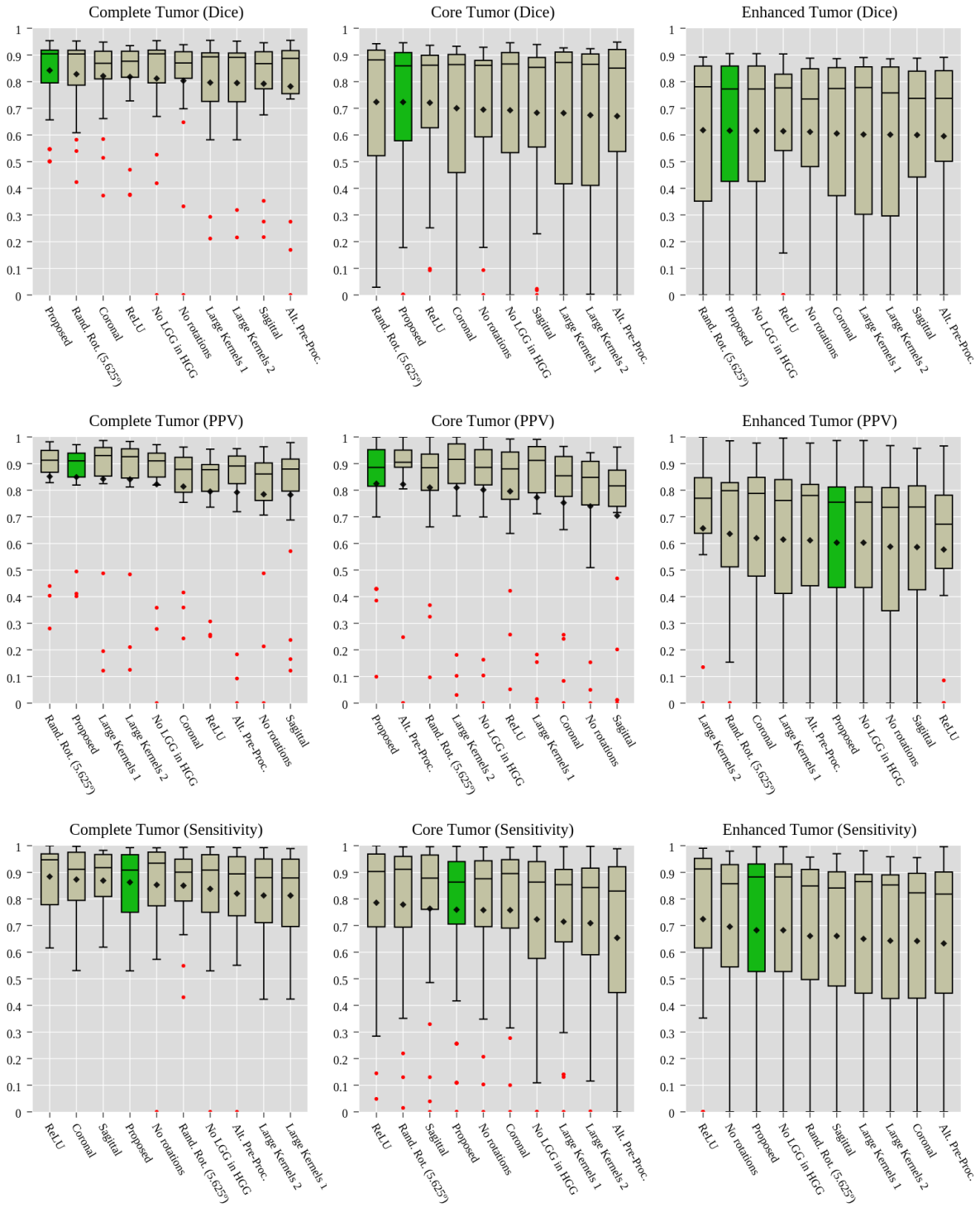


Figure 4.3: Boxplot for each of the experiments in Table 4.4 in the Leaderboard data set. The boxplot for the experiment of sampling training samples from HGG into LGG is not shown given the reduced number of subjects (4 LGG in 25 subjects for the Leaderboard data set). The diamond marks the mean.

filters at different scales, instead of point-wise properties as an intensity. Shah et al. (2011) presented a study regarding the segmentation of multiple sclerosis based on MRI images, showing that classifiers based on point-wise features, as intensity, improved after Nyúl normalization. This improvement was obtained by minimizing the data heterogeneity from multi-site multi-scanner MRI acquisitions. However, our experiment gives evidence that in MRI applications, CNN-based classifiers also improve after Nyúl

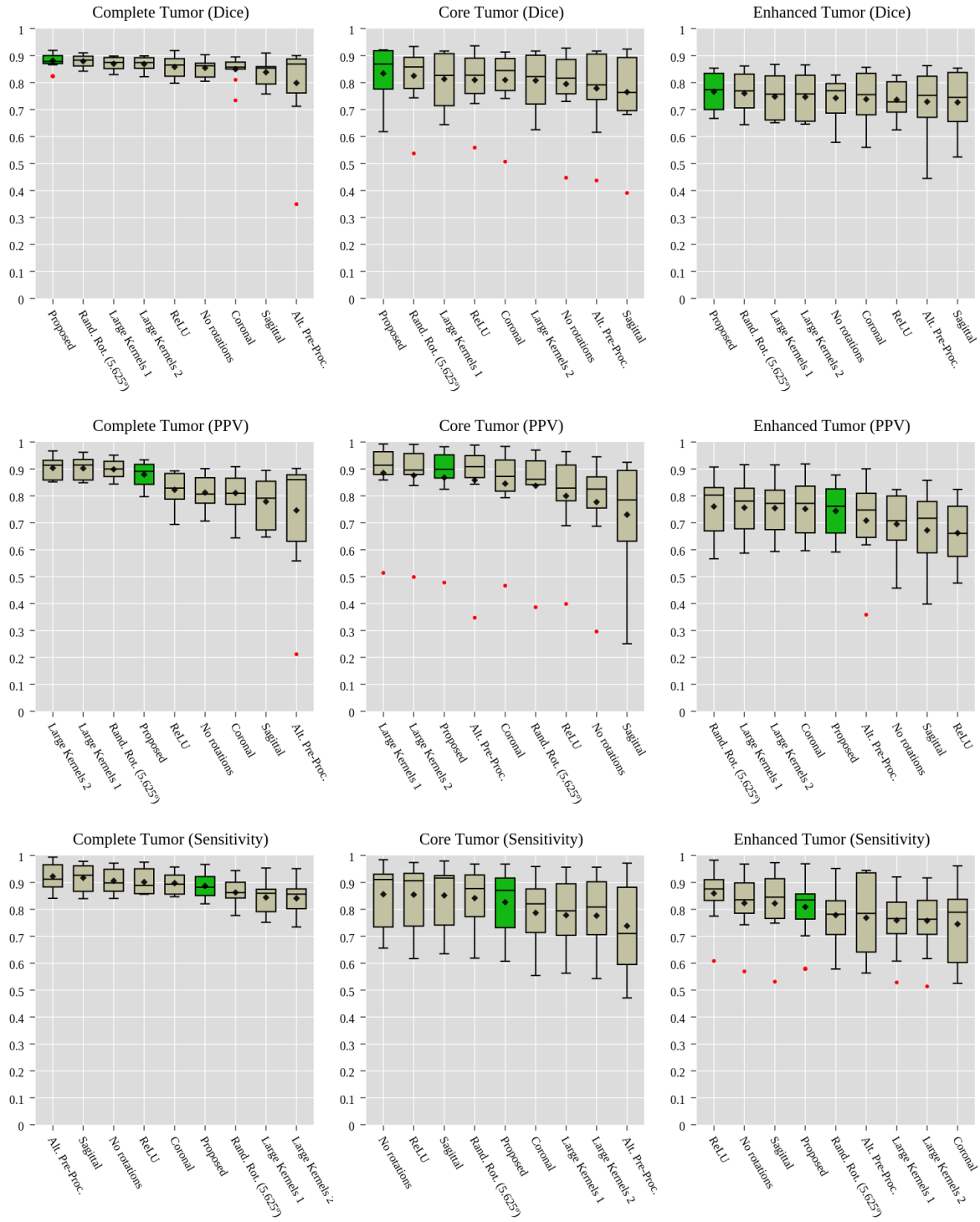
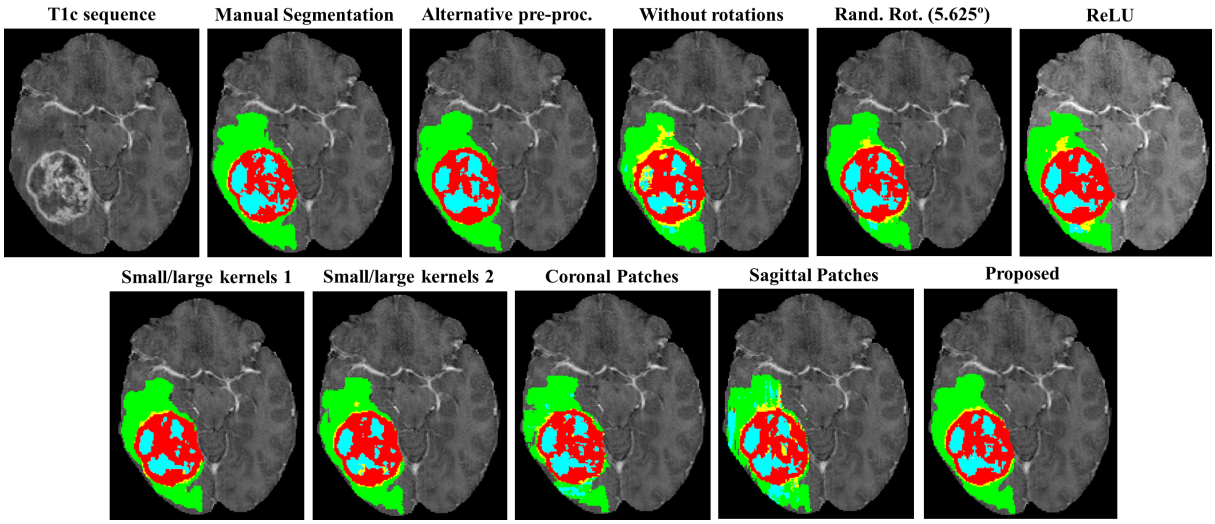
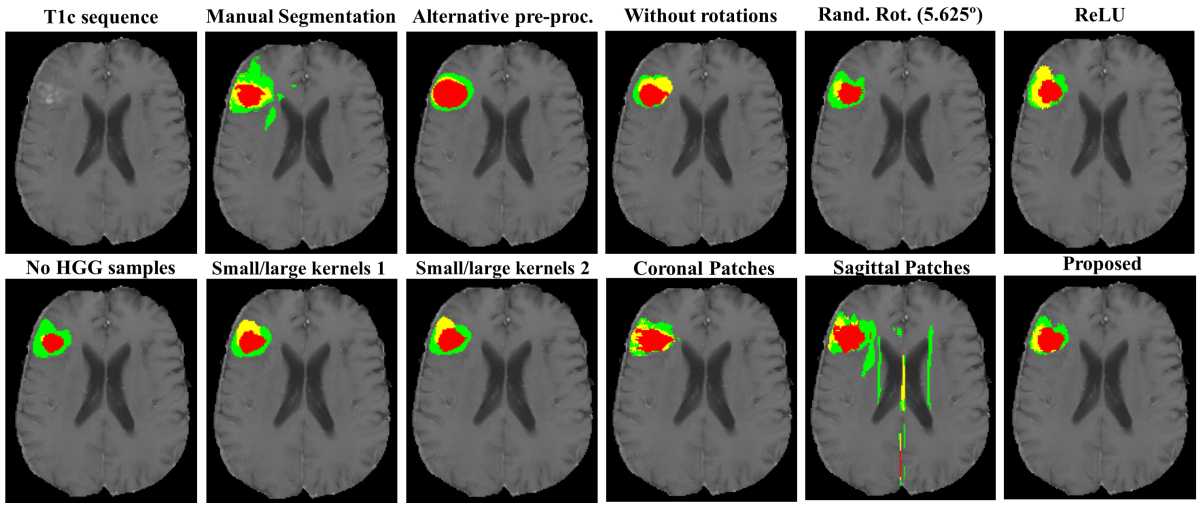


Figure 4.4: Boxplot for each of the experiments in Table 4.4 in the Challenge data set. The diamond marks the mean.

normalization, at least in the context of brain tumor segmentation. Additionally, we further investigated the effect of increasing the number of training epochs until 90 epochs, but we obtained no improvement with the simpler pre-processing. Referring to Fig. 4.5, we can observe that the proposed pre-processing enabled a better training of the CNN, such that the segmentation presented a better delineation of the non-enhancing and the necrosis regions in both data sets.



(a)



(b)

Figure 4.5: Examples of segmentations obtained with cross-validation, showing the effect of each component of the proposed method. In the first row, we have a HGG, and in the bottom row a LGG. Each color represents a tumor class: green – edema, blue – necrosis, yellow – non-enhancing tumor and red – enhancing tumor.

Data Augmentation Artificial data augmentation is a common procedure in the context of CNN, when the data set is relatively small. In the case of MRI images, we have a large number of samples for healthy and tumorous tissue, which may be the reason why most recent studies on brain tumor segmentation based on Deep Learning (Davy et al., 2014; Urban et al., 2014; Lyksborg et al., 2015) did not explore data augmentation. Havaei et al. (2017) considered its application, but found to be ineffective in their system.

We investigated two types of data augmentation. In the first case, we studied the effect of data augmentation by increasing the number of samples using rotations. In this study, we evaluated two variants. In the first, we used multiples of 90° (90°, 180° and 270°) for rotations (corresponding to the Proposed method), while in the second, we sampled three rotation angles from an array using an

uniform distribution, whose angles were equally spaced. The angle step was defined as $\pi \times 90°$ with $\pi \in \{1/8, 1/16, 1/32\}$. In this second variant, we consider $\pi = 1/16$. The case $\pi = 1/16$ was the best result, but the tests with other angles can be found in Table 4.5. In Table 4.4, we present the results with each variant and without rotations in both the Leaderboard and Challenge data sets. As can be observed, the rotations improved the performance in all regions for the DSC and PPV; but, we also note a decrease in the sensitivity for both variants in the Challenge data set. However, the mean gain obtained by including rotations was 2.6% (Leaderboard: 2.6%, Challenge: 2.7%) for the first variant (Proposed) and 2.3% (Leaderboard: 2.7%, Challenge: 2.0%) for the second variant. Comparing the two variants, we obtain a mean gain of 0.3% of the first variant in relation to the second. Also, the first variant has the advantage of being faster to compute. Observing Fig. 4.5, we conclude that the extra information provided by the rotations of the first variant in training the CNN resulted in segmentations with a better delineation of the complete tumor as well as of the intra-tumoral structures. In both grades, we have an excess of non-enhancing class, when we trained without data augmentation, and for HGG this class is even found inside the region formed by enhancing and necrotic structures, which does not happen in the manual segmentation.

Brain tumors are constituted by intra-tumoral structures with very different volumes, resulting in an imbalanced number of samples of each class. This underrepresentation of some classes impairs the performance of the CNN. So, we investigated a second type of data augmentation to balance the number of samples of each class, which consisted in extracting samples from necrosis and enhancing tumor regions in HGG to use as training samples in LGG. In Table 4.4, we compare the proposed approach, in which the number of samples of each tumor class in LGG were more balanced, with another experiment that uses only samples from LGG. We verify that the extra samples from HGG improved all metrics for the complete and core regions in the Leaderboard data set with a mean gain of 1.9%. Examining Fig. 4.5(b), we note that by sampling from HGG to LGG, we improved the training of the CNN. Observe that the tumor segmentation presented a better delineation of the enhancing and non-enhancing regions, although the sampling was only for enhancing and necrosis regions. The improvement of the non-enhanced region could be explained by the context introduced by the patches of the enhancing samples, since these two regions are next to each other.

Activation Function The gradients of ReLU are zero when the unit is not active, which may slow down the convergence during the optimization and lead to worst training. To avoid that problem, Maas et al. (2013) proposed LReLU as an alternative nonlinearity. So, we investigate the effectiveness of this activation function in brain tumor segmentation. In this experiment, only the activation function was changed in the proposed method. The results in both the Leaderboard and Challenge data sets are presented in Table 4.4. We verify that LReLU activation improved the performance of the proposed method in both data sets in the DSC and PPV, with the exception of the core in the DSC in the Leaderboard data set. ReLU activations presented better scores in the Sensitivity metric. However, the mean gain using LReLU instead of ReLU was 1.3% (Leaderboard: 0.44%, Challenge: 2.2%). Referring to Fig. 4.5, we find that using ReLU as an activation function resulted in an excessive segmentation of non-enhancing and necrosis regions outside the core for HGG.

Table 4.5: Study of artificial data augmentation using rotations of the patches. In each test, just the referred component was modified in the Proposed method. Results in bold represent metrics with p-value < 0.05 computed with the two-sided paired Wilcoxon Signed-Rank Test when comparing the results with each component of the Proposed method in each grade or combination of grades; underlined results represent the one with the highest metric for each region in each grade or combination of grades.

			DSC			PPV			Sensitivity		
Dataset	Method	Grade	Comp.	Core	Enh.	Comp.	Core	Enh.	Comp.	Core	Enh.
Leaderboard	**Proposed**	HGG	0.88	0.76	0.73	0.91	0.90	0.72	0.86	0.74	0.81
		LGG	0.65	0.53	0.00	0.54	0.42	0.00	0.86	0.86	0.00
		Combined	0.84	0.72	0.62	0.85	0.82	0.60	0.86	0.76	0.68
	Using no rotations	HGG	0.87	0.77	**0.73**	**0.86**	**0.83**	**0.70**	**0.89**	**0.78**	0.83
		LGG	0.47	0.31	0.00	0.39	0.25	0.00	0.68	0.66	0.00
		Combined	**0.80**	0.69	**0.61**	**0.78**	**0.74**	**0.59**	**0.85**	0.76	0.70
	Totally random rotations	HGG	0.86	0.76	0.73	**0.92**	0.84	**0.75**	**0.83**	0.75	**0.77**
		LGG	0.63	0.49	0.00	0.50	0.39	0.00	0.92	0.84	0.00
		Combined	**0.84**	0.70	0.60	0.83	0.79	**0.58**	**0.87**	0.72	**0.69**
	Random rotations $(90°/4)$	HGG	0.87	0.77	0.74	**0.93**	0.90	**0.76**	**0.84**	0.76	**0.78**
		LGG	0.64	0.54	0.00	0.51	0.42	0.00	0.91	0.89	0.00
		Combined	0.83	0.73	0.62	**0.86**	0.82	**0.64**	**0.85**	**0.78**	**0.65**
	Random rotations $(90°/8)$	HGG	0.87	0.77	0.74	**0.92**	**0.88**	**0.79**	**0.84**	**0.76**	**0.78**
		LGG	0.57	0.45	0.00	0.43	0.34	0.00	0.92	0.89	0.00
		Combined	**0.82**	0.72	0.62	0.84	**0.80**	**0.67**	**0.85**	**0.78**	**0.65**
	Random rotations $(90°/16)$	HGG	0.87	0.77	0.74	**0.92**	**0.89**	**0.76**	**0.84**	**0.76**	**0.79**
		LGG	0.62	0.49	0.00	0.49	0.38	0.00	0.91	0.87	0.00
		Combined	**0.83**	0.72	0.62	0.85	**0.81**	**0.64**	**0.85**	**0.78**	**0.66**
	Random rotations $(90°/32)$	HGG	0.86	0.76	0.73	**0.93**	**0.84**	**0.74**	**0.82**	0.75	**0.77**
		LGG	0.54	0.42	0.00	0.41	0.32	0.00	0.92	0.84	0.00
		Combined	**0.81**	0.71	0.61	0.84	**0.76**	**0.62**	**0.84**	0.76	**0.64**
Challenge	**Proposed**	HGG	0.88	0.83	0.77	0.88	0.87	0.74	0.89	0.83	0.81
	Using no rotations	HGG	**0.85**	**0.79**	0.74	**0.81**	**0.78**	**0.70**	**0.91**	**0.86**	0.82
	Totally random rotations	HGG	0.87	0.83	0.77	**0.85**	**0.87**	**0.74**	**0.89**	0.82	**0.82**
	Random rotations $(90°/4)$	HGG	0.88	0.82	0.76	**0.90**	0.85	**0.77**	**0.86**	0.83	**0.77**
	Random rotations $(90°/8)$	HGG	0.88	0.82	0.75	**0.89**	**0.83**	0.76	0.87	0.83	**0.77**
	Random rotations $(90°/16)$	HGG	0.88	0.82	0.76	**0.90**	**0.84**	**0.76**	**0.86**	0.84	**0.78**
	Random rotations $(90°/32)$	HGG	0.88	0.83	0.76	**0.90**	**0.84**	**0.76**	**0.86**	0.84	**0.78**

Deeper architectures/small kernels Using cascaded layers with small 3×3 kernels has the advantage of maintaining the same effective receptive field of bigger kernels, while reducing the number of weights, and allowing more non-linear transformations on the data. To evaluate the real impact of this technique on brain tumor segmentation, we changed the cascaded convolutional layers before each max-pooling of the proposed architecture by one layer with larger kernels with the equivalent effective receptive field. So, in HGG we changed the groups of layers 1, 2, 3 and 5, 6, 7 (Table 4.1) by one convolutional layer with 7×7 kernels each, while in the LGG we changed the groups of layers 1 and 2, and 4 and 5 (Table 4.2) by one layer with 5×5 kernels each. Using these architectures, we experimented two variants for both grades: 1) we maintained the 64 feature maps in the first convolutional layer and 128 in the second; 2) we increased the capacity of the CNN by using wider layers, namely, 128 feature maps in the first convolutional layer and 256 in the second. We present the results obtained in the Leaderboard and Challenge data sets in Table 4.4 and the boxplots in Fig. 4.3 and 4.4. In relation to variant 1, the mean gain was 2.4% (Leaderboard: 3.1%, Challenge: 1.6%), while for variant 2 it was 2.1% (Leaderboard: 2.4%, Challenge: 1.8%). In the majority of metrics, the proposed method obtained higher scores than both variants with bigger kernels, with some of them with statistical significance, while the variants achieved better scores in PPV (HGG in both data sets). In the boxplots, both variants seem to have larger dispersion and more outliers. In the segmentations of Fig. 4.5, although the segmentations by the variants appear with good quality, the proposed method can capture more details, and variant 2 classified some non-enhanced tumor inside the enhancing ring, which does not happen in the manual segmentation in HGG; in LGG the architecture with bigger kernels also identified an excess of non-enhancing tumor.

4.1.4.2 Patch Extraction Plane

The use of 2D patches in a MRI image requires that we define a plane perpendicular to an axis to extract patches. So, following the procedure defined in the previous subsection, we investigated the use of patches extracted in a plane perpendicular to the Axial, Coronal, and Sagittal axis. The results in both the Leaderboard and Challenge data sets are presented in Table 4.4. As can be observed, extracting patches in the plane perpendicular to the Axial axis presented the best overall performance with a mean gain of 2.33% relative to the Coronal plane (Leaderboard: 1.89%, Challenge: 2.78%) and 4.00% relative to the Sagittal plane (Leaderboard: 3.56%, Challenge: 4.44%). The Axial plane presented better DSC and PPV scores for both data sets than the Sagittal plane, but worst sensitivity for the Challenge data set and for the complete region in the Leaderboard data set. Considering Fig. 4.5, this can be explained by an over-segmentation of the tumor, which is corroborated by the lower PPV score. A similar pattern is found for the Coronal plane, which was better in the enhanced region for the PPV score and in the complete region for the Sensitivity score. The better performance obtained using patches extracted in the Axial plane can be explained by some acquisitions having lower spatial resolution in the Coronal and Sagittal planes, which can be considered a limitation of the BRATS databases.

4.1.4.3 General

Finally, as an overall analysis, we note some general trends across all experiments. Considering the boxplots, Fig. 4.3 and 4.4, we verify a lower dispersion for the complete region, presenting also a higher mean value for the same region. This lower dispersion is less expressive in the Leaderboard than in the Challenge data set, which may be explained by the worst performance of the algorithms on LGG subjects in this data set. Another general trend is found in Table 4.4 that shows that none of the algorithms found presence of enhanced region among the LGG subjects[5].

4.1.4.4 Global Validation

In Table 4.6, we compile the results of the top 5 methods in the Leaderboard and Challenge data sets of BRATS 2013 (including the proposed method). We also include the proposals by Havaei et al. (2017), Davy et al. (2014), and Urban et al. (2014) that are based on CNN. Appraising the results in Table 4.6, we conclude that no method is yet able to achieve the first place in all metrics and regions for brain tumor segmentation; but, the proposed method obtained the first position in DSC in the three regions (Challenge data set), according to the online evaluation platform (VirtualSkeleton, 2013). Also, based on the same evaluation, the proposed method obtained the overall first position in both data sets, outperforming the other methods.

Assessing the CNN-based methods, we observe that two of those methods (Davy et al., 2014; Urban et al., 2014) had modest performances, comparing with others not based on CNN; however, the method proposed by Havaei et al. (2017) exhibit higher metrics. They propose a novel and elaborated training and concatenation of two CNNs to capture more context into the training. Contrasting the two methods, while our method is better in Sensitivity, their method is in PPV in the complete region. According to Menze et al. (2015), the most difficult tasks in brain tumor segmentation are the segmentation of the core region for LGG and the enhancing region for HGG. In these two tasks our method outperformed Havaei et al. (2017). We note a larger difference in the core region in the Challenge data, which is considered an easier region to segment according to Menze et al. (2015). Based on the analysis of the key components in the previous section, we conclude that although our architecture is simpler, those components permitted a better training of our CNN classifier, compensating the lack of information of a larger context, which according to the experiments reported by Havaei et al. (2017) was found to be relevant.

The method proposed by Kwon et al. (2014b) is ranked in the second place in both data sets. They perform a joint segmentation and registration using a tumor growth model to transform an atlas of healthy patients into one with the tumor and its intra-tumoral structures. Given the complex shape of tumors, they refine the initial solution using the Expectation Maximization algorithm. Comparing their approach with ours in the Challenge data set (HGG tumors), our method obtained higher DSC in the enhancing region and in Sensitivity in the three regions, and their method was better in PPV in the complete and core regions. For the Leaderboard data set (LGG and HGG tumors), our method obtained higher metrics in the enhancing region in DSC and Sensitivity, and their method was better in the complete region in PPV and DSC, and in the core region in the three metrics. Another strong contender in the Leaderboard

[5]We currently are not able to confirm if these tumors contain enhanced tumor, since the expert segmentation is private.

Table 4.6: Results in the Leaderboard and Challenge data sets of BRATS 2013. The relative rank refers to the combination of the ranking in each metric for the referred class, while the position is the global ranking, as provided by the online evaluation platform (VirtualSkeleton, 2013).

	Methods	DSC			PPV			Sensitivity			Relative Rank			Position
		Comp.	Core	Enh.	Comp.	Core	Enh.	Comp.	Core	Enh.	Comp.	Core	Enh.	
Leaderboard	**Proposed**	0.84	0.72	0.62	0.85	0.82	0.60	0.86	0.76	0.68	3.67	3.33	1.67	1
	Kwon et al. (2014b)	0.86	0.79	0.59	0.88	0.84	0.60	0.86	0.81	0.63	3.33	1.67	5.00	2
	Zhao et al.† (Menze et al., 2015)	0.83	0.73	0.55	0.77	0.67	0.46	0.94	0.89	0.78	4.67	4.00	9.33	3
	agnm1‡	0.83	0.71	0.54	0.85	0.73	0.59	0.84	0.82	0.58	6.00	4.33	10.33	4
	havam2‡	0.82	0.69	0.56	0.83	0.77	0.62	0.83	0.69	0.58	7.67	7.00	8.00	5
	Urban et al. (2014)‡	0.70	0.57	0.54	0.65	0.55	0.52	0.87	0.67	0.60	14.00	18.67	12.33	17
	Havaei et al. (2017)††	0.84	0.71	0.57	0.88	0.79	0.54	0.84	0.72	0.68	–	–	–	–
	Davy et al. (2014)	0.72	0.63	0.56	0.69	0.64	0.50	0.82	0.68	0.68	–	–	–	–
Challenge	**Proposed**	0.88	0.83	0.77	0.88	0.87	0.74	0.89	0.83	0.81	7.00	3.33	5.33	1
	Kwon et al. (2014b,a)	0.88	0.83	0.72	0.92	0.90	0.74	0.84	0.78	0.72	9.33	5.00	13.00	2
	Tustison et al. (2015)	0.87	0.78	0.74	0.85	0.74	0.69	0.89	0.88	0.83	10.33	11.67	9.00	3
	havam2‡	0.88	0.78	0.73	0.89	0.79	0.68	0.87	0.79	0.80	8.33	10.67	13.33	4
	al-ss1‡	0.87	0.78	0.70	0.89	0.83	0.75	0.86	0.78	0.70	9.67	8.67	14.67	5
	Urban et al. (2014)†	0.86	0.75	0.73	0.82	0.75	0.79	0.92	0.79	0.70	11.67	16.00	11.67	12
	Havaei et al. (2017)	0.88	0.79	0.73	0.89	0.79	0.68	0.87	0.79	0.80	–	–	–	–
	Davy et al. (2014)	0.85	0.74	0.68	0.85	0.74	0.62	0.85	0.78	0.77	–	–	–	–

† Results retrieved from (VirtualSkeleton, 2013) using the cited method.

‡ Results retrieved from (VirtualSkeleton, 2013), but the method or author are unknown.

†† Results provided by the author using the cited method.

is the method proposed by Zhao (Menze et al., 2015). His method was better in the core region in DSC and outperformed all methods in the complete and core regions in sensitivity; however, since we note a significant drop in performance in PPV in the same regions, we may infer that probably the method by Zhao over-segmented the tumor. Also, we note another trend, both our and Kwon methods drop from the Challenge to the Leaderboard data set in most metrics (Kwon improved in the complete and core regions in sensitivity); however, appraising the separated metrics for HGG and LGG in the Leaderboard, Table 4.6, we observe that the performance of our method was similar in the complete and enhanced regions but dropped more significantly in the core region in DSC and in sensitivity; therefore, given the lower metric of LGG, we hypothesize that the general drop in both methods from the Challenge data set to the Leaderboard was mainly due to the LGG subjects; however, the method proposed by Kwon dropped less in the core region. Considering the performance in both data sets, we argue that both methods were similar in segmenting the complete tumor, the method proposed by Kwon was in general better in the core region and our method in delineating the enhanced structure. Assessing the running times, Kwon reports an average running time of 85 min. on an Intel Core i7 3.4 GHz machine, while our full pipeline presents an average running time of 8 min. using a GPU NVIDIA GeForce GTX 980 equipped on an Intel Core i7 3.5 GHz machine. This difference in running times is explained by our method performing an optimization only during training, which permits a fast segmentation during normal use.

Considering the state of the art, we verify that current CNN-based approaches (Zikic et al., 2014; Urban et al., 2014; Davy et al., 2014; Havaei et al., 2017; Lyksborg et al., 2015; Rao et al., 2015; Dvorák and Menze, 2015) have used larger filters and shallow architectures, with some using features computed by the CNN as input to a RF (Rao et al., 2015), or employing the network for structured prediction (Dvorák and Menze, 2015). Also, these works did not explore the stacking of several layers to apply more non-linearities on the data, which we showed to be important. In the CNN, these authors have used more

common non-linearities, as hyperbolic tangent or ReLU; however, our experiments indicate that LReLU is a strong alternative to ReLU and do not suffer of the limitations of the hyperbolic tangent (Maas et al., 2013). Although some authors found no advantage in using data augmentation (Havaei et al., 2017), we have shown that data augmentation and the adequate pre-processing have a significant impact on performance. Based on these facts, our conclusion is that the contributions in this article are orthogonal to current state of the art, existing potential for further improvement in brain tumor segmentation using MRI images by looking for synergies with the techniques studied by current works.

In Fig. 4.6, we present the segmentation of two patients with HGG and LGG, respectively, from the Leaderboard data set. Fig. 4.7 shows a patient with two tumors that were correctly detected and segmented from the Challenge data set.

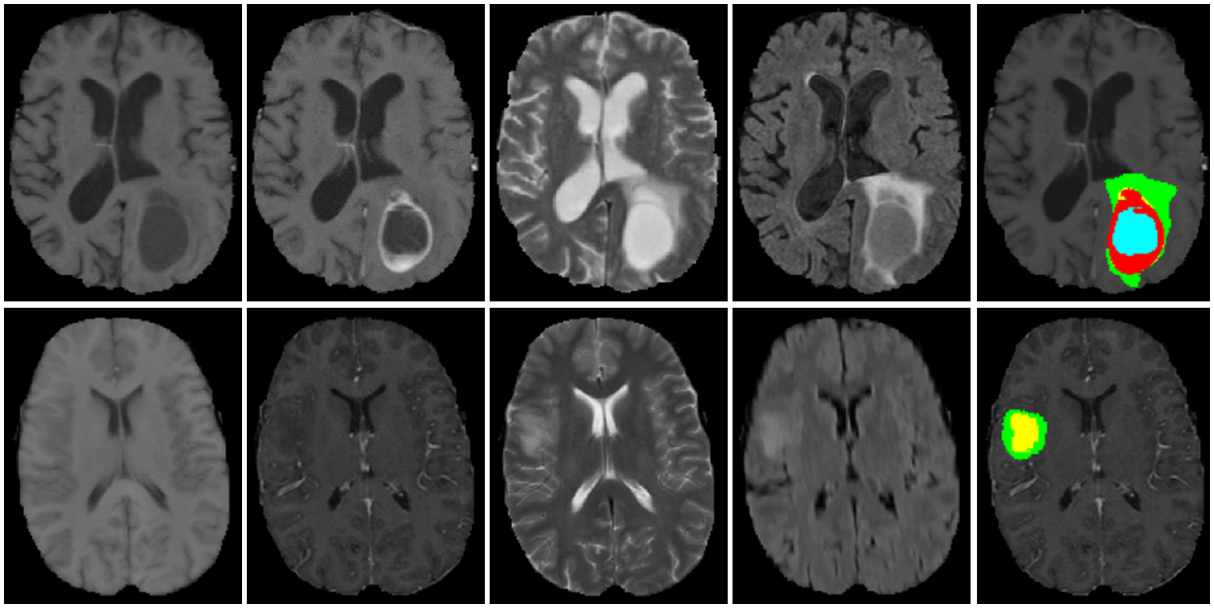


Figure 4.6: Examples of segmentations in BRATS 2013 Leaderboard data set, showing a HGG in the first row (subject id: 210) and a LGG in the bottom row (subject id: 105). From left to right: T1, T1c, T2, FLAIR, and the segmentation. Each color represents a tumor class: green – edema, blue – necrosis, yellow – non-enhancing tumor, and red – enhancing tumor.

