José Adriano Azevedo da Silva Ribeiro Pinto **Automatic Prediction of Ischemic Stroke from MRI images using Deep Learning**

UMinho | 2020

Universidade do Minho
Escola de Engenharia

José Adriano Azevedo da Silva Ribeiro Pinto

Automatic Prediction of Ischemic Stroke from MRI images using Deep Learning

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

Universidade do Minho

universidade de aveiro

U.PORTO

agosto de 2020

Universidade do Minho
Escola de Engenharia

José Adriano Azevedo da Silva Ribeiro Pinto

Automatic Prediction of Ischemic Stroke from MRI images using Deep Learning

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

agosto de 2020

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Acknowledgements

First of all, I would like to express my profound gratitude towards Professor Carlos A. Silva for guiding and conducting this work. During these years, I admired the constant eagerness in reaching new milestones, the dedication, the critical and methodical thinking and the knowledge sharing. Thank you for all the brainstorms, and for always challenge me and push me beyond what I thought I would be able to achieve. It was a privilege being supervised by Professor Carlos Silva. I also would like to thank Professor Victor Alves for all the support, availability and feedback during these years. I admired Professor Victor Alves' sense of practically with scientific rigour, and the capacity of abstraction. For all this, I am very grateful, and I wish them all the best.

I thank Professor Mauricio Reyes for the research collaboration and welcoming me in his Medical Imaging Analysis group at the Institute for Surgical Technologies & Biomechanics of the University of Bern. I thank the openness in integrating me in the group, and making me feel as I was part of the team. This was a memorable experience, which certainly allowed to grow professionally and personally.

A special thanks goes to Sérgio Pereira for his friendship, and with whom I had the privilege of brainstorming and debating fracturing research topics. I am confident that Sérgio will achieve his life goals, and I wish him the best. Also, I like to thank Joana Amorim, Patrícia Lopes and Alexandrine Ribeiro. It was nice working and collaborate with them during their Master's thesis. I wish them all the best.

The success of a PhD. can be dictated by several factors. Having good working environment can be considered one of those factors. Hence, I was privileged to be part of a lab of cool people, where I made good friends. Also, a special thanks goes to my dear and closest friends for their support and friendship. The time passed together was truly refreshing.

Finally, I need to refer two foundation stones. My parents and brother for their love, support, comprehension and belief in my capacities. My girlfriend Francisca for her love, cheering up, comprehension, support and confidence. Thank you for granting me the peace of mind needed to face these years, and making me realize to enjoy and seize the daily-life. This thesis is also theirs.

This work was supported by a scholarship awarded to José Adriano Azevedo da Silva Ribeiro Pinto by Fundaçao para a Ciência e Tecnologia (FCT), Portugal with scholarship reference PD/BD/113968/2015.

STATEMENT OF INTEGRITY

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

Resumo

Previsão automática de AVCs através de imagens de RMN usando Deep Learning

O Acidente Vascular Cerebral é uma das principais causas de morte, constituindo a segunda causa de morte nos países desenvolvidos. Representa também uma das principais causas de incapacidade funcional a nível mundial, tendo um grande impacto na sociedade. O Acidente Vascular Cerebral pode ser classificado em hemorrágico ou isquémico, sendo este último o subtipo mais frequente. O estudo imagiológico é fundamental na abordagem e planeamento do tratamento, onde a Tomografia Computorizada é o método de imagem mais comummente utilizado devido aos baixos custos de operação e acessibilidade. Contudo, quando disponível, a Ressonância Magnética é o método preferido, dada a sua capacidade na detecção de estadios precoces de isquemia cerebral. Desta forma, o estudo imagiológico permite não só a distinção do tipo de lesão e a sua localização, mas também uma melhor discriminação das áreas com enfarte das áreas de penumbra, onde existe a possibilidade de recuperação do tecido cerebral. Uma rápida ponderação dos riscos e benefícios associados à intervenção é necessária, que tem por base delineações grosseiras da lesão e a experiência clínica, havendo por isso, variabilidade intra- e inter-médico. Assim, ferramentas automáticas, permitem orientar e facilitar o processo de ponderação. Não obstante, o desenvolvimento destas ferramentas não é trivial, dada a variabilidade das lesões, dos fenómenos de perfusão e difusão cerebrais que ocorrem ao longo do tempo, bem como da variabilidade dos aparelhos de aquisição médica e a sua fraca resolução.

A Aprendizagem Automática compreende um vasto número de algoritmos, todos eles com o intuito de aprender padrões para realizar um dado objectivo ou tarefa. Uma categoria específica da Aprendizagem Automática é a Aprendizagem de Características, onde os algoritmos têm a capacidade de aprender e extrair automaticamente características através dos dados de entrada. Por sua vez, dentro dos métodos de Aprendizagem de Características, existem algoritmos de Aprendizagem Profunda, onde vários níveis são utilizados para uma maior capacidade de abstracção sobre os dados de entrada e, consequentemente, uma maior discriminação. Assim sendo, foram estudadas e aplicadas Redes Neuronais Convolucionais e Recorrentes, em três diferentes tópicos de investigação. No primeiro tópico, os mapas convencionais usados na prática clínica são combinados com os dados responsáveis por gerar os mapas convencionais. Com esta proposta foi possível demonstrar a vantagem em considerar ambos os tipos de dados em arquitecturas específicas. Uma segunda linha focou-se na conjugação dos dados clínicos do paciente com os dados imagiológicos. Para tal propôs-se uma função de custo, com o intuito de guiar o processo de aprendizagem da rede profunda. Mais ainda, a informação clínica, não imagiológica, foi introduzida como canal de entrada extra, garantido que informação específica de cada paciente é tida em consideração. Por último, explorou-se a aprendizagem não supervisionada, na caracterização da distribuição dos dados que descrevem a capacidade de perfusão e difusão e a hemodinâmica cerebral. Foram ainda validados vários componentes fulcrais da rede, nomeadamente as Redes Neuronais Recorrentes-Fechadas. Ao considerar a etapa de aprendizagem não supervisionada, demonstrou-se a capacidade em obter características representativas das propriedades supra-referidas, alcançando-se resultados estado da arte.

Palavras-chave: AVC, Aprendizagem Profunda, RMN

Abstract

Automatic Prediction of Ischemic Stroke from MRI images using Deep Learning

Stroke is a leading cause of death worldwide, being the second major cause of death in developed countries. Furthermore, it is also a major cause of disability, having a huge burden in society. World Health Organization predicts that a stroke event occurs at each two seconds. Stroke is categorized either as haemorrhagic or ischaemic, being the latter the most common type of stroke. Neuroimaging acquisitions play an important role during clinical assessment, evaluation and treatment planning. The most commonly used imaging technique is the Computerized Tomography, due to its availability and operational costs. Nonetheless, when available, Magnetic Resonance Imaging is preferred due to its higher capability in characterizing soft tissues, and capacity to detect early levels of ischemia. Onset neuroimaging acquisitions allow the physicians to locate and assess the brain tissue that can be recovered, which plays an important role during the treatment planning and follow-up. However, in a context where time equates to the loss of healthy brain tissue, physicians need to ponder the benefits and risks of performing clinical intervention, based on rough manual delineations and on clinical experience to predict the infarct growth across time. These tasks are time-consuming and prone to intra- and inter-physician variability. Hence, automatic prediction of stroke lesions based on onset neuroimaging acquisitions is needed to help and guide the physicians during the decision making process. The development of automatic methods is however an intricate task, due to the variety of stroke lesions, the underlying brain perfusion and diffusion processes, as well as the variability of Magnetic Resonance scans, their poor resolution and fast acquisitions.

Machine Learning comprehends a vast number of algorithms that aim to learn patterns from data, in order to achieve a specific goal or perform a specific task. One category of Machine Learning is the Representation Learning, where algorithms learn how to extract discriminative features directly from the input data. Among these methods, Deep Learning is a group of Representation Learning, which employs several levels of abstraction that characterize the input data. Thus, Convolutional and Recurrent Neural Networks were studied and applied for predicting the final stroke lesion. Three different lines of research were conducted. One research line focus on combining raw imaging data with the standard maps used in clinical practice. We demonstrate the added value of considering both data types in dedicated learning paths. Furthermore, we provide evidence on the impact of performing temporal pre-processing without hindering the performance of our method. A second line of research focused on studying and proposing methods that merge imaging with non-imaging data. To consider the latter clinical data we propose a custom loss function, to guide the learning process of the Deep Learning neural network, as well as an additional input channel, to consider patient-specific data. Lastly, we consider an unsupervised learning approach with the goal of characterizing the underlying distribution of the data. Considering the unsupervised learning block allowed us to demonstrate its discriminative power, and ground-breaking results. Additionally, we demonstrate the added value of considering Gated-Recurrent Neural Networks embedded in a Fully Convolutional Network.

All the methods developed during this thesis were trained and evaluated in publicly available datasets. This allows a fair comparison among state of the art proposals, and future comparisons with the different proposals contained in this thesis.

Keywords: Deep Learning, MRI, Stroke

Contents

Acronyms

T_{max} Time to Maximum.

ADC Apparent Diffusion Coefficient.

AIF Arterial Input Function.

ANN Artificial Neural Networks.

ASPECTS Alberta Stroke Programme Early CT Score.

ASSD Average Symmetric Surface Distance.

BPTT Back-Propagation Through Time.

CBF Cerebral Blood Flow.

CBV Cerebral Blood Volume.

CNN Convolutional Neural Network.

CPP Cerebral Perfusion Pressure.

CT Computed Tomography.

CTA Computed Tomography Angiography.

CTP Computed Tomography Perfusion.

DSC Dice Similarity Score.

DSC-MRI Dynamic Susceptibility Contrast Magnetic Resonance Imaging.

DWI Diffusion Weighted Imaging.

FCN Fully Connected Network.

FCNN Fully Convolutional Neural Network.

FLAIR Fluid Attenuation Inversion Recovery.

FN False Negatives.

FP False Positives.

GRU Gated Recurrent Unit.

GT Ground Truth.

HD Hausdorff Distance.

ISLES Ischaemic Stroke Lesion Segmentation.

LSTM Long-Short Term Memory.

MCA Middle Cerebral Artery.

MI Mutual Information.

MRI Magnetic Resonance Imaging.

mRS modified Rankin Scale.

MTT Mean Time to Transit.

NIHSS National Institutes of Health Stroke Scale.

PWI Perfusion Weighted Imaging.

RBM Restricted Boltzmann Machine.

rCBF relative Cerebral Blood Flow.

rCBV relative Cerebral Blood Volume.

ReLU Rectified Linear Unit.

RF Random Forest.

RNN Recurrent Neural Network.

ROI Region of Interest.

rt-PA recombinant tissue Plasminogen Activator.

scSE Spatial Concurrent Squeeze-Excitation.

SE Squeeze-Excitation.

SegSE SEGmentation Squeeze-Excitation.

SVM Support Vector Machine.

TanH Hyperbolic Tangent.

TICI Thrombolysis in Cerebral Infarction.

TOAST Trial Organon in Acute Stroke Treatment.

TSS Time since Stroke.

TTP Time to Peak.

TTT Time-to-Treatment.

List of Figures

List of Tables

Chapter 1

Introduction

The main goal of this thesis was to automatically predict the final infarct stroke lesion from onset neuroimaging acquisitions, namely MRI. Henceforth, our work focus on ischaemic stroke image analysis in functional MRI and Machine Learning, more particularly Representation Learning. This chapter starts by providing an overall context and motivation of these two topics, and its interplay on the work developed in this thesis. Section 1.2 presents the main objectives, while Section 1.3 describes the structure of the remaining document.

1.1 Context and Motivation

Stroke emerges from a sudden disruption of cerebral blood flow supply, deprecating the normal functioning of the brain. This condition can emerge from two different events: blockage or rupture of a blood vessel. The former is known as ischaemic stroke, while the latter is designated haemorrhagic stroke. Ischaemic stroke is the most common type of stroke, with an incidence of approximately $80 - 85$%, while the remaining $10 - 15$% are haemorrhagic. Due to its incidence and the possibility to perform therapeutic intervention, medical research and treatment is mainly focused on ischaemic stroke. The occlusion of a brain vessel caused by an ischaemic stroke deprecates the supply of nutrients and oxygen to proximal brain structures. Consequently, a cascade of haemodynamic events occurs to preserve regions in a hypoperfused state. However, as time passes, necrosis or apoptosis starts to occur, leading to the appearance of permanently damaged tissue, the infarct core. Therefore, rapid restoration of brain perfusion plays an important role in recovering tissue destined to infarct, the ischaemic penumbra, which generally surrounds the ischaemic core (González et al., 2011; Sandercock and Willems, 1992; Lopez et al., 2006).

Worldwide stroke is the second leading cause of death. Annually, 15 million people suffer a stroke, leading to the death of 5 million and the permanent brain damage of other 5 million. Permanently brain damaged people caused by stroke accounts for the third cause of disability worldwide and the fourth contributor to years of healthy life lost. In this context, stroke has a huge burden in society, having higher incidence in developed countries.