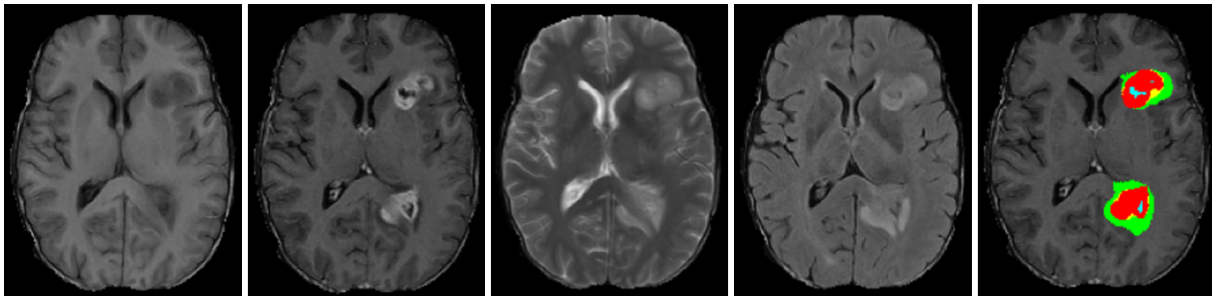


Figure 4.7: Examples of segmentations in BRATS 2013 Challenge data set (subject id: 310). From left to right: T1, T1c, T2, FLAIR, and the segmentation. Each color represents a tumor class: green – edema, blue – necrosis, yellow – non-enhancing tumor, and red – enhancing tumor.

4.1.4.5 Evaluation in BRATS 2015

The proposed architecture was further evaluated in BRATS 2015 database, both in the Training and Challenge sets. The differences when comparing with the models trained in BRATS 2013 were the number of samples for training, given the bigger size of the Training set, and in Dropout (p) that was increased to 0.5 in the HGG architecture.

BRATS 2015 Training set Some segmentation examples obtained in the Training data set are illustrated in Figure 4.8, where we can observe the necrosis, edema, non-enhanced, and enhanced tumor classes; quantitative results in the same set are presented in Table 4.7 and Figure 4.9. These results were obtained by 2-fold cross-validation and 3-fold cross-validation in HGG and LGG, respectively.

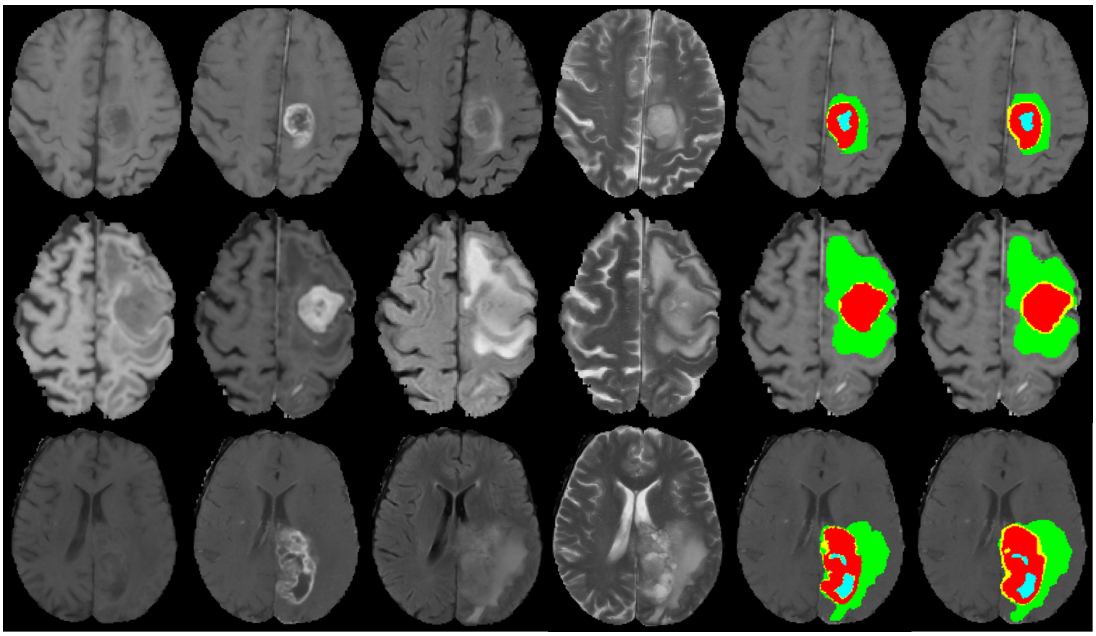
Observing Table 4.7, metrics in the Core and Enhanced regions of LGG are lower than in HGG, which may be due to the lower contrast of the former. In fact, the contrast in the Core region is lower in LGG (Menze et al., 2015) than in HGG. Additionally, although brain tumors are very heterogeneous, LGG tend to be smaller than HGG, with less Core tissues, as observed from the first and third rows of Figure 4.8(b). Another issue with LGG is the smaller number of training patients, when compared to HGG. From the boxplots in Figure 4.9, we can observe the larger dispersion in the Core region of LGG compared to HGG; in the enhanced tumor in LGG the boxplots range almost the full scale of the metrics, possibly because some of these tumors do not possess enhancing tumor. However, the results for the Complete region are similar in LGG and HGG, with similar dispersion as observed in the boxplots. There are some outliers in Figure 4.9, mainly in HGG, which may be due to the high variability of brain tumors and to the bigger amount of patients with HGG. Following the results in Table 4.7, in Figure 4.8 the boundaries of the complete tumor seem well defined, both in LGG and HGG. However, from the second and third rows in Figure 4.8(b) it seems that we are over-segmenting the Core classes in LGG; nevertheless, the second example looks particularly difficult with a big portion of tumor Core tissues in a very heterogeneous distribution, sharp shapes and details.

Participation on BRATS 2015 Challenge The CNNs were then re-trained for the participation in the on-site BRATS 2015 segmentation challenge, where the task was to segment the 53 Challenge subjects.

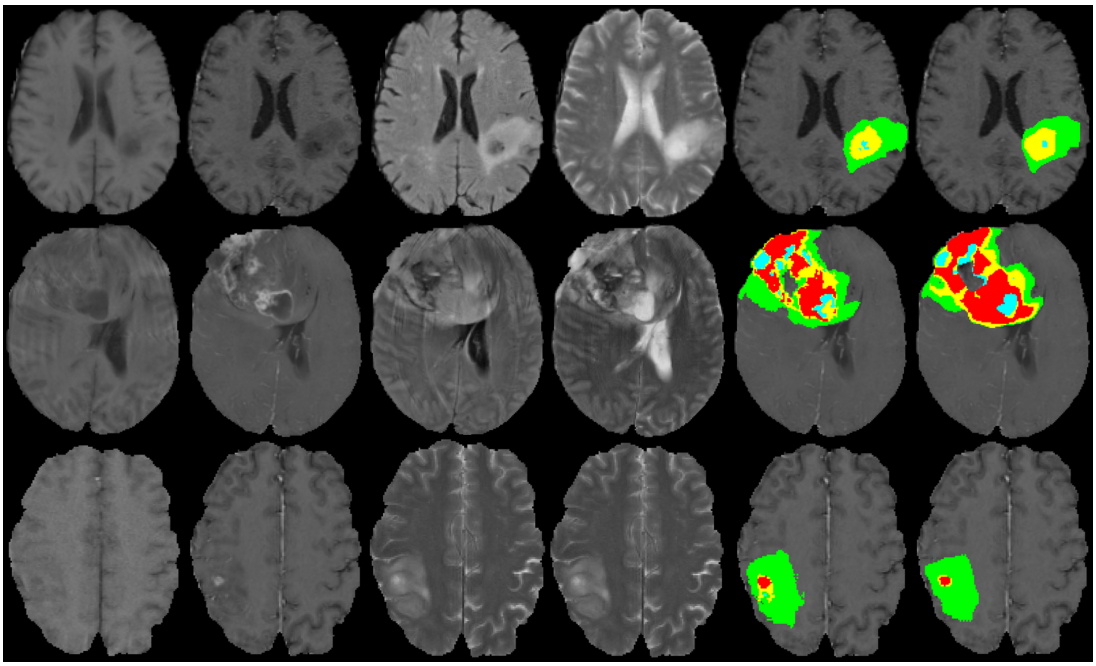
Figure 4.10 presents segmentation examples obtained in the Challenge data set, while Table 4.8 and Figure 4.11 present the quantitative results. In this case, all subjects in each grade of the Training data set were used for training the CNN, with the exception of six validation patients in each grade. To train the CNNs we extracted around 4,000,000 training patches of HGG and 1,800,000 of LGG, and we used

Table 4.7: Results (mean) obtained with BraTS 2015 Training data set.

	DSC			PPV			Sensitivity		
	Complete	Core	Enhanced	Complete	Core	Enhanced	Complete	Core	Enhanced
LGG	0.86	0.64	0.40	0.86	0.67	0.39	0.88	0.71	0.51
HGG	0.87	0.75	0.75	0.89	0.76	0.80	0.86	0.79	0.75
LGG + HGG	0.87	0.73	0.68	0.89	0.74	0.72	0.86	0.77	0.70



(a)

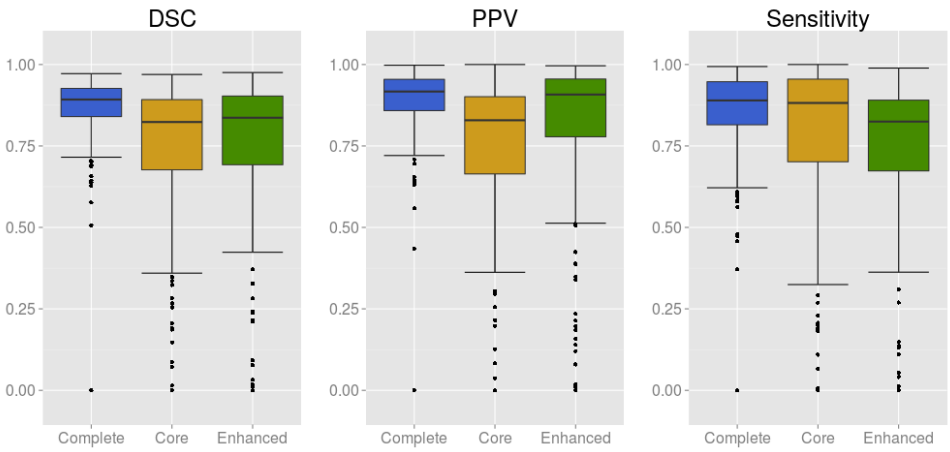


(b)

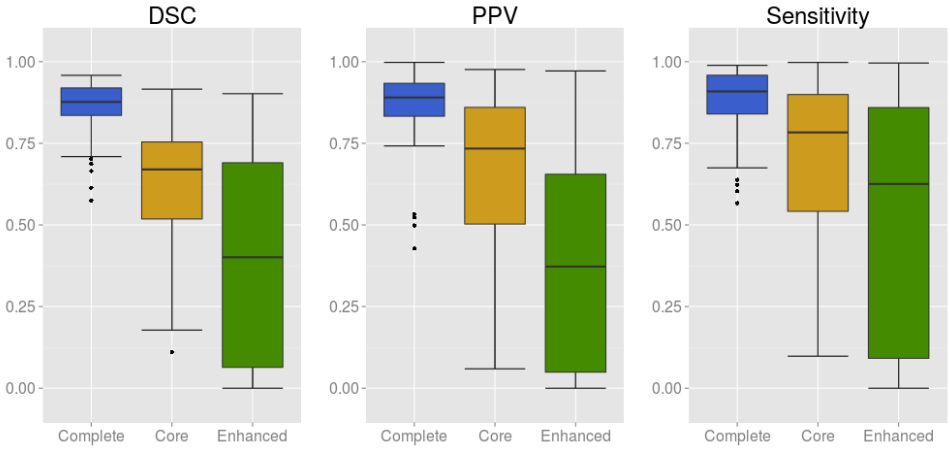
Figure 4.8: Segmentation examples on the Training data set from a) HGG and b) LGG. From left to right: T1, T1c, FLAIR, T2, manual segmentation and obtained segmentation. Colors in the segmentations represent: blue - necrosis, green - edema, yellow - non-enhanced tumor, red - enhanced tumor.

mini-batches of 128 training samples. However, the number of training patches was 4 times bigger due to the data augmentation. Observing Figure 4.10, the segmentations seem coherent with the expected tumor tissues, for example, the enhanced tumor portions appear delineated following the enhancing parts in T1c. Also, the complete tumor appears to be well delineated, when comparing with the FLAIR and T2 sequences, where the edema is hyperintense.

The proposed approach achieved the 2nd position among all participating teams.



(a)



(b)

Figure 4.9: Boxplot of the results in each of the evaluated brain tumor regions using the Training data set in a) HGG and b) LGG; black dots represent outliers

4.1.5 Conclusions

In summary, we propose a novel CNN-based method for segmentation of brain tumors in MRI images. We start by a pre-processing stage consisting of bias field correction, intensity and patch normalization. After that, during training, the number of training patches is artificially augmented by rotating the training patches, and using samples of HGG to augment the number of rare LGG classes. The CNN is built over convolutional layers with small 3×3 kernels to allow deeper architectures.

In designing our method, we address the heterogeneity caused by multi-site multi-scanner acquisitions of MRI images using intensity normalization as proposed by Nyúl et al. (2000). We show that this is important in achieving a good segmentation. Brain tumors are highly variable in their spatial localization and structural composition, so we have investigated the use of data augmentation to cope with such variability. We studied augmenting our training data set by rotating the patches as well as by sampling from classes of HGG that were underrepresented in LGG. We found that data augmentation was also quite effective, although not thoroughly explored in Deep Learning methods for brain tumor segmentation. Also, we investigated the potential of deep architectures through small kernels by comparing our deep

Table 4.8: Results (mean) using the Challenge data set of BraTS 2015.

DSC			Robust Hausdorff		
Complete	Core	Enh.	Complete	Core	Enh.
0.78	0.65	0.75	15.83	26.54	6.99

CNN with shallow architectures with larger filters. We found that shallow architectures presented a lower performance, even when using a larger number of feature maps. Finally, we verified that the activation function LReLU was more important than ReLU in effectively training our CNN.

We evaluated the proposed method in BRATS 2013 and 2015 databases. Concerning 2013 database, we were ranked in the first position by the online evaluation platform. Also, it was obtained simultaneously the first position in DSC metric in the complete, core, and enhancing regions in the Challenge data set. Comparing with the best generative model (Kwon et al., 2014b), we were able to reduce the computation time approximately by ten-fold. Concerning the 2015 database, we obtained the second position among thirteen contenders in the on-site challenge. We argue, therefore, that the components that were studied have potential to be incorporated in CNN-based methods and that as a whole our method is a strong candidate for brain tumor segmentation using MRI images.

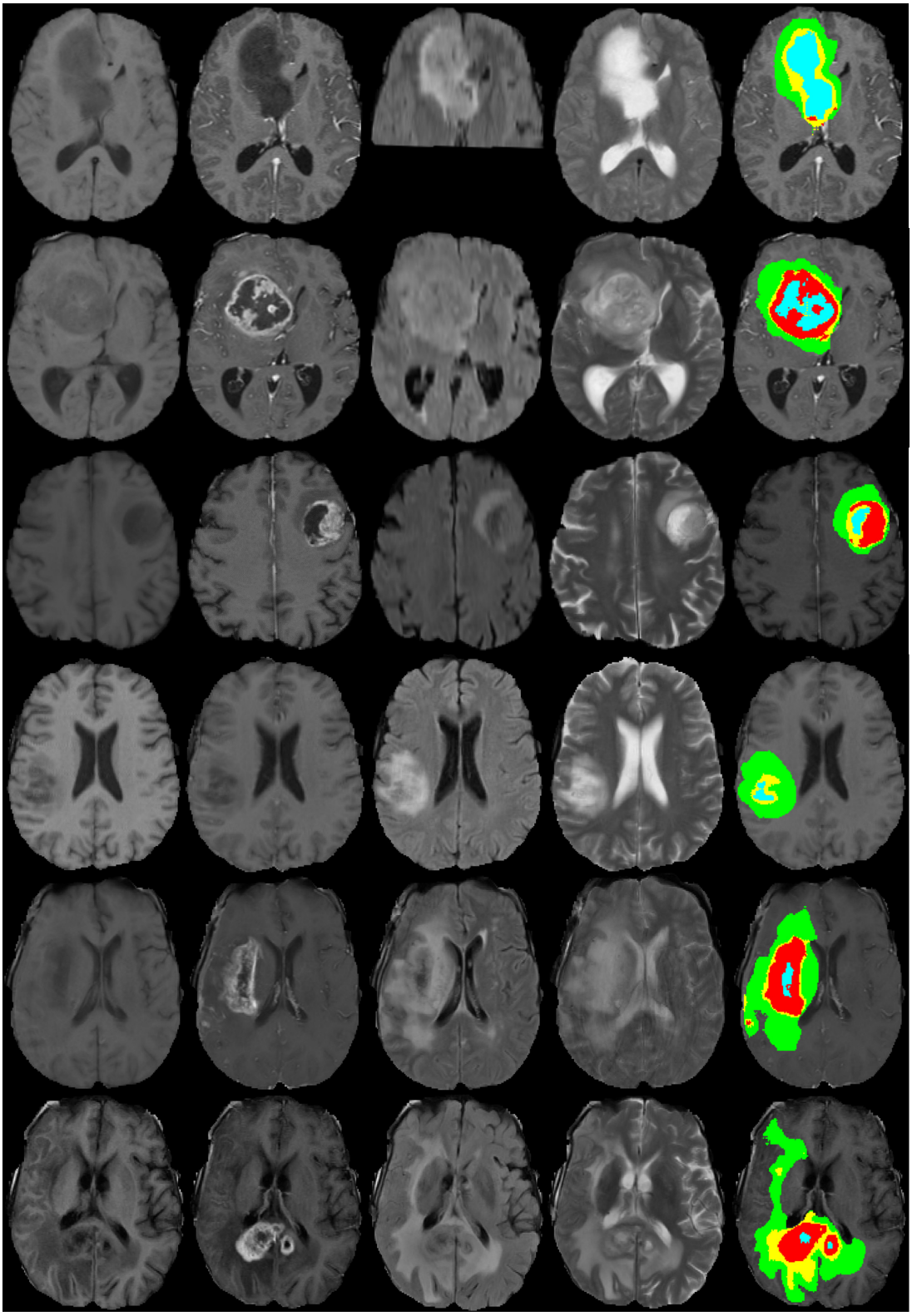


Figure 4.10: Segmentation examples on the BRATS 2015 Challenge data set. From left to right: T1, T1c, FLAIR, T2 and obtained segmentation. Colors in the segmentations represent: blue - necrosis, green - edema, yellow - non-enhanced tumor, red - enhanced tumor.

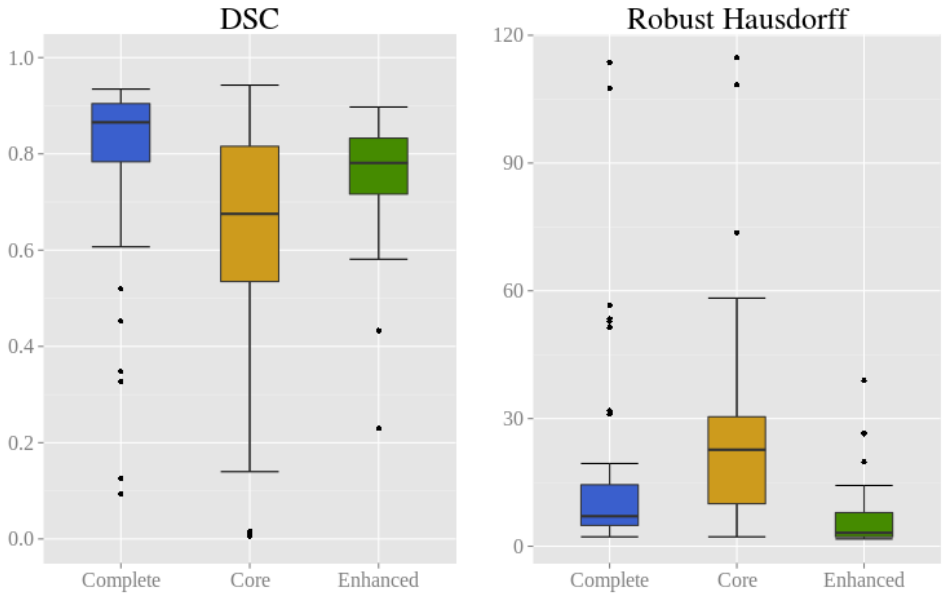


Figure 4.11: Boxplots of DSC and Robust Hausdorff Distance obtained using the Challenge data set of BraTS 2015.

4.2 Hierarchical segmentation with Fully Convolutional Networks

In this section, we tackle some of the disadvantages of the Classification CNN by conceiving and adopting a FCN architecture. In these networks, fully-connected layers are changed by convolutional layers with 1×1 kernels, which allows dense predictions by segmenting a full patch, instead of classifying just the central voxel.

Although with the previous Classification CNN we achieved state-of-the-art results at the time, this kind of approach has some disadvantages:

1. It is computationally demanding. We need one forward pass to classify each voxel in the image.
2. It has many parameters. The parameters of the fully-connected layers account for most of the parameters of the network (roughly 79% in our deeper network). This increases the chances of overfitting.
3. It has less context into account. Patches were relatively small, since bigger ones would require even more parameters. This can contribute for a higher number of false positive detections.

Contrasting to Classification CNN, FCN have the following advantages, or improvements:

1. It is computationally more efficient. In one forward pass, we can classify a patch of voxels/pixels.
2. It allows to drastically decrease the number of learnable parameters, by changing the fully-connected layers into convolutional layers with 1×1 kernels.
3. It can take more context into account. Since it is much more efficient, we can have bigger patches as input.
4. It takes into account more training data in each step, since it computes the loss in relation to a full patch.

Furthermore, in this section we explore a hierarchical segmentation approach. Therefore, we employ a binary FCN for segmenting the whole tumor. Then, a bounding box is defined as a ROI. Finally, a multi-class FCN segments all the tumor tissues inside the ROI. This approach was motivated by the fact that brain tumor segmentation is a highly imbalanced problem, hence, the ROI is helpful in that regard. Although other previous works used a hierarchical approach in the context of brain tumor segmentation, to the best of our knowledge this was the first time that it was applied to FCNs. This kind of approach was, then, largely followed by participants of the 2017 edition of the BRATS Challenge.

So, the contributions in this section are: 1) we study and develop an encoder-decoder-based FCN and apply it to brain tumor segmentation, and 2) we employ hierarchical brain tumor segmentation using FCN.

This section is based on the following publication:

- Pereira, Sérgio, et al. "On hierarchical brain tumor segmentation in MRI using fully convolutional neural networks: A preliminary study." IEEE 5th Portuguese Meeting on Bioengineering (ENBENG), 2017.

4.2.1 Introduction

Although brain tumors are not among the most common type of cancers, their mortality rate is quite alarming. Gliomas are originated from the glial cells of the brain, being the most frequent of these neoplasms. Magnetic Resonance Imaging (MRI) is the most used imaging modality to assess brain tumors. Segmentation of these lesions allows a better diagnosis and treatment planning. However, although MRI provides good tissue contrast, it also creates a lot of data (due to its 3D nature). Thus, manually labeling each voxel for volumetric delineation is infeasible in clinical practice and physicians normally just take rough measures from the tumor, such as diameter. So, these measures are prone to intra- and inter-rater variability, and use the rich information provided by the MRI image sub-optimally (Menze et al., 2015; Bauer et al., 2013). There is, thus, a need for reliable automatic and semi-automatic segmentation tools. However, brain tumor segmentation is a challenging task due to its variable size, shape, and location, as well as its heterogeneous tissue content. Moreover, MRI image analysis may be difficult since intensities can vary a lot among acquisitions, even for the same subject and scanner (Nyúl et al., 2000).

Discriminative machine learning-based methods for brain tumor segmentation are quite successful (Menze et al., 2015). In these approaches, there are two main stages: feature computation and classification. The first approaches of this kind relied on hand-crafted features followed by a classifier (Pinto et al., 2015a; Meier et al., 2014b; Bauer et al., 2011; Pinto et al., 2015b; Tustison et al., 2015; Zikic et al., 2012). Some of the features used for brain tumor segmentation are based on appearance (Pinto et al., 2015a; Meier et al., 2014b; Bauer et al., 2011; Pinto et al., 2015b), context (Pinto et al., 2015a; Zikic et al., 2012; Meier et al., 2014b; Bauer et al., 2011; Pinto et al., 2015b), edge detection (Pinto et al., 2015a; Meier et al., 2014b), or brain symmetry (Tustison et al., 2015; Meier et al., 2014b). After feature computation, most of these methods use ensembles of randomized trees (Pinto et al., 2015a; Meier et al., 2014b; Pinto et al., 2015b; Tustison et al., 2015; Zikic et al., 2012). Contrasting with the previous methodologies, more recent proposals based on Convolutional Neural Networks (CNN) learn the features automatically from data (Lyksborg et al., 2015; Pereira et al., 2016a, 2015; Havaei et al., 2017; Kamnitsas et al., 2017b). These methods have the advantage of bypassing the feature engineering stage, which may require high domain knowledge. Additionally, features and classifier are trained end-to-end. Lyksborg et al. (2015) and Pereira et al. (2016a, 2015) employed CNN using fully-connected layers as classifier, performing segmentation by classifying the central voxel of each image patch. These methods have the disadvantage of considering each voxel as independent entities, which may lead to misclassified small clusters. Havaei et al. (2017) proposed a cascaded CNN architecture where the output of one network is fed into the following one. With this, the authors include a dependence on the output labels of the previous network, as an alternative of a Conditional Random Field for segmentation regularization. Additionally, the fully-connected layers are changed to convolutional layers with 1×1 kernels that allow more efficient inference. Kamnitsas et al. (2017b) developed a 3D multi-scaled CNN. As Havaei et al. (2017), the last layer is convolutional, but a set of voxels is simultaneously segmented. Fully-convolutional networks (FCN) (Shelhamer et al., 2016; Ronneberger et al., 2015) employ a contracting pathway responsible for aggregating features into more complex representations and increase the effective field of view. Then, the expanding pathway upsamples the features maps to the original pixel space in order to

semantically segment a patch of voxels. In this way, FCN are very efficient and the segmentations are more regularized than pixel-by-pixel approaches.

Since brain tumors are made of several tissues, it is possible to devise hierarchical segmentation strategies where first the complete tumor is segmented, and later the tumor sub-tissues are differentiated, as in (Bauer et al., 2011; Meier et al., 2014b; Lyksborg et al., 2015). Neural Networks may be impaired by high class imbalance. In brain tumor segmentation this is particularly problematic, as the tumor tissues encompass just a small portion of the brain. Hierarchical segmentation may alleviate this problem by considering first the whole tumor, and then constraining the segmentation to a region of interest. In this paper, we study a hierarchical segmentation approach using FCN to segment patches of voxels. The remaining of this paper is organized as follows. In Section 4.2.2 we present the methodology. Results and discussion are shown in Section 4.2.3. Finally, in Section 4.2.4, we outline the main conclusions.

4.2.2 Materials and Methods

4.2.2.1 Dataset

The proposed approach was validated in the Brain Tumor Segmentation Challenge (BRATS) dataset (Menze et al., 2015). For each patient, there are available four MRI sequences: T1-weighted (T1), post-contrast gadolinium enhanced T1 (T1c), T2-weighted (T2), and FLAIR. All images are already co-registered to the T1c, skull-stripped, and interpolated to isotropic resolution of $1\ mm$. The manual segmentation of the training set distinguishes four tumor tissue classes: necrosis, edema, non-enhancing tumor, and enhancing tumor. The 2013 Challenge set consists of 10 subjects without publicly available manual segmentations. Thus, evaluation is performed by the online platform[6]. The segmentation task evaluates three structures: enhancing tumor, core (necrosis + non-enhancing + enhancing), and complete tumor (all tumor tissues).

4.2.2.2 Segmentation Method

The hierarchical segmentation approach has two main stages: whole tumor binary segmentation, followed by intra-tumoral tissue segmentation. Thus, after identifying the complete lesion, we use it as region of interest for the next CNN. Prior to each network, we pre-process the MRI sequences; additionally, post-processing is only used after the binary whole tumor segmentation.

Pre- and post-processing Similarly to Pereira et al. (2016a), we first pre-process the MRI images by standardizing their histograms (Nyúl et al., 2000). Then, 2D patches are extracted in the axial plane, and normalized with zero mean and unit variance in each sequence. After the whole tumor binary segmentation stage we also impose a volumetric constrain on connected components, similarly to Pereira et al. (2016a).

Fully Convolutional Network The implemented FCN is based on the U-net proposed by Ronneberger et al. (2015). Thus, there is a contracting and an expanding pathway. In the contracting part, the feature

[6] https://www.smir.ch/BRATS/Start2013

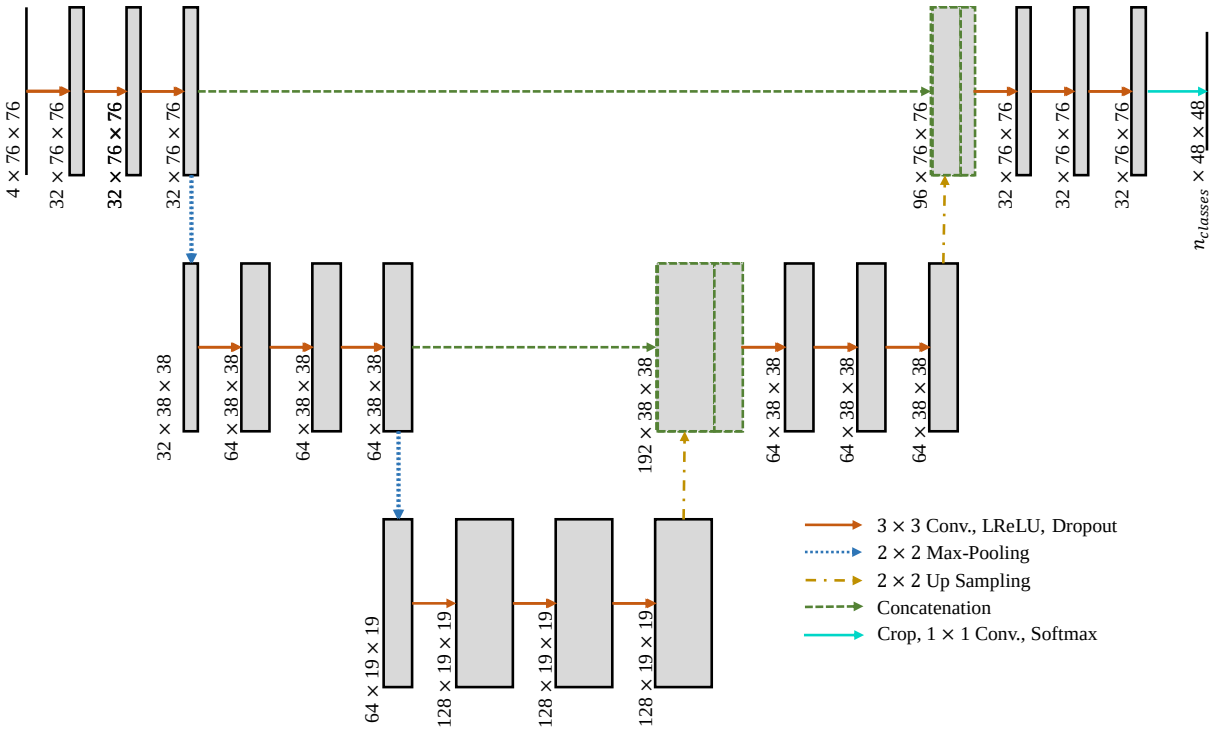


Figure 4.12: Architecture of the implemented FCN.

maps are subjected to pooling in order to summarize neighboring features and create higher level representations. Then, the expanding path upsamples the feature maps to the original resolution, in order to have correspondence between features and each pixel. However, when up-scaling deeper representations to the pixel space, we may lose the finer details, such as small objects, or sharper edges. Thus, lower level feature maps are concatenated with the up-scaled ones. This is then followed by convolutional layers that learn how to combine the data. The implemented FCN architecture is presented in Fig. 4.12. As non-linear activation, instead of Rectifier Linear Units (ReLU), we use the Leaky ReLU (LReLU) (Maas et al., 2013), defined as $f(x) = \max(0, x) + \alpha \min(0, x)$, where x is the input and α is the leakiness parameter; we set $\alpha = 0.3$. Contrasting to ReLU that imposes 0 for negative inputs, LReLU has a negative component defined by α that allows the gradients to propagate better. Additionally, after each convolutional layer with 3×3 kernels and LReLU activation, we use Dropout (Srivastava et al., 2014) with probability of $p = 0.2$. It works by removing random nodes with probability p in each training step. In this way, it acts as training regularization, preventing overfitting.

4.2.2.3 Setup

We used the FCN architecture in Fig. 4.12 for both the binary whole tumor segmentation and the multi-class intra-tumoral segmentation networks. The only difference is in the number of kernels of the last layer, which is related with the number of classes. We used 20 subjects during the training stage: 18 for training and 2 for validation. For selecting the training samples, we adopt the sampling scheme of Kamnitsas et al. (2017b), by selecting patches that are centered on normal or tumor tissue with probability of 0.5. No artificial data augmentation was used. The Cross Entropy was defined as the loss function to be minimized in order to train the CNN. To that end, we used Stochastic Gradient Descent with learning

rate of 0.01 as optimizer.

The neural networks were implemented using Keras (Chollet, 2015) with Theano backend, and cuDNN 5.1.

4.2.3 Results and Discussion

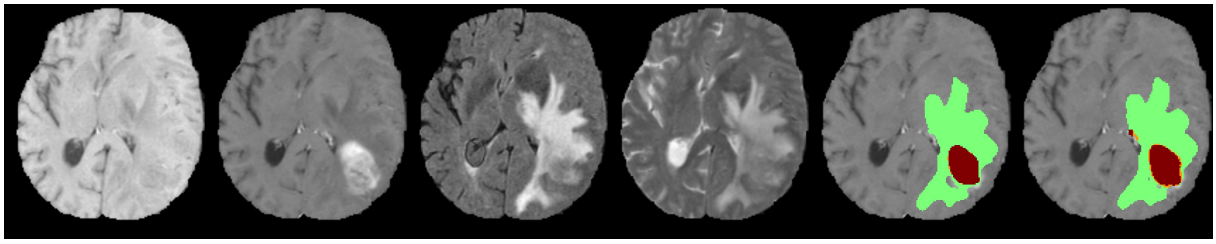
The described method was used to segment the BRATS 2013 Challenge dataset. As baseline, for comparison reasons, we also trained another FCN with the same architecture depicted in Fig. 4.12, but following a single stage (segmenting all tissues at once) approach. We report the results in Table 4.9, using the metrics Dice Score Coefficient (DSC), Positive Predictive Value (PPV), and sensitivity. Additionally, we conducted a statistical significance test with the paired Wilcoxon Signed-Rank Test (significance level: 0.05). We can observe from Table 4.9 that the complete tumor segmentation results improved in all metrics, when using the hierarchical approach. This reinforces the idea that a first pre-segmentation of the whole tumor is beneficial also for FCN. In the hierarchical approach, we separate the segmentation task into two simpler sub-tasks. First, the binary complete tumor FCN just needs to learn two classes distribution and how to detect the whole tumor. In this way, it can focus on MRI sequences where the complete tumor and its outer border is more conspicuous, such as the FLAIR and T2. Then, the second network needs to learn a multi-class segmentation task. Nevertheless, it does not need to learn how to differentiate perfectly the whole tumor from the normal tissue, since it just focus on a region of interest. We applied the same post-processing to both the binary stage of the hierarchical approach and the single stage approach. The segmentation of both the core and the enhancing tumor benefited from the hierarchical approach, too. We can observe from Table 4.9 the improvement in DSC and, especially, in sensitivity. The DSC in the enhancing tumor, and the sensitivity of core and enhancing, were found to be statistically different from the results of the single stage approach ($p - value < 0.05$). To some extent, this improvement in sensitivity appears at the expense of decreasing core and enhancing PPV; still, these decrements were not found to be statistically significant from the single stage approach, and the absolute gain in sensitivity is higher than the loss in PPV. Since there was an improvement in DSC, this means that more core tissues are being correctly detected and segmented. This may be explained by a better data balance of training a CNN on a region of interest, which allows it to focus more on tumor tissues, and less on an excellent delineation of the whole tumor. Brain tumor segmentation suffers from high data imbalance, as the tumor occupies just a small fraction of brain, and tumor sub-tissues are even less represented.

Segmentation examples from two subjects of BRATS 2013 Challenge dataset obtained with the single-stage and hierarchical approaches are shown in Fig. 4.13. We can observe that the tumor appears to be well delineated, and the segmentation is smooth. Moreover, the single-stage approach can hardly detect necrosis in tumor core (segmenting it mostly as edema), whereas the hierarchical approach identifies it better.

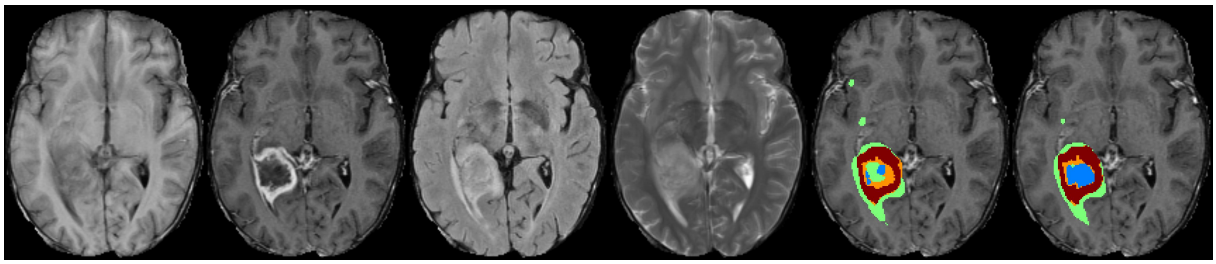
All tests were conducted on a desktop equipped with a NVIDIA GeForce GTX 970 GPU, an Intel Core i7-3930k 3.2GHz (x12) processor, 48 GB of RAM, and running Linux Mint 18 OS. A full MRI image segmentation with the hierarchical approach takes approximately 40 seconds.

Table 4.9: Results obtained in BRATS 2013 Challenge. We compare the hierarchical approach with the single stage approach (segmenting all tissues at once). Bold metrics were found to have $p-value < 0.05$ when comparing the two approaches.

	DSC			PPV			Sensitivity		
Method	Complete	Core	Enhancing	Complete	Core	Enhancing	Complete	Core	Enhancing
Single stage	0.83 ± 0.06	0.75 ± 0.17	0.72 ± 0.10	0.77 ± 0.10	0.80 ± 0.19	0.76 ± 0.11	0.91 ± 0.05	0.73 ± 0.19	0.72 ± 0.18
Hierarchical	$\mathbf{0.85 \pm 0.05}$	0.76 ± 0.17	$\mathbf{0.74 \pm 0.08}$	0.80 ± 0.09	0.78 ± 0.20	0.74 ± 0.11	0.92 ± 0.06	$\mathbf{0.79 \pm 0.17}$	$\mathbf{0.78 \pm 0.15}$



(a)



(b)

Figure 4.13: Examples of segmentations obtained in BRATS 2013 Challenge dataset with the subjects: a) 0308, and b) 0310. From left to right, we show the T1, T1c, FLAIR, and T2 sequences, followed by the tumor segmentation obtained by the single stage and hierarchical approaches, respectively. Each color represents a tumor class: green – edema, blue – necrosis, orange – non-enhancing tumor, and dark red – enhancing tumor.

4.2.4 Conclusion

Brain tumor segmentation is a multi-class problem, where we need to discriminate several tumor tissues. Thus, methods need to learn how to correctly detect the whole tumor and its sub-tissues. Previously, hierarchical approaches where the whole tumor is firstly binary segmented, and later sub-divided, were proposed. In this paper, we study a hierarchical brain tumor segmentation approach in the context of FCN. These networks allow to segment a full patch, achieving better regularization. It was found that this approach is beneficial for brain tumor segmentation using FCN, not only for detecting the tumor as a whole, but also for a better delineation of the core and the enhancing tumor. The improvements over the single stage approach were found to be statistically significant in DSC of complete and enhancing tumor, as well as sensitivity of core and enhancing tumor.

4.3 Adaptive feature recombination and recalibration

In the previous section we developed a hierarchical brain tumor segmentation approach using FCNs. In this section, we employ the same approach, but we explore adaptive feature recombination and recalibration. Additionally, our networks in this section are trained with MRI images from gliomas of both grades mixed.

Feature recalibration for classification consists in modeling the channel-wise relationships of feature maps and suppress the least informative ones (Hu et al., 2018). However, this approach, as proposed for object recognition problems, suppresses a feature map as whole. Hence, this is not adapted for semantic segmentation with FCN, where certain regions of the feature map may be relevant for the voxels in the same location, while others may be less important. Therefore, we propose adaptive spatial-wise feature recalibration.

Feature recombination consists in combining the feature maps using convolutional layers with 1×1 kernels. Before, this was employed for reducing the number of feature maps. However, we found that mixing the feature maps into a higher dimension followed by restoration of the number of channels leads to more complex features.

Therefore, the contributions in this section are: 1) we propose feature recombination by means of linear expansion and compression. 2) We adapt and propose feature recombination for FCNs. 3) We show experimentally that whole map feature recalibration is not well-suited for semantic segmentation with FCNs. Finally, 4) we evaluate these methodologies in brain tumor segmentation, and achieve competitive single-model results in BRATS 2017, and state-of-the-art results in BRATS 2013.

This section is based on the following publication:

- Pereira, Sérgio, et al. "Adaptive feature recombination and recalibration for semantic segmentation: application to brain tumor segmentation in MRI." Medical Image Computing and Computer Assisted Intervention (MICCAI), 2018.

4.3.1 Introduction

Brain tumor segmentation plays an important role during treatment planning and follow-up evaluation. But, it is time-consuming and prone to inter- and intra-rater variability. Therefore, automatic and reliable methods are desirable. However, brain tumor segmentation is a challenging task, due to their irregular shape, appearance, and location (Pereira et al., 2016a; Zhao et al., 2018). Recently, CNN-based approaches have achieved state-of-the-art results (Pereira et al., 2016a; Zhao et al., 2018; Kamnitsas et al., 2018). With enough data, CNNs can learn complex patterns, such as brain tumor attributes, which are, otherwise, difficult to capture by feature engineering.

Despite being used in many applications, many CNN developments are first evaluated in image classification. In VGGNet (Simonyan and Zisserman, 2014) it was shown that replacing a layer with large kernels by blocks of several layers with 3×3 kernels results in deeper and more powerful CNNs. Later, He et al. (2016) proposed residual learning using identity-based skip connections that allow better gradient

flows and training of very deep CNNs. Other studies explored the recombination of feature maps, either by compression with convolutional layers with 1×1 kernels (Lin et al., 2013), or by dividing a stack of feature maps into smaller groups with grouped convolutions (Xie et al., 2017). More recently, Hu et al. (2018) proposed feature map recalibration with the Squeeze-and-Excitation (SE) block. This block is inspired by the intuition that not all feature maps are informative for all classes. Therefore, the SE block learns how to adaptively suppress the least discriminative feature maps (recalibration). They showed that the simple addition of this block to state-of-the-art CNNs increased their representational power.

Semantic segmentation is one of the domains where CNNs have been pervasively used. Although one can use conventional CNNs with fully connected layers (Pereira et al., 2016a), FCNs (Ronneberger et al., 2015) are arguably one of the most important advancements regarding CNNs for semantic segmentation. In these architectures, fully-connected layers are replaced by convolutional layers, usually, with 1×1 kernels. In this way, a set of voxels from an image patch can be efficiently classified in just one forward pass. In FCNs, there is a direct spatial correspondence between units in the feature maps and the classified voxels. Most of the advancements in CNN design can be easily incorporated into FCN. For instance, the principles of VGGNet (Simonyan and Zisserman, 2014) and residual learning (He et al., 2016) were incorporated in (Pereira et al., 2016a) and (Kamnitsas et al., 2018), respectively. However, although the SE block has the attractive property of re-calibrating feature maps, it was conceived to weight whole feature maps, which is not optimal for FCN. Since there is a spatial correspondence between units in feature maps and the voxels, it is desirable to emphasize or suppress certain regions of the feature maps, instead of the whole feature map.

In this paper, we explore the recombination and recalibration of feature maps. In recombination, instead of reducing the feature maps number only, we employ linear expansion followed by compression for mixing the information. Additionally, we study how to incorporate recalibration into FCN. In SE block, global average pooling captures the whole contextual information in a feature map. Instead, we argue that dilated convolution (Yu and Koltun, 2016) is better suited for the recalibration block in FCN. Hence, the contribution of this paper is threefold. First, we propose recombination of feature maps by linear expansion and compression. Second, we explore feature maps recalibration in the context of FCN. We observe that the original SE block is not optimal for FCN, and we propose a better-suited alternative. Third, we evaluate our proposal on brain tumor segmentation, using publicly available data.

4.3.2 Methods

We follow a hierarchical FCN-based brain tumor segmentation approach. Thus, we start by roughly segmenting the whole tumor with a binary FCN (WT-FCN). Using this segmentation, we define a bounding box around the tumor with a margin of 10 extra voxels in each side. Finally, a second multi-class FCN (MC-FCN) is responsible for segmenting the multiple tumor structures inside the region of interest (ROI). The proposed FCNs are inspired by an encoder-decoder architecture with long skip connections (Ronneberger et al., 2015). The input for the FCNs are image patches extracted from all the available MRI sequences. In this section, we start by defining the baseline FCNs; then, we present the proposed recombination and recalibration (RR) block. This block is evaluated in the more challenging multi-class segmentation prob-

lem. The WT-FCN is fixed across all experiments to isolate and make it easier to compare improvements introduced by the RR block.

4.3.2.1 Baseline segmentation approach

The architecture of the 3D WT-FCN is depicted in Fig. 4.14. We used both regular blocks of convolutional layers, and blocks with residual connections and pre-activation (He et al., 2016). This network segments 3D patches, with the three pooling layers providing a large field of view. These two characteristics contribute for reducing the number of voxels with false positive tumor detections. The baseline MC-FCN architecture can be perceived from Fig. 4.15(a), by not considering the RR block. We design the MC-FCN as a 2D network, as a proof of concept to evaluate the proposed component, which makes it computationally cheaper than the WT-FCN. Hence, 2D image patches are extracted in the axial plane. Additionally, we observed no benefits from using residual connections, or from being as deep as the WT-FCN.

4.3.2.2 Recombination and recalibration

We propose recombination and recalibration of feature maps as complementary operations. Recombination consists in mixing the information across feature maps channels to create new combined features. In the past, convolutional layers with 1×1 kernels were proposed as cross channel parametric pooling (Lin et al., 2013) to decrease (compress) the number of feature maps. Also, bottleneck blocks in ResNet compress the channels, process them, and expand to more channels. Instead, we propose linear recombination of feature maps to increase the number of feature maps (expansion), followed by compression to the original number. This operation is done by convolutional layers with 1×1 kernels (Fig. 4.15(b)). Experimentally, we found an expansion factor of 4 to work well. Linear recombination of units in a given spatial location of the feature maps results from the weighted sum of the units in the same location of all feature maps. Hence, expansion combines features into a higher dimension, while compression learns how to compress features and suppress the least discriminative ones.

We propose the RR block (Fig. 4.15(c)) that combines both feature map recombination and recalibration. Feature map recalibration with the SE block, as proposed in (Hu et al., 2018), is shown in Fig. 4.15(c) – SE block. First, global average pooling summarizes each feature map into its average value to capture contextual information. Then, two fully connected (FC) layers[7] capture cross-channels relations. The first one is a compression layer with a factor of r, followed by ReLU activation. The second FC layer restores the original dimension, and is followed by the sigmoid activation function. Finally, this vector is channel-wise multiplied with the input feature maps, i.e., each feature map is multiplied by a corresponding scalar value, resulting in each feature map being scaled as a whole. Ideally, less discriminative feature maps are suppressed. This approach was shown to improve learning in image classification. In this problem, a feature map may have a strong response for a given class or subset of classes. However, in semantic segmentation with FCN, a patch of voxels is segmented at once. So, there is a correspondence between segmented pixels and units in feature maps. In this scenario, some regions of the feature maps may

[7] Equivalently implemented as convolutional layers with 1×1 kernels.

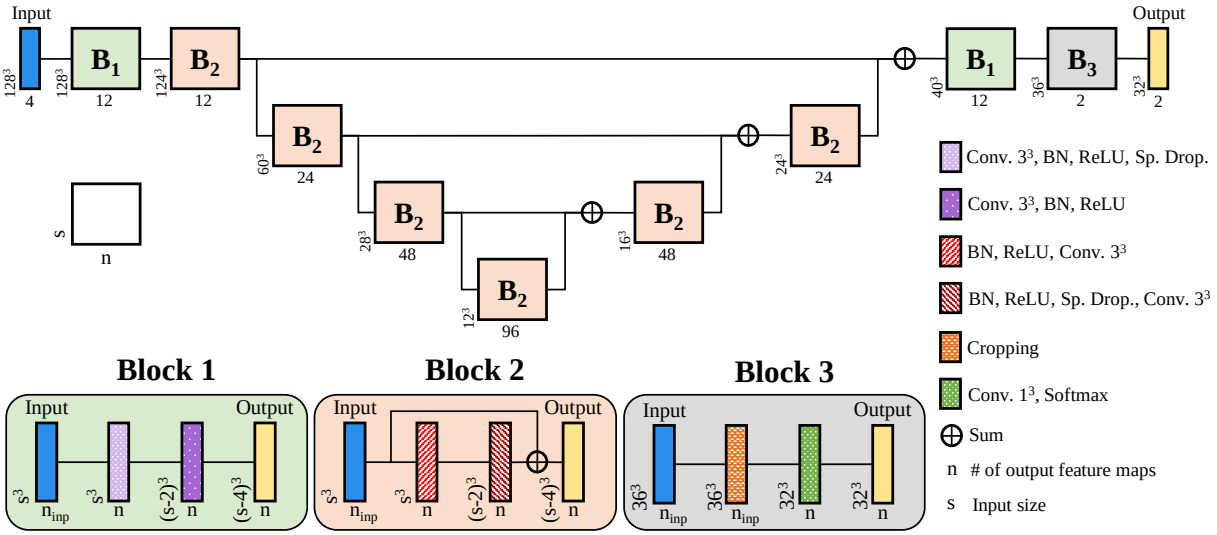


Figure 4.14: Architecture of the WT-FCN. Downsampling is obtained by max-pooling. We use upsampling to increase the feature maps size, and $1 \times 1 \times 1$ convolutional layers to adjust the number of feature maps, before addition. BN stands for batch normalization, and Sp. Drop. for spatial dropout.

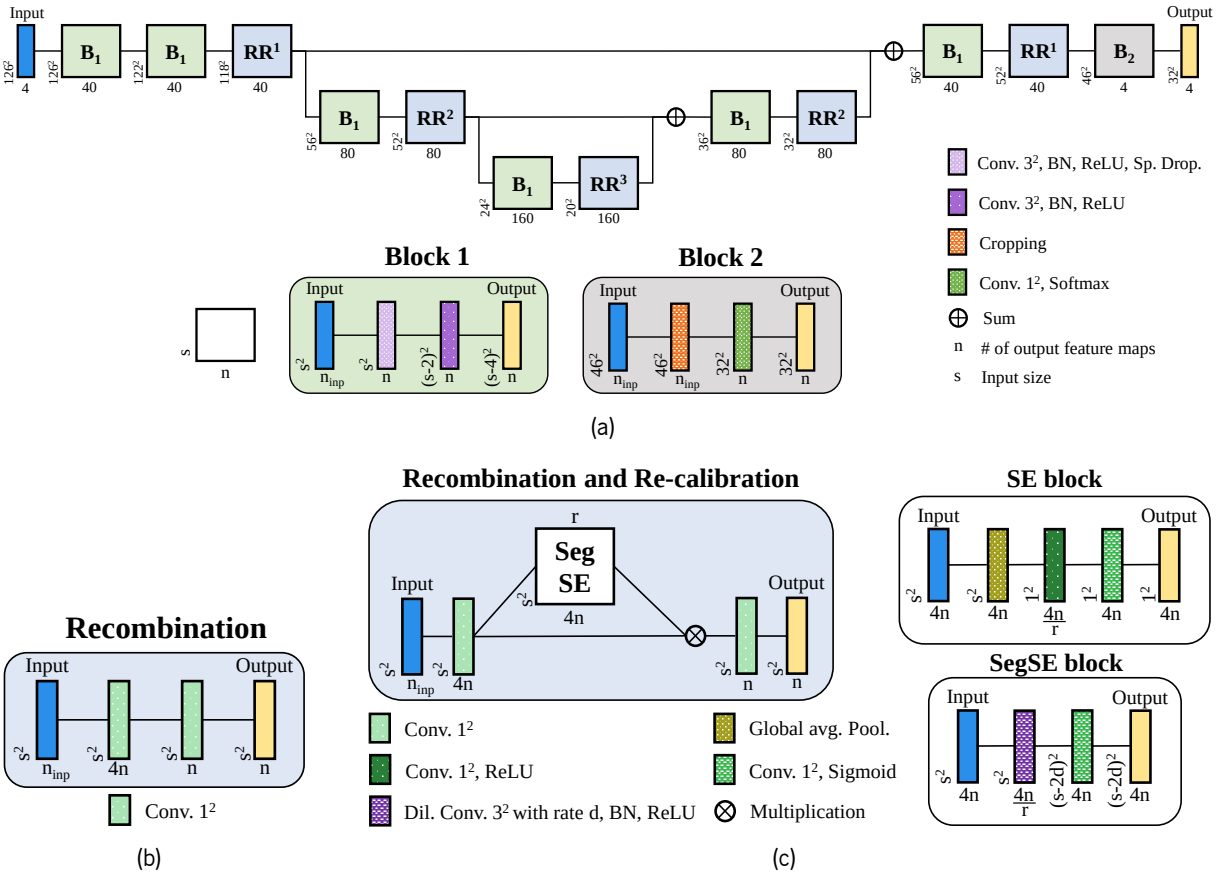


Figure 4.15: Architecture of the MC-FCN. a) Overview of the architecture with the RR block. Depicted input sizes correspond to the RR block with SegSE. Downsampling is obtained by max-pooling. We use upsampling to increase the feature maps size, and 1×1 convolutional layers to adjust the number of feature maps, before addition. b) Recombination block. c) RR block, and the SE and SegSE blocks.

have strong activations that are relevant for the structure that is being segmented in that spatial location. Hence, the SE block may be not optimal for semantic segmentation, since it collapses the whole feature

map into a single value, regardless of the regions. Thus, the proposed RR block includes a segmentation adapted SegSE block. A straightforward approach for adapting the SE block for semantic segmentation is by simply removing the global average pooling layer. In this way, the spatial correspondence among units and voxels is maintained. However, this is also not optimal, since contextual information is important to evaluate the spatial importance of a given feature. In preliminary experiments, this approach resulted in worse performance. Therefore, we propose our SegSE block (Fig. 4.15(c)) that uses a convolutional layer with 3×3 kernels with dilation d for context aggregation. Simultaneously, this layer is responsible for the compression stage. Experimentally, we found that the best dilation rates depend on the resolution of the feature maps. This is due to the fact that deeper layers already have a larger field of view. Hence, we set the rate in $\{RR^1, RR^2, RR^3\}$ (c.f. Fig. 4.15(a)) to $\{3, 2, 1\}$. In preliminary experiments, we evaluated using spatial average pooling followed by convolutional and transposed convolutional layers, but, we obtained worse performance. The reason is probably due to the checkerboard artifacts that appear with this combination of layers, but not with dilated convolution.

4.3.3 Experimental Setup

We evaluate the proposed blocks in the Brain Tumor Segmentation Challenge (BRATS) 2017 and 2013 databases (Bakas et al., 2017; Menze et al., 2015). BRATS 2017 has two publicly available datasets: Training (285 subjects) and Leaderboard (46 subjects). In BRATS 2013 we use Training (30 subjects) and Challenge (10 subjects). For each subject, there are four MRI sequences available: T1, post-contrast T1 (T1c), T2, and FLAIR. All images are already interpolated to $1mm$ isotropic resolution, skull stripped, and aligned. Only the Training sets contain manual segmentations. In BRATS 2017 it is distinguished three tumor regions: edema, necrotic/non-enhancing tumor core, and enhancing tumor. In BRATS 2013 the manual segmentations have necrosis and non-enhancing tumor separately, although we fuse these labels to be similar to BRATS 2017. Evaluation is performed for the whole tumor (all regions combined), tumor core (all, excluding edema), and enhancing tumor. Since annotations are not publicly available for 2017 Learderboard and 2013 Challenge, the evaluation is computed by the CBICA IPP and SMIR online platforms[8]. The development of the RR block was conducted in the larger BRATS 2017 Training set, which was randomly divided into training (60%), validation (20%), and test (20%)[9]. However, networks tested in BRATS 2013 Challenge were trained in the 2013 Training set.

Image pre-processing included bias field correction, and standardization of the intensity histograms of each MRI sequence, as in (Pereira et al., 2016a). During training, we use the crossentropy loss, the Adam optimizer with learning rate of 5×10^{-5}, weight decay of 1×10^{-6}, and spatial dropout probability of 0.05. Since we used convolution without padding, during multiplication or sum of feature maps with different sizes, we cropped the center part of the biggest one. The compression factor r in the SE and SegSE blocks was set to 10. Data augmentation included sagittal flipping and random rotations of 90°. For training the binary whole tumor FCN, all tumor regions in manual segmentations were fused into a single label. All the hyperparameters were found using the validation set, before evaluation in the test set.

[8] `https://ipp.cbica.upenn.edu/` and `https://www.smir.ch/BRATS/Start2013`

[9] Subjects id in each set are available: `https://github.com/sergiormpereira/rr_segse`.

The FCNs were implemented using Keras and Theano.

Metrics provided by the online evaluation platforms differ in BRATS 2017 and 2013. Hence, in BRATS 2017 we use Dice and the 95th percentile of the Hausdorff Distance (HD_{95}). In BRATS 2013, the online platform computes Dice, Sensitivity, and Positive Predictive Value (PPV).

4.3.4 Results and Discussion

We evaluate the effect of recombination and recalibration of feature maps using the SegSE block in the test set (20% of BRATS 2017 Training), and compare it with the baseline and RR with SE block. Quantitative results are presented in Table 4.10, and segmentation examples in Fig. 4.16. When we include the recombination by expansion followed by compression stage to the baseline FCN, we observe that Dice improves in all tumor regions. Although this improvement is negligible in the enhancing region, it is substantial in the whole tumor and in the tumor core. In fact, it achieves the highest Dice for tumor core of all the blocks. However, the HD_{95} is higher in all classes, when compared with the baseline.

In Table 4.10 we can find results for the recalibration stage, when joined with recombination to form

Table 4.10: Results (average) obtained in the test set (20% of BRATS 2017 Training). We evaluate recombination (Recomb.) of feature maps, and RR using both SE and SegSE blocks. Bold results show the best score for each tumor region.

	Dice			HD_{95}		
Method	**Whole**	**Core**	**Enh.**	**Whole**	**Core**	**Enh.**
Baseline	0.857	0.739	0.682	8.645	10.761	6.672
Baseline + Recomb.	0.865	**0.769**	0.687	9.720	11.453	7.790
Baseline + RR SE	0.859	0.756	0.672	8.939	13.306	7.319
Baseline + RR SegSE	**0.866**	0.766	**0.698**	**8.475**	**10.513**	**6.131**

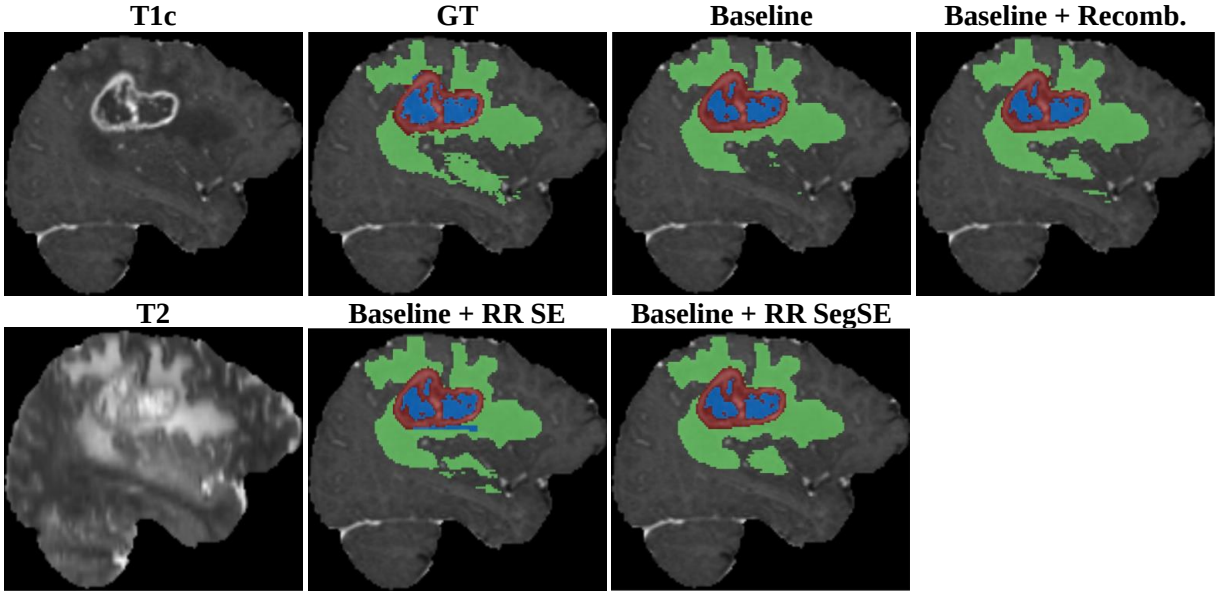


Figure 4.16: Examples of the segmentation obtained with each of the evaluated RR blocks. The colors in segmentations mean: green – edema, blue - tumor core, and red – enhancing tumor.

the RR block. We observe that the SE block, proposed in (Hu et al., 2018), leads to worse Dice, when compared to the baseline with recombination layers. Actually, the Dice of the enhancing tumor is even lower than the baseline. This may be due to enhancing region usually being a smaller part of the whole tumor volume. Additionally, finer details are needed to define this region, hence its contribution to a whole feature map response may be less strong than the other tumor regions and end up being suppressed by the SE block. Therefore, we conclude that the SE block, acting as whole feature map recalibration is not optimal for segmentation. Finally, it is possible to observe from Table 4.10 that the RR block with the proposed SegSE recalibration achieves the best scores, both in Dice (excepting core by a small margin) and HD_{95}. We note that the SegSE block is the only approach that substantially improves the Dice of enhancing tumor over the baseline. Moreover, the HD_{95} suggests that besides achieving good overlap scores, it also obtains the best contour definition. The SegSE stage comes at the cost of more parameters. In order to evaluate if its performance is due to these extra capacity, we proportionally increased the width of the baseline, such that its parameters number becomes similar to the network with RR + SegSE. The results obtained with this larger network in terms of Dice/HD_{95} were 0.852/9.049, 0.751/10.647, and 0.678/7.065 for the complete, core, and enhancing regions, respectively. So, the RR SegSE block improvements are due to better learning, and not directly to the higher capacity.

We compare with the state of the art in BRATS 2017 Leaderboard and BRATS 2013 Challenge in Tables 4.11 and 4.12, respectively. In BRATS 2017, most of the top performing methods are ensembles of FCN. In principle, taking a CNN or FCN and building an ensemble will certainly lead to better results. Since we are evaluating the effect of the SegSE block, we need to assess it in a single model. So, we compare our results with other single CNN approaches, such as Islam and Ren (2018) and Jesson and Arbel (2018), for the sake of fairness. Nevertheless, we present results obtained by the multi-model and multi-training settings ensemble proposed by Kamnitsas et al. (2018), the winner of BRATS 2017 Challenge. In the single network approaches, Islam and Ren (2018) employed a hypercolumns-inspired CNN. Jesson and Arbel (2018) used a FCN with multiple prediction layers and loss functions in different scales. Additionally, the authors employed a learning curriculum to deal with class imbalance. From Table 4.11, we observe that the baseline achieves competitive results, when compared with the single-model approaches. The Dice is comparable with Islam and Ren (2018), while the HD_{95} scores are smaller. Regarding the RR SegSE block, we confirm that it improves the baseline performance. Indeed, the results are competitive with Jesson and Arbel (2018), with better Dice for core and enhancing regions, and HD_{95} in the enhancing region. In BRATS 2013 Challenge (Table 4.12), the proposed FCN with the RR SegSE block improves over the baseline, again. The other compared methods are all recent and top performing CNN-based approaches. Pereira et al. (2016a) uses a plain CNN with fully-connected layers. Shen et al. (2017) uses a FCN enhanced by input symmetry maps and a boundary-aware loss function. Zhao et al. (2018) also proposes a FCN followed by a Conditional Random Field trained as Recurrent Neural Network and a sophisticated post-processing stage. We note that the proposed method achieves the highest Dice and Sensitivity scores. In fact, the baseline with the RR SegSE block is ranked 1^{st} by the online evaluation platform.

Table 4.11: Results (average) obtained in BRATS 2017 Leaderboard set. Bold results show the best score for each tumor region. Underlined scores are the best among single-model approaches (excluding Kamnitsas).

	Dice			HD_{95}		
Method	**Whole**	**Core**	**Enh.**	**Whole**	**Core**	**Enh.**
Islam and Ren (2018)	0.876	0.761	0.689	9.820	12.361	12.938
Jesson and Arbel (2018)	0.899	0.751	0.713	**4.160**	8.650	6.980
Kamnitsas et al. (2018)	**0.901**	**0.797**	**0.738**	4.230	**6.560**	**4.500**
Baseline	0.878	0.760	0.692	6.597	11.915	5.978
Baseline + RR SegSE	0.884	0.771	0.719	6.202	10.215	6.702

Table 4.12: Results (average) obtained in BRATS 2013 Challenge set. Bold results show the best score for each tumor region.

	Dice			PPV			Sensitivity		
Method	**Whole**	**Core**	**Enh.**	**Whole**	**Core**	**Enh.**	**Whole**	**Core**	**Enh.**
Pereira et al. (2016a)	0.88	0.83	0.77	0.88	**0.87**	0.74	0.89	0.83	0.81
Shen et al. (2017)	0.88	0.83	0.76	0.87	**0.87**	0.73	0.9	0.81	0.81
Zhao et al. (2018)	0.88	**0.84**	0.77	**0.9**	**0.87**	**0.76**	0.86	0.82	0.8
Baseline	0.87	0.83	0.77	0.81	0.81	0.71	**0.94**	0.88	0.87
Baseline + RR SegSE	**0.89**	**0.84**	**0.78**	0.86	0.83	0.71	0.93	**0.89**	**0.88**

4.3.5 Conclusion

Recalibration of feature maps has the power to adaptively emphasize discriminative feature maps and suppress the uninformative ones. However, this is not optimal in the context of FCN for segmentation. In this work, we propose recombination and recalibration of feature maps for semantic segmentation. The former employs linear expansion followed by compression of feature maps for mixing features, while the later adaptively recalibrates regions of the feature maps. We show that both recombination and recalibration improve over a competitive baseline. Although we opted for a simple U-net inspired network, the proposed block can be used in other more complex FCN. Still, our FCN with the RR SegSE block achieves competitive results in BRATS 2017 Leaderboard, when compared with other single-model approaches, and superior results in BRATS 2013 Challenge.

4.4 Summary

Convolutional Neural Networks possess the capability of efficiently learning features directly from the data. These models were proved to be extremely powerful for Computer Vision, getting wide attention due to their capabilities for object recognition.