Assessment and treatment planning of ischaemic stroke relies heavily on neuroimaging acquisitions. Computed Tomography (CT) is the standard imaging protocol for assessing stroke, due to its availability and operational costs. However, ideally MRI should be the preferred choice, since it presents higher sensitivity in detecting early stages of ischaemia, but also it grants a good contrast for soft tissues. Nonetheless, regardless of the neuroimaging acquisition performed, clinicians need to ponder the risks and benefits of performing clinical intervention. This decision-making process occurs in an environment where elapsed time equates to the loss of healthy brain tissue, hence demanding a high availability of human resources. Furthermore, the neuroimaging techniques are tuned for fast and short acquisition times, which increases the difficulty of the task. To that end, understanding and predicting how the stroke lesion will evolve over time is a crucial step. However, it is a time-consuming task, roughly performed and highly prone to intra- and inter-rater variability, raising the need for semi- or fully-automatic computerized methods.

Currently, predicting ischaemic stroke evolution across time in a clinical environment is still dominated by qualitative assessment of neuroimaging acquisitions based on clinical expertise. Despite the urging need to translate this clinical expertise knowledge into a qualitative scenario, stroke tissue prediction was a small area of research. Hence, the Ischaemic Stroke Lesion Segmentation (ISLES) dataset was created in 2015. Initially focused on stroke lesion segmentation, ISLES 2016 edition changed the research direction of the community by proposing to perform stroke tissue outcome prediction, increasing the importance of this field of research. Several methods were proposed to tackle this challenge (Winzeck et al., 2018). In an era dominated by Machine Learning methods, more specifically Representation Learning, the most promising proposals were based on deep neural networks, being the majority of them based on supervised learning. Nonetheless, classical approaches (e.g. Random Forests) alongside hand-crafted feature extractors were still proposed (Winzeck et al., 2018).

Machine Learning algorithms aim to learn from a given input data, and can be broadly divided into supervised and unsupervised methods. In one hand, supervised learning requires a label, or target, so that the learning process occurs in order to perform a task. On the other, unsupervised learning does not require a target variable, and the algorithms learn distributions of the data. Nonetheless, both types of learning demand a representation of the data as input. These representations can be obtained by either developing a pipeline of feature extractors or by learning how to extract features directly from the data. The former comprehends a process of feature engineering, where the resulting features are designated hand-crafted features. Despite effective, feature engineering may be a time demanding process, which requires expert domain knowledge. Learning how to extract features directly from the data comprehends Representation Learning algorithms. Deep Learning belongs to the group of Representation Learning that learns series of hierarchical dependencies in the underlying data. These methods characterize Deep Artificial Networks encompassing a stack of several learning blocks, allowing higher levels of abstraction and higher separability of the data. Recently, a vast number of deep neural networks architectures have been proposed. The interest emerged after the achievements of Krizhevsky et al. (2012) in ImageNet challenge, with Convolutional Neural Network (CNNs) playing an important role. Hence, Representation Learning has been a recent topic with high relevance in the scientific community. In the Medical Imaging Analysis field there are already interesting and state-of-the-art achievements obtained by Deep Learning-based methods (Pereira et al., 2016; Kamnitsas et al., 2017b). This work investigates the potential of Representation Learning approaches for stroke tissue outcome prediction, culminating with the proposal and evaluation of stroke tissue outcome prediction methods based on CNNs and RNNs.

The review paper on stroke tissue outcome prediction of Winzeck et al. (2018) demonstrates that pipelines based on hand-crafted features require a high amount of features. Besides the clinical imaging motivation, the high number of features aims to depict the high variability in lesion size, location and shape. Hence, Deep Learning-based methods pose as an appealing approach to learn data-driven discriminative features. These properties allowed Choi et al. (2016) to win the competition of ISLES 2016 (Winzeck et al., 2018). However, the key to a successful prediction of the final stroke lesion resides on a robust understanding of cerebral perfusion, diffusion and other patient specific factors (Liebeskind, 2003; Wardlaw, 2010). The work of McKinley et al. (2016) provides some evidence that in fact considering different clinical scenarios allows a better tissue outcome prediction. Nonetheless, encoding this clinical knowledge into Machine Learning approaches is not straightforward. Furthermore, one needs to study which architecture is better suited for prediction of final stroke lesion. Another issue emerges with the high lesion variability and high imbalanced data, where most of the voxels belong to healthy tissue.

1.2 Research Objectives

The main objective of this thesis is to study methods for prediction of stroke tissue outcome using Deep and Representation Learning methods applied to MRI data. Deep Learning-based methods have shown its potential in object recognition tasks, specially algorithms based in CNNs. However, for the task of prediction it is not obvious if deep neural networks will provide results capable of being used in clinical practice. Furthermore, this prediction task differs from typical prediction problems, such as forecasting, where the current and posterior time-point information is known, as well as the labels from the current time-point. Predicting stroke lesion demands a high level of detail, since the delineation of the final lesion is needed, as well as the knowledge of the underlying brain phenomena which varies through time.

The first line of research of this work is the investigation of the pre-processing methods responsible for generating the standard parametric perfusion MRI maps. Currently, the spatio-temporal MRI acquisitions are subdued to a pipeline of signal processing methods, which have been recognized as ill-posed mathematical problems. Hence, we aim to study how one could retrieve additional information from this raw data and combine them with the standard parametric maps, overcoming the potential loss of information.

The second line of research is to explore and propose an automatic stroke tissue outcome prediction method that simultaneously incorporates non-imaging clinical information with imaging information. This allows the clinician to assess and evaluate different outcome scenarios and ponder on the risks and benefits of performing clinical intervention.

As final line of research, it conducts an investigation on the underlying cerebral haemodynamic with unsupervised learning methods. During the ischaemic stroke assessment clinicians tend to retrieve information from specific parametric maps, in order to characterize important factors that might dictate the overall progress of the ischaemic lesion over time. Hence, we aim to include this knowledge into an unsupervised deep neural network with the goal of providing new insights on the cerebral perfusion and diffusion.

This research aims to provide evidences that stroke tissue outcome prediction can be successfully tackled with Machine Learning methods, contributing at the same time to the field of computer science, mainly to Machine Learning. Additionally, we bear in mind that these advances might, at the long term, impact the quality of life of stroke patients and the general acceptance of Artificial Intelligence in clinical practice.

1.3 Overview

This thesis is organized in seven chapters. It starts by providing a clinical overview of stroke and a theoretical description of some concepts in Machine Learning. Afterwards, it describes the three different lines of research pointed out in Section 1.2.

Chapter 2 addresses the clinical foundations of stroke, and how different neuroimaging acquisitions are currently used to assess stroke, with special focus on perfusion and diffusion principles. Afterwards, it discusses the need for automatic methods capable of predicting the final infarct stroke lesion from onset MRI imaging. The motivation behind the development of such approaches and how they can impact clinical practice is also presented. In addition, it considers the challenges and main lines of research for stroke tissue outcome prediction. Lastly, Chapter 2 provides insight on the main trends observed on automatic prediction of final stroke lesion. These trends are sustained by a state-of-the-art review presented in Section 2.2.2.

Machine Learning foundations are described in Chapter 3. Then, the focus turns to the Representation Learning methods, with emphasis in the algorithms employed in this work, namely CNNs, Recurrent Neural Networks (RNNs) and Restricted Boltzmann Machines (RBMs).

The first line of research is presented in Chapter 4. We propose the combination of spatio-temporal perfusion imaging with the standard parametric perfusion maps. Standard parametric maps of perfusion are obtained from spatio-temporal perfusion imaging data. Hence, we propose a dedicated Machine Learning approach capable of dealing with 4D data, aiming to improve the level of information present in the standard parametric maps. Furthermore, we investigate a temporal pre-processing algorithm capable of decreasing the number of temporal acquisitions needed to characterize patients' brain perfusion, without the loss of performance.

Another line of research is described in Chapter 5. In this Chapter we investigate the combination of non-imaging data, which is commonly generated in a clinical context. We identify and suggest two distinct levels where non-imaging data can be considered: population- and patient-level. In the former we propose to codify non-imaging data into a custom loss function, investigating how it can drive the learning process of a Deep Learning-based method. In the latter level we propose to encode clinical information as an extra-channel, so that in the testing phase each patient's specific measures are considered.

Chapter 6 investigates the use of unsupervised learning, namely RBMs, applied to standard parametric maps of diffusion and perfusion to characterize different cerebral haemodynamic, which do not correlate directly with stroke lesion delineation. Furthermore, based on the clinical expertise when assessing ischaemic stroke lesions, we propose to codify such knowledge into a two-pathway unsupervised learning block.

Lastly, Chapter 7 performs an overall summary of the main research findings achieved during this work, as well as a perspective on open lines of research.

Chapter 2

Current research in neuroimaging Ischaemic Stroke

Stroke is the most common cerebral vascular disease. Among brain strokes, the ischaemic stroke is the most common one, having higher prevalence in developed countries (Feigin et al., 2014). Hence, ischaemic stroke is an active and growing research problem with huge impact on society. Assessment and diagnosis of ischaemic stroke is performed using neuroimaging acquisitions, since these allow the characterisation of the underlying brain vascular phenomena (Feigin et al., 2014).

2.1 Ischaemic Stroke: the problem

This section addresses the main neuroimaging acquisition and analysis techniques employed in ischaemic stroke. More specifically, it will focus on parametric and non-parametric MRI, which are the preferred image acquisitions for predicting final infarct core lesion, when available. In addition, it considers the challenges and opportunities in predicting final infarct lesion.

2.1.1 Overview

Cerebrovascular diseases occur in vessels that supply or drain the blood from the brain. The most common factors with the potential of modifying the risk of cerebral vascular disease are: high blood pressure, high cholesterol, smoking, abusive alcohol consumption, physical inactivity, overweight and dietary factors (Brainin and Heiss, 2014). One of the most common events originated from cerebral vascular diseases is the stroke, which is the second leading cause of death worldwide, occurring at a pace of every two seconds, with six people dying of stroke at each ten seconds (Feigin et al., 2014; World Health Organization et al., 2007). Moreover, stroke is the fourth leading cause of disease burden, measured in disability-adjusted life years, following heart disease, HIV and unipolar depressive disorders (Lopez et al., 2006).

Epidemiologically, stroke is a rapid development of focal (or global) neurologic deficit lasting more than 24 hours, and with no apparent cause other than a vascular disruption of blood supply to the correspondent area of the brain (WHO MONICA Project et al., 1990). In cases where such event lasts less than 24 hours, it is designated transient ischaemic attack (Albers et al., 2002), which is usually a predictor of an upcoming stroke event (Johnston et al., 2000).

Stroke can be characterised either as haemorrhagic or ischaemic. Haemorrhagic stroke emerges from a rupture of a blood vessel and can occur within the brain, being designated intra-cerebral haemorrhage, or between the interior and exterior spaces that cover the brain, called subarachnoid haemorrhage. As for ischaemic stroke, it consists of an occlusion of a blood vessel (Grysiewicz et al., 2008). In developed countries, about 15% of all strokes are haemorrhagic, and 85% are ischaemic stroke (Lopez et al., 2006). Based on the underlying pathophysiology there are several classification schemes for ischaemic stroke. The most widely used is the Trial Organon in Acute Stroke Treatment (TOAST) classification, which divides ischaemic stroke into five subtypes: large-artery atherosclerosis (hardening and thickening of arteries), cardiogenic embolism (cardiac blockage), small vessel occlusive disease, stroke of other known cause, and stroke of unknown cause; where the first three subtypes are the most common ones (Adams Jr et al., 1993).

Treatment of ischaemic stroke patients consists of thrombolytic therapy (e.g. thrombectomy or thrombolysis), that aims to re-establish the perfusion deficit, and hemicraniectomy, to relieve intracranial pressure (Brainin and Heiss, 2014). However, regardless of stroke prevalence and impact, the difficulty in understanding and controlling such rapidly evolving event has led to an underfunded research area (Luengo-Fernandez et al., 2015; Pendlebury, 2007), where there are still key-factors that need to be addressed. Consequently, current clinical intervention procedures are restrained to a short time-window of applicability (Saver et al., 2016; dela Peña et al., 2017).

2.1.2 Clinical perspective

Ischaemic stroke is responsible for compromising the blood flow and energy supply to specific areas of the brain (Ga, 2008). The absence of nutrients and exchange of CO_2 and O_2 triggers a series of neurochemical mechanisms referred to as ischaemic cascade. The ischaemic cascade is a complex and heterogeneous process, where cells can suffer from different levels of ischaemia, encompassing five major phenomena: excitotoxicity and ionic imbalance, oxidative stress, inflammation, apoptosis and peri-infarct depolarization and final cell death. These phenomena occur in neurons, glial or endothelial cells, regardless of the cell type (Brouns and De Deyn, 2009). In areas closer to the vessel occlusion, where the blood flow is severely reduced or inexistent, excitotoxic and necrotic cell death occurs within minutes. These areas are designated the core of the ischaemic territory, consisting of irreversibly damaged tissue (Hossmann, 1994; Fisher and Garcia, 1996). Moreover, depending on the location of the occlusion, it can constrain the blood flow of the peripheral region, being designated as the ischaemic penumbra. This area encompasses functionally impaired tissue, but structurally intact. Here several mechanisms of the ischaemic cascade are triggered leading to a progressive cellular injury and eventual cell death. As time passes, all ischaemic penumbra can become infarct core (Markus et al., 2004). The viability of the penumbra tissue is dependent on various conditions, such as the type of brain tissue and the location of the thrombus. In terms of type of brain tissue, when a stroke event occurs in white matter tissue instead of grey matter tissue, it leads to severe ischaemia and tissue oedema, since the susceptibility of the white matter cells to ischaemia is lower and the normal rates of blood flow are low, which shortens the viability of penumbra tissue. Additionally, it causes extensive neurological deficits (Stys, 1998; Petty and Wettstein, 1999). As for the location factor, the presence of a secondary structure of vessels that grants cerebral blood flow near the penumbra area, decreases the cell death pace. These vessel structures are responsible for one of the most important factors to consider in ischaemic stroke: the collateral circulation. Identifying collateral circulation provides useful knowledge in treatment and recovery planning (Liebeskind, 2003).

Ischaemic stroke pathophysiological mechanisms evolve temporally and spatially, enduring from hours to days, even after clinical intervention (Zivin, 1998). Temporally, an ischaemic stroke can be divided into four major time-windows: hyper-acute (0-6 h), acute (6-24 h.), sub-acute (from 24 h. to 2 weeks) and chronic (more than 2 weeks) (Allen et al., 2012). The hyper-acute phase comprehends the early moments where the occlusion or reduction of blood flow leads to the appearance of an infarct tissue area. During this phase, begins the shift of water from extracellular to the intracellular space. The successive increase of intracellular water content leads to the appearance of brain swelling, also called oedema, in the sub-acute phase. The chronic phase corresponds to a stage where the oedema and the inherent mass effect decreases, giving room to tissue loss and gliosis (i.e. damage to central nervous system cells). Fig. 2.1 illustrates the dynamic process of ischaemic stroke. Among the four stages, the crucial one is the acute stage, where after clinical diagnosis the treatment has higher chances of success in interrupting or decreasing the initial processes of an ongoing ischaemic cascade (Allen et al., 2012).

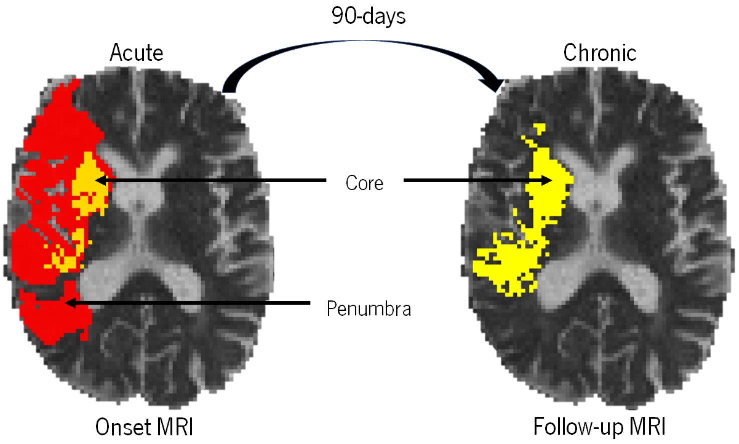


Figure 2.1: Infarct growth of a patient case with ischaemic stroke, in a 90-day temporal window.

Clinical evaluation of patients with signs of ischaemic stroke is of utmost importance, since cerebrovascular diseases are typically associated with cardiovascular or systemic diseases. These clinical conditions can impact the success of the therapeutic procedure, for example, a cardiovascular impairment increases the risks of haemorrhage or vascular injury. Additionally, in these clinical conditions the success of the rehabilitation therapy and outcome can also be diminished, since the applicable treatment options are short (González et al., 2011). To measure stroke neurological impairment the two most common scales are the modified Rankin Scale (mRS) (Quinn et al., 2008) and the National Institutes of Health Stroke Scale (NIHSS) (Harrison et al., 2013). The NIHSS is a 15-item scale that standardizes and quantifies the neurological state of a patient, with focus on intrinsic aspects that characterise a stroke event (e.g. aphasia)

(Brott et al., 1989). Various neurological functions, directly impaired by stroke, are assessed numerically, providing an ordinal, non-linear scale. The scale ranges from 0 (no impairment) to a maximum of 42 and characterises the language capabilities, motor function, sensory loss, consciousness, visual fields, extraocular movements, coordination, neglect and speech. Within the scale, scores greater than 21 are usually considered as severe impairment of neurological functions (Brott et al., 1989). As for the mRS, it is a hierarchical scale that only aims to characterise global neurological deficits related to mobility, having a maximum value of six, which denotes death (Quinn et al., 2008). The mRS scale is generally applied to evaluate recovery from stroke at a 90-day follow-up, being highly correlated with NIHSS assessment (Muir et al., 1996; Young et al., 2005). However, the limited range of scores made it harder to assess smaller changes, when compared to other scales (Young et al., 2005).

Treatment of ischaemic stroke, designated as intra-arterial therapy, aims to salvage the penumbra region (Brouns and De Deyn, 2009). Intra-arterial therapy can be divided into two major approaches: chemical, and mechanical. On one hand, chemical intra-arterial therapy either aims to dissolve the clots with thrombolytic agents or interferes in the cellular phenomena occurring during the ischaemic cascade. On the other, in mechanical intra-arterial therapy the goal is to remove clots by performing a surgical intervention. The first steps on intra-arterial therapy of ischaemic stroke began in 1995, with early intravenous administration of recombinant tissue Plasminogen Activator (rt-PA) (National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995). Even nowadays rt-PA remains the standard approved drug therapy for ischaemic stroke, due to its safeness and effectiveness (Wardlaw et al., 2012). However, rt-PA administration still has low utilization rates, due to the short three hour time-window of applicability, after symptoms onset, with lower risks of haemorrhagic complications (Emberson et al., 2014). Furthermore, even with administration of rt-PA half of the patients do not recover completely or die (Wardlaw et al., 2012). Hence, novel therapeutic strategies emerged to encompass a larger number of ischaemic stroke patients eligible for clinical intervention. One of such strategies is the mechanical therapy, granting reperfusion by performing recanalisation of the occlusion. Such endovascular therapies provide higher levels of reperfusion, and therefore better long-term outcomes (Smith et al., 2008; Rha and Saver, 2007; Lisboa et al., 2002). Mechanical intervention can be divided into three types: recanalisation or antegrade reperfusion, global reperfusion, and transvenous retrograde reperfusion; being the most common practice the recanalisation by endovascular thrombectomy. Nonetheless, the applicability of these therapeutics is still low. In the majority of the specialized centres less than 10% of acute ischaemic strokes are subdued to intravenous thrombolysis, whereas 7% – 15% are qualified for endovascular intervention (Henninger and Fisher, 2016).

The infarct location and size impacts patient's outcome and recovery (Chen et al., 2000; Beloosesky et al., 1995). The majority of studies show that motor recovery and functional outcome after stroke are worst for proximal infarctions with large lesion volumes, contrarily to distal and low volume infarctions. However, when occurring in non-sensorimotor areas characterising stroke outcome is difficult (Chen et al., 2000; Beloosesky et al., 1995).

Neuroimaging acquisitions play an important role on diagnosing and assessing stroke. Besides confirming the clinical distinction between ischaemic and haemorrhagic strokes, neuroimaging acquisitions map the responsible vascular infarction and the affected surrounding tissues, which is crucial for prognosis on the outcome of a patient (Wardlaw, 2010).

2.1.3 Neuroimaging

Image acquisition of ischaemic stroke patients is of utmost importance since besides providing information about lesion location and extent, it helps the physicians in the treatment decision-making process and in the recovery planning (Schellinger et al., 2003; Lev and Nichols, 2000). Neuroimaging in acute ischaemic stroke aims to provide information at four different levels of clinical relevance (Warach, 2001):

- The presence of haemorrhage.
- The applicability of thrombolysis in the presence of an intravascular thrombus.
- Identification of irreversibly damaged tissue, also known as infarct core.
- Identification and characterisation of the penumbra area with higher potential of being salvaged.

The latter two are the focus in ischaemic stroke medical imaging analysis. By identifying and locating the ischaemic core and penumbra tissue, it is possible to assess the risks and potential benefits of reperfusion therapies, such as thrombolysis and mechanical clot removal (e.g. thrombectomy). Patients with large penumbra tissue are the ones who benefit the most from reperfusion therapies, whereas patients with a small penumbra area have low reperfusion gains. Additionally, patients with large core lesions are the ones with higher risks of haemorrhage (Warach, 2001; Wardlaw, 2010).

In a context where time equates to the loss of brain tissue, and the available time window for treatment is short, ischaemic stroke neuroimaging acquisitions are tuned for speed, to identify as quickly as possible the patients that benefit from thrombolysis or other therapies (Selim et al., 2002). Therefore, the following section will present an overview of the standard acquisition protocols used for ischaemic stroke. Afterwards, it will focus on predicting final infarct volume from standard parametric maps of MRI, addressing the main advantages of MRI acquisitions, with focus on diffusion and perfusion imaging.

2.1.3.1 Computed tomography

Neuroimaging in ischaemic stroke is still commonly achieved by CT due to its availability, speed and cost (Wardlaw et al., 2014; Powers et al., 2018).

CT measures the X-ray signal attenuation, which is proportional to the tissue density (Hounsfield, 1973). Such attenuation values are measured in Hounsfield Units, which form a linear density scale where water has an arbitrary value of zero (Lev and Gonzalez, 2002). In the presence of an ischaemic stroke, hypoattenuated regions in CT characterise regions of severe and irreversible ischaemic damage, since the cell water content is absent. CT can simultaneously identify ischaemic regions likely to infarct, as well as predict the functional outcome and assess the success of the most used clinical therapies (von Kummer et al., 1994). Moreover, conventional CT allows the physicians to exclude the presence of intracranial haemorrhage, and also identify a large infarction core, both of which are contraindications for treatment (rt PA Stroke Study Group et al., 1998; Hacke et al., 2008). To grant a standard approach of quantifying the ischaemic extension of hypodense areas in CT, the Alberta Stroke Programme Early CT Score (ASPECTS) was presented in 2000 (Barber et al., 2000). ASPECTS divides the Middle Cerebral Artery (MCA) territory in ten regions, on two CT axial slices. The score is obtained by subtracting, from a top score of 10, one point for each region with signs of ischaemic hypodensity. A score of 10 depicts a totally normal perfused MCA, whereas a score of 0 corresponds to a complete infarcted MCA territory (Barber et al., 2000).

However, conventional CT has some limitations in identifying with accuracy hypo-perfused tissues (Wardlaw et al., 1999). In early ischaemic stroke (3-6 h), the CT signal attenuation caused by tissue oedema due to early infarction is minimal and often imperceptible to the human eye. Additionally, conventional CT lacks the capacity to detect large vessel occlusions and subsequently in predicting patients who benefit from thrombolysis (Pressman et al., 1987). Even for small bleeds, CT acquisitions have lower sensitivity when compared to MRI (Fiehler et al., 2007). To overcome these limitations two different CT modalities emerged: CT Angiography (CTA) and CT Perfusion (CTP). Due to the technological advances in scanners and their availability in emergency settings, CTA is becoming the first-line diagnostic for patients with signs of acute ischaemic stroke (Lev and Nichols, 2000; Powers et al., 2018). CTA consists of the acquisition of CT images during the administration of a contrast agent. CTA besides being fast, simple and accurate, has the capability of excluding patients who have acute stroke but do not benefit from thrombolytic therapy, due to the absence of large vessel occlusions (Lev et al., 1995). Contrarily to CTA, which depicts bulks of vessel flow, CTP is sensitive to capillary and tissue blood flow, providing insight of the blood flow delivery (Villringer et al., 1988). Therefore, CTP complements the CTA allowing simultaneously fast acquisition times, availability and affordability (Lev et al., 2001; Smith et al., 2003; Gleason et al., 2001). However, there are still drawbacks, since both imaging acquisitions require the administration of contrast agent and the exposure to radiation (Mullins et al., 2004). In addition, the degree of coverage depends highly on the available equipment, and the post-processing of CTA and CTP demands a considerable amount of computational and technical resources (Schaefer et al., 2008).