In this chapter, we explored the capabilities of CNNs for brain tumor segmentation. We started from a Classification CNN-inspired network. We employed convolutional layers with small 3×3 kernels that allowed a deeper network. Also, we experimentally show evidences that careful pre-processing with histogram standardization in the context of CNNs is beneficial. This was novel, since the mindset at the time was that Deep Learning-based models could learn directly from the data. Additionally, we proposed inter-grade data augmentation, which was beneficial for the segmentation of LGGs.

Novel developments in CNN's design led to the development of FCNs for semantic segmentation. These architectures are much more efficient than Classification CNNs. First, they can process a set of voxels in just a forward pass. Second, it allows CNNs to process more data, enabling taking larger contexts into account. Finally, during training time, we can backpropagate the errors in relation to much more voxels. Because of these advantages, we adopted FCNs for our brain tumor segmentation pipeline. We further proposed and employed a hierarchical brain tumor segmentation approach using FCNs. This was pivotal in decreasing the amount of false positive detections, as well as in balancing the class distributions during training time. In fact, it allowed us to reduce the need for post-processing, which was a weakness of the Classification CNN-based approach.

Finally, having tuned the FCN-based hierarchical brain tumor segmentation approach, we proposed feature recalibration and recombination. Although CNNs can optimize the representations for the task at hand, some features are obviously more important for detecting some classes than others. Recalibration consists in suppressing the less informative ones. However, in FCNs a feature map may contain relevant data in some spatial locations, whereas in others it may be less important. Hence, recalibration must be adaptive in relation to location. This was shown to be, indeed, better-suited than whole-maps recalibration. Additionally, feature maps recombination by means of linear expansion to higher dimensions followed by restoration of the number of feature maps was shown to be beneficial. The reason for this observation may be because it enables the mixture and generation of complex features.

Chapter 5

Interpretability of Machine Learning System for Segmentation

In this chapter we investigate the interpretability of a Machine Leaning-based system. The proposed system is composed of a Restricted Boltzmann Machine for unsupervised feature learning, and a Random Forest classifier. The case studies at hand are brain tumor segmentation and penumbra estimation in ischemic stroke lesions. We are interested in interpreting the system in relation to the input MRI sequences. To that end, we define two levels of interpretation: global and local.

This chapter is based on the following publication:

- Pereira, Sérgio, et al. "Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation." Medical Image Analysis, 2018.

Contribution The work in this chapter resulted from an internship of the author of this thesis in the Medical Image Analysis group of the Institute for Surgical Technology and Biomechanics at the University of Bern, Switzerland with Professor Doctor Mauricio Reyes. The development of the ideas and methodologies was done jointly with Doctor Raphael Meier. The author implemented the methodologies, the experiments, and the collection of results. Finally, the author was the main writer of the manuscript.

Contents

5.1 Introduction

Machine learning approaches can be broadly divided into two categories: those using hand-crafted features, and those relying on Representation Learning techniques. Representation Learning refers to a set of general machine learning methods for automatic learning and extraction of features directly from data. By contrast, hand-crafted features require expert knowledge on the problem, hence making them more problem-dependent (LeCun et al., 2015). Notwithstanding, there is usually a data representation mapping stage that takes the input data and transforms it into a more discriminative representation.

Despite the success of Representation Learning-based methods (Salakhutdinov et al., 2007; Krizhevsky et al., 2012; Pereira et al., 2016a; Kamnitsas et al., 2017b), they are often regarded as uninterpretable "black boxes". This is due to the large number of layers, or nodes, which makes it difficult to unveil the relations between inputs and outputs. In fact, this undesirable characteristic is shared with other machine learning models, such as Random Forests (RFs) with many trees, or linear models with thousands of features (Ribeiro et al., 2016a,b; Lipton, 2016). Thus, in general, there is a trade-off between the capacity/complexity of the model and its interpretability. Nevertheless, while much of the focus on machine learning has been dedicated to solving complex problems with high performance, the possibility of interpreting a decision of a model is still a very desirable property of a predictive system, but has not received much attention so far. This is especially important with the pervasive adoption of machine learning-based models in critical areas such as in radiology (Wang and Summers, 2012), where a prediction should not be blindly followed. "Black box" models may look untrustworthy in the sense that a decision cannot be explained, especially in case of failure.

However, as stated by Lipton (2016), trust and interpretability may be ill-defined. We can understand trust as the confidence in the model itself, or its prediction (Ribeiro et al., 2016b; Freitas, 2014). In the

former, we can trust a model if we have confidence that it will behave as expected after deployment. We can base this trust on the measured performance. Yet, adversarial examples (Szegedy et al., 2014; Nguyen et al., 2015) show us that unpredictable, or bizarre, behaviors may arise, even in highly accurate systems. Trusting a prediction means that we have enough confidence that a given prediction is correct (Ribeiro et al., 2016a,b; Lipton, 2016). Interpretability is a way to enhance trust in a system. Understanding the predictions and how information is encoded in a model can help us to comprehend as to why it fails, and avoid the undesirable trial and error development procedure (Zeiler and Fergus, 2014). Interpretability can appear as an interpretation of the model itself, or as a post-hoc interpretation of the model through its predictions. The former is hindered by the complexity of the model, which is usually proportional to the difficulty of the task and model performance. On the other hand, post-hoc interpretability is based on a qualitative explanation of an already trained system, by means of visualization, or study of examples. In this way, there is less need to sacrifice model's performance/complexity for the sake of interpretability (Ribeiro et al., 2016a,b; Lipton, 2016). To further contextualize these ideas, the state of the art in model interpretation is presented in the following section.

5.1.1 Previous work

We can broadly differentiate between two approaches for model interpretation: *global* and *local* interpretation. The global interpretation of a machine learning model is aimed at understanding *how* information extracted from the input data is used by the model to perform predictions. Local interpretation is aimed at understanding *why* a certain decision was made by the model at hand.

In order to enable global interpretability, some authors proposed to simplify, or transform the models (Tibshirani, 1996; Olden and Jackson, 2002; Craven and Shavlik, 1996; Hara and Hayashi, 2016). Tibshirani (1996) proposed Lasso to force some weights of the model to be exactly 0, which enhances its interpretability. Craven and Shavlik (1996) converted a previously trained neural network into a more interpretable decision tree. Olden and Jackson (2002) proposed a method to remove unimportant connections from a neural network. Hara and Hayashi (2016) interpreted tree ensembles by approximating a simpler model derived from the minimization of the KL-divergence between that simpler model and the more complex model. Gallego-Ortiz and Martel (2016) deduced rules from RFs and presented them for interpretation. These proposals have been defined for particular models and attempt to simplify models to make them globally interpretable. While these models provide information about how the model learned the training data, they can be less practical when there are high-dimensional feature vectors.

Another group of methods that are more model-agnostic treat the model as "black boxes", and bring understanding about their decisions. This is accomplished either by perturbation of the features (Cortez and Embrechts, 2011; Krause et al., 2016; Ribeiro et al., 2016b), or by fitting a simpler model to the predictions of the more complex one (Baehrens et al., 2010; Ribeiro et al., 2016a,b). These approaches are post-hoc in the sense that they do not explain the inner workings of the model itself, but its predictions. Particularly, perturbing features (Cortez and Embrechts, 2011; Krause et al., 2016; Ribeiro et al., 2016b) and observing its impact on the decision may provide an estimate of the feature importance, but it does not take into account correlations among features. Additionally, it may be impractical for high-dimensional fea-

ture vectors. Other approaches try to interpret the learning algorithm locally, i.e., its behavior in the vicinity of the test samples (Baehrens et al., 2010; Ribeiro et al., 2016a,b). To this end, Baehrens et al. (2010) approximated the predictions of the model under analysis with another model. Then, an explanation vector was defined as the derivative of the probabilistic output in relation to the data point. Explanation vectors provide information about which features would affect more the prediction of that sample. Although the method by Baehrens et al. (2010) may provide some insight about the model, a human may not extract any interpretation from it, if the explanation vector is high-dimensional. Ribeiro et al. (2016b) also pursued local interpretability as a way to achieve model agnostic interpretation that could be applied even for very complex deep networks. For a given sample in the feature space of the model, a set of synthesized examples in the vicinity of the sample is created, whose prediction is obtained from the model under analysis. Then, a simpler model is fitted to these samples and interpreted. The simpler model does not represent the original complex model globally, but it is an approximation of its behavior in the vicinity of the given sample. Nevertheless, this approach is agnostic to the model being interpreted and robust to its complexity. Most of the previous proposals relied on some sort of visualization to present the data for human interpretation, focusing on approximations of the model under analysis. Other approaches purely relied on visualizing the topology of neural networks (Hinton et al., 1986; Wejchert and Tesauro, 1989; Tzeng and Ma, 2005). Zrihem et al. (2016) used t-SNE (Maaten and Hinton, 2008) in the context of deep reinforcement learning to reduce the dimensionality of neural activations of a Deep Q Network, in order to study the policies of the agent at hand. Visualization in the context of Convolutional Neural Networks (CNN) comes in the form of saliency maps that inform which region of the image was important for a given class (Simonyan et al., 2013), or deconvolving an activation and projecting it in the image space (Zeiler and Fergus, 2014). The later involves coupling a deconvolutional neural network to the CNN under analysis.

High-dimensional feature vectors increase the computational load and complexity of a machine learning model, as well as the risk of overfitting due to irrelevant features having spurious correlations with the target variable. Furthermore, it may render the interpretability of a model more difficult (Tibshirani, 1996). Hence, feature selection may be seen as a prerequisite for enabling model interpretation. Univariate feature selection methods evaluate the relationship of each feature with some condition of interest, but cannot detect interactions among features, which is the advantage of multivariate methods. Some of the latter approaches are wrappers around a learner that iteratively evaluate subsets of features in relation to their predictive power (Ganz et al., 2015). However, these recursive feature elimination methods may be unpractical for very large datasets. Random Forests also stand as a multivariate approach for feature selection (Konukoglu and Ganz, 2014), due to their capability to measure feature importances through the mean decrease impurity (MDI), allowing us to rank features. The drawback is that one still needs to choose a user-defined threshold on the ranking, or, as proposed by Konukoglu and Ganz (2014), on an upper bound on false positive rates in selecting unimportant features as relevant ones.

Taking into account the aforementioned interpretability-related studies, one can draw some conclusions. 1) Post-hoc approaches provide tools to potentially interpret more complex models. This copes with modern trends favoring powerful, yet complex, approaches, such as methods based on Deep Learning. 2) A model can be agnostically, but locally, interpreted, providing insights regarding each individual decision.

This may allow us to infer about its coherence and to reason about mistakes. 3) On the other hand, approaches that focus on a global interpretation can provide clues about how the model learned to look at the data, however, they are model-specific and lack the local interpretability necessary to understand individual predictions. 4) High-dimensional feature representations may pose difficulties for interpretation. Thus, effective feature selection methods might be required. 5) Visualization tools are natural human understandable data exploration tools for enhancing interpretability. 6) Interpretation is ultimately performed by the human expert. For example, in radiology, clinical experts tend not to trust machine decisions as the interpretability and ultimately the trustworthiness of automatic algorithms tend to be low (Wang and Summers, 2012). If we want to increase trust, we should devise methodologies for interpretability that are simple and understandable by humans. Additionally, with interpretability methodologies, we expect to retrieve hints to answer the following questions: 1) How does a system use the input data to solve the task at hand? 2) When a system fails, why does it fail? 3) Is the system capturing the relevant relations in the data? For instance, Ribeiro et al. (2016b) found that a system that was accurately detecting Husky dogs in pictures was basing its prediction in the presence, or not, of snow and not in the dog itself. This is an example of a system that bases its predictions on a feature that does not coherently fit into the concept of the target variable (Husky dog).

5.1.2 Motivation and contributions

Motivated by the current trend for favoring complex models at the expense of interpretability, in this paper we propose to interpret a machine learning system both globally and locally. In contrast to previous studies, we hypothesize that both global and local interpretability provide complementary insights about the machine learning system at hand. The focus of this work is on the interpretation of automatically learned features for lesion segmentation in medical images. Particularly, we will study the interpretation of features stemming from Magnetic Resonance Imaging (MRI) sequences. We propose to drive feature selection for the task at hand by coupling a Restricted Boltzmann Machine-based data representation model with a RF classifier, to jointly consider existing correlations between imaging data, features, and target variables. The contributions in this work are the following: i) We explore a strategy for global interpretability by inferring which parts of the input data contributed the most for highly important features in different segmentation tasks, thus indirectly interpreting how a RBM encoded the data. ii) We interpret image segmentations locally through assessing the spatial relevance of the features distributed in the image space. Finally, iii) in order to leverage interpretability, we also evaluate a joint Mutual Information and RF feature importance strategy for automatically selecting important features. We evaluate the proposed approaches in brain tumor segmentation and penumbra estimation in ischemic stroke lesions. In an ischemic stroke, the closest tissues to the blocked blood vessels are at high risk of infarction. Those tissues around this core that suffer from the reduced blood supply, but are still salvageable, form the penumbra region. The databases used in this study for brain tumor segmentation and penumbra estimation are publicly available and being actively used in recent research (Menze et al., 2015; Maier et al., 2017); thus, enabling future comparison with our proposal.

The remainder of this paper is organized as follows. In Section 5.2, we introduce the two basic

components used for our interpretable machine learning system. The proposed system, feature selection, and interpretability methodologies are presented in Section 5.3. In Section 5.4 we describe the databases and experimental setup. Results are presented in Section 5.5. Then, in Section 5.6, we discuss our results. Finally, we draw the main conclusions in Section 5.7.

5.2 Preliminaries

Our machine learning system is based on a Restricted Boltzmann Machine (RBM) (Smolensky, 1986) to learn features, and a RF classifier (Breiman, 2001) as discriminative model. RBMs are generative unsupervised Representation Learning techniques that learn the intrinsic representation of the data without information regarding any target variable (Hinton et al., 2006). Hence, RBMs can be trained on large unlabeled data sets, as typically found in the clinics. This represents an advantage over supervised learning approaches, as manual labeling of large data sets is expensive, time-consuming, and prone to intra- and inter-observer variability. Features extracted by these unsupervised models have proven to be useful for texture classification (van Tulder and de Bruijne, 2016) and volume estimation (Zhen et al., 2016). Since RBMs are an unsupervised method, after feature learning it is necessary to employ a task-related supervised learning algorithm (van Tulder and de Bruijne, 2016, 2015; Zhen et al., 2016), or supervised fine-tuning stage (Nair and Hinton, 2010) that learns how to map the learned features to the desired task. RFs are one of the possible supervised learning algorithms, which have shown advantages as a classifier, such as being robust to overfitting, and successfully dealing with high-dimensional feature vectors by identifying and ranking relevant features (Criminisi and Shotton, 2013). While this capability of RFs may be used to find which features better explain the target variables, or for feature selection (Konukoglu and Ganz, 2014), getting unbiased feature importance measures is not trivial (Louppe et al., 2013). Due to its advantages, RFs have been successfully used in medical image analysis, e.g. (Criminisi and Shotton, 2013; Meier et al., 2014c; Zhen et al., 2016; Menze et al., 2016; Pereira et al., 2016b; Maier et al., 2017; Meier et al., 2016; McKinley et al., 2016).

Since high-dimensional feature vectors may impair the interpretability of a model, feature selection is required (Tibshirani, 1996). Filter-based feature selection methods have the advantages of scaling well and being computationally efficient. Mutual Information (MI) is an information measure that can be used to assess feature relevance. Moreover, it has the advantage of measuring any kind of relation among variables, even non-linear ones (Bennasar et al., 2015; Vergara and Estévez, 2014). Methods based on this measure evaluate the MI between features and target variables (Peng et al., 2005; Battiti, 1994). However, there is no involvement of the learning algorithm that is supposed to employ the selected features in the actual selection procedure. In this paper, however, we use MI both between data and features, as well as between features and task-related classes through a RF.

5.3 Methods

In this work, we consider two main blocks: the machine learning system, and the interpretability system (Fig. 5.1). Taking into account the previous work (Subsection 5.1.1), feature selection is required to enable an effective model interpretation. Thus, we will first introduce our machine learning system and propose a methodology for feature selection. Subsequently, we will present methods for model interpretation, which exploit the feature selection.

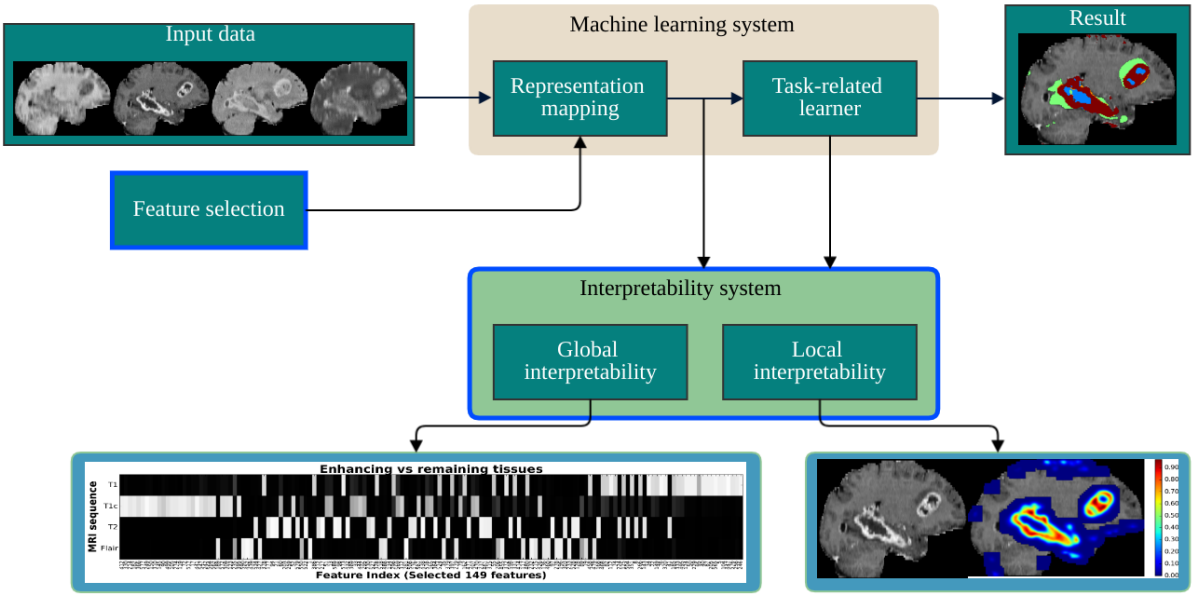


Figure 5.1: Proposed system. The machine learning system is composed of a representation mapping stage that generates the input features for a task-related learner, which computes the prediction. Feature selection is performed to enable an effective interpretation of the machine learning system. In order to enhance model interpretability, the combined use of global and local interpretability is proposed. Blue colored frames mark the modules representative of the main contributions in this paper. The visualization of the training stage of the machine learning system and feature selection is omitted for simplicity. We show an example application in brain tumor segmentation.

5.3.1 Machine learning system

There are two main stages in a machine learning system: representation mapping and the task-related learner. The former corresponds to the feature computation stage, which can be performed by representation learning or feature engineering. The latter is the predictive model, which is task-dependent since it is a supervised learning algorithm.

5.3.1.1 Representation mapping

We use RBM (Smolensky, 1986) to realize the representation mapping stage of our machine learning system. This is an undirected graphical Representation Learning model. The nodes are organized into one visible and one hidden layer, whose states are represented by the vectors $\boldsymbol{v} = [v_i : i = 1, \ldots, m]$,

and $\boldsymbol{h} = [h_j : j = 1, \ldots, n]$, respectively. All nodes in one layer are connected to all nodes in the other layer with weights represented by the matrix $\boldsymbol{W} = [w_{ij}]$. No intra-layer connections exist. In this work, the inputs to the visible layer are patches of shape $d \times d \times d$ extracted from the set of available MRI sequences $\mathcal{C}$. Then, the patches are represented as a 1D vector and fed into the visible layer; thus, $m = d \cdot d \cdot d \cdot |\mathcal{C}|$ (Fig 5.2). Originally, RBMs were proposed to model binary data in both layers. However, image patches are represented by a continuous range of values. Thus, the visible units are defined as linear units with independent Gaussian noise, which allows us to model continuous-valued inputs (Hinton, 2012). Noisy Rectifier Linear Units (NReLU) are used to represent the hidden units, since they proved to be suitable for feature extraction (Nair and Hinton, 2010). Thus, after receiving the input in the visible layer, the RBM can compute the activations in the hidden layer, thus mapping the input into a feature vector. The joint configuration of the states of the visible and hidden units is represented by an energy function defined as

$$E(\boldsymbol{v}, \boldsymbol{h}) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij}, \tag{5.1}$$

where a_i is the bias of the visible unit i, b_j is the bias of the hidden unit j, and σ_i represents the standard deviation of the Gaussian noise of v_i (Hinton, 2012; Nair and Hinton, 2010). Having the energy function, the joint probability distribution over $\boldsymbol{v}$ and $\boldsymbol{h}$ is

$$P(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} e^{-E(\boldsymbol{v}, \boldsymbol{h})}, \tag{5.2}$$

where Z represents the partition function. Computing Z is impractical, but we can still sample in parallel the state of all the units in one layer conditioned on the other layer, given that there are no intra-layer connections. Thus, we sample the hidden and visible units, respectively, as (Nair and Hinton, 2010; van Tulder and de Bruijne, 2015)

$$P(h_j|\boldsymbol{v}) = \max\left(0,\ \sum_i w_{ij} v_i + b_j + \mathcal{N}\left(0, sigm\left(\sum_i w_{ij} v_i + b_j\right)\right)\right), \tag{5.3}$$

$$P(v_i|\boldsymbol{h}) = \mathcal{N}\left(\sum_j w_{ij} h_j + a_i, \sigma_i\right), \tag{5.4}$$

where $\mathcal{N}$ represents the Gaussian distribution and $sigm$ the sigmoid function.

We use Contrastive Divergence (Hinton, 2002) with one step of alternating Gibbs sampling to train the model. Since learning σ is difficult, following Hinton (2012), we normalize each component of the data with zero mean and unit variance and consider $\sigma_i = 1$. We also employ momentum, and both L1 and L2 weight-decay. L1 enforces sparsity, which leverages interpretability (Hinton, 2012), similarly to Lasso (Tibshirani, 1996). After training, we compute features as noise-free activations of the NReLU units. These units exhibit intensity equivariance if they are noise free and have zero biases ($b_j = 0$) (Nair and Hinton, 2010).

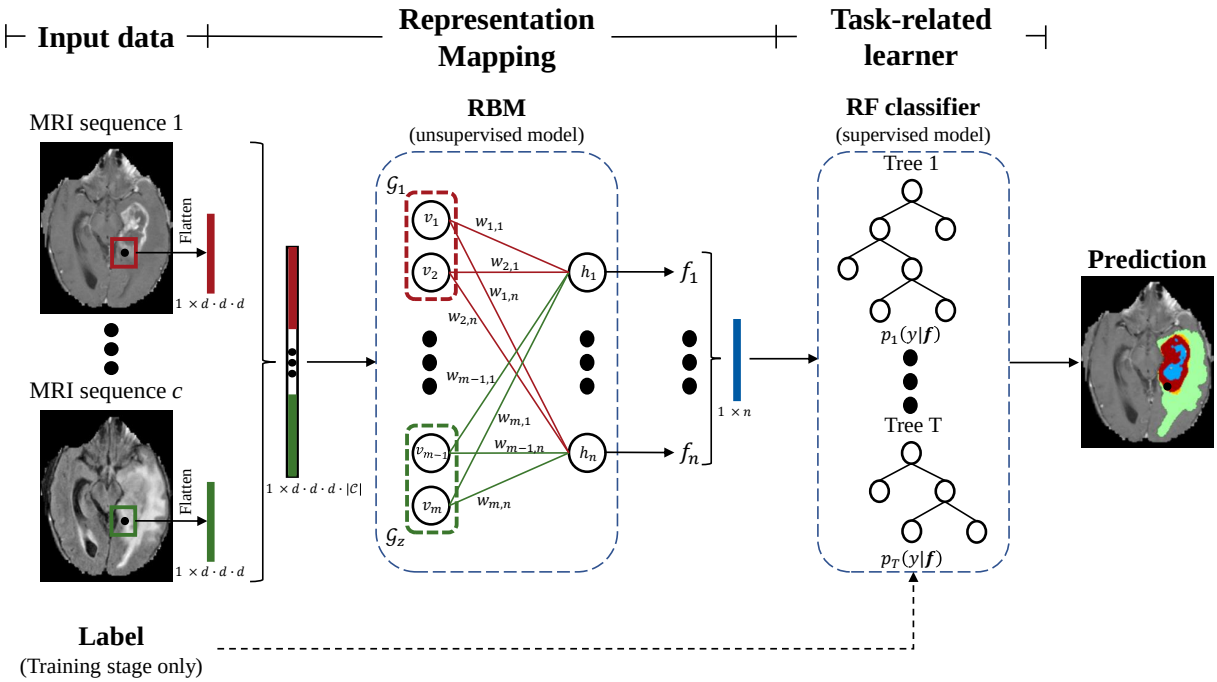


Figure 5.2: Machine learning system. We use RBM as Representation Mapping and RF classifier as Task-related Learner. Patches are extracted from each MRI sequence, flattened, and concatenated into one single 1D vector. The RBM receives the imaging data in the visible layer, and maps it into a feature vector, as activations of the hidden units. $\mathcal{G}_y$ identify meaningful groups of visible units that receive data from a distinct MR sequence. The color in the connections identify weights that are linked to a given MRI sequence. The feature vector is fed into the RF classifier, which outputs a prediction for the central voxel of the patch (black dot).

5.3.1.2 Task-related learner

The RBM learns features from data in an unsupervised way, and without knowing the task for which the features will be used for. So, we need a supervised learning algorithm to learn how to make predictions out of those features. We use a RF classifier (Breiman, 2001) as task-related learner, by training it in a supervised way. This model is an ensemble of Decision Trees, each one trained on a randomly selected subset of the training set with replacement (bootstrap). Each training and testing sample is represented by the activations of the hidden layer of the RBM. So, each sample is represented by a *n*-dimensional feature vector and fed into the RF (Fig. 5.2). As the samples traverse the trees, a subset of features (randomly chosen during training) is evaluated in each node. This characteristic, together with the number of trees, allows the algorithm to deal with high-dimensional feature vectors. The randomness in the algorithm allows it to be robust to overfitting. At the same time, although RF can estimate the feature importance, getting unbiased measures is not trivial (Louppe et al., 2013; Criminisi and Shotton, 2013). For the reader interested in more details regarding the RF classifier, we refer to Breiman (2001) and Criminisi and Shotton (2013).

5.3.1.3 Joint RBM-RF Mutual Information approach for feature selection

Usually, a feature represents the response of a feature detector applied over the data. Having noise is harmful since the learning algorithm may capture spurious correlations between the features and the labels. Additionally, feature noise, arising from detectors that enhance spurious variations in the data, may be adverse for the learning algorithm (Zhu and Wu, 2004). So, we hypothesize that a good feature should correlate with the class labels, and represent the data from which it was computed; hence, effectively connecting data with class labels. If it holds true, then interpreting a prediction may be more feasible in terms of which input caused it. Since MI is a measure of statistical dependence, we employ it to quantify the quality of the mapping between data and labels, through the features. Our procedure consists of the following steps: First, after training the RBM, we compute n features (activations of the hidden units) for each of the s training samples. Thus, for each $k \in \{1, \cdots, n\}$ feature, we define a vector $\boldsymbol{f}_k = [f_r : r = 1, \cdots, s]$ that represents the values of feature k in all the training samples. Then, in the case of multisequence MRI data, which is typically used in many clinical scenarios such as the ones presented here, we measure MI between each feature $\boldsymbol{f}_k$ and the intensities of each c MRI sequence ($\boldsymbol{i}_c = [i_r]$) to quantify the statistical dependence between features and each sequence. Finally, for each feature, we combine the MI measure between $\boldsymbol{f}_k$ and each MRI sequence, as

$$MI_k\left(\boldsymbol{f}_k, \boldsymbol{i}_c\right) = \sum_c H\left(\boldsymbol{f}_k\right) + H\left(\boldsymbol{i}_c\right) - H\left(\boldsymbol{f}_k, \boldsymbol{i}_c\right), \tag{5.5}$$

where H corresponds to the Shannon entropy. We will refer to MI_k as RBM-MI in order to express that the features are calculated by the RBM.

In RF, the contribution of each feature to decrease the impurity of training samples as they traverse the RF nodes can be evaluated with the MDI metric (Louppe et al., 2013). Although this estimate may be biased, it is still recommended, as obtaining unbiased feature importance estimations from tree-based ensemble methods is quite impractical (Louppe et al., 2013). MDI is computed using the Information Gain as splitting criteria in the nodes, which is equivalent to measure MI between the decision in the nodes and the class of the samples (Nowozin, 2012). We will denote this second component as RF-MDI.

The key idea in our proposal is to unify RBM-MI and RF-MDI into a common metric, in order to evaluate the overarching mapping between data and labels. Hence, for feature selection, we link MI measures between features and data (through RBM-MI) with MI measures between features and classes (through RF-MDI). From experiments, we observe that when we plot the RF-MDI and RBM-MI for feature k in descending order (Fig. 5.3, left) the curves are similar in shape, with a steep initial decrease. Then, features are gradually less important. We want to find the point that corresponds to the transition between important and unimportant features, both in terms of data representation and class label characterization. To this end, we measure the Pearson correlation coefficient between the sorted RF-MDI and RBM-MI measures for increasingly larger feature subsets; we note that the Pearson correlation coefficient is computed over the RF-MDI and RBM-MI measures, not the features themselves. Finally, the maximum in the Pearson correlation coefficient curve indicates the transition point γ between important and unimportant features (Fig. 5.3(b)). The final subset of selected features corresponds to the union of the best γ features sorted

according to RF-MDI and RBM-MI (c.f. Fig. 5.3 and Algorithm 5). As noted in (Ganz et al., 2015), using intersection instead of union can result in an empty set, although it can be prevented by including domain-specific knowledge. The proposed approach has the advantage of not requiring any pre-defined threshold (e.g. on the number of features). Still, if desired, one can define a minimum percentage or number of features to be evaluated through MI and MDI, and retrieve the union of those subsets.

5.3.2 Interpretability system

The interpretability of a machine learning model can be defined as the *human reasoning* on *how* a model captures the input data, and *why* a certain decision is made. In contrast to previous works, we acknowledge the relevance of both global and local interpretability as providers of complementary information.

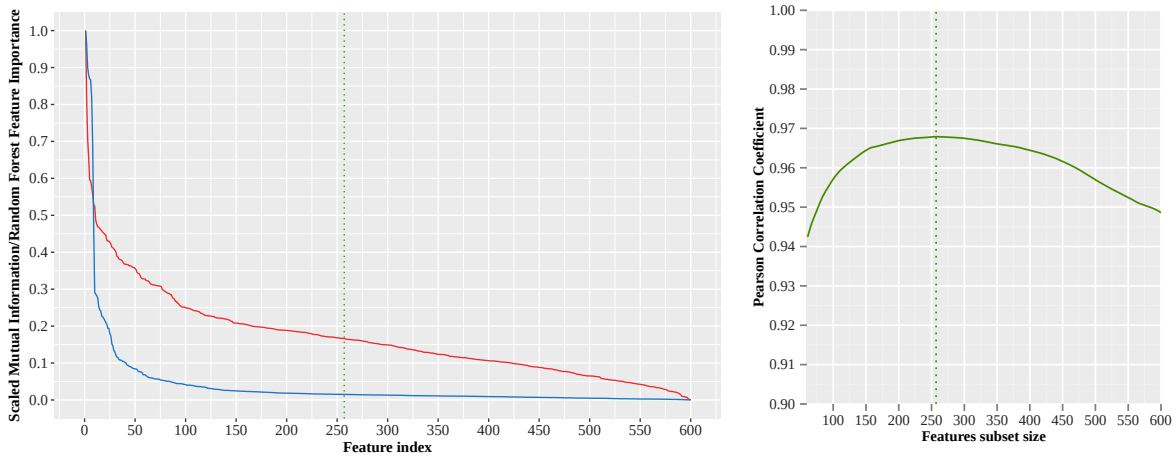


Figure 5.3: Feature selection is based on RF-MDI and RBM-MI. a) RBM-MI (red) and RF-MDI (blue) of each feature are plotted in descending order. b) Pearson correlation coefficient between accumulating subsets of features MDI and MI. The dotted green vertical line marks the maximum of the Pearson correlation coefficient.

Algorithm 5 Joint RBM-RF Mutual Information feature selection

function FeatureSelection $(\boldsymbol{mi},\ \boldsymbol{mdi})$

Input: The vector sorted in descending order of RBM-MI values between each feature and the MRI sequences $\boldsymbol{mi}$; the vector sorted in descending order of RF-MDI $\boldsymbol{mdi}$

Output: The subset $\mathcal{S}' \subseteq \mathcal{S}$ containing the index of selected features

$\boldsymbol{pc} \leftarrow initialize\_zeros()$
for $j \in \mathcal{S}$ **do**
  $\boldsymbol{pc}_j \leftarrow pearson\_correlation(\boldsymbol{mi}_{0,\cdots,j},\ \boldsymbol{mdi}_{0,\cdots,j})$
end for
$\gamma \leftarrow argmax(\boldsymbol{pc})$
$\mathcal{S}' \leftarrow get\_features(\mathcal{S}, \boldsymbol{mi}, \gamma)\ \cup\ get\_features(\mathcal{S}, \boldsymbol{mdi}, \gamma)$
return $\mathcal{S}'$

Global interpretability is defined as the reasoning and identification of relevant inputs for the machine learning system as a whole. Although less model agnostic, it may help us answering *how* a model captures the (training) input. In the context of medical imaging, it is a description of what the system learns from a *population*. This may be used as a sanity check before deployment of the given machine learning model, since the expert can infer whether the way the model is capturing the input data is coherent with his prior knowledge on the problem. In addition, it is a valuable tool to unveil possible biases in the datasets introduced by the population subjects' selection or data processing.

We define local interpretability as the reasoning and identification of the relevant inputs for a given decision, on a per-sample basis. Thus, locality is more related to *why* a decision is made, which becomes more relevant after deployment, especially to study failures. In the context of medical imaging, it can retrospectively help in understanding model decisions for a particular *subject*. Since our application is on image segmentation, local interpretability is at the voxel level and thus also patient-level.

Our approach is post-hoc in the sense that interpretation comes at a later stage to model training, instead of being embedded in the system itself. In this way, we avoid sacrificing complexity/performance of the model in favor of interpretability. In the following subsections we describe how global and local interpretability are implemented in the proposed approach.