2.1.3.2 Magnetic resonance imaging

Conventional MRI has low sensitivity in detecting acute ischaemic lesions. In addition, when performing image acquisitions, it presents more impairments than CT (Wardlaw, 2010). However, with the new technologies, capable of characterising brain perfusion and diffusion, MRI applied to ischaemic stroke gained a lot of interest in the medical field. Mainly, due to its rapidity and accuracy in detecting the infarct core and the penumbra regions, and as a predictor of stroke outcome (González et al., 2011). Moreover, its fast echo planar scanning grants to diffusion and perfusion MRI acquisitions some level of resistance to patient motion. Furthermore, the acquisition is performed within seconds to two minutes (González et al., 2011). These advanced imaging techniques provide higher notions of acute stroke pathophysiology by characterising the cerebral vasculature and haemodynamics (González et al., 2011). Being so, diffusion and perfusion MRI have become the preferred imaging modalities for assessing and diagnosing ischaemic stroke onset, but also to predict clinical outcome and the final infarct core lesion at a 90-day follow-up (Powers et al., 2018). Fig. 2.2 illustrates some examples of MRI images acquired in a clinical emergency setting, to assess an ischaemic stroke patient.

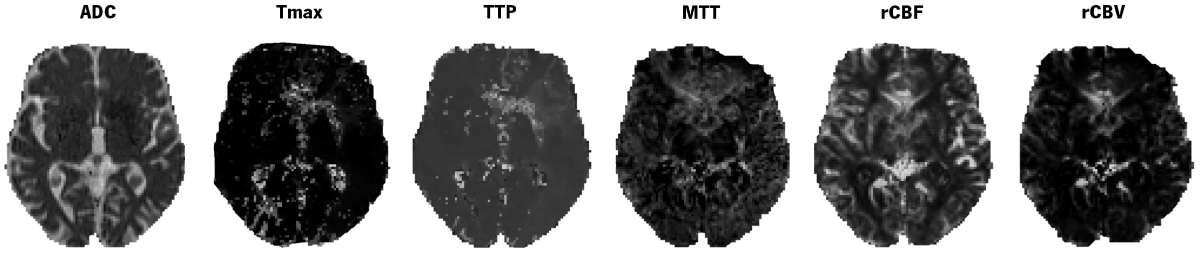


Figure 2.2: Standard parametric MRI diffusion (ADC) and perfusion maps (T_{max}, TTP, MTT, rCBF, rCBV).

Diffusion weighted MRI

Diffusion Weighted Imaging (DWI) is the most reliable method for detecting the hyperacute and acute infarct core tissue and distinguish the infarct core from other diseases (Gonzalez et al., 1999). This acquisition is based upon the principles of diffusion, which consist on the gradient of a particular substance from high to low concentrations. However, in biological processes, such as the ones that occur on the brain, DWI characterises the motion of molecules present in the water, generally referred as Brownian motion, or self-diffusion (Cooper et al., 1974). One of the most common DWI acquisitions is the Apparent Diffusion Coefficient (ADC), as can be seen in Fig. 2.2. The ADC parametric map is typically used to define the permanently damaged tissue apart from surrounding tissue where the infarct tends to extend. The core tissue is already visible in the hyperacute and acute phases as tissue characterised by hypointense regions in the ADC parametric maps (Grant et al., 2001). Due to its high contrast-to-noise ratio, DWI acquisitions can easily detect the diffusion of water that occurs even on early ischaemia (Mullins et al., 2002). Moreover, when available, the high sensibility of DWI provides useful insight in predicting clinical outcomes such as NIHSS and mRS (Thijs et al., 2000).

Perfusion weighted MRI

Perfusion Weighted Imaging (PWI) or perfusion weighted MRI, characterises haemodynamic conditions at a microvascular level (Østergaard, 2005). Hence, it allows the identification of brain tissue that is at risk of infarction due to impaired perfusion, but is not irreversibly damaged, and therefore may benefit from clinical treatment. PWI acquisitions can be totally non-invasive, by capturing brain haemodynamic based on the contrasting properties of specific particles in the blood (Williams et al., 1992). However, in clinical context the acquisition is typically performed with the injection of a bolus of exogenous contrasting material, being such technique called Dynamic Susceptibility Contrast MRI (DSC-MRI) (Rosen et al., 1990). DSC-MRI acquisitions encompass a time series of data capturing the bolus passage through the microvasculature of the brain. The most common contrast agent used in clinical practice for DSC-MRI is the gadolinium (Giesel et al., 2009). As the bolus passes along the cerebral vasculature it creates a magnetic susceptibility, which in turn causes a loss in the MRI signal translating to hypointense areas. The MRI signal attenuation depends highly on the vessel diameter. Moreover, in the presence of a vessel blockage the signal attenuation of a specific region is almost absent, unless there is collateral blood flow (Rosen et al., 1990; Liebeskind, 2003). Such temporal behaviour is illustrated in Fig. 2.3 in an acute ischaemic stroke patient, who was subdued to the DSC-MRI acquisition.

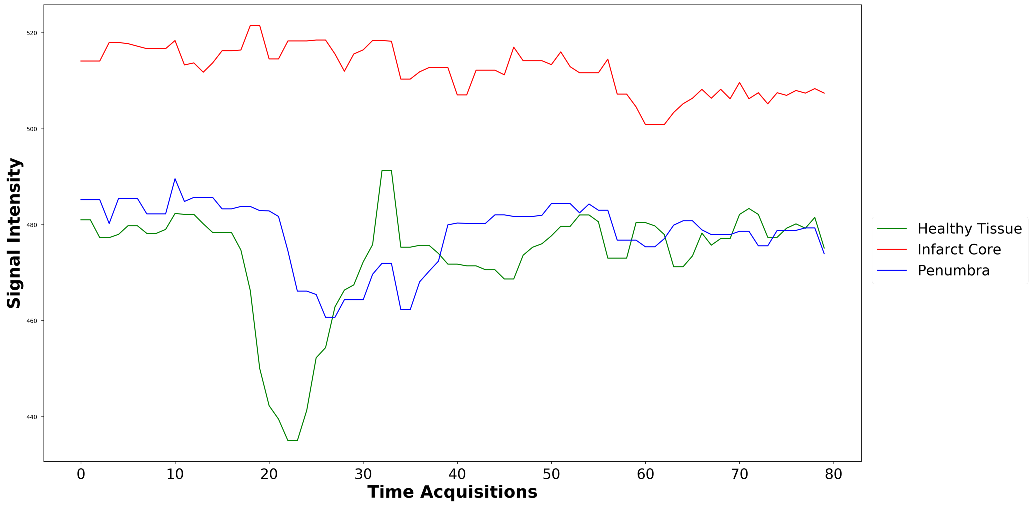


Figure 2.3: Time-attenuation curve of DSC-MRI acquisition of healthy and permanently damaged and penumbra tissues in a patient with ischaemic stroke.

As illustrated by Fig. 2.3, healthy tissue has normal perfusion capabilities, which leads to a decrease in the intensity signal as the contrast agent flows through it. Contrarily, in the penumbra tissue, since the brain blood flow surrounding the infarct area presents low perfusion rates, the variations of the intensity signal are smaller. Lastly, in the infarct core regions, the absence of brain blood flow translates into no changes in the intensities across time. These phenomena, depicted in Fig. 2.3, are the ones captured by DSC-MRI images. In order to depict all these phenomena, DSC-MRI is acquired before, during and after bolus injection (Hosseini and Liebeskind, 2018). Acquisitions prior to bolus passage allow a definition of a baseline signal intensity, which then starts to drop as the contrast bolus arrives to the brain and returns to baseline values as the contrast agent is washed out from the brain (Rosen et al., 1990). However, in an emergency room setting, DSC-MRI presents itself as a time-consuming and an impractical task in assessing and diagnosing ischaemic stroke (Wardlaw, 2010). Fig. 2.4 illustrates a DSC-MRI acquisition, in clinical practice, containing a considerable amount of imaging data, with intensity changes barely perceptive to the human eye. To surpass such disadvantages, temporal postprocessing of DSC-MRI is responsible to summarize the cerebral blood flow properties present in the DSC-MRI acquisitions into regional perfusion parametric maps (González et al., 2011).

In the DSC-MRI, the concentration of contrast agent depicted in each MRI voxel is linearly correlated

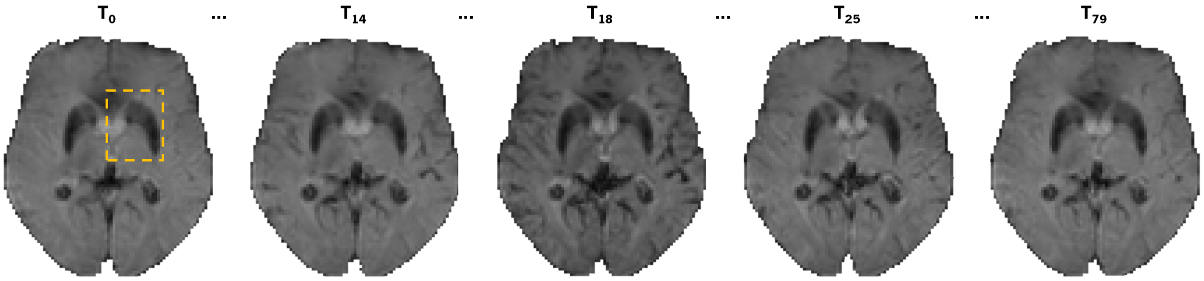


Figure 2.4: DSC-MRI acquisitions on a patient with ischaemic stroke. The dashed box comprehends the region where the final infart stroke lesion was delineated at the 90-day follow-up.

with the change in the rate of the $T2^*$ relaxation (baseline MRI signal used by default in PWI acquisitions), when compared to the baseline acquisitions. Since the $T2^*$ is an exponential process, the relation between the concentration of gadolinium and the signal intensity at a given time t is mathematically given by 2.1 (Østergaard, 2005):

$$C_{gd}(t) = -kln\left(\frac{S_t}{S_0}\right) \tag{2.1}$$

In Equation 2.1, $C_{gd}(t)$ denotes the concentration of gadolinium at time t after bolus arrival, S_t the signal intensity at such time, and S_0 represents the signal intensity before bolus arrival. The variable k denotes a constant whose values relate to the MRI acquisition setup, namely the Excitation Time (TE) of the MRI radio frequency pulse, $k = \frac{1}{TE}$. From this knowledge it is possible to obtain a time-concentration curve of contrast agent, inversely related to a time-attenuation curve, responsible for the generation of standard parametric perfusion maps (Østergaard, 2005). A theoretical example of the time-concentration curve in Perfusion CT imaging is shown in Fig. 2.5. In addition, it displays the information captured by each surrogate parametric perfusion maps.

Despite Fig. 2.5 refers to the signal intensity variation in Perfusion CT (measured in Hounsfield Units), the fundamental principles present are kept equal in DSC-MRI acquisitions. The permeability slope shown characterises the presence of contrast agent that was accumulated by brain tissues. This phenomenon is also an indicative that in fact there is a disruption in the normal brain blood flow (Hosseini and Liebeskind, 2018). To characterise the vasculature of the brain, the three most common perfusion maps are the Cerebral Blood Volume (CBV), Cerebral Blood Flow (CBF) and Mean-Time-to-Transit (MTT). The CBV aims to characterise the amount of contrast agent present in the tissue over the acquisition time. This parametric map emerges by computing the area under the time-concentration curve for each voxel, after the bolus arrival, therefore requiring little post-processing. As for CBF it results from the first slope in the time-concentration curve, characterising the bolus arrival in the brain. Hence, a higher slope characterises a faster brain blood flow, meaning that the contrast agent reaches its maximum concentration sooner in time (Østergaard et al., 1996). For penumbra areas, as shown in Fig. 2.3, the perfusion slope is lower, since the inflow is slower and less robust, and may be provided by collateral circulation. Regardless of the parametric map of perfusion, an accurate measurement of CBV and CBF is a difficult task, since the slope of bolus