5.3.2.1 Global interpretability

We propose to study the relationship between the input data (in our example applications, the different MRI sequences) and the RBM features. We define first a set containing all available n feature indices, i.e. $\mathcal{S} = \{1, 2, \ldots, n\}$. Based on our feature selection method presented in section 5.3.1.3, we can generate a reduced feature set $\mathcal{S}' \subseteq \mathcal{S}$, which is more suitable for model interpretation than the complete set of features. Some visible units may be grouped into z subsets of meaningful groups $\mathcal{G}_y$, such that $\mathcal{G} = \{\mathcal{G}_y : y = 1, 2, \ldots, z\}$. Each $\mathcal{G}_y$ contains the indices of the visible units belonging to that meaningful group. For instance, when using image patches from multi-sequence MRI acquisitions we can group visible units belonging to the same sequence. In other words, the z-available image sequences define the groups $\mathcal{G}_y$ (see Fig. 5.2). In the extreme case where no meaningful groups can be defined, each visible unit v_i is a subset in itself. For interpretation, we compute the squared L2-norm of the weights connecting a hidden unit to the visible units of each group $\mathcal{G}_y$. This way, we determine the contribution of each group $\mathcal{G}_y$ to a hidden unit (i.e. feature). We repeat this procedure for all hidden units. We compute the squared L2-norm because a negative weight may still make a visible unit contribute positively to the hidden unit response, since the inputs are normalized with zero mean and unit variance. Additionally, it decreases the contribution of very small valued weights that would contribute to noise, which may impair the interpretability through visualization techniques. Algorithm 6 presents the complete procedure. Taking brain tumor segmentation as example, we identify the relevance of the different MRI sequences for a specific task (e.g. segmentation of the complete tumor vs. normal tissues) by studying the weights connecting the hidden units of the most and least important features.

To facilitate interpretability, we sort the selected features in descending order of their importance (we took the RF-MDI as a measure of feature importance for the classifier itself). Then, we plot the squared

Algorithm 6 Squared L2-norm computation for global interpretability

function GlobalInterpretability $(\boldsymbol{W},\ \mathcal{S}',\ \mathcal{G})$

Input: The weight matrix $\boldsymbol{W}$, the set $\mathcal{S}' \subseteq \mathcal{S}$ containing the index of selected features (=output of algorithm 5), and the set of meaningful groups $\mathcal{G}$

Output: Matrix with shape $[|\mathcal{G}|, |\mathcal{S}'|]$ of squared L2-norms $\boldsymbol{L} = [l_{y,j'}]$

```
for j' ∈ S' do
    for G_y ∈ G do
        l_{y, j'} ← Σ_{i'∈G_y} w_{i',j'}^2
    end for
end for
return L
```

L2-norm of the weights connecting the hidden units of the RBM to the subgroups of meaningful features (e.g. Fig. 5.4(c)).

5.3.2.2 Local interpretability

We start by selecting the most meaningful features of a sample x (e.g. voxel of interest) by examining its neighborhood. By randomly perturbing the feature vector, which is the activations of the hidden layer of the RBM, of the selected sample x, a set of π synthetically generated samples ($\mathcal{X} = \{x_p : p = 1, \cdots, \pi\}$), with feature vector $\boldsymbol{f}_{x_p}$, is created. These synthetic neighbors are close to the original sample in the feature vector space. Then, a classification (y_p) for each neighbor is calculated with the classifier under analysis $f(\cdot)$ (RF in our machine learning system), as $y_p = f(\boldsymbol{f}_{x_p})$. Afterwards, a Ridge regression model ($g(\cdot)$) is trained on the synthetic set, using the output provided by the classifier under analysis as target, to select a pre-defined number of explaining features, corresponding to those yielding the strongest responses to the input, i.e. highest value of the product between the weights of the Ridge regressor and features. The output of $g(\cdot)$ is a subset $\mathcal{S}''$ of the available feature indices contained in $\mathcal{S}'$. Neighbors are weighted according to their euclidean distance to the original test case (d_{x_p}). Ridge regression has the advantage of being simple and extremely efficient, which allows it to be applied voxelwise[1]. The procedure is depicted in Algorithm 7. Here, we consider $\mathcal{S}'$ provided by a previous feature selection procedure, which selects features that highly correlate both with labels and the data. However, there is nothing that prevents Algorithm 7 from being applied while taking the full set of features into consideration. Additionally, this procedure can be used both in binary and multi-class problems.

Having the selected features, we could generate L2-norm plots similar to the ones proposed for global interpretability for each test sample. However, since we are dealing with images, it is more insightful to observe which parts of an image are more important for a given task, such as segmenting a particular tumor compartment. After selecting the features that better explain a prediction, we assume that all of those features must be equally taken into account for interpretation. However, the weights may be in different ranges. So, we independently normalize the absolute value of weights connecting each hidden

[1] Instead of the proposed approach with Ridge, we could use stability selection with Lasso (Meinshausen and Buhlmann, 2010), which can select the optimal number of features. However, it still needs a threshold above which a feature is considered relevant. Additionally, it works by training several models on randomly chosen subsets of the training set, making it much more computationally demanding than Ridge.

unit to the visible units to $[0,\ 1]$. In our case, we predict the class of each voxel x based on the features computed by the hidden units of the RBM, which are extracted from a patch $\boldsymbol{p}$ centered on that voxel. Thus, we proceed by summing the weights connecting each selected hidden unit to the visible units of each sequence for all corresponding voxels contained in the patch centered on voxel x. By repeating this for all voxels in the image space $\mathcal{I}$, we obtain, for each MR sequence $c \in \mathcal{C}$, a corresponding image $\boldsymbol{H}_c$ that contains voxel-wise importance values. We denote this procedure as spatial feature relevance for local interpretability. The procedure is described in Algorithm 8.

The selection of locally relevant features is inspired by Ribeiro et al. (2016b) and motivated by some features being important for some classes, while other features may be important for some other classes. Thus, even though we selected a subset of important features before, some of them are more relevant depending on the sample under analysis. Contrasting with Ribeiro et al. (2016b), where a simpler, yet interpretable, model (Ridge regression in our case) serves as feature selector and explainer, we go further as to explain which parts of the input data mostly contributed to the features response of the sample under analysis. We realized this by projecting the MRI sequence relevance for a given feature back to the image space of the respective MR sequences. Additionally, we previously selected a subset of features

Algorithm 7 Local test case specific feature selection

function LocalFeatSel $(x,\ f(\cdot),\ n_{feat},\ n_{x_p},\ g(\cdot), \mathcal{S}')$

Input: A test sample x, a model $f(\cdot)$, the number of desired features n_{feat}, the number n_{x_p} of neighbors of x, and feature selection method $g(.)$

Output: Indices of the selected features $\mathcal{S}'' \subseteq \mathcal{S}'$

$\{x_p,\ d_{x_p}\} \leftarrow Perturb(x,\ n_{x_p})$
$y_p \leftarrow f(x_p)$
$\mathcal{S}'' \leftarrow g(x_p,\ y_p,\ d_{x_p},\ n_{feat})$
return $\mathcal{S}''$

Algorithm 8 Spatial feature relevance for local interpretability

function SpatialFeatureRelevance $(\mathcal{I},\ \mathcal{S}'',\ \mathcal{C}, \boldsymbol{W})$

Input: the image space $\mathcal{I}$, the selected features for each test sample $\mathcal{S}''$ (=output of algorithm 7), the set of (pre-aligned) MRI sequences $\mathcal{C}$ (e.g. $\mathcal{C} = \{T_1, T_{1c}, T_2, FLAIR\}$), and the weights matrix $\boldsymbol{W}$

Output: Images with the voxel-wise importance for each sequence $\mathcal{H} = \{\boldsymbol{H}_c\}$

$\mathcal{H} \leftarrow initialize\_zeros(\mathcal{I})$
$\boldsymbol{W} \leftarrow normalize\_weights(\boldsymbol{W})$
for $x \in \mathcal{I}$ **do**
  $\boldsymbol{p} \leftarrow get\_patch\_indices(x)$
  for $j'' \in \mathcal{S}''$ **do**
    for $c \in \mathcal{C}$ **do**
      $\boldsymbol{w}_c \leftarrow sequence\_weights(\boldsymbol{W}, j'', c)$
      for $e \in \boldsymbol{p}$ **do**
        $\boldsymbol{H}_c(e) \leftarrow \boldsymbol{H}_c(e) + \boldsymbol{w}_c(e)$
      end for
    end for
  end for
end for
return $\mathcal{H}$

that correlate both with the labels and the data, which differs from Ribeiro et al. (2016b).

5.4 Experimental Setup

5.4.1 Databases

The proposed methodologies were applied to two segmentation problems with multisequence MRI data: brain tumor segmentation and penumbra estimation in acute ischemic stroke, for which model predictions can be interpreted in the context of clinical expert knowledge and manual segmentation protocols (Menze et al., 2015; Maier et al., 2017).

5.4.1.1 Brain tumor segmentation

For this problem, we used the publicly-available BRATS 2013 database (Menze et al., 2015) of the MICCAI Brain Tumor Segmentation (BRATS) Challenge. The database has three sets with different number of subjects: Training (30), Leaderboard (25), and Challenge (10). The Training set contains manual ground truth segmentations, distinguishing four tumor tissues: necrosis, edema, contrast-enhancing tumor, and non-enhancing tumor. The evaluation of the Leaderboard and Challenge segmentation was performed via the online platform SMIR[2] for three tumor regions: complete (all tumor tissues combined), core (necrosis + enhanced + non-enhanced), and enhancing tumor. For each subject there are four MRI sequences available with interpolated isotropic resolution of $1\ mm$: T1-weighted (T1), gadolinium-enhanced T1 (T1c), T2-weighted (T2), and Fluid-attenuated Inversion Recovery (FLAIR). All sequences are already rigidly aligned, and skull-stripped. Further pre-processing included bias field correction (Tustison et al., 2010), and normalization of the intensities in each MRI sequence with a histogram standardization method (Nyúl et al., 2000). Finally, we normalized the intensities of brain voxels to zero mean and unit variance.

We chose the BRATS 2013 database due to two reasons: First, the ground truth data are manual segmentations obtained by the fusion of four expert raters. These expert raters followed a manual segmentation protocol (Menze et al., 2015) to acquire the ground truth data. Hence, it enables us to interpret our machine learning system with respect to this protocol. Second, the dataset contains preoperative brain tumor images only. In contrast to postoperative images, treatment-related imaging changes (e.g. radiation necrosis (Mullins et al., 2005)) are absent in preoperative images thus rendering an evaluation of our interpretation methodologies less complicated.

5.4.1.2 Penumbra estimation in acute stroke

For investigating the penumbra estimation in acute ischemic stroke, we employed the Stroke Perfusion Estimation (SPES) database of the MICCAI Ischemic Stroke Lesion Segmentation (ISLES) Challenge (Maier et al., 2017). The Training dataset includes 30 subjects with publicly available manual ground

[2] `https://www.smir.ch/BRATS/Start2013`

truth segmentations, while the Challenge set is composed of 20 subjects. As in BRATS, the results for the Challenge set are computed by an online platform[3]. Seven MRI sequences, comprising structural and physiological sequences, were available: T1c, T2, Diffusion Weighted Imaging (DWI), cerebral blood flow (CBF), cerebral blood volume (CBV), time-to-peak (TTP), and time-to-max (Tmax). All sequences are already rigidly registered to the T1c sequence with image resolution of $2\ mm$ and skull stripped. Further pre-processing included the bias field correction (Tustison et al., 2010), and normalization of the intensities with a histogram standardization method (Nyúl et al., 2000) for the T1c, T2, and DWI sequences. Additionally, we clipped Tmax intensity values above 60 (Tmax $>$ 6s threshold, as followed by the manual segmentation protocol experts (Maier et al., 2017)). Finally, we normalized the intensities of brain voxels to zero mean and unit variance.

5.4.2 Model training & parameters

Around 40,000 samples were extracted from each subject, and classes were balanced by having 50% of normal tissue and 50% of total lesion tissue (in BRATS we further approximately balance the sampling of tumor tissues). The RBM consisted of 600 hidden units (=features). For training the RBM, the learning rate (ϵ) was kept constant for the first 10 epochs, and then linearly decreased until the end of the training. In the case of momentum (η), it was kept constant until epoch 100, and then linearly increased until the end of training; no momentum was used in SPES. For penumbra estimation we thresholded the probabilistic output of the RF at 0.6 (empirically found in the validation set). The remaining hyperparameters of the RBM and RF are shown in Table 5.1. For training the RBM, we extracted the patches centered on a given voxel from all MRI sequences. Then, the patches are represented as a 1D vector and fed into the visible layer of the RBM. So, for brain tumor segmentation the visible layer has $4 \cdot 9 \cdot 9 \cdot 9 = 2916$ units, while for penumbra estimation it is $7 \cdot 5 \cdot 5 \cdot 5 = 875$ (see Fig. 5.2). For the local interpretability method, each image voxel's features were perturbed to generate 2400 synthetic samples, and the 10 most representative features were selected. We set the regularization parameter λ of the Ridge regression model to 1.0. We used the RF implementation in Scikit-learn (Pedregosa et al., 2011); the hyperparameters that are not defined in Table 5.1 were set to default values. We used LIME[4] for generating the neighborhood of the points and select the local relevant features for local interpretability. The implementation of the proposed algorithms is available online[5].

In this paper, we focus on enhancing the interpretability of the machine learning system at hand, thus a thorough performance evaluation is out of scope. Nevertheless, evaluating the segmentation to some extent is imperative to assess if the model is learning. Thus, we report the Dice Similarity Coefficient (Dice) for BRATS and SPES, as well as the Average Symmetric Surface Distance (ASSD) for the latter, as defined in (Menze et al., 2015; Maier et al., 2017).

[3] `https://www.smir.ch/ISLES/Start2015`
[4] `https://github.com/marcotcr/lime`
[5] `https://github.com/sergiormpereira/EIML`

Table 5.1: Hyperparameters of the RBM and RF of our machine learning system. In RF, when not indicated, default values were used.

Database	Algorithm	Hyperparameter	Value
BRATS SPES	RBM	Hidden units	600
		Mini-batch	32
		Gibbs samp. steps	1
		$\boldsymbol{W}$ init.	$\mathcal{N}(0,\ 0.0001)$
		$\boldsymbol{a}$, $\boldsymbol{b}$ init.	0
	RF	Trees	200
		Split. crit.	Info. gain
BRATS	RBM	Patch size	$9 \times 9 \times 9$
		Epochs	262
		Initial ϵ; Final ϵ	0.0001; 0.000 000 4
		Initial η; Final η	0; 0.5
		L1; L2	0.001; 0.02
SPES	RBM	Patch size	$5 \times 5 \times 5$
		Epochs	498
		Initial ϵ; Final ϵ	0.001; 0.000 01
		L1; L2	0.0002; 0.0002

5.5 Results

5.5.1 Feature selection

In Table 5.2, we present segmentation results using features selected with the proposed method in BRATS and SPES. In order to have our approach compared with other methods for automatic feature selection, we also present results when features are selected by Embedded or Wrapper feature selection-based approaches. Embedded methods are represented by Lasso, Elastic Net, and stability selection with Lasso (Meinshausen and Buhlmann, 2010). Recurrent feature elimination using a linear kernel Support Vector Machine (RFE L-SVM), a RF, or a Ridge are categorized as wrapper approaches. We note that the results reported in Table 5.2 were obtained using the RF classifier, but, with features selected by those feature selection approaches.

In BRATS, with the proposed approach, we obtained a set of 329 features. The remaining methods selected the following number of features: elastic net – 440, Lasso – 168, RFE L-SVM – 420, RFE RF – 350, RFE Ridge – 400, and Lasso stability selection – 572. For the case of SPES challenge data, the proposed approach yielded a subset of 117 features. In this application, the other methods selected the following number of features: elastic net – 383, Lasso – 376, RFE L-SVM – 200, RFE RF – 350, RFE Ridge – 550, and Lasso stability selection – 583. See # features BRATS/SPES in Table 5.2. It is important to note that these embedded and wrapper methods were executed in a cross-validation scheme to find the optimal parameters and features. The only statistical difference was observed when comparing the proposed method with RFE RF in SPES, under a paired Wilcoxon Signed-Rank Test with significance level of $\frac{0.05}{7}$ (Bonferroni-correction).

Table 5.2: Comparison of feature selection methods on BRATS and SPES data. The percentages indicate the fraction of features retained after feature selection. The metrics in the right correspond to the Dice on BRATS 2013 (Leaderboard and Challenge) and SPES.

				BRATS Leaderboard			BRATS Challenge			SPES
Selection approach	**Method**	**# features BRATS**	**# features SPES**	**Complete**	**Core**	**Enh.**	**Complete**	**Core**	**Enh.**	**Penumbra**
Embedded	Elastic Net	440 (73%)	383 (64%)	0.73 ± 0.22	0.60 ± 0.28	0.58 ± 0.33	0.81 ± 0.05	0.75 ± 0.14	0.72 ± 0.10	0.75 ± 0.14
	Lasso	168 (28%)	376 (63%)	0.73 ± 0.22	0.57 ± 0.28	0.56 ± 0.32	0.81 ± 0.05	0.73 ± 0.14	0.71 ± 0.10	0.74 ± 0.14
	Lasso stability selection	572 (95%)	583 (97%)	0.74 ± 0.21	0.61 ± 0.28	0.57 ± 0.33	0.80 ± 0.05	0.75 ± 0.14	0.72 ± 0.10	0.74 ± 0.14
Wrapper	RFE L-SVM	420 (70%)	360 (60%)	0.74 ± 0.22	0.61 ± 0.28	0.58 ± 0.33	0.80 ± 0.05	0.74 ± 0.14	0.72 ± 0.10	0.75 ± 0.14
	RFE RF	350 (58%)	200 (33%)	0.74 ± 0.21	0.60 ± 0.28	0.58 ± 0.33	0.80 ± 0.05	0.74 ± 0.14	0.72 ± 0.10	0.75 ± 0.14
	RFE Ridge	400 (67%)	550 (92%)	0.74 ± 0.21	0.60 ± 0.28	0.57 ± 0.33	0.80 ± 0.05	0.74 ± 0.13	0.72 ± 0.10	0.74 ± 0.14
	Proposed – All feat.	600 (100%)	600 (100%)	0.74 ± 0.21	0.61 ± 0.28	0.58 ± 0.33	0.81 ± 0.05	0.74 ± 0.13	0.72 ± 0.10	0.75 ± 0.14
	Proposed – Sel. feat.	329 (55%)	117 (20%)	0.73 ± 0.22	0.60 ± 0.28	0.58 ± 0.33	0.81 ± 0.04	0.74 ± 0.13	0.72 ± 0.09	0.74 ± 0.14

5.5.2 Comparison with other segmentation methods

Although the focus of this work is on interpretability of our machine learning system instead of performance, we compared our results with the contestants of the on-site BRATS 2013 (Table 5.3) and SPES 2015 (Table 5.4) challenges. The compared methods include proposals based on ensembles of randomized trees, such as Festa, Meier, Reza, and Tustison, in BRATS; or CH-Insel, DZ-Uzl, and BE-Kul2, in SPES. The method CA-Usher, in SPES, is built over a supervised Representation Learning algorithm (CNN).

We observed that in BRATS the set of important features changes accordingly to the task at hand (c.f. Subsection 5.5.3.1). For example, when we segment the complete tumor as a binary problem, or all tissues as a multi-label segmentation problem. Hence, motivated by this observation, and inspired by Meier et al. (2014a), we evaluated a hierarchical approach: First, we segmented the complete tumor, then we segmented the tumor tissues inside the previously defined region of interest. In Table 5.3 it is possible to observe that the detection of the complete tumor improved with the hierarchical approach, suggesting that different features are useful for different tasks. In BRATS (Table 5.3), the proposed model achieved a lower Dice compared to Tustison. However, it is on pair with the other top methods. In SPES (Table 5.4), the obtained results are comparable with the algorithms in the mid-table positions.

Table 5.3: Comparison with other methods on BRATS 2013 challenge set. Results obtained from (Menze et al., 2015).

	Dice		
Method	**Complete**	**Core**	**Enh.**
Cordier	0.84	0.68	0.65
Doyle	0.71	0.46	0.52
Festa	0.72	0.66	0.67
Meier	0.82	0.73	0.69
Reza	0.83	0.72	0.72
Tustison	0.87	0.78	0.74
Zhao	0.84	0.70	0.65
Proposed – All feat.	0.81	0.74	0.72
Proposed – Sel. feat.	0.81	0.74	0.72
Proposed – Hierarchical	0.84	0.74	0.71

Table 5.4: Comparison with other methods on SPES challenge set. Results obtained from (Maier et al., 2017).

	Penumbra	
Selection method	**Dice**	**ASSD**
CH-Insel	0.82 ± 0.08	1.65 ± 1.40
DE-Uzl	0.81 ± 0.09	1.36 ± 0.74
BE-Kul2	0.78 ± 0.09	2.77 ± 3.27
CN-Neu	0.76 ± 0.09	2.29 ± 1.76
DE-UKF	0.73 ± 0.13	2.44 ± 1.93
BE-Kul1	0.67 ± 0.24	4.00 ± 3.39
CA-Usher	0.54 ± 0.26	5.53 ± 7.59
Proposed – All feat.	0.75 ± 0.14	2.43 ± 1.93
Proposed – Sel. Feat.	0.74 ± 0.14	2.48 ± 2.04

5.5.3 Interpretability

We present two case studies for interpretability: brain tumor segmentation (Subsection 5.5.3.1) and penumbra estimation in acute ischemic stroke (Subsection 5.5.3.2). In both cases, we present global and local interpretation results.

5.5.3.1 Brain tumor segmentation

Since the segmentation of brain tumors and their sub compartments reflect a multi-label classification problem, we can define different segmentation tasks: all tissues at once, complete tumor vs. normal tissue, enhancing tumor vs. remaining tissues, or necrosis vs. remaining tissues. The first task is a multi-label classification problem, where the target labels are all tissues – normal, necrosis, edema, non-enhancing, and enhancing tumor. The other tasks are binary classification problems; in the case of complete tumor vs. normal tissue, we fuse all tumor tissues of the manual segmentation into just one class, and we use it for training. These different segmentation tasks serve the purpose of interrogating the machine learning system at hand on the usefulness of features extracted from the different MRI sequences. To leverage interpretability, we selected important features with the proposed feature selection method (Subsection 5.3.1.3), leading to the following number of selected features: 329 (all tissues at once), 403 (complete tumor vs. normal tissue), 149 (enhancing tumor vs. remaining tissues), and 92 (necrosis vs. remaining tissues). For each resulting feature set, we trained a RF model and performed global and local model interpretation analyses.

Global interpretation We can interpret the model from a global point of view by inspecting how it learned the input data. Fig. 5.4 shows the squared L2-norm plots for global interpretability, as well as some feature maps representative of each zone of importance alongside with the sequence to which they are more related to (in terms of mutual information). In Fig. 5.4 it is possible to observe that features

encoding information from the T1 sequence are mostly relegated to the tail of the important features. In contrast, features computed by hidden nodes that were strongly connected to T1c, T2, or FLAIR are given more importance. In very specific tasks, such as segmenting enhancing tumor (Fig. 5.4(c)), or necrosis (Fig. 5.4(d)), some particular sequences are preferred, such as T1c, or T2, respectively. On the other hand, the top features of more complex (multi-label) tasks, such as segmenting all tissues at once, have a higher mixture of features strongly connected to T1c, T2, and Flair. A similar behavior is observed when we segment the complete tumor (Fig. 5.4(b)), with the difference that T1c is less important, because FLAIR and T2 are sufficient to delineate the lesion as a whole. Interestingly, the hidden nodes of the RBM are more connected to one specific MRI sequence, instead of collecting information and combining multiple sequences. This is confirmed by the feature maps that can depict some specific tissues. For instance in Fig. 5.4(a), left, it is conspicuous for enhancing tumor, or in Fig. 5.4(b), second from left, the feature map appears to enhance edema.

Local interpretation From the local interpretability point of view, we studied the local spatial feature relevance for assessing how the input data was used for voxel-wise predictions in a given test subject (Fig. 5.5 and Fig. 5.6). Similarly to the global interpretability, we studied the local interpretability for the different tumor segmentation tasks (i.e. “all tissues at once”, “complete vs. normal tissue”, “enhancing vs. remaining tissues”, and “necrosis vs. remaining tissues”). From Fig. 5.5 it is observed that the identification of each class is attributed to a subset of the available MRI sequences coherent with observations from the global analysis. For instance, enhancing tumor is strongly linked to T1c, while necrosis extracts more information from T2, but also from T1c in some extent. In the case of complete tumor (Fig. 5.6(a)), FLAIR resembles to play an important role, although T1c contributes considerably in the region of enhancing tumor. In contrast to global interpretability, the local interpretability analysis allows us to better disentangle the relevance of the different sequences for a particular patient and image region as well as to study the cause of false positive segmentations. As an example, in Fig. 5.6(a) on the superior part of the brain there are some false positive tumor segmentations visible. We observed that they are more related to T1c and FLAIR than to the remaining sequences. As discussed below, the local interpretability analysis allowed us to find potential causes and pre-processing related issues that led to these false positives.

5.5.3.2 Penumbra estimation

Contrasting to the multi-class brain tumor segmentation, in SPES the task is binary and aims at segmenting the penumbra region. On the other hand, SPES contains seven MRI sequences including structural and physiological information, in contrast to the four structural MRI sequences in BRATS. In this dataset, the number of selected features by the proposed approach was 117.

Global interpretation Fig. 5.7 shows the squared L2-norm plots for the global interpretability of the model, as well as some feature maps and the MRI sequence to which the respective hidden unit is most connected. First, we can observe that some specific MRI sequences contribute much more than others to the most relevant features. The first three most important features come from the TTP sequence, while

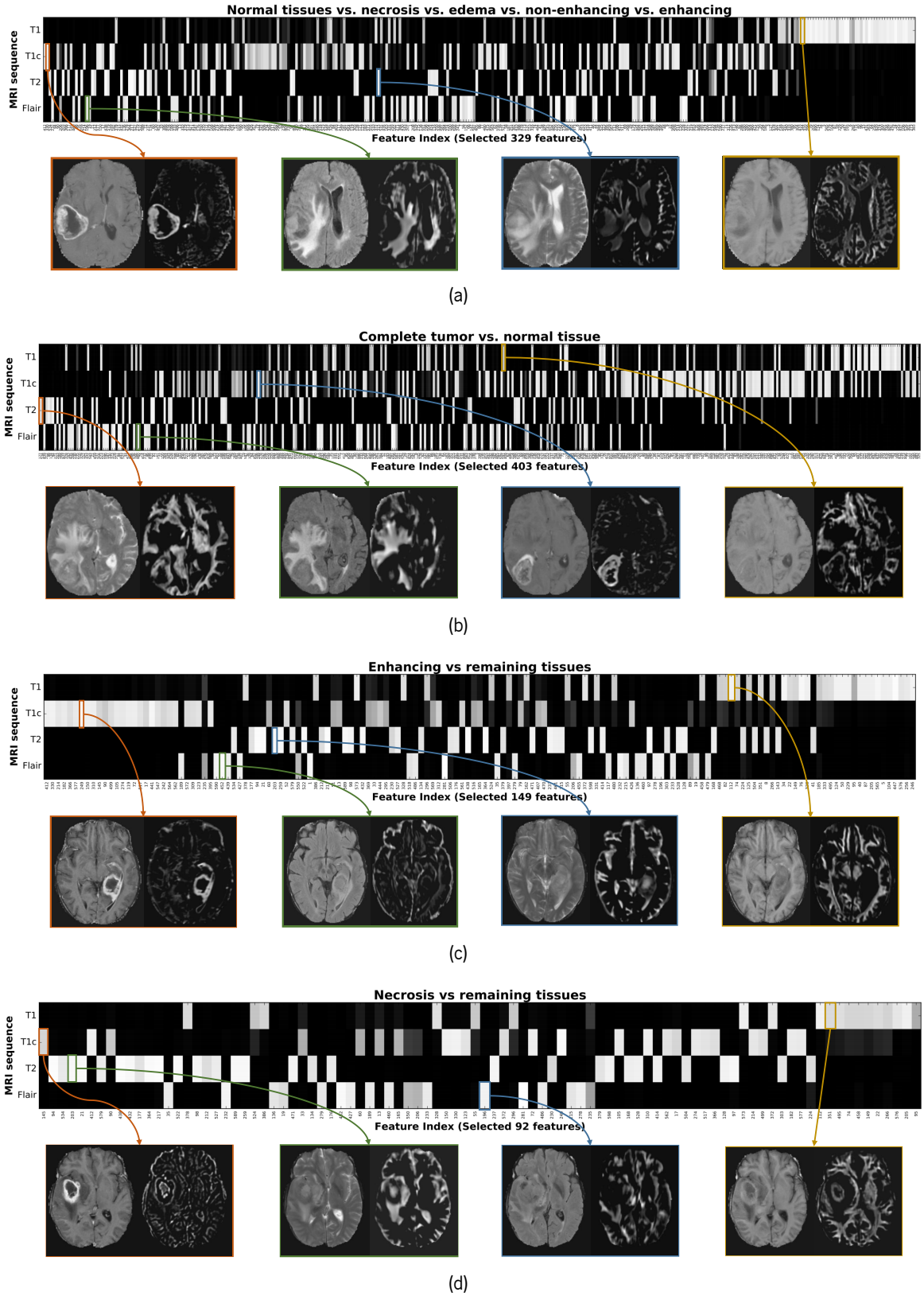


Figure 5.4: Global interpretability on the BRATS model. Several segmentation tasks are studied: a) all tissues at once (multi-label), b) complete tumor vs. normal tissues, c) enhancing tumor vs. remaining tissues, and d) necrosis vs. remaining tissues. For each task we show: top) squared L2-norm plots. Features are sorted from most to least important (left to right). Brighter means higher squared L2-norm of the weights connecting the hidden unit of a given feature to a given MRI sequence. Bottom) examples of pairs of MRI sequences (left) and feature maps (right).

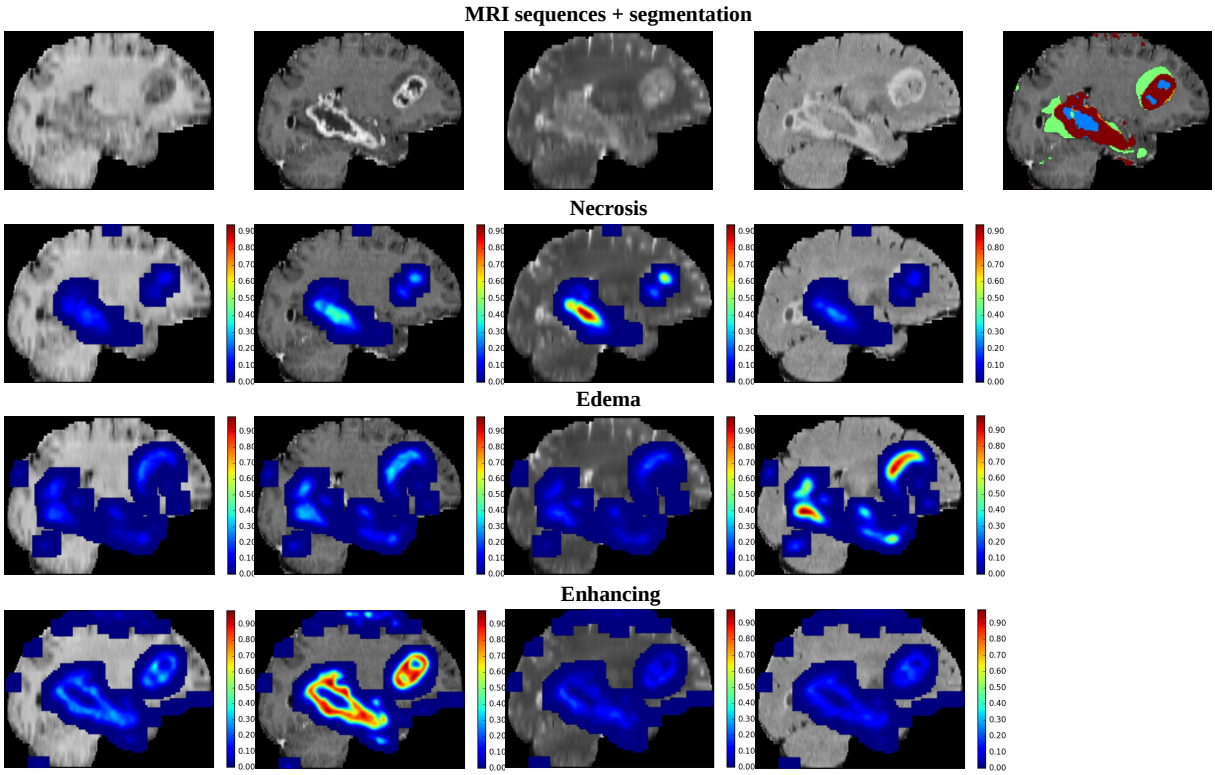


Figure 5.5: Spatial feature relevance for local interpretability of the Challenge subject 0310 in BRATS for the task of segmenting all tissues at once (multi-label classification). From left to right, we show the T1, T1c, T2, and FLAIR sequences, as well as the obtained segmentation in the first row. In the segmentation, the tumor tissues are: blue – necrosis, green – edema, orange – non-enhancing, and red – enhancing tumor. Each row corresponds to how the input data was used for predicting each class.

overall the Tmax sequence has the largest number of most important features. Observing the feature maps in Fig. 5.7, they characterize the stroke region either as a hypointense or hyperintense area. MRI sequences such as DWI, T1c, and T2, have some features strongly connected to them for the most important features, but appear mainly on the least important section of the ranked features in Fig. 5.7 top. Interestingly, contrasting to BRATS (Fig. 5.4) where features are mostly related to just one MRI sequence, in penumbra estimation some features are computed from both the DWI and T2 sequences. Finally, the CBF and CBV sequences are barely represented.

Local interpretation Local spatial feature relevance for penumbra estimation is presented in Fig. 5.8. It is possible to observe that Tmax and TTP are the sequences from which the model takes more information. TTP appears with higher magnitude for relevance, but in the posterior part of the segmentation it is lower compared to Tmax. Features related to the other MRI sequences are less preferred than Tmax and TTP, with T1c appearing with a larger contribution to approximate the overall stroke region segmentation, and CBV appearing to be the least important for voxel-wise predictions.