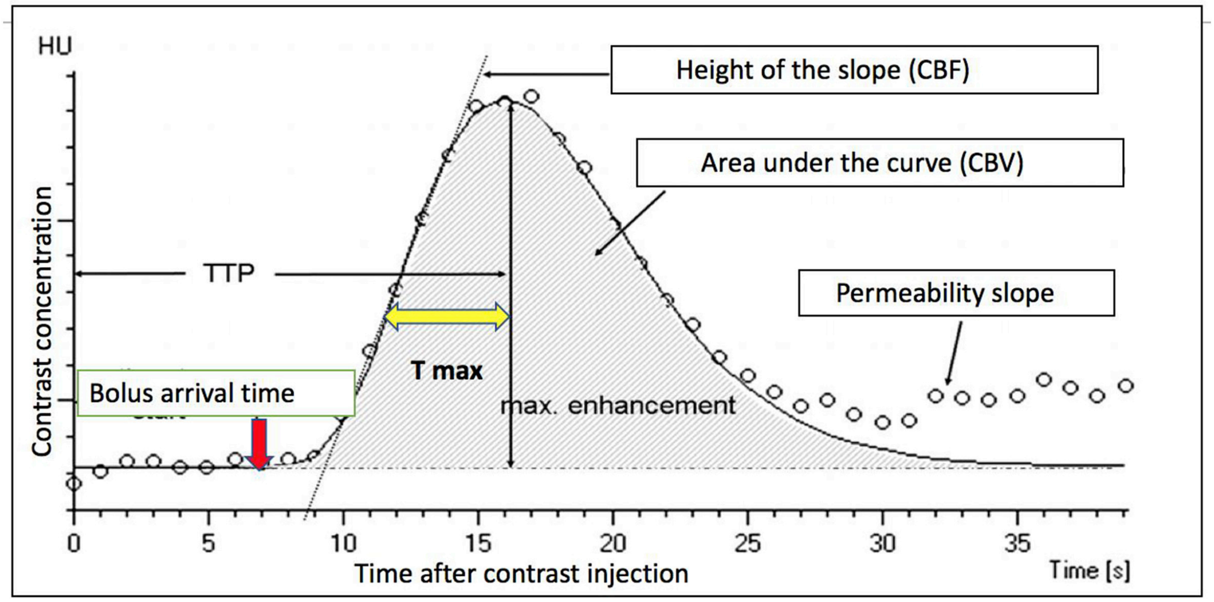


Figure 2.5: Time-concentration curve of a voxel, and the correspondent derived parametric perfusion maps in Perfusion CT imaging. Reproduced from Hosseini and Liebeskind (2018) with permission from Elsevier.

arrival can be influenced by the vascular system responsible to grant the brain blood flow (Østergaard et al., 1996). To surpass this limitation, the time-concentration curve of each voxel is convolved with an Arterial Input Function (AIF) that characterises the blood flow of a major feeding artery of the brain, typically the contra-lateral MCA, the internal carotid artery, or an average of multiple measurements in regions that contain multiple large brain arteries. Being so, a deconvolution algorithm receives as inputs two time-concentration curves, the curve of the targeted voxel and a reference curve, generating a residue function that serves as an estimation of CBF, in relation to a contra-lateral region, hence being designated relative-CBV (rCBV) and relative-CBF (rCBF). After computing rCBV and rCBF parametric maps, one can compute the MTT in accordance with the central volume theorem – MTT $= {}^{CBV}/_{CBF}$. Alongside the referred parametric perfusion maps other perfusion maps are typically generated, such as the Time to maximum (T$_{\text{max}}$) and the Time to Peak (TTP). The T$_{\text{max}}$ characterises the time at which the residue function reaches its maximum. T$_{\text{max}}$ perfusion parametric map allows a fairly and fast delineation of regions with suspicions of altered perfusion, being independent of rCBV, rCBF and MTT parametric maps. The TTP measures the time for signal intensity to reach its minimum (maximum concentration of contrast agent), being capable of delineating conspicuous regions of perfusion, without the need for a deconvolution algorithm or the definition of an AIF. However, TTP maps are correlated and dependent on other parametric maps, rCBV, rCBF, MTT, and T$_{\text{max}}$ (Østergaard et al., 1996).

The presence of a blood vessel occlusion, which leads to a regional decrease in cerebral perfusion pressure (CPP), causes the blood vessels to dilate as a mechanism to reduce vascular resistance and maintain blood flow. Visually, this mechanism can be captured by PWI as an increase in the signal intensity of rCBV and consequently the MTT parametric maps. However, if vasodilation is incapable of maintaining cerebral blood flow, besides the referred changes in intensity, the CBF parametric map may demonstrate an intensity decrease in area affected by the vessel occlusion. These mechanisms are key factors to understand the different conditions of ischaemia, where the T_{max} and TTP are of little impact, since they are indirectly related to cerebral haemodynamics and therefore tissue viability (Astrup et al., 1981). In Fig. 2.6, one can observe this behaviour in an acute ischaemic patient, within the ischaemic penumbra and ischaemic core.

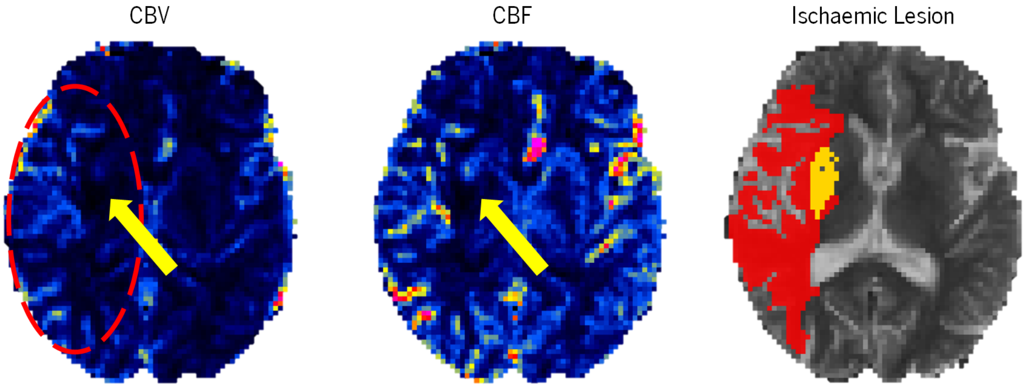


Figure 2.6: Example case of an ischaemic stroke patient, and the respective intensity findings in the CBV and CBF parametric maps. In addition, the penumbra (red) and the core (yellow) are delineated over a T2 sequence.

As can be observed in Fig. 2.6, the ischaemic core has low intensity values both in the CBV and CBF parametric maps. Whereas for the penumbra area, the CBV shows a slightly higher signal intensity in the right hemisphere, when compared to its contra-lateral portion. In the CBF map, it is observed a subtle decrease in the signal intensity in the right parietal lobe. Generally, these findings are summarized in Table 2.1, showing the intensity variations expected for each condition of the acute phase, in CBV, CBF and MTT parametric maps.

Table 2.1: Intensity increase (↑) or decrease (↓) observed in parametric perfusion maps in the presence of an acute haemodynamic conditioning, in the different tissue types.

	CBV	CBF	MTT
Compensated low CPP	↑	-	↑
Ischaemic Penumbra	↑	↓	↑
Ischaemic Core	↑↓	↓	↑↓

Even by generating standard parametric maps by temporally processing DSC-MRI, simultaneous identification of ischaemic core and penumbra is however an intricate task, being usually complemented with DWI acquisitions. The DSC-MRI struggles in identifying the core tissue that is destined to infarct, which can be performed by a quantitative analysis of CBV maps. However, in cases where the infarct core may show post-ischaemic hyperperfusion, CBF measures cannot identify tissue with conspicuous behaviour of ischaemia. Moreover, since DSC-MRI acquisitions are translated to relative-CBF maps (regional measurements in relation to a reference), they do not contain reliable information regarding the absolute cerebral blood flow (Rempp et al., 1994). Another example is the truncation of the time-concentration curves, due to an incomplete number of acquisitions, leading to the presence of artefacts when estimating CBV (de Ipolyi et al., 2010), and a delayed bolus passage, which can be interpreted as a false underperfused region (Calamante et al., 2000). Therefore, DSC-MRI acquisitions has some pitfalls.

On the overall, DSC-MRI acquisitions can be viewed as a useful tool for assessing the tissue with chances of being rescued, whereas DWI provides information on the non-salvageable tissue, hence focusing on the identification of the region nearby the thrombus occlusion.

2.1.4 Acute ischaemic stroke image analysis

Imaging of acute ischaemic stroke is a powerful tool in the clinical emergency setting, since it allows the physicians to detect the non-salvageable tissue, but also functionally impaired tissue with high chances of being salvageable by clinical intervention. To do so, the preferred image modalities are based on perfusion and diffusion MRI (Wardlaw, 2010; Powers et al., 2018). The capability to characterise the cerebrovascular dynamics allows the physicians the possibility to estimate the infarct core growth and predict the final infarct lesion. This estimation provides valuable knowledge, when assessing the risks and benefits associated with clinical intervention and posterior clinical outcome and recovery planning (Sorensen et al., 1996). Moreover, the development of these strategies can increase the eligibility of more patients to benefit from clinical intervention.

This section aims to provide insights on the intricate task of predicting stroke lesion outcome, but also the challenges and opportunities that machine learning-based methods can have in supporting the physicians decision-making process. Therefore, the focus will be in predicting stroke lesion outcome based on the onset perfusion and diffusion MRI acquisitions.

2.1.4.1 Predicting final stroke lesion from acute MRI acquisitions

When compared to conventional MRI or CT, DWI and PWI acquisitions have high sensitivity and specificity in diagnosing ischaemia, and in identifying early regions destined to infarct (Lövblad et al., 2015; Simonsen et al., 2015). By comparing both ischaemic stroke regions delineated in PWI and DWI modalities, physicians can define an area at risk of infarct with high chances of being salvageable (Rimmele and Thomalla, 2014), which is the main target of ischaemic stroke therapy (Lövblad et al., 2015). Such area is designated the Diffusion/Perfusion mismatch region, where various studies have shown correlations with the clinical outcome and final infarct lesion of a patient (Kane et al., 2007; Rimmele and Thomalla, 2014). Fig. 2.7 illustrates a clinical example, characterising this region.

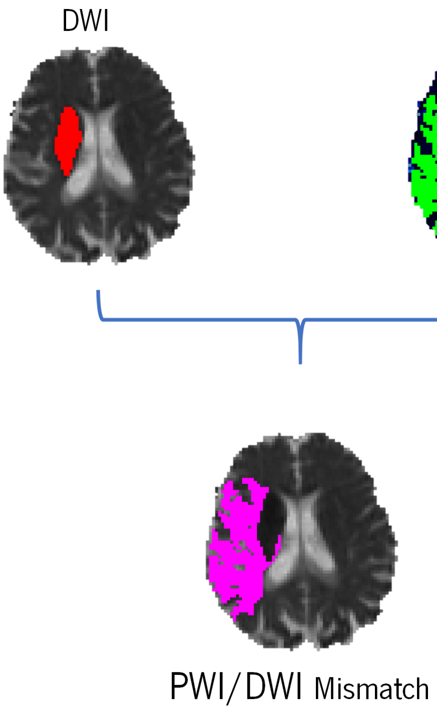


Figure 2.7: PWI/DWI mismatch in an ischaemic stroke patient.

According to the behaviour shown in Fig. 2.7, in the absence of successful clinical intervention (reperfusion) and collateral blood flow that maintains underperfused brain tissue viable for recovery, the hyperintense region delineated by the acute PWI will progress to permanently damaged tissue, leading to an infarct growth. As for the acute DWI, since it provides information on already irreversible damaged brain tissue, in the presence of a successful reperfusion the final infarct lesion will be smaller, when compared to the previous scenario, and will comprehend a portion or the totality of hypointense region in the DWI. Hence, the acute PWI acquisition characterises the worst clinical outcome scenario, whereas DWI designates a better outcome prognostic. Several studies have shown that PWI lesions larger than the DWI lesion is in fact an indicative of an infarct growth. However, in the opposite scenario, where the DWI is larger than the PWI, predicting the infarct growth is difficult. To conclude, across time, the stroke infarct can remain unchanged or it can either grow at a slow or fast pace, depending on several factors (Barber et al., 1998). Therefore, simultaneous acquisition of diffusion and perfusion MRI allows the physician to estimate robustly the clinical outcome of a patient but more importantly the final infarct volume (Rimmele and Thomalla, 2014; Kane et al., 2007; Schaefer et al., 2002). Estimating the final infarct core volume is an essential step when deciding whether clinical intervention should be performed (Lövblad et al., 2015; Rimmele and Thomalla, 2014).

Ischaemic stroke comprehends a high number of heterogeneous phenomena that can influence the brain vasculature behaviour and perfusion deficits. Hence, estimating the progress of the stroke lesion, and consequent decision on clinical intervention, from imaging acquisition remains a challenging task (Winzeck et al., 2018). Current clinical practice to estimate the Diffusion/Perfusion mismatch is either performed manually or by using semi-automatic tools (Dani et al., 2011). Although the latter approach reduces the clinical neuroimaging assessment time, they are still prone to intra- and inter-rater variability (Deng et al., 2019). In the light of this context, reliable automatic approaches can be of interest and help physicians in performing a better treatment and recovery planning.

2.1.4.2 Challenges in predicting final ischaemic stroke

Prediction of the final infarct core volume in ischaemic stroke poses as a complex task but of great value for treatment and recovery planning. Regardless of stroke lesion location and size at onset assessment, one must keep in mind a series of complex and correlated temporal processes that can influence the overall progress of ischaemia occurring in hypoperfused tissue. One of such aspects is the presence of collateral blood flow, which grants considerable perfusion rates, extending the viability of an underperfused brain tissue (Liebeskind, 2003). In the light of such phenomena neuroimaging allows a higher understanding of the undergoing processes occurring in the brain, with emphasis on MRI acquisitions of perfusion and diffusion (Barber et al., 1998).

In acute ischaemic stroke, clinical evaluation of the standard parametric maps (e.g. ADC and T_{max}) can identify infarct tissue and hypo-perfused tissue that will infarct in the absence of therapeutic intervention. Hypointense regions of the ADC map characterise regions with limited diffusion, which usually indicates irreversible tissue damage (i.e. infarct core) (Butcher and Emery, 2010a), while hyperintense regions of the T_{max} map indicate perfusion prolongation, which correlate to underperfused brain tissue (i.e. penumbra) (Butcher and Emery, 2010b). Besides considering the complexity of this time-evolving process, to correctly predict the final ischaemic stroke lesion, it is also necessary to consider the impact of the clinical intervention on the underlying brain perfusion and diffusion. To better understand the latter aspect, consider the two acute ischaemic patients illustrated in Fig. 2.8.

Analysing the selected cases, it is possible to draw two conclusions. From the standard parametric maps, illustrated in Fig. 2.8a, in this patient, the ADC does not present any hypointense region, so no infarct tissue may be identified, but the T_{max} delineates a region of low perfusion restriction. Although, the final infarct prediction should consider a small lesion, due to the inability to identify infarct tissue on the onset ADC, the follow-up delineation considered a large final lesion. This phenomenon is explained by an unsuccessful clinical reperfusion. Observing now Fig. 2.8b, this patient comprehends the opposite phenomena, where the final infarct lesion is smaller than the hypointense region present in the ADC (arrow). This indicates reversible diffusion restriction, which is a rare case (Labeyrie et al., 2012) delineated by the radiologist using a follow-up T2-weighted acquisition. So, the method has not only to capture the time-evolving process of diffusion and perfusion, but needs also to consider the success level of the clinical intervention, which may condition the final lesion either to be confined to the hypointense region of the ADC map, or to additionally grow to brain tissue areas that are hyperintense in the T_{max}. Thus, once again, it is possible to retrieve that predicting the final infarct stroke lesion is a challenging problem.

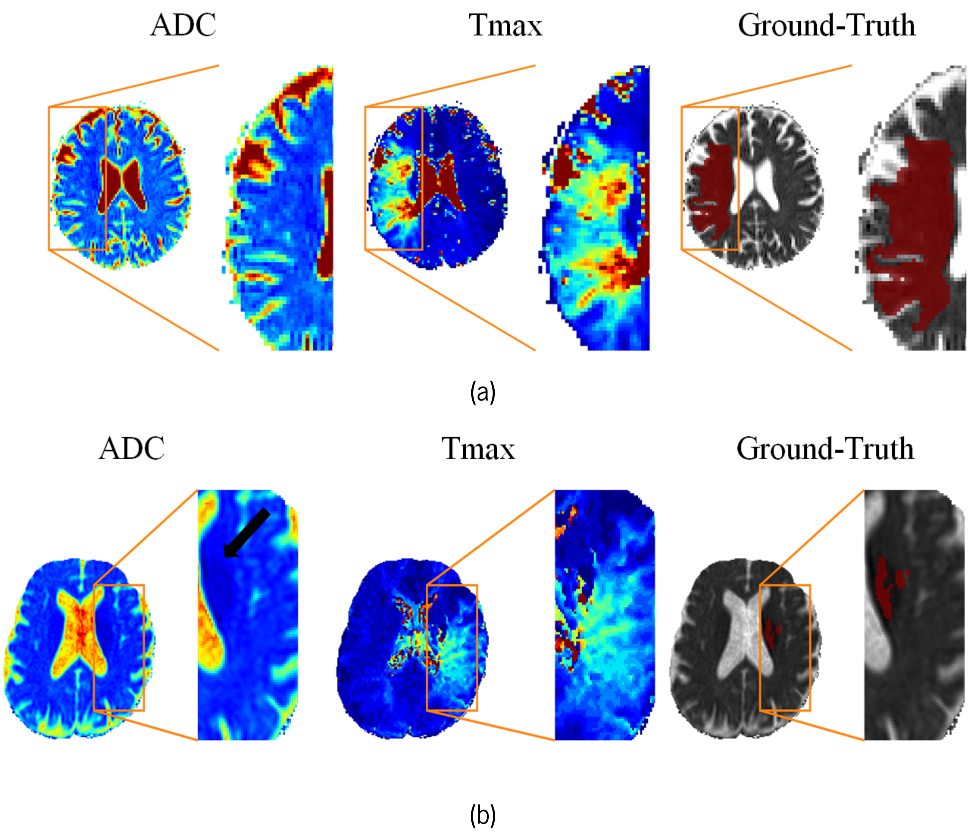


Figure 2.8: ADC and T_{max} parametric maps, and the final lesion delineated at a 90-day follow-up, overlapped with the onset ADC, of patient 0036 (Fig. 2.8a) with an unsuccessful reperfusion, and patient 0006 (Fig. 2.8b), where the clinical intervention was successful. Cases retrieved from ISLES 2017 training set.

In addition to the complex dynamic processes occurring in ischaemic stroke, neuroimaging acquisitions pose some challenges as well. In a clinical context usually known for the motto – " Time is brain " – neuroimaging acquisitions, regardless of being MRI or CT, are tuned for fast and short acquisitions (Nakamura et al., 2005). Besides the associated noise and artefacts that can occur in such circumstances, other events can impact the assessment of ischaemic stroke (Nakamura et al., 2005). For the specific case of MRI, the acquisitions can be influenced by the bias field artefact, which comprehends a smoothly variant tissue inhomogeneity that translates to different signal intensities characterising the same tissue in different locations (Vovk et al., 2007). Lastly, different acquisition protocols employed across different clinical centers makes it difficult to extrapolate clinical findings across patients (Winzeck et al., 2018).

2.2 State-of-the-art

Contrary to stroke lesion segmentation, where several methods have already been proposed (Rekik et al., 2012; Maier et al., 2017; Weinman et al., 2003), the complexity of stroke tissue outcome prediction has only been recently tackled by the machine learning and medical imaging analysis research communities. Nonetheless, the development of such proposals has already been recognized as an important research area to further explore treatment viability and posterior assessment of its success (Lou, 2019).

Automatic final infarct core prediction from onset ischaemic stroke acquisitions is a recent research field, where the majority of the proposed methods can be grouped into the class of supervised discriminative methods. Hence, there are proposals based on multivariate linear regression models (Rose et al., 2001; Scalzo et al., 2012), decision trees (McKinley et al., 2016), and CNN-based deep neural network architectures (Choi et al., 2016; Maier et al., 2017). There are also proposals based on generative unsupervised statistical models (Kemmling et al., 2015; Abulnaga and Rubin, 2018).

2.2.1 Main trends on automatic final infarct prediction

Supervised discriminative learning methods aim to characterise a conditional distribution $p(y|x)$ to be capable of predicting a target variable y from an input source x. To do so, these methods focus on learning the distribution space of the data, rather than learning the phenomenon responsible for the generation of such data. Hence, supervised discriminative methods avoid prior assumptions of the data, which can be incorrect (Murphy, 2012). Discriminative methods are mostly applicable in a supervised learning approach, where for each training instance there is a correspondent label (Murphy, 2012).

First steps on stroke final infarct prediction encompassed handcrafted features, which are fed to a discriminative classifier, generally a (RF) (Rose et al., 2001; McKinley et al., 2016). However, recent approaches are based in Representation Learning, where the learning paradigm shifts from developing handcrafted features to designing a suitable architecture that learns the best set of features. Hence, in Representation Learning, the method learns a feature space directly from data (LeCun et al., 2015).

In Ischaemic Stroke Lesion Segmentation (ISLES) 2016 and 2017 editions all the published methods, to the best of our knowledge are based on supervised discriminative approaches. Consequently, there are already published methods based on representation learning for predicting stroke final infarct core using MRI imaging (Choi et al., 2016; Winzeck et al., 2018). Nielsen et al. (2018) employed a well-known architecture, the SegNet, to predict the infarct core at 30-day follow-up after onset diagnosis of ischaemic stroke. The research contained in this thesis follows the same line of thought employing discriminative supervised and unsupervised learning methods based on representation learning techniques.

With the release of ISLES 2018 edition, new methods have been proposed (Dolz et al., 2018; Abulnaga and Rubin, 2018; Liu, 2018; Islam et al., 2018; Pinheiro et al., 2018). However, ISLES 2018 focus on segmenting the stroke lesion, instead of predicting the final infarct with on-set CT imaging acquisitions, which goes beyond the scope of this work. In the following section, the proposed methods for stroke tissue outcome prediction are reviewed.

2.2.2 Methods of stroke tissue outcome prediction

Rose et al. (2001) proposed an approach for stroke tissue prediction, with a two-stage method based on parametric perfusion and diffusion MRI maps. The first stage of the method defines a ROI based on the intensity signal of specific standard parametric maps, e.g. MTT, CBF, CBV, and DWI. The second stage performs stroke tissue outcome prediction by employing Gaussian mixture models trained on different sets of parametric maps, confined to the ROI.

Bauer et al. (2014) used RFs to segment the onset stroke lesion or to predict the final stroke infarct depending on the availability of onset acute imaging or follow-up imaging. Similarly, McKinley et al. (2016) also used RFs in a two-stage approach for lesion characterisation and lesion outcome prediction. Each stage encompasses two RFs classifiers. In the first stage the goal is to define a ROI, which contains the brain region with low levels of perfusion. To achieve so, the two RFs are trained with hand-crafted features extracted from different sets of MRI parametric maps. Afterwards, by having defined the location and extension of the brain tissue being affected by the perfusion deficit, a second set of two RFs performs tissue outcome prediction. At this stage the classifiers were trained on different sets of patients, stratified by the TICI score. One classifier is trained with unsuccessfully reperfused patients, whereas a second classifier is trained with successfully reperfused patients. The final prediction is obtained by combining the results of both classifiers, using a logistic regression model.

Scalzo et al. (2012) proposed a framework for stroke tissue outcome prediction, which characterises the state of the lesion four days after clinical intervention (thrombectomy). From the FLAIR MRI sequence, ADC and T_{max} MRI maps, the method applies a regression model that learns the behaviour of neighbouring voxels within a cuboid.

Kemmling et al. (2015) employed a multi-modality approach of CT and MRI maps with non-imaging clinical meta-data, namely the TICI score and the time to treatment of each patient, to perform tissue outcome prediction. The authors employ a multivariate generalized linear model responsible for computing a voxel-wise probability, which estimates the final ischaemic infarct lesion. The generalized linear method is expanded to combine, at a voxel-wise level, the imaging information with non-imaging variables of time and degree of revascularization (TICI).

Recently proposed methods for stroke tissue outcome prediction have been using deep learning-based models (Choi et al., 2016; Winzeck et al., 2018; Nielsen et al., 2018). Choi et al. (2016), winner of the ISLES 2016 Challenge, proposed an ensemble of twelve CNN architectures, grouped into two sets of networks. The first group comprehends four 3D U-Net based architectures (Ronneberger et al., 2015) performing voxel-wise tissue outcome prediction. The second group of networks uses two-pathway Fully Connected Networks (FCNs) performing two types of patch-wise classification. One path classifies a patch as lesion if it includes any lesion voxel. The other FCN path classifies a patch as lesion if the central voxel is a lesion. After merging the two pathway FCN, the method incorporates meta-data by adding a dense layer of clinical predictors merged with the imaging output of each network. The final stroke lesion prediction results from a weighted merging of all models. At ISLES 2017 Challenge competition new models were proposed based on deep neural networks, which were analysed and compared in Winzeck et al. (2018). Mok and Chung (2017) applied deep adversarial training for stroke tissue outcome prediction in an ensemble of U-Net based architectures. Monteiro and Oliveira (2017) proposed a method based on the V-Net architecture (Milletari et al., 2016). The training was conducted with a custom loss function that computes the weighted sum between Dice score and cross entropy. Lucas and Heinrich (2017) proposed the application of a U-Net based architecture, where besides including the MRI maps from the same slice, it also includes patches from 3 neighbouring slices and 2 hemispheric flips. At the expanding tract of the architecture, each level computes a Dice loss after softmax activation. Afterwards, all losses are summed up, being the loss of foreground and background weighted accordingly to a prior probability. Robben and Suetens (2017) employed a CNN-based architecture inspired by Kamnitsas et al. (2017b). From two-pathway 3D networks, the MRI inputs combined with clinical meta-data are fed to two networks. A first one that keeps the resolution of the data, and a second one where the resolution was lowered by a factor of 3. The output of each network is then transformed to the same scale and merged by a fully connected layer. Similarly, Niu et al. used multiple scales of overlapping 3D patches to capture local and global spatial information. Rivera et al. also built on the work of Kamnitsas et al. (2017b) and Milletari et al. (2016), by proposing a scheme to extract different patch resolutions, independent of each other, that are feed into four different paths. Afterwards, a fully connected layer combines all the outputs to perform stroke tissue outcome prediction. Pisov et al. (2017) employed an ensemble strategy by combining different CNN-based architectures to overcome the strong anisotropy of the data. Yoon et al. provided a two-stage gated CNN. In a first stage, the authors perform lesion detection and delineation. Afterwards, based on the probability maps of the first stage, a second CNN architecture intervenes on regions where the probability maps of background and foreground are similar.