Removal of MRI sequences Given the observations from the global and local interpretations that CBV and CBF play a minor role in penumbra estimation, we experimented to train the system without those MRI sequences. The system trained without the CBV sequence results in ASSD of 2.58 and Dice

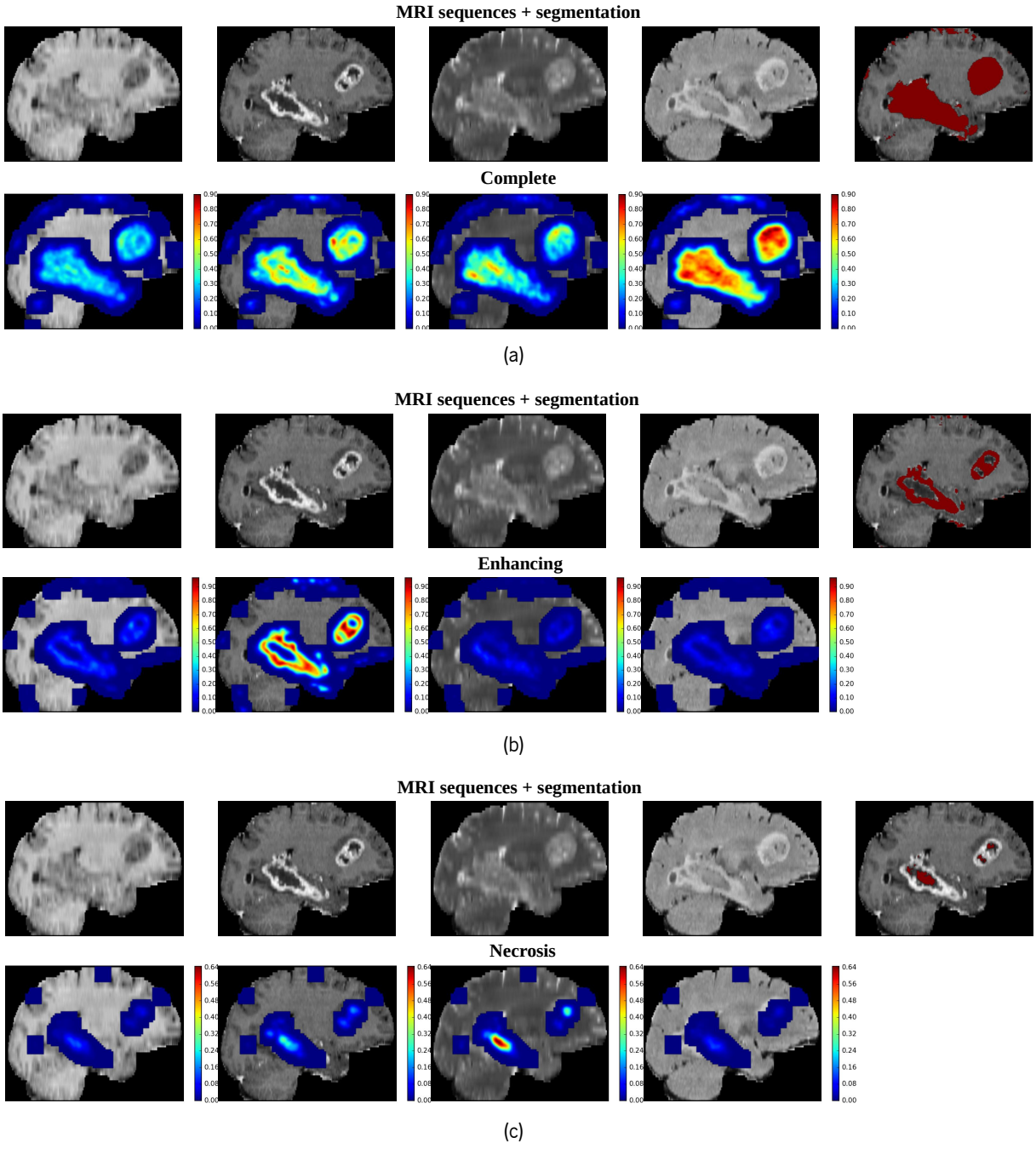


Figure 5.6: Spatial feature relevance for local interpretability of the Challenge subject 0310 in BRATS for the tasks: a) complete tumor vs. normal tissues, b) enhancing tumor vs. remaining tissues, and c) necrosis vs. remaining tissues. From left to right we show the T1, T1c, T2, and FLAIR sequences, as well as the obtained segmentation in the first row of each task.

of 0.74. When we further removed both the CBV and CBF sequences the results were: ASSD – 2.54 and Dice – 0.74. Comparing with Table 5.4, we can observe that the results are equivalent to those using all the available MRI sequences.

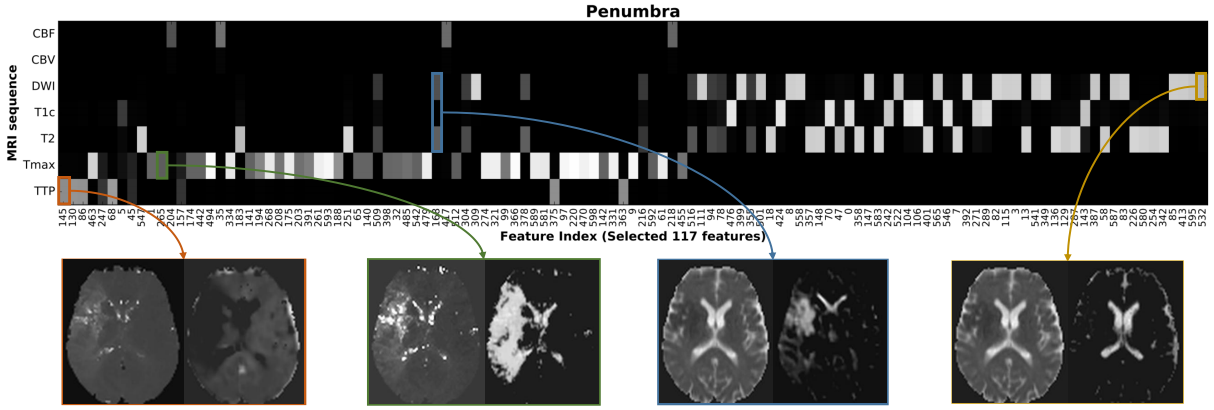


Figure 5.7: Global interpretability on the SPES model. Top) squared L2-norm plots. Features are sorted from most to least important (left to right). Brighter means higher squared L2-norm of the weights connecting the hidden unit of a given feature to a given MRI sequence. Bottom) examples of pairs of MRI sequences (left) and feature maps (right).

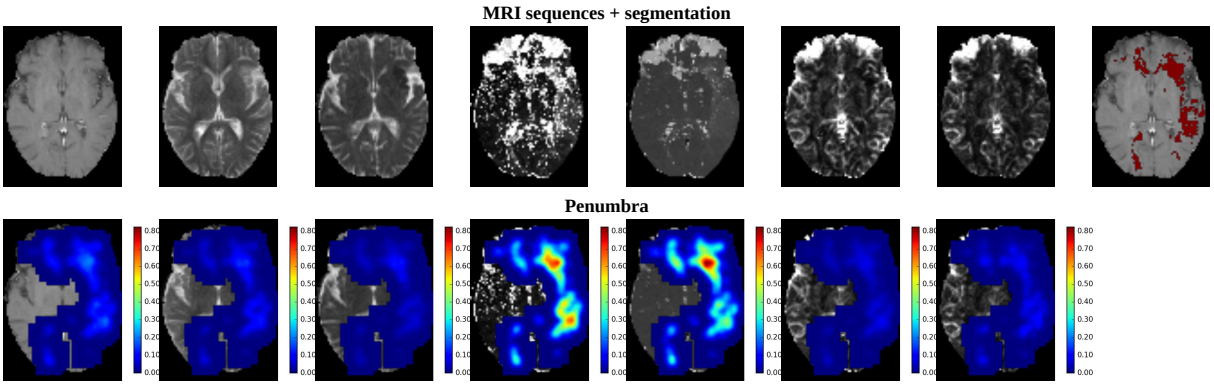


Figure 5.8: Spatial local interpretability of the SPES Challenge subject 1. Top) MRI sequences and segmentation. From left to right we show the T1c, T2, DWI, Tmax, TTP, CBV, and CBF sequences, as well as the obtained segmentation. Bottom) local interpretability maps.

5.6 Discussion

Machine learning systems are pervasively being adopted as decision-support systems in critical fields, like the medical domain. At the same time, the models are increasingly complex for the sake of performance. This may pose a problem in their adoption due to trust reasons, as it is difficult to explain the model and/or the respective predictions. Thus, there is a need to understand *how* a model learned the input data, and *why* a certain prediction was made. In this paper, we propose a novel methodology to enhance the interpretability of automatically extracted machine learning features. We also investigated the notion of global and local interpretability. Global interpretability provides insights as to *how* the model learned the data from a population. This allows us to infer if the studied model is coherent with the experts knowledge. On the other hand, local interpretability leverages the understanding on *why* a prediction on a subject-specific level was made. A limitation of interpretability systems is that there is no quantitative metric available to measure interpretability of a machine learning system yet. Thus, we base our discussion on extensive comparison with medical domain knowledge. The proposed machine learning system encompasses a RBM as representation mapping stage and a RF as task-related learner. High-dimensional

feature vectors may impair interpretability. Thus, we propose a MI-based feature selection scheme that simultaneously take into account the mapping between the input data and its representation (features), and from the representations to the the task at hand (labels).

5.6.1 Joint RBM-RF approach for feature selection

From Table 5.2 we observed that all feature selection methods perform similarly, both in mean and standard deviation; the only statistical difference was found when comparing the proposed approach with RFE RF in SPES. However, just by looking at the metrics, some embedded wrapper methods seem to achieve slightly better performances. However, they were evaluated in a cross-validation scheme to find the best parameters and number of features. Moreover, Wrapper recursive feature elimination schemes require training several models with progressively smaller feature vectors. The embedded methods do not require the recursive feature elimination scheme. However, the selected features may be optimal for the used learner, but not for the model that we must use in the end. Contrarily, by combining RBM-MI and RF-MDI, the proposed approach offers the advantage of automatically selecting the optimal number of features and does not require a threshold to be defined, nor a recursive feature elimination scheme. The main motivation for feature selection in the context of this paper is to choose features that both correlate with the data and the labels, and to leverage interpretability. In that sense, the proposed approach provides more satisfactory results than the other methods, by decreasing the dimension of the feature vector to almost half in BRATS, and to around 20% in SPES (only Lasso in BRATS achieves a more compact feature subset). Moreover, the proposed feature selection approach does not impact the segmentation performance, when comparing results to a model using all 600 RBM-derived features. Thus, the hypothesis that selecting features with high mutual information both with labels and the data is viable. In this experiment, we proposed to use Pearson Correlation Coefficient computed over the RBM-MI and the RF-MDI measures to detect the transition between important and unimportant features. This choice came from the observation of the decreasing regime of both metrics, and its empirical nature represents the main limitation of the feature selection approach.

At the same time, this big reduction in the number of features imply that although RBM can automatically compute features, many of them can be useless in the presence of more powerful ones. This may be caused, in part, by the unsupervised nature of RBM, since it does not know for which task the features are going to be employed (Larochelle and Bengio, 2008). Although models using a high number of features can be prone to overfitting, in our experiments we did not observe such tendency, probably due to the robustness of RF (Criminisi and Shotton, 2013).

Observing that the subset of important features changes with the task at hand, we devised a hierarchical approach in BRATS. This allowed us to improve the segmentation of complete tumor. With this approach, we are on pair with the methods of the on-site results of BRATS 2013 Challenge, with the exception of the winner of that edition. We note that despite these methods being from 2013, they still remain as representative of RF-based approaches. All of them rely on hand-crafted features, while ours is based on an unsupervised representation mapping algorithm. In SPES, the proposed machine learning system is positioned on pair with the mid-table methods. However, top methods incorporate expert prior knowl-

edge. CH-Insel includes atlases information, and presence of the voxel in the ipsi- or contralesional side. Additionally, both CH-Insel and De-Uzl compute features of symmetry in relation to the mid-sagittal plane. This kind of information cannot be captured by a representation learning algorithm. Even so, although the proposed system is based on an unsupervised model, it achieved better metrics than a CNN-based proposal (CA-Usher). We note, however, that the CA-Usher team achieved high performances in terms of metrics in the training set (similarly as all the other teams). This behavior may be related to overfitting on the training by this team, which may have been alleviated by the unsupervised nature of our approach.

5.6.2 Interpreting automatically extracted features in brain tumors

We can define several segmentation tasks in brain tumor image analysis. This allows us to interrogate and interpret the machine learning system and assess if it is learning well, according to clinical expert knowledge on the problem. Observing the squared L2-norm plots of the RBM weights connecting the hidden and visible units representing each MRI sequence in Fig. 5.4, we obtained insight into which sequences are more important for the different tasks. When we segment all tissues at once (Fig. 5.4(a)), the most important features extract information from the T1c, T2, and FLAIR sequences. As expected, since T1 adds less information to the other ones, the features connecting strongly to this sequence are the least important, or appear sparsely represented in the most important features. From a clinical point of view, this result is valid for pre-operative brain tumor images as contained in the BRATS 2013 data set. In the feature maps, we note that specific patterns of enhancing tumor and edema were extracted from T1c and FLAIR. Fluid-filled compartments are mostly enhanced in T2, while T1 encodes the fatty tissue, mainly. This global interpretation is in line with clinical domain knowledge, and, hence, allows us to conclude that the model is correctly utilizing the imaging information. From the local interpretation results, in Fig. 5.5, we can observe not only which sequence contributes the most for each label class, but also study this contribution with respect to image regions. For example, segmentation of edema is mainly based on FLAIR. However, according to Menze et al. (2015), segmentation of edema is also based on T2. The reason for the model to prefer FLAIR may be because it can differentiate edema from the cerebrospinal fluid, which is contained e.g. in the ventricles. T1c is the second choice in predicting edema. Although one could expect T2 to appear after Flair, the enhancing rim reflects a strong prior on the extent of the tumor core (hence limiting the extent of edema, too), at least in the images of BRATS 2013, which are mostly high-grade glioma. Thus, since we deal with patches in our system, the model may learn that close to edema, features with high response on T1c may exist. As expected, T1c clearly dominates the predictions of enhancing tumor.

5.6.2.1 Interpreting the binary brain tumor segmentation tasks

Regarding the binary brain tumor segmentation tasks, we observed that among the most important features for complete tumor (Fig. 5.4(b)) there are features strongly related to all sequences, contrasting, for instance, with segmentation of the enhancing tumor, or necrosis. This is due to the higher variability of the tumoral tissues included in the “complete tumor” region, than in the other binary tasks where the tissue type shows much lower intensity variability across sequences. For the binary task of segmenting

the complete tumor as a whole, the T2 and FLAIR sequences are more important. This is in accordance with the manual segmentation protocol used in BRATS, where the complete tumor was firstly defined based in those same MRI sequences (Menze et al., 2015). From the first (i.e. most-left) feature map in Fig. 5.4(b), it can be observed that the hypointense portion of the T2 image is encoded in the respective feature map, including mainly areas of white matter and solid tumor tissue. Interestingly, some T1c-based features appear to capture intra-tumoral regions. This can be observed in the third (from left to right) feature map of Fig. 5.4(b), and in the local explanation of predictions in Fig. 5.6(a), where T1c is important in the enhancing rim region. This observation shows that the Ridge regression is able to identify locally important features, indeed. However, following results from both global and local analysis, features derived from FLAIR appear to be the most dominant for defining the complete tumor, in general. When we segmented the enhancing tumor against all the other tissues, the most important features are provided by hidden units with their weights strongly connected to the T1c visible units. In Fig. 5.6(b), we also observed that locally the T1c sequence completely dominates over the other ones. This was expected, since enhancing tumor region is characterized by the T1c image. Finally, in the necrosis segmentation task, features that are strongly connected to the T2 sequence appeared more often among the most important features in the global squared L2-norm plots (Fig. 5.4). However, some T1c-based are also ranked among the top. For the patient case 0310, shown in Fig. 5.6(c), we can see two lesions, each with necrotic tissue. Interestingly, for segmenting the larger necrotic core of the more central lesion, the T2 sequence seems to be more relevant than T1c. In contrast, for segmenting necrosis in the smaller lesion both sequences appear equally important. This observed complementarity of T1c and T2 is in line with the manual segmentation protocol (Menze et al., 2015), since both T1c and T2 are used for defining the tumor core and for differentiating the non-enhancing part of the tumor from necrosis and enhancing tumor. Necrotic regions are mostly surrounded by the enhancing tumor, hence the importance of the T1c for this task. Indeed, in the feature map related to this sequence, shown in Fig. 5.4(d), the enhancing rim appears completely dark, while the inner part, corresponding to necrosis, is enhanced. In this way, we are able to verify and conclude that the system learned the correct relations in the data. We emphasize, however, that these observations apply for pre-operative acquisitions only. In a post-operative setting the relevance of the different MRI sequences for tumor segmentation is different (Meier et al., 2017).

5.6.2.2 Further considerations

Apart from studying the relevance of the different sequences for tumor tissue predictions, we observed for patient case 0310 in Fig. 5.6(a) some misclassified tumor regions on the top of the brain. The dominant source for this misclassification are features related to the T1c or FLAIR sequences. Hence, these misclassifications are probably caused by errors of the skull-stripping procedure, which was insufficient for this image region (remaining extra-cerebral tissue such as e.g. meningeal tissue that typically appears enhanced in T1c, similarly to tumor tissue).

All the previously mentioned observations suggest that, despite being a completely unsupervised algorithm, the RBM is able to identify tissue patterns and the most important spatial features of the imaged pathologies at hand. Moreover, by inspecting the strength of the weights we can identify which inputs

were considered more important for each feature. Taking these results into account, it is clear that RBM computes representations that are more important depending on the task at hand. This is confirmed by our hierarchical approach yielding improved results over the "segment all at once" scheme, and more similar to those obtained by hand-crafted features.

Although some sequences are clearly less important than others, we observed a degradation in performance if some MRI sequence is removed. This is in accordance with Havaei et al. (2016), since the authors observed a performance drop when some sequence is missing, too. Brain tumors are characterized by being a heterogeneous kind of lesion, with some portions conspicuous only in some MRI sequences. As we can observe in Fig. 5.4, 5.5, and 5.6, all the four MRI sequences have some importance for some task.

5.6.3 Interpreting automatically extracted features in acute ischemic stroke

We also evaluated the proposed methodologies in a penumbra estimation problem, using the SPES database of the ISLES2015 MICCAI challenge. During acute ischemic stroke, the most severely affected region, the core lesion, consists of irreversibly damaged tissue. Penumbra refers to the larger region of dysfunctional, but salvageable, tissue at risk of infarction. Hence, it is imperative to treat this region as fast as possible. The core lesion is conspicuous in DWI, but penumbra is not (Copen et al., 2011; Straka et al., 2010). Additionally, the core lesion is smaller than the penumbra region. Possibly because of this, features computed by hidden nodes that are strongly connected to the DWI sequence were not ranked among the most important features in the squared L2-norm plot for global interpretability, Fig. 5.7. If the system heavily relied on DWI, it could underestimate the penumbra extension (observe third feature map in Fig. 5.7, which relies both on DWI and T2). Conversely, the penumbra is better visualized in Perfusion Weighting Imaging. CBV corresponds to the blood volume in the area. Regions of low CBV correlate with the core and final outcome of the infarction; however, we are interested in predicting the complete penumbra. In turn, CBF is related to the supply of oxygen and nutrients to the tissue, being more correlated to the salvageable tissue. Thus, CBF allows us to study which regions are underperfused. Still, both CBV and CBF have some disadvantages: they are heterogeneous between gray and white matter, even in normal tissue; they are susceptible to errors caused by signal clipping, and they are affected if the blood-brain barrier is not intact (Copen et al., 2011; Straka et al., 2010). Moreover, the penumbra in those sequences is underestimated if the bolus is delayed and in short acquisitions. Nevertheless, CBF suffers less from this problem than CBV (Copen et al., 2011; Straka et al., 2010). While the machine learning system selected just a few hidden nodes that compute CBF-related features, it did not select any for CBV, Fig. 5.7. This may be related with CBF being more related to the penumbra, while CBV appears to identify regions closer to the core. Additionally, the disadvantages of these sequences, pointed out above, may contribute for a more heterogeneous data, hence being harder to capture relations in it. On the contrary, TTP and Tmax are the sequences with the highest importance, according to the squared L2-norm plots of Fig. 5.7. In fact, these sequences are not directly measuring perfusion, but correlate well with hypoperfusion. TTP and Tmax are more independent of the tissue type (less heterogeneous in gray and white matter) and the acquisition time, than CBV and CBF. Moreover, the lesions are conspicuous in

these sequences (Copen et al., 2011; Straka et al., 2010). For this reason, Straka et al. (2010) proposed a method for penumbra estimation based on thresholding the Tmax $>$ 6 s, and further removal of small clusters. Furthermore, the manual segmentation protocol of the SPES database starts by thresholding the Tmax sequence to have a first segmentation of the hypoperfused region. The other MRI sequences are then used to refine it, by removing the sulci, non-stroke pathologies, and previous infarcts (Maier et al., 2017). These considerations show the importance of Tmax for estimating penumbra. The feature maps shown in Fig. 5.7 appear to identify penumbra patterns, by appearing hypointense in the TTP image and hyperintense on the Tmax, in the area of interest. Some features based on T1c and T2 also appear as important, which may be related to the suppression of sulci, similarly to the manual segmentation protocol. In the spatially distributed explanation of the predictions (Fig. 5.8), the importance of TTP and Tmax is confirmed. Interestingly, one can note that those two sequences change their importance according to the location. The importance of CBV is not zero everywhere because, although no features are strongly linked to that sequence, there are some residual weights that accumulate during the local interpretability algorithm, even though we employed L1-norm to turn the least important weights to 0. This may be an artifact of the algorithm, although the magnitude of importance of CBV is much lower than the strongest responses, thus can be considered as negligible.

Although the system learned to rely on Tmax and TTP at the expense of CBV and CBF, we know from the literature that the latter sequences should have some discriminative power (Copen et al., 2011; Straka et al., 2010). Still, as previously mentioned, their importance is reduced, as computed by our methodologies. However, upon confirmation with a clinical expert, this may point to a bias introduced during the manual segmentation step, since it heavily relies on the Tmax perfusion sequence that is thresholded at 6 s. The machine learning system may learn to recognize its importance, since the expert similarly applies an intensity threshold on the Tmax image. Moreover, this may account for the success of the top-2 methods in Table 5.4, since both thresholded Tmax. So, interpreting a model may unveil potentially imperceptible biases on the training data. In fact, the possibility to disclosure problems in the data is pointed out by Ribeiro et al. (2016b) as one of the advantages of developing interpretation methodologies for machine learning systems.

When we trained a system without the CBV and CBF sequences, the results were similar to using all the MRI sequences. Of course, as mentioned before, this may be due to a bias in the manual segmentation procedure. Nevertheless, the interpretation of the system allowed us to identify MRI sequences that are less important and remove them from the system. Thus, interpretation may help to identify unimportant MRI sequences for some tasks, which can be helpful for reducing acquisition time and cost.

5.7 Conclusion

In conclusion, we propose a machine learning system based on a RBM as representation mapping and a RF as task-specific learner. Furthermore, we propose methodologies and definitions for the machine learning system interpretability, both globally and locally. Despite being a shallow model, a RBM can still learn meaningful features in an unsupervised way that are useful for segmentation. Indeed, although being

unsupervised, it learned to compute tissue specific features, as observed in the feature maps. The fact that it is shallow, however, makes it simpler to find useful information in its weights. This suggests that despite being regarded as "black boxes", we can still interpret the behavior of these models. We observed that the most important features could extract sequence- and task-specific knowledge. Contrasting with the common belief that these models mix all the information in their weights, in fact these findings suggest that it is employed in an organized fashion. Furthermore, we could verify that the system was able to capture information coherent with expert knowledge, such as the manual segmentation protocol used for BRATS 2013 (Menze et al., 2015) and penumbra estimation for SPES (Maier et al., 2017). This was observed both globally and locally (spatially distributed in the image space). For instance, in the complete tumor vs. normal tissue in BRATS, it was observed through the local interpretability methodology that FLAIR is the most important sequence, as expected. However, for the regions overlapping with the enhancing tumor the system still recognized the importance of the T1c sequence. Also, we could suggest a possible bias towards the importance of the Tmax sequence introduced by the manual segmentation protocol in SPES. With our interpretability methodologies, we aimed at improving the transparency and interpretability of Representation Learning-based methods to increase their acceptance in clinical applications. Finally, we proposed a strategy for feature selection combining RBM features and RF MDI. Summarizing, we present a methodology joining RF and RBM for data understanding, and feature extraction and selection. This approach opens opportunities to understand how MRI sequences are being used for each segmentation task; thus, it can potentially be useful to refine imaging protocols for a given segmentation task, with an impact both in acquisition time and cost. As well, it may be helpful to understand and take advantage of sequence specific features. In the future, we want to investigate how these findings can be extended for other models (for instance, Deep Belief Networks) and applications. Additionally, we will investigate more principled approaches for feature selection using both RBM-MI and RF-MDI.

5.8 Summary

Machine learning systems are achieving better performances at the cost of becoming increasingly complex. However, because of that, they become less interpretable, which may cause some distrust by the end-user of the system. This is especially important as these systems are pervasively being introduced to critical domains, such as the medical field.

Representation Learning techniques are general methods for automatic feature computation. Nevertheless, these techniques are regarded as uninterpretable "black boxes". In this chapter, we propose a methodology to enhance the interpretability of automatically extracted machine learning features. The proposed system is composed of a Restricted Boltzmann Machine for unsupervised feature learning, and a Random Forest classifier, which are combined to jointly consider existing correlations between imaging data, features, and target variables. We define two levels of interpretation: global and local. The former is devoted to understanding if the system learned the relevant relations in the data correctly, while the latter is focused on predictions performed on a voxel- and patient-level. In addition, we propose a novel feature importance strategy that considers both imaging data and target variables, and we show the ability of the

approach to leverage the interpretability of the obtained representation for the task at hand.

We evaluated the proposed methodology in brain tumor segmentation and penumbra estimation in ischemic stroke lesions. We show the ability of the proposed methodology to unveil information regarding relationships between imaging modalities and extracted features and their usefulness for the task at hand. In both clinical scenarios, we show that the proposed methodology enhances the interpretability of automatically learned features, highlighting specific learning patterns that resemble how an expert extracts relevant data from medical images.

Chapter 6

Automatic Brain Tumor Grading From MRI Data Using Convolutional Neural Networks and Quality Assessment

In this chapter we investigate automatic glioma grading from structural MRI sequences using a CNN-based approach. Glioma grading is a crucial step for treatment planning. However, biopsy and histological studies are time demanding, and prone to sampling error. Therefore, automatic glioma grading may expedite treatment planning. Furthermore, we incorporated interpretability techniques in the developed CNN-based method for quality assurance. In this way, we verify if the model is taking into consideration correct patterns of the data. Finally, we show that interpretability is helpful in identifying learning problems, thus allowing correction strategies.

This chapter is based on the following publication:

- Pereira, Sérgio, et al. "Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment." Workshop on Interpretability of Machine Intelligence in Medical Image Computing, Lecture Notes in Computer Science, 2018.

Contribution The author of this thesis was responsible for conceiving and implementing the ideas and methods described in this chapter. Furthermore, he also conducted the experiments, and analyzed the resulting data. Finally, the author was the writer of the manuscript listed above.

Contents

6.1 Introduction

Gliomas are the most common primary brain tumors, being graded according to their malignancy. The most aggressive one is Glioblastoma Multiforme. These HGGs proliferate and infiltrate the surrounding tissues at a very fast pace. In fact, patients have a very short life expectancy, even if under treatment (Van Meir et al., 2010). Lower grade gliomas are less aggressive, and patients have a better prognosis. Nevertheless, LGGs can evolve into HGGs, hence, follow-up is required (Grier and Batchelor, 2006). Glioma grading is crucial when deciding the treatment procedure, which can range from surgery followed by chemo- and radiotherapy, to a "wait and see" approach. The latter avoids invasive procedures and is more common with LGGs (Grier and Batchelor, 2006; Menze et al., 2015).

Histopathological diagnosis of biopsy specimens is the gold standard for glioma grading. However, it is time consuming, invasive, and prone to sampling error (Zacharaki et al., 2009). MRI is the standard imaging technique for brain tumor diagnosis in clinical practice. In general, attributes of HGG in MRI include the contrast enhancing tumor tissue, necrotic core, edema, non-enhancing tumor, and mass effect. LGGs are usually more diffuse, non-enhancing, smaller, and cause less mass effect. Nonetheless, some HGGs may have some attributes of LGGs, and vice versa (Grier and Batchelor, 2006; Steed et al., 2018; Van Meir et al., 2010). Tumor grading from imaging data would be useful in clinical practice, since it would avoid the sampling error, and expedite treatment planning by anticipating the histopathological results (Zacharaki et al., 2009). Additionally, it would avoid the invasive biopsy procedures during follow-up. Studies suggest that perfusion MRI is more informative for glioma grading than structural MRI sequences (Zacharaki et al., 2009). Still, perfusion MRI is not widely acquired in clinical practice (Essig et al., 2013); in fact, perfusion MRI is seen as a plus, while structural MRI is part of the current consensus recommendations for standardized brain tumor imaging (Ellingson et al., 2015). Computer-based tumor grading from MRI is relatively unexplored. Zacharaki et al. (2009) predict the grade of gliomas from MRI images using a Support Vector Machine classifier. The method requires radiologists to manually define four ROIs in the tumor. Khawaldeh et al. (2017) use CNNs in a semi-automated approach where the tumor grade is predicted from 2D slices selected by radiologists, which may result in multiple and possibly ambiguous predictions for the same patient.

Convolutional Neural Networks offer the potential for learning tumor grading directly from imaging data without human-defined ROIs. However, these methods may fall into overfitting, and learn spurious patterns in the data. Hence, a quality assurance stage before deployment of these methods is desirable. As shown by Pereira et al. (2018c), interpretability of machine learning methods, through explanations of their predictions, allows one to assess which parts of the MRI image are more important for a prediction. In this way, one can evaluate if a model is trustworthy. Moreover, explanations may provide hints on undesirable behaviors, and allow one to devise improving strategies.

The contributions in this section are the following. i) We propose to use 3D CNN for automatic glioma grading from conventional multisequence MRI, either from the whole brain, or an automatically defined tumor ROI. ii) We assess the predictions by means of visual explanations. In this way, we were able to assess the predictions' trustworthiness and, as shown in the experiments, detect a problem in pre-processing. Finally, iii) we validate our approach on a publicly available database, making it more easily comparable with future proposals.

6.2 Methods

The proposed grading system has two main stages: ROI extraction, and glioma grade prediction. Additionally, we have an interpretation of predictions stage that serves as prediction quality assessment, and we use it for two purposes. First, to evaluate if regions indicative of tumor grade are the most relevant ones for classification. Second, to identify possible problems with the method (e.g. focus on spurious patterns) and devise strategies to obtain better classifiers.

6.2.1 Extraction of the region of interest

We consider and evaluate glioma grading from two ROI: the whole brain, and the tumor region. First, we automatically identify these regions in the image, and define a bounding box around them. Second, these volumes are extracted, resized to a fixed size, and fed into the tumor grade classification CNN. We note that an independent CNN is trained for each of the ROI. Regarding the whole brain region, in a skull-stripped image a bounding box can be easily defined from the brain mask.

For the tumor ROI, a bounding box is defined after segmenting the whole tumor. In order to account for segmentation mistakes, we give a margin of 10 voxels in each side of the bounding box, while maintaining the aspect ratio of the tumor.

Segmentation of the whole tumor from multisequence MRI is achieved with a 3D U-net-inspired (Ronneberger et al., 2015) fully convolutional network; the network architecture is depicted in Fig. 6.1 (top). A 3D patch is extracted from each MRI sequence, stacked as channels, and fed into the network. The encoder path is responsible for learning the higher order features. Max-pooling layers increase the field of view, but downsample the feature maps. Features computed by higher (deeper) convolutional layers are more abstract. However, these features lack fine details that are important for segmentation. Since the feature maps are downsampled, we need to map the lower resolution feature maps back to the input patch resolution. This is done by upsampling. As we upsample feature maps, we sum them with the feature maps of equivalent size of lower layers of the encoder path. Further convolutional layers fuse the lower and higher-level features. We also employ residual blocks with pre-activations (He et al., 2016) that make training of deep networks easier. To deal with overfitting, we used a variant of Dropout called Spatial Dropout (Tompson et al., 2015). The last layer is a $1 \times 1 \times 1$ convolutional layer, with sofmax activation.