Clèrigues et al. (2018) proposed a deep neural network architecture based on the U-Net (Ronneberger et al., 2015), designating it SU-Net. The authors developed a 3D network to enforce tissue differentiation and to capture collateral blood flow. Recognizing that predicting the final infarct core is an intricate task, the authors employ local residual connections and local residual blocks to allow a better gradient flow when performing the backpropagation step. More interestingly, the authors proposed an asymmetric encoder-decoder design, which resulted in a decoder path with fewer convolutional layers and consequently a smaller parameter footprint. Training such network was performed under sampling strategies that enforce an equal distribution of the two population data (Kamnitsas et al., 2017b), healthy and lesion tissues, with soft Dice loss (Milletari et al., 2016). Recognizing that highly complex deep neural network architectures struggle in capturing the underlying haemodynamic phenomena of stroke tissue evolution across time.

Lucas et al. (2018) proposed a combination of a 3D U-Net based architecture with a Convolutional Auto-Encoder for stroke tissue outcome prediction with CT. The latter aims to mimic clinical expertise when predicting the final infarct core lesion by capturing the anatomical latent space behind stroke lesions. Hence, in a first stance, the 3D U-Net architecture predicts the tissue core and penumbra, which is then fed to the auto-encoder that enforces spacial regularization.

Nielsen et al. (2018) evaluated different CNN architectures in stroke tissue outcome prediction. The authors show that deeper architectures, based on the SegNet architecture (Badrinarayanan et al., 2015), perform better when compared against shallow CNNs and standard thresholding techniques. The method predicts on a 30-day follow-up acquisition achieving robust and precise results. Nonetheless, contrarily to all the deep learning-based methods reviewed so far, the prediction is performed in a temporal window with fewer changes in hypo-perfused volume and salvaged tissue, instead of predicting for a 90-day follow-up.

The majority of the proposed methods for stroke tissue outcome prediction only considers the standard parametric maps (Winzeck et al., 2018). Only recently, perfusion DSC-MRI has been considered. Although not applied to stroke tissue outcome prediction, there are approaches that aim to achieve a higher level of abstraction from the perfusion DSC-MRI (McKinley et al., 2018; Hess et al., 2018). Hess et al. (2018) developed a deep neural network architecture to avoid the need for a deconvolution step. The method aims to generate new standard parametric maps from an automatic machine learning approach, independent of the underlying mathematical foundations and drawbacks of the deconvolution. Robben et al. (2018) also focused on spatio-temporal data to predict the final infarct volume, but having as neuroimaging acquisition the CTP data. The proposed network is inspired by the work of Kamnitsas et al. (2017b), where the authors showed the added value of avoiding the deconvolution step and provided the temporal data directly to a deep neural network architecture. However, the work proposed by Robben et al. (2018) is still dependent on the manual definition of an AIF, which can be prone to inter- and intra-observer variability. In addition, Robben et al. (2018) combined imaging with clinical meta-data increasing the performance of their method.

From the reviewed literature, the proposed methods for stroke tissue outcome prediction lack in the capacity of considering the DSC-MRI spatio-temporal imaging information into deep neural network architectures. Moreover, even with the recent proposal of Robben et al. (2018) for spatio-temporal images of CTP, the infarct area predicted has a short time window of evolution and still requires manual intervention for the definition of an AIF. Another open area of research resides in the inclusion of non-imaging clinical information. There are already proposals that combine non-imaging information alongside imaging data gathered at the onset time (Robben et al., 2018; Choi et al., 2016). However, to the best of our knowledge, none of the proposed approaches has the capability to encode and characterise the non-imaging clinical information at a population-level. Lastly, in the literature, the cerebral blood flow haemodynamic has either been indirectly considered by dichotomization of the training data (McKinley et al., 2016) or directly considered by specific non-imaging clinical information of reperfusion (e.g. TICI score). However, from clinical expertise, particular standard parametric maps provide useful information regarding the onset perfusion deficits that characterise the tissue which can be salvageable. The referred three lines of research are the ones addressed in this thesis.

2.3 Summary

Stroke accounts for one of the deadliest causes of death worldwide, having one of the heaviest economical burdens on the society. Two types of stroke can be diagnosed: haemorrhagic or ischaemic, where the latter is the most common type. Neuroimaging of stroke, besides allowing a distinction between haemorrhagic or ischaemic, allows the evaluation and assessment of such events. CT still remains the most widely used due to its availability and operation costs. Nonetheless, to help clinicians assess the infarct core tissue and tissue at-risk advanced techniques of MRI provide a better understanding of brain blood flow and lesion delineation. Regardless of the neuroimaging technique used, assessing ischaemic stroke is an intricate task. This task, which needs to be performed in a short period, is a dynamic process that evolves over time and is influenced by a series of physiological properties. Therefore, one of the open problems in ischaemic stroke resides in the automatic and robust infarct core prediction based on the onset acquisitions, to better guide the clinicians on the selection of the best therapeutic. Given the complexity of this task, alongside the recent advances in the Machine Learning techniques, these approaches pose as interesting and viable for predicting stroke tissue outcome. Automatic prediction of the final ischaemic stroke lesion is of clinical relevance, since it helps the physicians in assessing the risks and potential benefits of intra-arterial thrombolysis or mechanical clot removal, but also in planning the recovery process. Moreover, the fast processing times of these algorithms do not impair the clinical evaluation.

Automatic prediction of stroke tissue outcome is a recent and growing research field in the medical imaging community. Despite being far from being applicable in clinical practice, the development of methods, capable of predicting the final infarct core, have been already recognized as of great importance in ischaemic stroke (Lou, 2019). By providing important information to the physicians about the underlying dynamic process of a stroke lesion, it may also guide them in the time-critical decision-making process, which ponders the risks and benefits of performing clinical intervention. The majority of the proposals for predicting the final stroke infarct core are based in discriminative supervised learning methods, more specifically in deep neural network architectures, namely FCNNs (Winzeck et al., 2018). Regardless of the model scheme, predicting final infarct core is still a challenging and intricate task, that needs to consider scenarios of successful and unsuccessful reperfusion. Furthermore, in each reperfusion scenario, predicting the infarct growth, and consequently the final stroke lesion, needs to be aware of various haemodynamic factors (e.g. location and collateral circulation), which hinders the learning process.

Chapter 3

Machine learning concepts

Machine learning methods are the main foundations for all the conceived methods contained in this work. Hence, this chapter provides an overview on machine learning starting by distinguishing the different types of learning, then followed by the important problem of representation learning, giving special emphasis on deep neural networks, which are the key algorithms of the proposed methods for predicting the final infarct stroke lesion.

3.1 Overview

Machine learning encompasses a broad range of algorithms with the main goal of learning patterns from data, to posteriorly perform a certain task. These two phases are performed sequentially, where the first stage is designated as training, and the second one corresponds to testing for evaluation or model usage after deployment. Training occurs so the model can learn different and distinct patterns, by optimizing parameters to achieve a goal. The optimization of the parameters occurs under constrained conditions, also called hyper-parameters, which define the overall behaviour of the learning phase. Testing applies the learned model to unseen data. To conduct both phases, it is possible to identify three distinct sets of data. The training data, used for the training phase, and the validation and testing data, which are used for the testing phase. The validation data allows the evaluation of the machine learning model, enabling the search for the best set of hyper-parameters, while evaluating the model on the testing data characterises its robustness and generalization capacity to unseen data. (Murphy, 2012; LeCun et al., 2015).

Regardless of the phase, machine learning methods require data, which consists of a set $\mathcal{D}$ comprehending several records or samples. Generally, each record is characterised by quantitative measures, commonly designated as features or attributes. Thus, a given record, indexed by i is described by a feature vector of M features, $x_i = [f_j : j = 1, \cdots, M]$ with $x_i \in \mathbb{R}^M$. Considering all the records, N, of the set $\mathcal{D}$, the data can be described by a matrix $\mathcal{X} \in \mathbb{R}^{N \times M}$. However, in applications where the records have a structural meaning (e.g. audio, image, video), x_i benefits from keeping the same structure, being commonly designated as feature maps (Murphy, 2012; LeCun et al., 2015).

3.1.1 Learning methods

Machine Learning methods can be distinguished in three different types, depending on how the learning process is conducted, being designated as: supervised, unsupervised and reinforced. The latter will not be considered, since it is out of scope of this work.

Supervised learning occurs when each input record x_i has a correspondent output label y_i forming an input-output pair. From this data distribution, one can formulate an unknown function f that maps the input to the output, as shown in Equation 3.1.

$$y = f(x) \tag{3.1}$$

During training of a supervised learning algorithm, the set, $\mathcal{D}$, with N records can be defined as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$. The training of the machine learning model aims to obtain an estimation of the mapping function, $\hat{f}$, capable of predicting a label $\hat{y}$ based on the input: $\hat{y} = \hat{f}(x)$. Depending on the type of the output data, one can identify two different supervised learning tasks. If the output is a categorical variable, the problem is designated classification, else, if the output is a real value, it is designated regression (Murphy, 2012). One of the drawbacks of supervised learning is the need for labelled data, which is an expensive task to perform and requires high usage of human resources (LeCun et al., 2015). However, when available, supervised learning methods are capable of achieving competitive and state-of-the-art results (LeCun et al., 2015).

In unsupervised learning there is no information regarding the labels. The learning process only occurs given the inputs. Therefore, during the training phase the dataset is defined as $\mathcal{D} = \{x_i\}_{i=1}^{N}$. Due to the absence of an output label that constrains the learnable mapping function $\hat{f}$, the learning process aims to retrieve the structure and representation of the data by inspection alone. The method learns a density estimation of the data, $p(x|\theta)$, where θ represents the parameters of the model. Unsupervised learning is a complex task, where one cannot define an error metric that evaluates the learning capability of the model. Therefore, the learning stage is conditioned to constrains due to two main reasons. The first is to avoid copying the input data in the inferred variables of the model. The second is to ensure that these inferred variables represent the input data with fewer features, avoiding noise and redundancy. Examples of unsupervised learning algorithms include: Principal Component Analysis (Jolliffe, 2011), Restricted Boltzmann Machines (RBMs) (Smolensky, 1986), and Autoencoders (Hinton and Zemel, 1994).

3.1.2 Classification

There are several supervised learning algorithms. However, in this research we focus on algorithms for classification tasks, which includes well-known techniques, i.e. Support Vector Machines (Cortes and Vapnik, 1995), Random Forests (Breiman, 2001), and Deep Neural Networks (LeCun et al., 1998).

In classification problems, methods learn a mapping function, $\hat{f}$, that provides an estimation $\hat{y} \in \{1, \cdots, k\}$ of a given input data, where k consists of the number of classes. When $k = 2$, the learning problem is designated a binary classification task, whereas for $k > 2$ it is designated a multi-class task.

Some methods, instead of predicting the class directly, learn a probabilistic distribution based on the input x, conditioned to the training data, $\mathcal{D}$, and the model parameters, θ, which is mathematically given by: $p(y|x;\mathcal{D},\theta)$. Only after, $\hat{f}$ computes the most probable class of a record x_i from the maximum posteriori estimation applied to the probabilistic estimation of each class, as shown in Equation 3.2 (Murphy, 2012; LeCun et al., 2015).

$$\hat{y} = \hat{f}(x) = \underset{c\in\{1,\cdot,k\}}{\text{argmax}}\, p(y = c|x) \tag{3.2}$$

Note that probabilistic classification does not predict the class directly, which is one of the major advantages of these learning methods. The capability to know the magnitude of the probabilities, provides useful knowledge on the degree of confidence of a given classifier. In addition, these classifiers can be combined with other classifiers (Murphy, 2012; LeCun et al., 2015).

3.2 Representation learning

The machine learning methods described so far do not extract features directly from the data. Hence, regardless of the learning process, these methods are highly dependent on the domain knowledge of the data. Furthermore, the discriminative power of the input data allows a higher success in performing the task at hand. Therefore, data scientists focus their effort on developing data extraction processes to obtain discriminative features. Generally, these processes are achieved through transformation and/or context representation of the source data. The development of discriminative features is called feature engineering and the extracted features are designated handcrafted features. However, it requires domain and prior knowledge, which often leads to problem-dependent features and high expertise from the data scientist (LeCun et al., 2015; Bengio et al., 2013).

Opposite to feature engineering, representation learning encompasses algorithms that can learn how to extract representations (features), directly form input data. Hence, the paradigm resides in designing the best suitable architecture to extract discriminative features. The most common approaches are based on Artificial Neural Network (ANN), which employs layers that output non-linear representations of its input. Connecting several layers leads to a network structure with the capability of learning high order and complex features, therefore having a higher level of abstraction from the input data. Due to the complexity and depth of the networks, these are designated Deep Artificial Neural Networks (LeCun et al., 2015; Goodfellow et al., 2016). Consequently, the learning process of Deep Artificial Neural Networks is commonly referred in the machine learning field as Deep Learning.