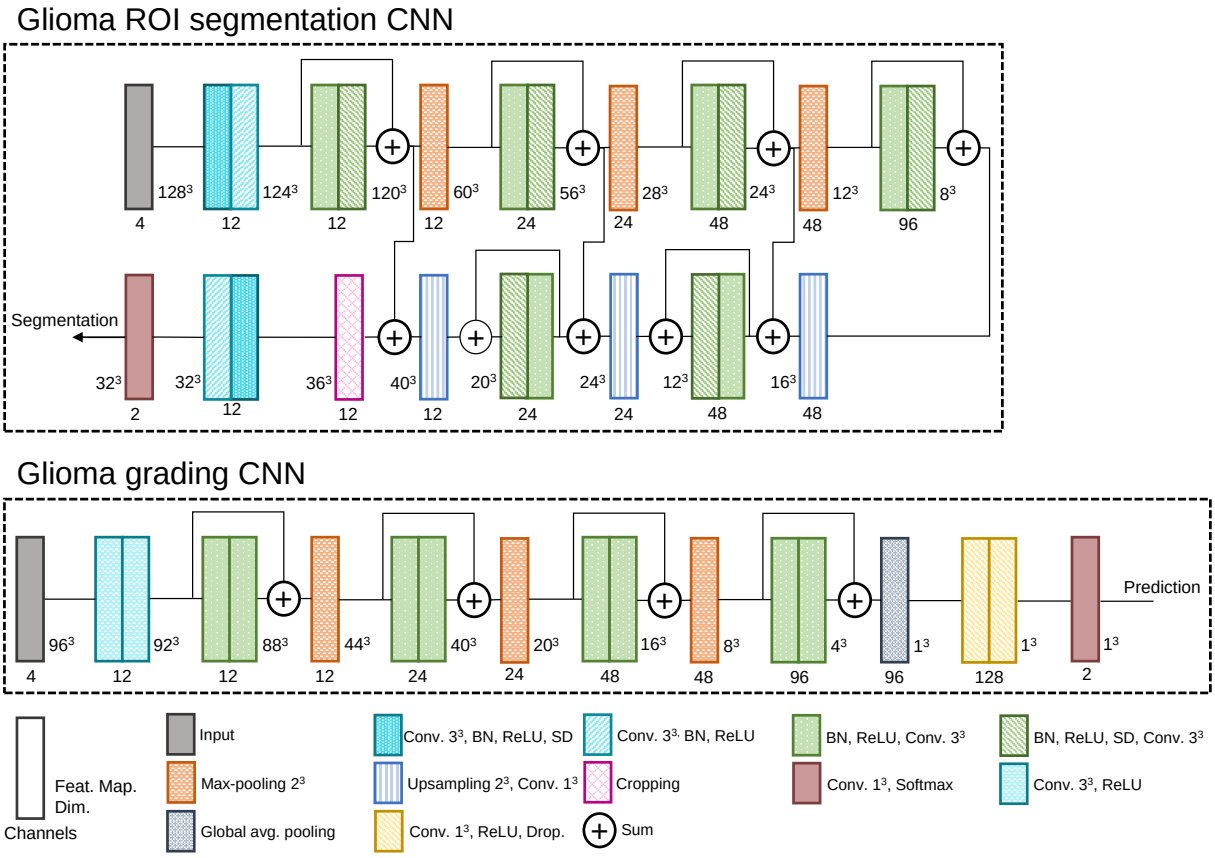


Figure 6.1: Architectures of the CNNs used for glioma segmentation (top), and tumor grade classification (middle). Description of each block can be found in the bottom. BN stands for batch normalization, SD for spatial dropout, and Drop. for dropout.

6.2.2 Glioma grading CNN

We train a glioma grading CNN with similar architecture for each ROI (Fig. 6.1, middle). The ROI is extracted from each MRI sequence and resized to 96^3, before feeding it to the CNN. In these architectures, we also employ residual convolutional blocks with pre-activations (He et al., 2016), which contribute for better learning. After the convolutional feature computation layers, we use Global Average Pooling to summarize each feature map. Then, a cascade of $1 \times 1 \times 1$ convolutional layers act as fully-connected layers. Finally, the last layer outputs a probabilistic prediction of the tumor grade. Given the amount of available data, we use aggressive on-the-fly data augmentation during training. The data augmentation procedures were: sagittal flipping, rotation of $[-20°, 20°]$, $90°$ rotation, and exponential intensity transformation with random $\gamma \in [0.85, 1.15]$.

6.2.3 Grade prediction interpretability

To perform quality assessment of tumor grade prediction, we use the interpretability methods Guided Backpropagation (Springenberg et al., 2014) and Gradient-weighted Class Activation Mapping (GradCAM) (Selvaraju et al., 2017), after extending them to 3D. This is done at prediction time.

Guided Backpropagation (Springenberg et al., 2014) is based on the idea that the gradient with respect to the input image, visualized in the image space, is informative of which parts of the image are more

discriminative for the neurons activation. It starts by computing a forward pass through the network layers. During backpropagation, the true gradient is not calculated. Instead, a variation that results in better explanations of ReLU activations is used. This is performed by zeroing both the gradients in the units with 0 value after ReLU activation, and the negative gradients. In this way, the backward signals of neurons that contribute for decreased activation are discarded. Although visually discriminative, Guided Backpropagation has the disadvantage of not being discriminative in relation to the predicted class (i.e. it can highlight areas of interest to the network but not to which class).

In contrast to Guided Backpropagation, GradCAM is class discriminative, but the explanation maps may have lower resolution. GradCAM tries to explain how the feature maps $\boldsymbol{F}$ of a layer l support the class prediction y^c. To that end, the gradient of the unit predicting the class with respect to the feature maps of the layer of interest $\frac{\partial y^c}{\partial \boldsymbol{F}^l}$ is backpropagated. Then, the weight α_l of each feature map for the class prediction is computed as the global average pooling of the gradients. Being i, j, k the indices of each of the N elements of the gradient, the weights are given by

$$\alpha_l{}^c = \frac{1}{N} \sum_i \sum_j \sum_k \frac{\partial y^c}{\partial \boldsymbol{F}^l_{ijk}}. \tag{6.1}$$

Finally, the explanation map E^c for the class is generated by the sum of F^l weighted by $\alpha_l{}^c$, as

$$E^c = \max\left(\sum_l \alpha_l{}^c F^l, 0\right). \tag{6.2}$$

The $\max(\cdot, 0)$ function discards information contributing for decreased activation for the class. The explanation map has the same resolution as the feature maps of interest, thus, interpolation is typically needed to map results to the original image space.

6.3 Experimental Setup

The proposed methods were evaluated using BRATS 2017 Training set (Bakas et al., 2017; Menze et al., 2015), which has the particularity that subjects are organized according to the tumor grade into HGG (Glioblastoma Multiforme) and LGG. There are 285 pre-operative acquisitions: 210 HGG, and 75 LGG. For each subject there are 4 MRI sequences available with $1\ mm$ isotropic resolution: T1, T1c, T2, and FLAIR. All sequences are already aligned, and skull stripped. We randomly divided the 285 subjects into 60% training, 20% validation, and 20% testing[1]. The manual segmentations of the different tumor compartments were merged into a single label to train the whole tumor segmentation network.

Two pre-processing steps are applied: bias field correction (Tustison et al., 2010), and standardization of the image intensities inside the brain mask to zero mean and unit variance. All networks were trained with the Adam optimizer and crossentropy loss. For the whole tumor segmentation, learning rate was set to 5×10^{-5}, spatial dropout probability to 0.05, and weight decay to 1×10^{-6}. Regarding the CNNs

[1]Grades' proportions were maintained in each set. The subjects' id in each set are available online: `https://github.com/sergiormpereira/brain_tumor_grading`.

for tumor grade prediction, the hyperparameters of the network were: learning rate — 1×10^{-4}, dropout probability — 0.4, and weight decay — 1×10^{-4}. We used convolutional operations without padding, therefore, in skip connections, we cropped the feature maps to the same size of the smaller ones, before summing. During training, the bounding box of tumor ROI was defined using the manual segmentations. The grading CNNs were implemented with PyTorch and experiments were conducted using a NVIDIA GeForce Titan Black GPU.

For evaluation, we computed precision, recall, and f1-score. Since these metrics are influenced by class imbalance, we provide them for both LGG and HGG. Additionally, we compute the accuracy (acc) and the area under the receiver operating characteristic curve (ROC-AUC), which provide insights on the general ability of the classifier to distinguish between the classes.

6.4 Results and Discussion

Table 6.1 shows quantitative results for tumor grade prediction from each of the ROI (whole brain, and tumor). We note that it is expected to achieve lower f1-score, precision, and recall for LGG, since it is the minority class. Before feeding the images to the CNNs, we standardize the image intensities with zero mean and unit variance. Common approaches in the computer vision domain compute these statistics from the whole image. However, in MRI images, the background region is usually filled with 0 intensity values after skull stripping. When we standardize the intensities in the whole image, we achieve acc. of 0.895 (whole brain) and 0.877 (tumor ROI). However, from the Guided Backpropagation maps (Fig. 6.2), we observe that the CNN considers the border of brain as discriminative, which for our data should not be a predictor of tumor grade. This is probably due to high gradients, since background has negative values, after standardization. Hence, we changed our pre-processing strategy by standardizing the image intensities inside the brain mask, only. After this approach, we observed that, mostly, the CNN does not consider the brain border as relevant for tumor grading. More interestingly, this simple change considerably boosted the metrics of tumor grade prediction from the tumor ROI (Table 6.1). For instance, acc. and ROC-AUC improved from 0.877 and 0.8841 to 0.9298 and 0.9841, respectively. This shows an advantage of the interpretability stage, since it allowed us to identify a systematic problem and correct it; we note that the border problem would otherwise have gone unnoticed, as results were already competitive.

Focusing on the variant with the standardization in the brain mask, we observe in Table 6.1 that grade prediction from the tumor ROI (acc – 0.9298, ROC-AUC – 0.9841) achieves better scores than grade prediction from the whole image (acc. – 0.895, ROC-AUC – 0.8913). Despite this, we note that tumor grade prediction from the whole brain achieves an acc. of 0.895, f1-score of 0.9286, precision of 0.9286, and recall of 0.9286 for HGG. Fig. 6.3 shows interpretability maps for some examples. We note that GradCAM provides maps with the same resolution as the feature maps of the layer of interest. We compute GradCAM maps with the output of the third (Res3) and fourth (Res4) residual blocks (Fig. 6.1). Fig. 6.3(a) shows interpretability maps for grade predictions from the whole tumor. In the first row, the CNN was able to correctly grade it as HGG. From the two GradCAM maps we observe that the

Table 6.1: Tumor grade results for LGG and HGG in the two ROI: whole brain, and tumor. We show results for each variant of the image intensities standardization procedure.

Region	Standardization	Grade	F1-score	Precision	Recall	Acc	ROC-AUC
Whole brain	Whole image	LGG	0.8000	0.8000	0.8000	0.8950	0.8857
		HGG	0.929	0.929	0.929		
	Brain mask	LGG	0.8000	0.8000	0.8000	0.8950	0.8913
		HGG	0.9286	0.9286	0.9286		
Tumor ROI	Whole image	LGG	0.7879	0.7222	0.8667	0.8770	0.8841
		HGG	0.9136	0.9487	0.881		
	Brain mask	LGG	0.8667	0.8667	0.8667	0.9298	0.9841
		HGG	0.9524	0.9524	0.9524		

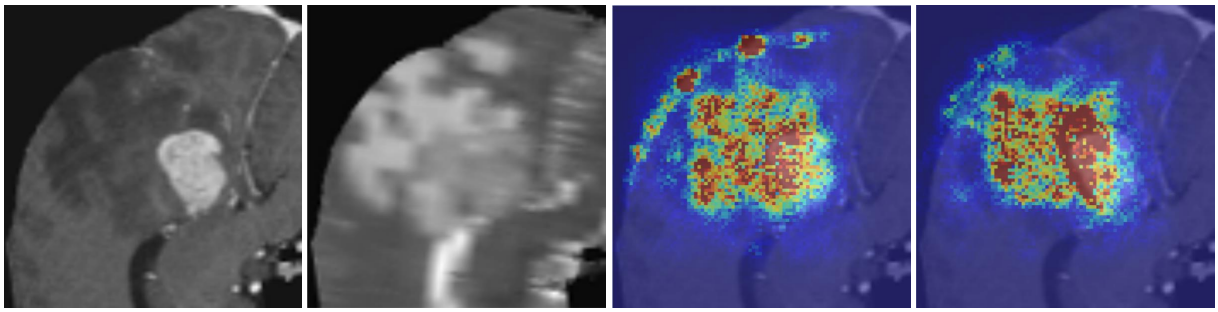
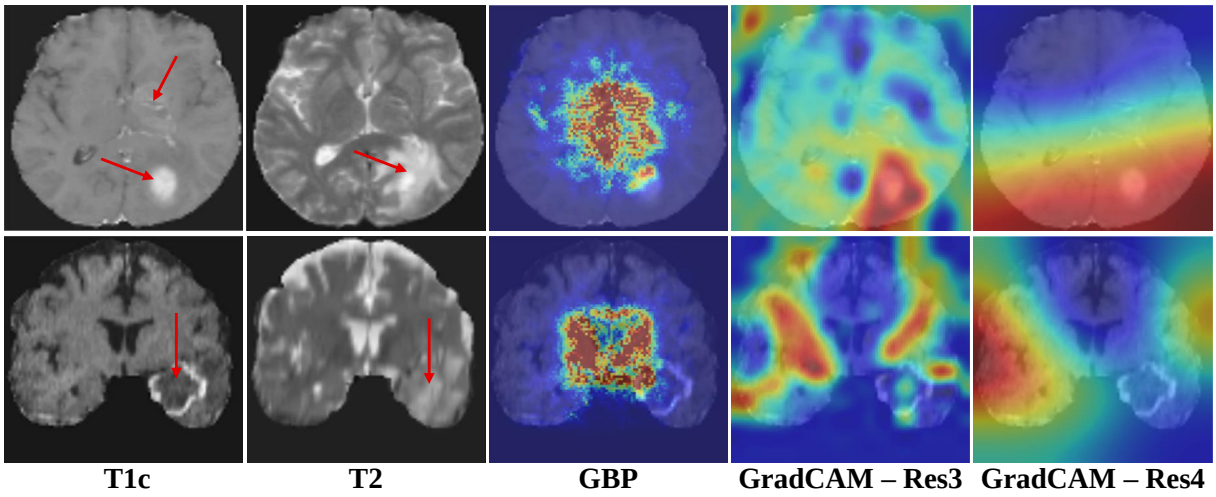
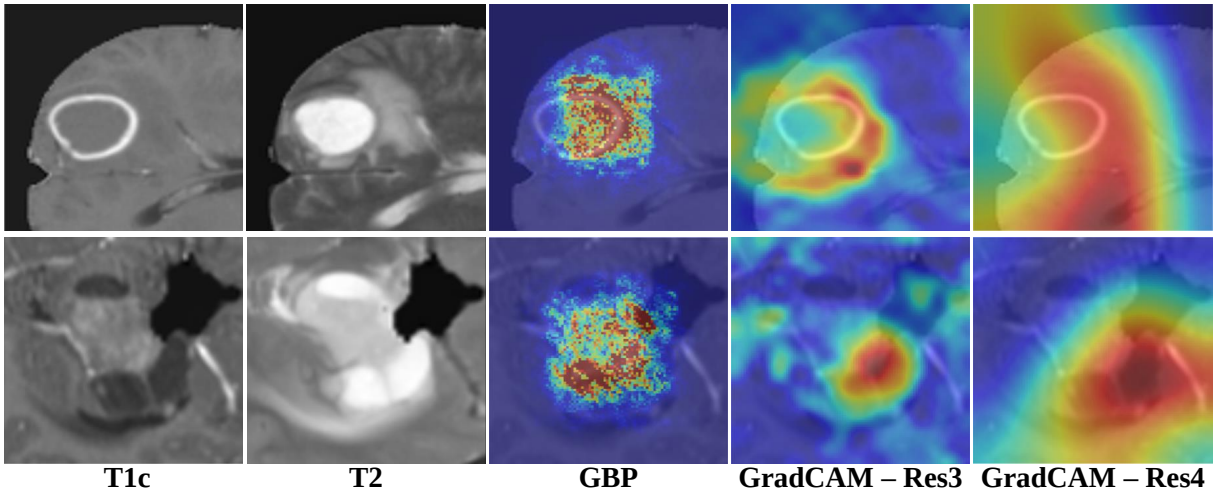


Figure 6.2: Example of the effect of intensity standardization on the Guided Backpropagation maps. Warmer colors represent stronger responses. From left to right: T1c, T2, Guided Backpropagation map on image standardized over the whole image, and Guided Backpropagation map on image standardized in the brain region only.

region of tumor was considered the most discriminative. The Guided Backpropagation shows focus on the ventricles, but, more interestingly, on both tumor lesions. In the second row, a HGG was mistakenly classified as LGG. The GradCAM maps are dispersed across the brain, instead of focusing in the tumor. We note that GradCAM is class discriminative, so, we show maps for LGG class. The Guided Backpropagation map concentrates in the ventricles. We observe that the CNN for tumor grading from the whole image focus on the ventricles frequently. We know that mass effect is a feature of HGG, and the ventricles are largely affected by it (Steed et al., 2018). Hence, the CNN may have learned that it is a predictor of malignancy. Actually, the subventricular zone is thought to be the origin of glioma cells, and nearby brain tumors are associated with worse prognosis (Liu et al., 2016). The focus on ventricles may explain why the example in the second row is misclassified as LGG, since its effect on ventricles is smaller than the first row example. Fig. 6.3(b) shows examples of tumor prediction from the tumor ROI. In the first row, a HGG is correctly classified. From the GradCAM maps, we observe that the CNN correctly locates the tumor. Additionally, the Res3 and Guided Backpropagation maps appear to focus on the transition from necrosis to enhancing tumor and edema. This is in accordance with domain knowledge, as such an enhancing rim is characteristic for HGG. The second row of Fig. 6.3(b) is a LGG misclassified as HGG. In this case, it is a LGG with enhancing tumor. For this reason, the GradCAM maps for HGG and the Guided Backpropagation map seem to indicate that the enhancing tissues were responsible for the prediction, as it is a feature of



(a)



(b)

Figure 6.3: Interpretability maps for grade predictions from a) the whole brain, and b) the tumor ROI. Warmer colors represent stronger responses. In a) the arrows indicate the tumor lesions; example on top is correctly classified as HGG, while example in the bottom is a HGG misclassified as LGG. In b), the top example is a correctly classified HGG, while in the bottom a LGG is misclassified as HGG.

HGG. It is possible that this is an evolving LGG that requires monitoring.

From the previous discussion, we see that GradCAM and Guided Backpropagation maps provide insights into the factors that contribute for a classification. So, we can see this interpretability stage as a quality assurance that enables us to check if the generated explanations are according to clinical knowledge. For instance, in the first row of Fig. 6.3(a) the explanations are focused on the tumor region. However, in the second row, the interpretability maps have high responses in regions that do not contain tumor. Thus, it may be a sign of an unreliable prediction, since it was based on regions of the image that are probably irrelevant. Additionally, the border effect problem, detected from the Guided Backpropagation maps, was a spurious pattern learned by the CNN.

6.5 Conclusion

Tumor grading from imaging data offers a fast and non-invasive approach for anticipating tumor grading, compared with histopathological diagnosis of biopsy specimens. We propose CNN for automatic brain tumor grading from MRI images, without the need of expert ROI definition. When we predict the grade from the whole brain, we achieve acc. of 0.895, while the prediction from the tumor ROI reaches an acc. of 0.9298. Therefore, our results show that grading is possible from both ROIs, although the latter achieves substantially better scores. Additionally, we employed interpretability approaches for prediction assessment, which allowed us to improve the pre-processing stage. Moreover, it may help in assessing if a decision is trustworthy by observing if it was actually based on the tumor region, or regions that are coherent with clinical knowledge.

6.6 Summary

Glioblastoma Multiforme is a high grade, very aggressive, brain tumor, with patients having a poor prognosis. Lower grade gliomas are less aggressive, but they can evolve into higher grade tumors over time. Patient management and treatment can vary considerably with tumor grade, ranging from tumor resection followed by a combined radio- and chemotherapy to a “wait and see” approach. Hence, tumor grading is important for adequate treatment planning and monitoring.

The gold standard for tumor grading relies on histopathological diagnosis of biopsy specimens. However, this procedure is invasive, time consuming, and prone to sampling error. Given these disadvantages, automatic tumor grading from widely used MRI protocols would be clinically important, as a way to expedite treatment planning and assessment of tumor evolution. In this chapter, we propose to use Convolutional Neural Networks for predicting tumor grade directly from imaging data. In this way, we overcome the need for expert annotations of regions of interest. We evaluate two prediction approaches: from the whole brain, and from an automatically defined tumor region. Finally, we employ interpretability methodologies as a quality assurance stage to check if the method is using image regions indicative of tumor grade for classification.

Chapter 7

Conclusions

The aim of this thesis was the automatic segmentation and classification of brain tumors on multi-sequence MRI images with Deep and Representation Learning-based methods. Our contributions were detailed in the previous three chapters, following three main topics: segmentation, glioma grading, and interpretability of opaque Machine Learning methods. In each topic we have presented the developed methods and the obtained results, as well as discussion wrapping up our observations and conclusions. Therefore, in this last chapter, we sum up the main contributions and conclusions. Finally, we close this document with our perspectives on opened lines of research that we consider deserving further investigation.

Contents

7.1 General conclusions

Brain tumors are not the most frequent cancers. Yet, they have high mortality rates, especially gliomas, which are the most common ones. The most used imaging technology for assessing these tumors in clinical practice is MRI. Multi-sequence MRI provides rich information, such as volumetric data and sequences with different contrasts. However, this richness comes at the expense of requiring more time for its analysis. Indeed, physicians often rely on simplified measurements in clinical practice, which may not exploit the full potential of MRI, while at the same time being prone to intra- and inter-rated variability. Therefore, we focused the work in this thesis on glioma MRI image analysis, with the ultimate objective of enabling the development of automated computerized solutions.

Gliomas are very heterogeneous in appearance, size, shape, and location. Moreover, MRI acquisitions vary a lot with the acquisition site and the equipment. This makes it technically very challenging to develop image analysis tools. In this thesis, we focused our research on learning from data using Machine Learning, especially on Deep and Representation Learning techniques.

Representation Learning has the capability of learning features directly from the data. It was a transversal objective to all the developed methods to understand its capabilities. Indeed, we observed that it can learn powerful and discriminative features. Actually, we have achieved state-of-the-art results in brain tumor segmentation using a CNN-based approach. This allowed us to surpass hand-crafted based and probabilistic modeling approaches, which require domain expert knowledge. Also, using RBMs, we managed to achieve competitive results, even though being an unsupervised shallow model. Therefore, we confirmed the capabilities of Representation Learning to learn complex and powerful features. Additionally, it tells us quite a lot about the challenges of brain tumor image analysis and the difficulties of its modeling by human reasoning.

In the following subsections we overview our main conclusions and contributions regarding each of the topics of this thesis.

7.1.1 Image segmentation

Arguably, our segmentation approach was the topic that mutated the most during this thesis work. This is a reflection of the fast pace in Deep Learning research observed in the last years.

We proposed classification CNNs for glioma segmentation in Section 4.1 (Pereira et al., 2015, 2016a). In this first approach, only the central voxel of a patch was classified. Inspired by Simonyan and Zisserman (2014), we investigated the use of small 3×3 kernels. By stacking several layers with small kernels, we can achieve the same field of view as larger kernels, but with the advantage of having a deeper network with less parameters. Indeed, this proved beneficial, and we may observe that the use of small kernels is now the preferred choice in CNN models (Litjens et al., 2017; He et al., 2016; Hu et al., 2018; Xie et al., 2017; Pereira et al., 2018b; Kamnitsas et al., 2016). Using this approach, we verified how sampling is important for achieving a balanced segmentation system. In segmentation, there are often some classes that are much less represented, raising a problem of class imbalance. Hence, care must be taken during sampling. The issue with class imbalance was especially observed when we needed to sample from HGGs to overcome the lack of enhancing samples in LGGs.

In the literature using CNNs for brain tumor segmentation, we commonly find simple pre-processing stages (Kamnitsas et al., 2017b; Havaei et al., 2017), e.g., only standardization of the image intensities with zero mean and unit variance. The reason is the belief that CNNs, and Deep Learning techniques in general, can deal well with the raw data. However, we observed that careful pre-processing with histogram normalization is very beneficial for CNNs. This is due to the variability of multi-site multi-scanner acquisitions of MRI images. Indeed, we observed smoother training and better results. Note that we kept this pre-processing during the development of all segmentation methods.

A classification CNN can be quite effective for segmentation, but it has some issues, such as being heavy on learnable parameters. This not only makes it prone to overfitting, but it makes the classification

CNN computationally demanding for voxel-wise segmentation. Therefore, in Section 4.2 (Pereira et al., 2017), we explore FCNs with an encoder-decoder topology. In this approach, fully-connected layers are discarded and substituted by convolutional layers. In his way, we could significantly reduce the number of learnable parameters. Moreover, it can process a patch of voxels in just one forward pass, which pushes computational efficiency even further.

While in a classification CNN we can more or less accurately enforce a training data sampling scheme, in FCNs approaches this is not straightforward. The reason is that in the former, each patch is mapped to a single class, while in the latter a patch is mapped to a patch of labels. Observing that the number of voxels of normal tissue largely exceeds the number of tumor voxels, we devised a hierarchical FCN-based scheme. Therefore, we first roughly segment the whole tumor, then we define a cuboid ROI, and finally we segment all tissues (normal + tumor tissues) inside the ROI. In this way, by training the multi-class FCN inside a ROI, the class imbalance problem is mitigated. To our knowledge, this was the first hierarchical brain tumor segmentation approach using FCNs. Similar and related approaches have since then been successfully used, such as the method proposed by Wang et al. (2018).

An important topic in Deep and Representation Learning is related with improving the discriminative power of features. In Section 4.3 (Pereira et al., 2018b) we explored both feature recombination and recalibration in FCNs for semantic segmentation. In feature recombination we linearly recombine the features. This alone demonstrably improved the results. Actually, the combination of features results into more complex ones. Not only recombination is beneficial, but also recalibration was shown to be. In semantic segmentation using FCNs, a unit of a feature map is directly related with the prediction of the corresponding voxel class. Therefore, we have proposed and shown how to adaptively and spatially recalibrate features maps. Consequently, in some locations of the feature maps, the features that are more irrelevant for the predicted class are suppressed. Therefore, we may also conclude that improving the performance of a CNN can be done through design strategies, and not only by increasing depth.

7.1.2 Glioma grading

Gliomas are very heterogeneous in appearance and characteristics, in such a way that some features may be observed in both HGGs and LGGs. Of course, some of the features are more associated with one of the grades. This aspect makes automatic glioma grading a difficult task.

We tackled automatic glioma grading from conventional MRI using a 3D CNN in Chapter 6 (Pereira et al., 2018a). The whole volume 3D assessment is important for two main reasons. First, a tumor mass must be classified as just one grade. And, second, because a slice-wise approach would require the implementation of some slice selection criteria such as manual selection by a radiologist. This not only may introduce intra- and inter-rater variability, but may also result in different grade classifications depending on the chosen slice.

An advantage of using CNNs for classification is that they can efficiently process large amounts of data, such as a whole volume. In the case of grading, the CNN was able to learn the task with surprisingly good results while being fed with the whole brain. However, we also employ grading with the binary segmentation stage used in Section 4.3. In this way, we define a ROI around the tumor, which we feed

into the CNN. In this case, we verify that it achieves better results. Yet, it is more sensitive to pre-processing. In fact, we verify again the importance of pre-processing. Finally, we concluded that these algorithms offer the opportunity for accurate tumor grading from MRI data acquired in clinical routine. Furthermore, they leverage automatic glioma grading without the need for an accurate segmentation.

7.1.3 Interpretability of Machine Learning

Deep and Representation Learning-based systems are often regarded as "black boxes" due to their lack of explainability. This is especially true for large and complex models. However, starting from a RBM + RF classifier system, in Chapter 5 (Pereira et al., 2018c), we showed that it is possible to extract intelligible information about how the model encoded learning. We may define two levels of interpretability: global and local. In the former, we aim at explaining what the model learned. Specifically, how input data is encoded by relevant features. In local interpretability, we aim at explaining which data was more relevant for a prediction. In this regard, we use a surrogate and much simpler model that learns the predictions of the model under analysis in the neighborhood of the sample being predicted. This simpler model is more interpretable. It represents the decision function of the main model only locally. Nevertheless, it provides satisfactory insights into the reasons behind predictions. Using Mutual Information, we were able to devise a feature selection strategy. This was motivated by the fact that very large feature vectors are less interpretable.

In the tasks of brain tumor segmentation and penumbra estimation we were able to verify that the system learned and made predictions in a way that was coherent with human expert knowledge. Moreover, the proposed interpretability methodologies were able to show intriguing findings. We observed that some misclassifications in brain tumor were due to defective skull stripping, especially in the T1c and FLAIR sequences. Also, interpretability provided hints that in penumbra estimation there may exist a bias in the manual segmentations towards the Tmax sequence. Therefore, we concluded that interpretability helps in detecting causes of failure, as well as problems in the data.

In the case of CNNs one may explore the gradients for explainability purposes. In glioma grading, it was possible to observe the impact of the brain ventricles, as it shows the mass effect of the tumor, typically found in HGGs. Enhancing tumor was also correctly found to be a predictor of the glioma grade. Furthermore, we observed again that by using interpretability, it was possible to detect causes of failures, and improve the development of the methods. Indeed, we managed to correct the pre-processing used for glioma grading, which significantly boosted the accuracy.

Summarizing, interpretability is crucial for increasing trust in machine learning-based systems, but may also positively contribute for more efficient development cycles.

7.2 Perspectives on Opened Research Lines

During the course of this thesis work we identified interesting lines of research for future work, which will be presented in the following paragraphs.

The return of domain knowledge One often cited advantage of Deep Learning models over more conventional Machine Learning models is their capability of handling raw data (LeCun et al., 2015). This is actually true, as state-of-the-art results are being achieved with Deep Learning-based approaches in several applications, from pure object recognition (He et al., 2016; Hu et al., 2018), to biomedical and medical image segmentation (Pereira et al., 2018b, 2016a; Kamnitsas et al., 2018; Ronneberger et al., 2015). Nevertheless, the fast progress in performance observed in recent years seems to be slowing down. Taking brain tumor segmentation into consideration, our proposal (Pereira et al., 2016a) considerably improved over the winner of BRATS 2013 (Tustison et al., 2015). Although we were able to achieve a new state of the art in the same dataset in our recent work (Pereira et al., 2018b), when we compare with other top performing methods (Zhao et al., 2018; Shen et al., 2017; Pereira et al., 2016a) we observe a smaller magnitude on the improvement than before. Similar observations can be found for BRATS 2015, as in (Qin et al., 2018). The winner of BRATS 2017 (Kamnitsas et al., 2018) achieved that rank by employing a large ensemble of CNNs with different architectures, and trained with different settings (optimizer, hyper-parameters, or loss functions), which shows the difficulty in coming up with an architecture that outperforms in every object and metric. However, experts in the medical field, such as radiologists, have strong domain knowledge. In fact, in the "pre-Deep Learning era", segmentation methods were dominated by approaches including prior knowledge, either through probabilistic priors (Menze et al., 2010), or domain-specific handcrafted features, such as symmetry (Tustison et al., 2015). Recently, we verified that a CNN may benefit from the addition of channels from the wavelet transform (Oliveira et al., 2018), which are commonly used as handcrafted features. Therefore, we envision that in the near future the most successful models will join both learned features and domain knowledge.

The inclusion of more sequences In this work we focused the developed methods in brain tumor image analysis using conventional structural MRI. These sequences are routinely acquired in clinical practice, being part of the consensus for standardized brain tumor imaging protocol (Ellingson et al., 2015). Therefore, the developed methods present higher potential of having practical value. Nevertheless, some other MRI sequences may provide extra information. For instance, perfusion MRI was found to encode information that may be useful for brain tumor classification and grading (Zacharaki et al., 2009; Essig et al., 2013). Perfusion and diffusion acquisitions may be indicative of the physiology and microsctructure of brain tumors. Although not conclusive, they may potentially be helpful in assessing the extension of the tumor beyond what is visible in T2 and FLAIR sequences (Milchenko et al., 2014; Guo et al., 2016; Svolos et al., 2014; Essig et al., 2013). It is still to be investigated if these sequences can boost the performance of Deep Learning-based brain tumor segmentation and classification methods. At the same time, research is needed regarding how to exploit it. For instance, perfusion acquisitions are 4D because of its time component. Recently, Pinto et al. (2018a) encoded the time slices as channels in a CNN after defining a window of interest for lesion outcome prediction in stroke. Hence, one still needs to verify if a similar approach would be feasible in brain tumors. Another approach would be to encode that data using sequence modeling.

The search for reliability Despite the rapid improvement observed in brain tumor segmentation performances, an issue is still to be solved – reliability. We refer to reliability as the ability of the system to maintain its performance regardless of the difficulty of the tumor being segmented or graded. For example, in the boxplots of Figures 4.3, 4.4, and 4.11 we can observe large variabilities in metrics. In some subjects we achieve excellent performance, while in others the performance is poor. This is generally observed in brain tumor segmentation, such as in Figure 10 of (Havaei et al., 2017), or Table 2 of (Wang et al., 2018). Note, however, that it may be partially caused by existing LGGs not having some of the tumor tissues. Furthermore, the large heterogeneity in the appearance, shape, and size of gliomas may also influence. More data may be necessary to deal with this issue. Also, this means that brain tumor segmentation and classification are not closed topics, and we predict that future winners of the BRATS challenge will be the ones with better reliability, instead of methods that may be almost perfect in some subjects, while significantly degrading in others. It may be more desirable to achieve a slightly lower average performance, as long as the variance across subjects is much smaller.

Something that may especially hinder Machine Learning-based methods is the differences between the conditions during training and after deployment. This poses an obvious reliability challenge. In the case of MRI a source of variability may be the differences observed in acquisitions from different scanners and sites. In our proposals, we mitigated this issue by pre-processing the images with a histogram matching approach. Therefore, more effective pre-processing may be researched. Recently, Karani et al. (2018) shared the parameters of a CNN across several scanners and protocols, while shortly retraining the Batch Normalization blocks in new settings. A downside is the need of annotated training samples. Kamnitsas et al. (2017a) proposed an unsupervised domain adaptation approach using adversarial training. Its unsupervised nature obviates the need for manual segmentations. However, the network needs a training stage for adaptation. The advantage of our pre-processing is that it maps the test image into a learned histogram, hence no re-training is required. However, it obviously depends on the learning stage of the histogram. Therefore, we point out that research into dealing with data from different sources is crucial for improving reliability, not only to improve performance, but also to make models better suited for deployment.

Another issue is the distribution of the labels in the training set. Class imbalance is a serious problem that may impact the performance of the deployed solutions. In (Pereira et al., 2016a) we carefully sampled patches to balance the classes, since we only classified central voxels. This approach is not feasible in FCNs, therefore, in (Pereira et al., 2017) we devised a hierarchical segmentation scheme. Other approaches were proposed, such as a two-stage sampling scheme during training (Havaei et al., 2017), or enforcing half of the samples in a batch to be selected from the foreground (Kamnitsas et al., 2017b). More elaborated approaches try to balance the classes through the loss function by weighting classes differently (Sudre et al., 2017), or with the Dice loss function (Milletari et al., 2016). However, the former introduces new hyper-parameters that are highly dependent on the training labels distribution, which may not faithfully represent the distribution that is observed in real clinical practice. The Dice loss is effective, but in prior experiments we observed that it leads the probabilistic predictions to concentrate either in 0 or 1, which makes it useless to regard the probabilities as confidence measures. The same was observed by Kamnitsas et al. (2018). In tumor grading, we balance training by sampling LGGs twice, but better

approaches are sure to exist. Therefore, we believe that class imbalance is still an open problem, both in segmentation and classification tasks.

The quality and safety assurance Partially related with reliability is the quality and safety assurance of the methods. We observed that sometimes Machine Learning-based methods may produce huge mistakes. In this case, is it safe to provide them, especially in critical domains, such as the medical field? Should a physician extract measurements from a poor quality segmentation? We believe that research is needed regarding quality assurance. Interpretability is certainly important for explaining predictions, therefore providing the tools for checking if a prediction is trustworthy is necessary. We made efforts in this direction in segmentation (Pereira et al., 2018c) and grade classification (Pereira et al., 2018a). Uncertainty is another strong line of research. For instance, people can visualize the regions of the segmentations with low certainty, and decide to trust it or not, or even target those regions for manual correction (Jungo et al., 2018). In another direction, Valindria et al. (2017) try to estimate the performance of a segmentation algorithm in a test subject without available manual segmentation through Reverse Classification Accuracy. A setback of this approach is its assumption that a similar subject will be available in a reference database with manual segmentations.

In summary, we identify three lines of research for quality and safety assurance: interpretability, uncertainty, and performance estimation.

The enhancing of architectures through adaptive modules For several years, improvements in performance were based on increasing the depth of Deep Learning architectures (Simonyan and Zisserman, 2014; He et al., 2016). However, really deep and big models, like ResNet-1001 (He et al., 2016) or GoogLeNet (Szegedy et al., 2015), may not be feasible in the case of medical imaging because of scarcer labeled data. Notwithstanding, there are other architecture options in neural networks besides depth that deserve research. Examples are the cardinality explored in ResNeXt (Xie et al., 2017), dilated convolution (Yu and Koltun, 2016), and feature maps recalibration (Hu et al., 2018).

In the case of brain tumor segmentation, although we have witnessed the proposal of deeper architectures over time (Pereira et al., 2016a; Kamnitsas et al., 2017b; Wang et al., 2018), there is no evidence that going to extreme depths is beneficial. We also observed this in our experiments. Hence, recently, other architectural designs are being tested, without the explicit purpose of increasing the depth *per se*, but to extract more discriminative features. In Pereira et al. (2018b), we explored feature recombination and recalibration for semantic segmentation. On another direction, Qin et al. (2018) investigated the automatic choice of scale. Both approaches extract features in an adaptive manner, i.e., in such a way that they are more useful for the sample under inference. Therefore, adaptive feature learning is a potential research direction.

The increasing importance of interpretability As discussed along this document, interpretability of Machine Learning-based methods is very important, especially in critical domains.

We proposed global and local interpretability approaches in (Pereira et al., 2018c), and we showed them in image segmentation context. Further research is needed to explore if and how similar approaches

can be used in other models, such as Deep Belief Networks. In (Pereira et al., 2018a) we used interpretability for assessing the regions of the image that are important for a grade prediction. This allowed us to detect and correct a pre-processing issue. Therefore, something that deserves further research is the inclusion of such methods in the development cycle, and study how it helps in guiding the development. GradCAM was one of the used methods. It provides an interpretation on the regions of the image that contributed the most for the predictions. However, in MRI, there are several sequences that may contribute differently. So, methods are needed that explain predictions in terms of sequences. Moreover, all of the interpretability methods explored in this thesis deal with single predictions. In other words, an input is mapped to a single output. To our knowledge, interpretability approaches for the case of FCN for segmentation, where we have a set of pixels/voxels predicted at once, is still to be researched.

A major challenge when developing and using interpretability methods is their evaluation. How do we know that an explanation is actually correct? How can we compare the explanations provided by different approaches? One possibility would be to compare the explanations with the ones given by human experts (Doshi-Velez and Kim, 2017; Pereira et al., 2018c). Doshi-Velez and Kim (2017) propose a taxonomy for the evaluation of interpretability, but more work needs to be done regarding this issue.

Finally, the term interpretability may be ill-defined (Lipton, 2016). Different authors use it with different meaning. Hence, an important task in the domain of interpretability is to formalize it. Furthermore, since interpretability of Machine Learning-based methods is very important in critical domains, it is crucial to develop it further in order to enhance trust by their end users. We believe that more work on applying interpretability methods is needed to show without doubt their capabilities.

References

Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al., 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications 5, 4006.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., MÃžller, K.R., 2010. How to explain individual classification decisions. Journal of Machine Learning Research 11, 1803–1831.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data 4, 170117.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I.J., Bergeron, A., Bouchard, N., Bengio, Y., 2012. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on neural networks 5, 537–550.

Bauer, S., Fejes, T., Slotboom, J., Wiest, R., Nolte, L.P., Reyes, M., 2012. Segmentation of brain tumor images based on integrated hierarchical classification and regularization. Proceddings of MICCAI-BRATS , 10–13.

Bauer, S., Nolte, L.P., Reyes, M., 2011. Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 354–361.

Bauer, S., Wiest, R., Nolte, L.P., Reyes, M., 2013. A survey of mri-based medical image analysis for brain tumor studies. Physics in medicine and biology 58, R97.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 1798–1828.

Bengio, Y., et al., 2009. Learning deep architectures for ai. Foundations and trends® in Machine Learning 2, 1–127.

Bennasar, M., Hicks, Y., Setchi, R., 2015. Feature selection using joint mutual information maximisation. Expert Systems with Applications 42, 8520–8532.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y., 2010. Theano: a CPU and GPU math expression compiler, in: Proceedings of the Python for Scientific Computing Conference (SciPy).

Boureau, Y.L., Ponce, J., LeCun, Y., 2010. A theoretical analysis of feature pooling in visual recognition, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 111–118.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Breiman, L., Friedman, J., Stone, C., Olshen, R., 1984. Classification and Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series, Taylor & Francis.

Chollet, F., 2015. Keras. `https://github.com/fchollet/keras`.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G.B., LeCun, Y., 2015. The loss surfaces of multilayer networks, in: Artificial Intelligence and Statistics, pp. 192–204.

Chow, D., Qi, J., Guo, X., Miloushev, V., Iwamoto, F., Bruce, J., Lassman, A., Schwartz, L., Lignelli, A., Zhao, B., et al., 2014. Semiautomated volumetric measurement on postcontrast mr imaging for analysis of recurrent and residual disease in glioblastoma multiforme. American Journal of Neuroradiology 35, 498–503.

Ciresan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2012. Deep neural networks segment neuronal membranes in electron microscopy images, in: Advances in Neural Information Processing Systems (NIPS), pp. 2843–2851.

Copen, W.A., Schaefer, P.W., Wu, O., 2011. Mr perfusion imaging in acute ischemic stroke. Neuroimaging Clinics of North America 21, 259–283.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.

Cortez, P., Embrechts, M.J., 2011. Opening black box data mining models using sensitivity analysis, in: Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, IEEE. pp. 341–348.

Craven, M.W., Shavlik, J.W., 1996. Extracting tree-structured representations of trained networks. Advances in Neural Information Processig Systems (NIPS) , 24–30.

Criminisi, A., Shotton, J., 2013. Decision forests for computer vision and medical image analysis. Springer Science & Business Media.

Davy, A., Havaei, M., Warde-Farley, D., Biard, A., Tran, L., Jodoin, P.M., Courville, A., Larochelle, H., Pal, C., Bengio, Y., 2014. Brain tumor segmentation with deep neural networks. MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS) , 31–35.

DeAngelis, L.M., 2001. Brain tumors. New England Journal of Medicine 344, 114–123.

Dempsey, M.F., Condon, B.R., Hadley, D.M., 2005. Measurement of tumor "size" in recurrent malignant glioma: 1d, 2d, or 3d? American Journal of Neuroradiology 26, 770–776.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S.K., Nouri, D., Maturana, D., Thoma, M., Battenberg, E., Kelly, J., Fauw, J.D., Heilman, M., diogo149, McFee, B., Weideman, H., takacsg84, peterderivaz, Jon, instagibbs, Rasul, D.K., CongLiu, Britefury, onas Degrave, 2015a. Lasagne: First release. URL: http://dx.doi.org/10.5281/zenodo.27878, doi:10.5281/zenodo.27878.

Dieleman, S., Willett, K.W., Dambre, J., 2015b. Rotation-invariant convolutional neural networks for galaxy morphology prediction. Monthly Notices of the Royal Astronomical Society 450, 1441–1459.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning .

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., Wood, A., 2017. Accountability of ai under the law: The role of explanation. arXiv preprint arXiv:1711.01134 .

Drevelegas, A., 2005. Extra-axial brain tumors. European Radiology 15, 453–467.

Dvorák, P., Menze, B., 2015. Structured prediction with convolutional neural networks for multimodal brain tumor segmentation. MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS) , 13–24.

Dziurzynski, K., Blas-Boria, D., Suki, D., Cahill, D.P., Prabhu, S.S., Puduvalli, V., Levine, N., 2012. Butterfly glioblastomas: a retrospective review and qualitative assessment of outcomes. Journal of Neuro-oncology 109, 555–563.

Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al., 2009. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). European Journal of Cancer 45, 228–247.

Ellingson, B.M., Bendszus, M., Boxerman, J., Barboriak, D., Erickson, B.J., Smits, M., Nelson, S.J., Gerstner, E., Alexander, B., Goldmacher, G., et al., 2015. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. Neuro-oncology 17, 1188–1198.

Ellingson, B.M., Bendszus, M., Sorensen, A.G., Pope, W.B., 2014. Emerging techniques and technologies in brain tumor imaging. Neuro-oncology 16, vii12–vii23.

Essig, M., et al., 2013. Perfusion mri: the five most frequently asked technical questions. American Journal of Roentgenology 200, 24–34.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., Bray, F., 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. International Journal of Cancer 136, E359–E386.

Fox, S.I., 2006. Human Physiology 12th Editon. McGraw-Hill press, New York, USA.

Freitas, A.A., 2014. Comprehensible classification models: A position paper. SIGKDD Explor. Newsl. 15.

Gallego-Ortiz, C., Martel, A.L., 2016. Interpreting extracted rules from ensemble of trees: Application to computer-aided diagnosis of breast mri, in: ICML Workshop on Human Interpretability in Machine Learning (WHI). ArXiv:1606.08288.

Ganz, M., Greve, D.N., Fischl, B., Konukoglu, E., 2015. Relevant feature set estimation with a knock-out strategy and random forests. NeuroImage 122.

Geremia, E., Menze, B.H., Ayache, N., 2013. Spatially adaptive random forests, in: 2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1344–1347.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: International Conference on Artificial Intelligence and Statistics, pp. 249–256.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 315–323.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Goodman, B., Flaxman, S., 2016. Eu regulations on algorithmic decision-making and a "right to explanation", in: ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv. org/abs/1606.08813 v1.

Gooya, A., Pohl, K.M., Bilello, M., Cirillo, L., Biros, G., Melhem, E.R., Davatzikos, C., 2012. Glistr: glioma image segmentation and registration. IEEE Transactions on Medical Imaging 31, 1941–1954.

Grier, J.T., Batchelor, T., 2006. Low-grade gliomas in adults. The oncologist 11, 681–693.

Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., Giannotti, F., 2018. A survey of methods for explaining black box models. arXiv preprint arXiv:1802.01933 .

Guo, L., Wang, G., Feng, Y., Yu, T., Guo, Y., Bai, X., Ye, Z., 2016. Diffusion and perfusion weighted magnetic resonance imaging for tumor volume definition in radiotherapy of brain tumors. Radiation Oncology 11, 123.

Hara, S., Hayashi, K., 2016. Making tree ensembles interpretable, in: ICML Workshop on Human Interpretability in Machine Learning (WHI). ArXiv:1606.05390.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference and prediction. 2 ed., Springer.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Medical Image Analysis 35, 18–31.

Havaei, M., Guizard, N., Chapados, N., Bengio, Y., 2016. Hemis: Hetero-modal image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 469–477.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks, in: European Conference on Computer Vision (ECCV), pp. 630–645.

Hinton, G.E., 2002. Training products of experts by minimizing contrastive divergence. Neural computation 14, 1771–1800.

Hinton, G.E., 2012. Neural Networks: Tricks of the Trade: Second Edition. Springer Berlin Heidelberg. chapter A Practical Guide to Training Restricted Boltzmann Machines.

Hinton, G.E., McClelland, J.L., Rumelhart, D.E., 1986. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1, MIT Press, Cambridge, MA, USA. chapter Distributed Representations, pp. 77–109.

Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Computation 18, 1527–1554.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580v1 .

Hinton, G.E., Zemel, R.S., 1994. Autoencoders, minimum description length and helmholtz free energy, in: Advances in Neural Information Processing Systems (NIPS), pp. 3–10.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24, 417.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) .

Hu, L.S., Ning, S., Eschbacher, J.M., Gaw, N., Dueck, A.C., Smith, K.A., Nakaji, P., Plasencia, J., Ranjbar, S., Price, S.J., et al., 2015. Multi-parametric mri and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma. PloS One 10, e0141506.

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

Islam, A., Reza, S.M., Iftekharuddin, K.M., 2013. Multifractal texture estimation for detection and segmentation of brain tumors. IEEE Transactions on Biomedical Engineering 60, 3204–3215.

Islam, M., Ren, H., 2018. Multi-modal pixelnet for brain tumor segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pp. 298–308.

Jarrett, K., Kavukcuoglu, K., LeCun, Y., et al., 2009. What is the best multi-stage architecture for object recognition?, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE. pp. 2146–2153.

Jesson, A., Arbel, T., 2018. Brain tumor segmentation using a 3d fcn with multi-scale loss, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.

Jones, H.R., Hreib, K., 2012. Netter\'s neurology .

Jungo, A., McKinley, R., Meier, R., Knecht, U., Vera, L., Pérez-Beteta, J., Molina-García, D., Pérez-García, V.M., Wiest, R., Reyes, M., 2018. Towards uncertainty-assisted brain tumor segmentation and survival prediction, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al., 2018. Ensembles of multiple models and architectures for robust brain tumour segmentation, in: International MICCAI Brainlesion Workshop.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017a. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks, in: International Conference on Information Processing in Medical Imaging (IPMI), Springer. pp. 597–609.

Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A.V., Criminisi, A., Rueckert, D., Glocker, B., 2016. Deepmedic for brain tumor segmentation, in: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer. pp. 138–149.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017b. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical Image Analysis 36, 61–78.

Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E., 2018. A lifelong learning approach to brain mr segmentation across scanners and protocols, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Khawaldeh, S., Pervaiz, U., Rafiq, A., Alkhawaldeh, R.S., 2017. Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks. Applied Sciences 8, 27.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR) .

Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R., Büchler, P., 2013. The virtual skeleton database: An open access repository for biomedical research and collaboration. Journal of Medical Internet Research 15.

Kononenko, I., 2001. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine 23, 89–109.

Konukoglu, E., Ganz, M., 2014. Approximate false positive rate control in selection frequency for random forest. arXiv:1410.2838 .

Kotrotsou, A., Zinn, P.O., Colen, R.R., 2016. Radiomics in brain tumors: an emerging technique for characterization of tumor environment. Magnetic Resonance Imaging Clinics 24, 719–729.

Krause, J., Perer, A., Ng, K., 2016. Interacting with predictions: Visual inspection of black-box machine learning models, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM. pp. 5686–5697.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105.

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S.A., Schabath, M.B., Forster, K., Aerts, H.J., Dekker, A., Fenstermacher, D., et al., 2012. Radiomics: the process and the challenges. Magnetic Resonance Imaging 30, 1234–1248.

Kwon, D., Akbari, H., Da, X., Gaonkar, B., Davatzikos, C., 2014a. Multimodal brain tumor image segmentation using glistr, in: MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS), pp. 18–19.

Kwon, D., Shinohara, R.T., Akbari, H., Davatzikos, C., 2014b. Combining generative models for multifocal glioma segmentation and registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 763–770.

Lambin, P., Leijenaar, R.T., Deist, T.M., Peerlings, J., de Jong, E.E., van Timmeren, J., Sanduleanu, S., Larue, R.T., Even, A.J., Jochems, A., et al., 2017. Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology 14, 749.

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

Larjavaara, S., Haapasalo, H., Sankila, R., Helén, P., Auvinen, A., 2008. Is the incidence of meningiomas underestimated? a regional survey. British Journal of Cancer 99, 182.

Larjavaara, S., Mäntylä, R., Salminen, T., Haapasalo, H., Raitanen, J., Jääskeläinen, J., Auvinen, A., 2007. Incidence of gliomas by anatomic location. Neuro-oncology 9, 319–325.

Larochelle, H., Bengio, Y., 2008. Classification using discriminative restricted boltzmann machines, in: Proceedings of the 25th International Conference on Machine Learning (ICML), ACM. pp. 536–543.

Law, M., Yang, S., Wang, H., Babb, J.S., Johnson, G., Cha, S., Knopp, E.A., Zagzag, D., 2003. Glioma grading: sensitivity, specificity, and predictive values of perfusion mr imaging and proton mr spectroscopic imaging compared with conventional mr imaging. American Journal of Neuroradiology 24, 1989–1998.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Computation 1, 541–551.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Lee, C.H., Wang, S., Murtha, A., Brown, M.R., Greiner, R., 2008. Segmenting brain tumors using pseudo–conditional random fields, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 359–366.

Lin, M., Chen, Q., Yan, S., 2013. Network in network. International Conference on Learning Representations (ICLR) .

Lipton, Z.C., 2016. The mythos of model interpretability, in: ICML Workshop on Human Interpretability in Machine Learning (WHI). ArXiv:1606.03490.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88.

Liu, S., Wang, Y., Fan, X., Ma, J., Ma, W., Wang, R., Jiang, T., 2016. Anatomical involvement of the subventricular zone predicts poor survival outcome in low-grade astrocytomas. PloS One 11.

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

Louis, D.N., Perry, A., Reifenberger, G., Von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., Ellison, D.W., 2016. The 2016 world health organization classification of tumors of the central nervous system: a summary. Acta Neuropathologica 131, 803–820.

Louppe, G., Wehenkel, L., Sutera, A., Geurts, P., 2013. Understanding variable importances in forests of randomized trees, in: Advances in Neural Information Processing Systems (NIPS).

Lyksborg, M., Puonti, O., Agn, M., Larsen, R., 2015. An ensemble of 2d convolutional neural networks for tumor segmentation, in: Scandinavian Conference on Image Analysis, Springer. pp. 201–211.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proc. ICML.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-sne. Journal of Machine Learning Research 9, 2579–2605.

Mabray, M.C., Barajas, R.F., Cha, S., 2015. Modern brain tumor imaging. Brain Tumor Research and Treatment 3, 8–23.

Macdonald, D.R., Cascino, T.L., Schold Jr, S.C., Cairncross, J.G., et al., 1990. Response criteria for phase ii studies of supratentorial malignant glioma. Journal Clinical Oncology 8, 1277–1280.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. Medical Image Analysis 35, 250–269.

McKinley, R., Häni, L., Gralla, J., El-Koussy, M., Bauer, S., Arnold, M., Fischer, U., Jung, S., Mattmann, K., Reyes, M., et al., 2016. Fully automated stroke tissue estimation using random forest classifiers (faster). Journal of Cerebral Blood Flow & Metabolism .

Meier, R., Bauer, S., Slotboom, J., Wiest, R., Reyes, M., 2013. A hybrid model for multimodal brain tumor segmentation. Multimodal Brain Tumor Segmentation 31, 31–37.

Meier, R., Bauer, S., Slotboom, J., Wiest, R., Reyes, M., 2014a. Appearance-and context-sensitive features for brain tumor segmentation, in: MICCAI BraTS.

Meier, R., Bauer, S., Slotboom, J., Wiest, R., Reyes, M., 2014b. Appearance-and context-sensitive features for brain tumor segmentation. Proceedings of MICCAI BRATS Challenge , 020–026.

Meier, R., Bauer, S., Slotboom, J., Wiest, R., Reyes, M., 2014c. Patient-specific semi-supervised learning for postoperative brain tumor segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Springer International Publishing. pp. 714–721.

Meier, R., Knecht, U., Loosli, T., Bauer, S., Slotboom, J., Wiest, R., Reyes, M., 2016. Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. Scientific Reports 6, 23376.

Meier, R., Porz, N., Knecht, U., Loosli, T., Schucht, P., Beck, J., Slotboom, J., Wiest, R., Reyes, M., 2017. Automatic estimation of extent of resection and residual tumor volume of patients with glioblastoma. Journal of Neurosurgery , 1–9.

Meinshausen, N., Buhlmann, P., 2010. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 417–473.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2015. The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging 34, 1993–2024.

Menze, B.H., Leemput, K.V., Lashkari, D., Riklin-Raviv, T., Geremia, E., Alberts, E., Gruber, P., Wegener, S., Weber, M.A., Székely, G., Ayache, N., Golland, P., 2016. A generative probabilistic model and discriminative extensions for brain lesion segmentation with application to tumor and stroke. IEEE Transactions on Medical Imaging 35, 933–946.

Menze, B.H., Van Leemput, K., Lashkari, D., Weber, M.A., Ayache, N., Golland, P., 2010. A generative model for brain tumor segmentation in multi-modal images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 151–159.

Milchenko, M.V., Rajderkar, D., LaMontagne, P., Massoumzadeh, P., Bogdasarian, R., Schweitzer, G., Benzinger, T., Marcus, D., Shimony, J.S., Fouke, S.J., 2014. Comparison of perfusion-and diffusion-weighted imaging parameters in brain tumor studies processed using different software platforms. Academic Radiology 21, 1294–1303.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 3D Vision (3DV), 2016 Fourth International Conference on, pp. 565–571.

Mullins, M.E., Barest, G.D., Schaefer, P.W., Hochberg, F.H., Gonzalez, R.G., Lev, M.H., 2005. Radiation necrosis versus glioma recurrence: conventional MR imaging clues to diagnosis. Americal Journal of Neuroradiology 26, 1967–72.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814.

Nguyen, A., Yosinski, J., Clune, J., 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 427–436.

Nowozin, S., 2012. Improved information gain estimates for decision tree induction, in: International Conference on Machine Learning (ICML). ArXiv:1206.4620.

Nyúl, L., Udupa, J., 1999. On standardizing the mr image intensity scale. Magnetic Resonance in Medicine 42, 1072–1081.

Nyúl, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of mri scale standardization. IEEE Transactions on Medical Imaging 19, 143–150.

Ohgaki, H., Kleihues, P., 2013. The definition of primary and secondary glioblastoma. Clinical Cancer Research 19, 764–772.

Olah, C., 2018. Neural networks, manifolds, and topology. URL: `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology`.

Olden, J.D., Jackson, D.A., 2002. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling 154.

Oliveira, A., Pereira, S., Silva, C.A., 2018. Retinal vessel segmentation based on fully convolutional neural networks. Expert Systems with Applications .

Parmar, C., Velazquez, E.R., Leijenaar, R., Jermoumi, M., Carvalho, S., Mak, R.H., Mitra, S., Shankar, B.U., Kikinis, R., Haibe-Kains, B., et al., 2014. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PloS One 9, e102107.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analalysis and Machine Intelligence 27, 1226–1238.

Pereira, Sérgio Meier, R., Alves, V., Reyes, M., Silva, C.A., 2018a. Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment, in: Workshop on Interpretability of Machine Intelligence in Medical Image Computing, Lecture Notes in Computer Science.

Pereira, S., Alves, V., Silva, C.A., 2018b. Adaptive feature recombination and recalibration for semantic segmentation: application to brain tumor segmentation in mri, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Pereira, S., Meier, R., McKinley, R., Wiest, R., Alves, V., Silva, C.A., Reyes, M., 2018c. Enhancing interpretability of automatically extracted machine learning features: application to a rbm-random forest system on brain lesion segmentation. Medical Image Analysis 44, 228–244.

Pereira, S., Oliveira, A., Alves, V., Silva, C.A., 2017. On hierarchical brain tumor segmentation in mri using fully convolutional neural networks: A preliminary study, in: 5th Portuguese Meeting on Bioengineering, IEEE.

Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2015. Deep convolutional neural networks for the segmentation of gliomas in multi-sequence mri, in: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing. pp. 131–143.

Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2016a. Brain tumor segmentation using convolutional neural networks in mri images. IEEE Transactions on Medical Imaging 35, 1240–1251.

Pereira, S., Pinto, A., Oliveira, J., Mendrik, A.M., Correia, J.H., Silva, C.A., 2016b. Automatic brain tissue segmentation in mr images using random forests and conditional random fields. Journal of Neuroscience Methods 270, 111–123.

Pinto, A., Pereira, S., Correia, H., Oliveira, J., Rasteiro, D.M., Silva, C.A., 2015a. Brain tumour segmentation based on extremely randomized forest with high-level features, in: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 3037–3040.

Pinto, A., Pereira, S., Dinis, H., Silva, C.A., Rasteiro, D.M., 2015b. Random decision forests for automatic brain tumor segmentation on multi-modal mri images, in: 4th Portuguese Meeting on Bioengineering.

Pinto, A., Pereira, S., Meier, R., Alves, V., Wiest, R., Silva, C.A., Reyes, M., 2018a. Enhancing clinical mri perfusion maps with data-driven maps of complementary nature for lesion outcome prediction, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Pinto, A., Pereira, S., Rasteiro, D., Silva, C.A., 2018b. Hierarchical brain tumour segmentation using extremely randomized trees. Pattern Recognition .

Polyak, B.T., 1964. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics 4, 1–17.

Pope, W.B., Sayre, J., Perlina, A., Villablanca, J.P., Mischel, P.S., Cloughesy, T.F., 2005. Mr imaging correlates of survival in patients with high-grade gliomas. American Journal of Neuroradiology 26, 2466–2474.

Porz, N., Habegger, S., Meier, R., Verma, R., Jilch, A., Fichtner, J., Knecht, U., Radina, C., Schucht, P., Beck, J., et al., 2016. Fully automated enhanced tumor compartmentalization: man vs. machine reloaded. PloS One 11, e0165302.

Prastawa, M., Bullitt, E., Ho, S., Gerig, G., 2004. A brain tumor segmentation framework based on outlier detection. Medical Image Analysis 8, 275–283.

Qin, Y., Kamnitsas, K., Ancha, S., Nanavati, J., Cottrell, G., Criminisi, A., Nori, A., 2018. Autofocus layer for semantic segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI) .

Rao, V., Sharifi, M., Jaiswal, A., 2015. Brain tumor segmentation with deep learning. MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS) , 56–59.

Reza, S., Iftekharuddin, K., 2014. Multi-fractal texture features for brain tumor and edema segmentation, in: Medical Imaging 2014: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 903503.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016a. Model-agnostic interpretability of machine learning, in: ICML Workshop on Human Interpretability in Machine Learning (WHI). ArXiv:1606.05386.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016b. "why should i trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, pp. 1135–1144.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI), Springer. pp. 234–241.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533.

Russell, S., Dewey, D., Tegmark, M., 2015. Research priorities for robust and beneficial artificial intelligence. Ai Magazine 36, 105–114.

Salakhutdinov, R., Mnih, A., Hinton, G., 2007. Restricted boltzmann machines for collaborative filtering, in: International Conference on Machine Learning (ICML), ACM. pp. 791–798.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626.

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D.L., Collins, D.L., 2011. Evaluating intensity normalization on mris of human brain with multiple sclerosis. Medical Image Analysis 15, 267–282.

Shelhamer, E., Long, J., Darrell, T., 2016. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analalysis and Machine Intelligence .

Shen, H., Wang, R., Zhang, J., McKenna, S.J., 2017. Boundary-aware fully convolutional network for brain tumor segmentation, in: International Conference on Medical image computing and computer-assisted intervention (MICCAI), pp. 433–441.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034 .

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 .

Smolensky, P., 1986. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1, MIT Press, Cambridge, MA, USA. chapter Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 .

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

Steed, T.C., Treiber, J.M., Brandel, M.G., Patel, K.S., Dale, A.M., Carter, B.S., Chen, C.C., 2018. Quantification of glioblastoma mass effect by lateral ventricle displacement. Scientific Reports 8, 2827.

Straka, M., Albers, G.W., Bammer, R., 2010. Real-time diffusion-perfusion mismatch analysis in acute stroke. Journal of Magnetic Resonance Imaging 32, 1024–1037.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 240–248.

Suetens, P., 2017. Fundamentals of medical imaging. Cambridge university press.

Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning, in: International Conference on Machine Learning (ICML), pp. 1139–1147.

Svolos, P., Kousi, E., Kapsalaki, E., Theodorou, K., Fezoulidis, I., Kappas, C., Tsougos, I., 2014. The role of diffusion and perfusion weighted imaging in the differential diagnosis of cerebral tumors: a review and future perspectives. Cancer Imaging 14, 20.

Szegedy, C., Inc, G., Zaremba, W., Sutskever, I., Inc, G., Bruna, J., Erhan, D., Inc, G., Goodfellow, I., Fergus, R., 2014. Intriguing properties of neural networks, in: International Conference on Learning Representations (ICLR).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Tabatabai, G., Stupp, R., van den Bent, M.J., Hegi, M.E., Tonn, J.C., Wick, W., Weller, M., 2010. Molecular diagnostics of gliomas: the clinical perspective. Acta Neuropathologica 120, 585–592.

Therasse, P., Arbuck, S.G., Eisenhauer, E.A., Wanders, J., Kaplan, R.S., Rubinstein, L., Verweij, J., Van Glabbeke, M., van Oosterom, A.T., Christian, M.C., et al., 2000. New guidelines to evaluate the response to treatment in solid tumors. Journal of the National Cancer Institute 92, 205–216.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) , 267–288.

Tieleman, T., 2008. Training restricted boltzmann machines using approximations to the likelihood gradient, in: Proceedings of the 25th international conference on Machine learning, ACM. pp. 1064–1071.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C., 2015. Efficient object localization using convolutional networks, in: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 648–656.

van Tulder, G., de Bruijne, M., 2015. Why does synthesized data improve multi-sequence classification?, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 531–538.

van Tulder, G., de Bruijne, M., 2016. Combining generative and discriminative representation learning for lung ct analysis with convolutional restricted boltzmann machines. IEEE Transactions on Medical Imaging 35, 1262–1272.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4itk: improved n3 bias correction. IEEE Transactions on Medical Imaging 29.

Tustison, N.J., Shrinidhi, K., Wintermark, M., Durst, C.R., Kandel, B.M., Gee, J.C., Grossman, M.C., Avants, B.B., 2015. Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with antsr. Neuroinformatics 13, 209–225.

Tzeng, F.Y., Ma, K.L., 2005. Opening the black box - data driven visualization of neural networks, in: VIS 05. IEEE Visualization, 2005., pp. 383–390.

Urban, G., Bendszus, M., Hamprecht, F., Kleesiek, J., 2014. Multi-modal brain tumor segmentation using deep convolutional neural networks. MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS) , 1–5.

Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2017. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE Transactions on Medical Imaging 36, 1597–1606.

Van Meir, E.G., Hadjipanayis, C.G., Norden, A.D., Shu, H.K., Wen, P.Y., Olson, J.J., 2010. Exciting new advances in neuro-oncology: The avenue to a cure for malignant glioma. CA: a cancer journal for clinicians 60, 166–193.

Velazquez, E.R., Meier, R., Dunn Jr, W.D., Alexander, B., Wiest, R., Bauer, S., Gutman, D.A., Reyes, M., Aerts, H.J., 2015. Fully automatic gbm segmentation in the tcga-gbm dataset: Prognosis and correlation with vasari features. Scientific Reports 5, 16822.

Vergara, J.R., Estévez, P.A., 2014. A review of feature selection methods based on mutual information. Neural Computing and Applications 24, 175–186.

VirtualSkeleton, 2013. BRATS 2013. URL: https://www.virtualskeleton.ch/BRATS/Start2013[Accessed:September30,2015].

Vovk, U., Pernus, F., Likar, B., 2007. A review of methods for correction of intensity inhomogeneity in mri. IEEE Transactions on Medical Imaging 26, 405–421.

Wachter, S., Mittelstadt, B., Floridi, L., 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law 7, 76–99.

Wang, G., Li, W., Ourselin, S., Vercauteren, T., 2018. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries .

Wang, S., Summers, R.M., 2012. Machine learning and radiology. Medical Image Analysis 16, 933–951.

Wejchert, J., Tesauro, G., 1989. Neural network visualization., in: Advances in Neural Information Processig Systems (NIPS), pp. 465–472.

Wen, P.Y., Macdonald, D.R., Reardon, D.A., Cloughesy, T.F., Sorensen, A.G., Galanis, E., DeGroot, J., Wick, W., Gilbert, M.R., Lassman, A.B., et al., 2010. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. Journal of Clinical Oncology 28, 1963–1972.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks, in: Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995.

Yang, D., Rao, G., Martinez, J., Veeraraghavan, A., Rao, A., 2015. Evaluation of tumor-derived mri-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma. Medical Physics 42, 6725–6735.

Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions, in: International Conference on Learning Representations (ICLR).

Zacharaki, E.I., Kanas, V.G., Davatzikos, C., 2011. Investigating machine learning techniques for mri-based classification of brain neoplasms. International Journal of Computer Assisted Radiology and Surgery 6, 821–828.

Zacharaki, E.I., Wang, S., Chawla, S., Soo Yoo, D., Wolf, R., Melhem, E.R., Davatzikos, C., 2009. Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme. Magnetic Resonance in Medicine 62, 1609–1618.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: European Conference on Computer Vision (ECCV), Springer. pp. 818–833.

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., Fan, Y., 2018. A deep learning model integrating fcnns and crfs for brain tumor segmentation. Medical Image Analysis 43, 98–111.

Zhen, X., Wang, Z., Islam, A., Bhaduri, M., Chan, I., Li, S., 2016. Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation. Medical Image Analysis 30, 120–129.

Zhu, X., Wu, X., 2004. Class noise vs. attribute noise: A quantitative study. Artificial Intelligence Review 22, 177–210.

Zikic, D., Glocker, B., Konukoglu, E., Criminisi, A., Demiralp, C., Shotton, J., Thomas, O.M., Das, T., Jena, R., Price, S.J., 2012. Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel mr, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer. pp. 369–376.

Zikic, D., Ioannou, Y., Brown, M., Criminisi, A., 2014. Segmentation of brain tumor tissues with convolutional neural networks. MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS) , 36–39.

Zrihem, N.B., Zahavy, T., Mannor, S., 2016. Visualizing dynamics: from t-sne to semi-mdps, in: ICML Workshop on Human Interpretability in Machine Learning (WHI). ArXiv:1606.07112.