3.2.1 Artificial Neural Networks

An Artificial Neural Network is formed by a collection of small processing units, designated as nodes, linked by weighted connections in a structure of three distinct layers: input layer, X, hidden layer, H, and output layer, Y. The fundamental structure of an ANN is illustrated in Fig. 3.1 (Rosenblatt, 1958; Rummelhart and McClelland, 1986).

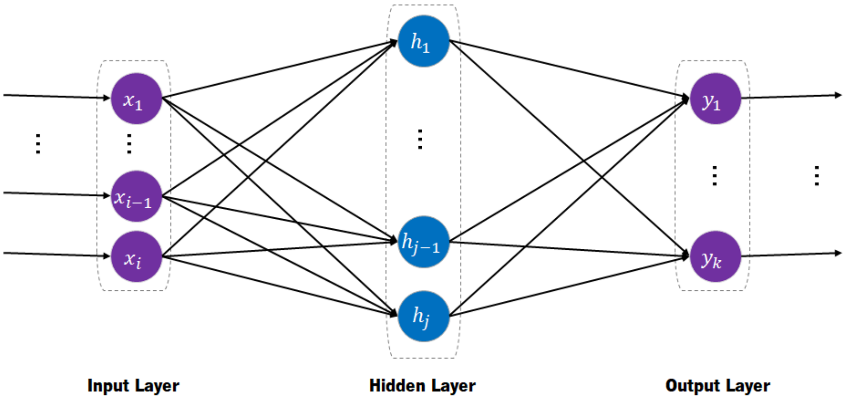


Figure 3.1: Structure of an Artificial Neural Network with one hidden layer.

Accordingly, to the connections established among nodes, two major types of ANNs can be identified: acyclic and cyclic. In the former, commonly designated feed forward neural networks, connections occur only from nodes in one layer to nodes in the next layer. When a node of a layer is connected to all nodes in the previous and following layers, the resultant structure is designated fully connected feed forward network. As for cyclic ANNs, also known as Recurrent Neural Networks (RNNs), the connections form a cycle so that the output, of a given layer, is fed back to the network, either to the same layer or previous layers (Graves, 2012).

In an Artificial Neural Network, stacking various layers turns the function f, that maps the input x to the output y, described in Equation 3.1, into a nested function, f_{NN}, as shown in the following Equation 3.3:

$$y = f_{NN}(x) \tag{3.3}$$

For demonstration purposes, let us consider a 3 layer neural network (excluding the input layer). The mapping function f is described as: $f_3(f_2(f_1(x)))$.

The training process of an ANN is constituted by a forward pass and a backward pass. The forward pass starts by presenting a pattern to the input layer, which is then propagated throughout the hidden layers until it reaches the output node. Based on the output of such pass, one can then perform a backward pass, where the parameters of the neural network are updated (Goodfellow et al., 2016). The following section provides a description of the fundamental component of ANNs, the hidden nodes, alongside the functions responsible for outputting non-linear representations of the data, the activation functions. Then, Section 3.2.3 details the learning process of ANNs and how they learn by changing the weight of connections, allowing the learning of higher representation levels of the input, and lastly, we delve on the factors that enabled Deep Learning algorithms.

3.2.2 Hidden Layers and Activation Functions

The hidden layer is the key factor that enabled ANNs to extract complex and abstract features. Consequently, having higher levels of abstraction from the input data allows the extraction of complex and discriminative features, making possible a better separability among classes. Thus, each node of the hidden layer receives a signal from nodes of the previous layer, or vector of inputs, computes an affine transformation, and then transform it non-linearly, using an activation function. Usually, the latter operation serves as distinction property among several hidden layers. For a given layer, l, the two operations, that occur along all hidden nodes, n, can be mathematically described by (Goodfellow et al., 2016; LeCun et al., 2015):

$$z_{h_l} \overset{\text{def}}{=} \mathcal{W}_l h_{l-1} + b_l \tag{3.4}$$

$$h_l \overset{\text{def}}{=} \phi_l(z) \tag{3.5}$$

In Equation 3.4, $\mathcal{W}_l$ denotes the weight matrix, that defines the connections between the nodes in the previous layer $n_{h_{l-1}}$ and the nodes in the current one n_{h_l}, hence, $\mathcal{W}_l \in \mathbb{R}^{n_{h_{l-1}} \times n_{h_l}}$. As for b_l, it denotes the bias vector, where $b_l \in \mathbb{R}^{n_{h_l} \times 1}$. The obtained result, z, is called a pre-activation, having as input vector the previous layer h_{l-1}, multiplied by the weighted connection, and influenced by the bias. The final output, h_l, in Equation 3.5 results from an element-wise non-linear activation function ϕ, which also outputs a vector (Goodfellow et al., 2016).

Non-linear activations allow the ANN to approximate non-linear functions. Otherwise, if all the activations were linear the ANN would be a chain of linear functions, since $W_l h_{l-1} + b_l$ is linear, and a linear function of a linear function is also linear. There are several options for activation functions. Nonetheless, all of them must fulfil the requirement of being differentiable, so that the optimization of the ANN is computable, more specifically such that it can find the best set of weights of each layer for the task at hand. The most popular activation functions are the sigmoid function (Equation 3.6), the hyperbolic tangent (TanH) (Equation 3.7), the rectified linear unit (ReLU) (Equation 3.8), and one of its most used variants – the Leaky ReLU (Equation 3.9) (Goodfellow et al., 2016).

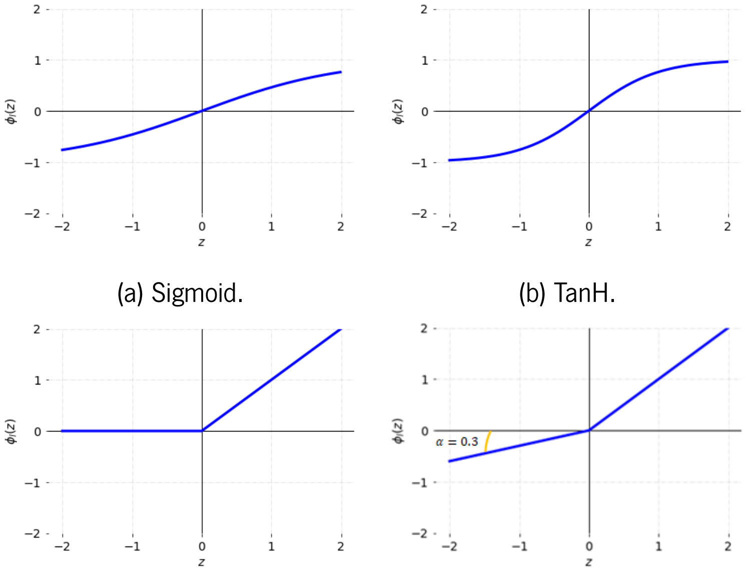
$$\phi_l(z) = \frac{1}{1 + e^{-z_l}} \tag{3.6}$$

$$\phi_l(z) = \frac{e^{z_l} - e^{-z_l}}{e^{z_l} + e^{-z_l}} \tag{3.7}$$

$$\phi_l(z) = \text{max}(0, z_l) \tag{3.8}$$

$$\phi_l(z) = \text{max}(0, z_l) + \alpha\ \text{min}(0, z_l) \tag{3.9}$$

From Equation 3.9, the Leaky ReLU consists of expanding the ReLU activation function for the negative domain, which is controlled by a parameter α. When α is learned, the Leaky ReLU is designated Parametric ReLU (PReLU). Notice that for the specific case of the ReLU activation function, in Equation 3.8, it is not differentiable at the 0 value, which is a peculiar activation function that goes against the rule that all the functions must be differentiable in an ANN. In fact, to assure the optimization of an ANN, the activation functions need to be differentiable in its domain or in a majority of its domain. Fig. 3.2 illustrates the output of the referred activation functions between an input range of $[-2, 2]$.



(a) Sigmoid.

(b) TanH.

(c) ReLU.

(d) Leaky ReLU.

Figure 3.2: Most commonly used non-linear activation functions. The Leaky ReLU for the negative portion of its input as a learnable parameter, α, which is learned during training. For demonstration purposes α was set to 0.3.

The sigmoid function, besides being used as activation function, can be employed as a final layer to output a probabilistic distribution, in binary classification problems. However, for multi-class problems, where there is the need of one output for each class, a specific activation function should be employed, namely the softmax function. The softmax function normalizes the input across the output of all output nodes as shown in Equation 3.10 (Goodfellow et al., 2016).

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}, \tag{3.10}$$

In Equation 3.10, i denotes the index of each node, z in the softmax, with a total number of nodes K, which corresponds to the total number of classes. By considering all the output nodes, each component is within the interval of $[0, 1]$ corresponding to a probability of the input belonging to the class of index

i. The softmax function can be viewed as a generalization of the sigmoid function for multidimensional outputs (Goodfellow et al., 2016).

3.2.3 Training neural networks

To perform updates to its weights, each node of an ANN needs to be differentiable to minimize the magnitude of an error. The error is computed by a cost function or loss function, which measures the difference between the prediction and the target. Hence, in each node the goal is to update its weights to minimize the loss function. This minimization problem is solved by employing gradient descent optimization algorithms. Gradient descent computes the partial derivative of the loss function with respect to the weights of a node. From the magnitude of the partial derivative, the weights of the node are updated in the direction of the negative of the gradient. However, due to the presence of non-linear operations in neural networks, the optimization criterion can become highly non-convex, with several local minima. The optimization of these networks is an intricate task to perform, generally trained in an iterative process designated as gradient descent optimization. Contrarily to logistic regression and SVMs, where the optimization algorithms are convex, in gradient descent there is no global assurance of global minimum (Goodfellow et al., 2016). Nonetheless, as the network grows deeper, higher are the similarities among local minima, providing a robust and competitive trained neural network (Choromanska et al., 2015).

3.2.3.1 Loss function

The loss function relates the task at hand with the weights of an ANN, for example in classification or regression. From this relation a penalty value is computed for wrongly learned weights that need to be updated properly during training of an ANN. There are a variety of loss functions that can be employed.

In classification tasks, the most common loss functions are the categorical cross-entropy loss and the soft-dice loss, shown in the Equation 3.11 and Equation 3.12, respectively (Milletari et al., 2016; Goodfellow et al., 2016).

$$\mathcal{L}(y, \hat{y}) = -\sum_{k=1}^{K} y_k \log(\hat{y}_k) \tag{3.11}$$

$$\mathcal{L}(y, \hat{y}) = 1 - 2 \times \frac{\sum_k (y_k \times \hat{y}_k)}{\sum_k (y_k^2 + \hat{y}_k^2)} \tag{3.12}$$

In both equations, k denotes the class index, while $\hat{y}$ and y designates the predicted class by the model, and the true class, respectively. The categorical cross-entropy loss has also a binary variant, designated binary cross-entropy loss, for $k = 2$. There are several other loss functions that can be used in classification, but are out of the scope of this thesis.

The data scientists are responsible for selecting and studying the best suited loss function for the learning task at hand. Categorical cross entropy penalizes heavily high probability values attributed to wrong classes (Goodfellow et al., 2016). However, for datasets where the data is severely unbalanced the penalization factor can lead to a non-trainable network. In an unbalanced dataset one commonly used loss function is the soft dice. Additionally, for the majority of segmentation problems, the soft dice loss has a direct correlation between the optimization and the performance metric (Dice score) (Milletari et al., 2016).

Another approach to deal with class imbalance can be achieved by an asymmetrical loss function, for example by weighting classes when computing the categorical cross entropy (Goodfellow et al., 2016). Another example that takes advantage of the particularities of the categorical cross entropy and of the soft dice loss is the one proposed by Taghanaki et al. (2019), the combo loss. The combo loss results from a weighted sum between the categorical cross entropy and the soft dice, as shown in Equation 3.13.

$$\begin{aligned}\mathcal{L}(y,\hat{y}) = &\alpha\left(2\frac{\sum_k(y_k \times \hat{y}_k)}{\sum_k(y_k^2+\hat{y}_k^2)}\right)\\ &+(1-\alpha)\times\left(\frac{1}{K}\sum_{k=1}^{K}\beta(y_k - \text{ln}(\hat{y}_k)+(1-\beta)[(1-y_k)\,\text{ln}(1-\hat{y}_k)]\right)\end{aligned} \tag{3.13}$$

In Equation 3.13, the term α controls the contribution between the categorical cross-entropy and the soft dice, whereas the term β controls the influence of the false positives and false negatives. Values of β bellow 0.5 promotes a higher importance of the false positives when computing the categorical cross entropy. When β is higher than 0.5, the presence of false negatives results in a higher penalization by the loss function.

3.2.3.2 Back-propagation

Before performing the gradient descent optimization, one needs to compute the gradient itself. This step is designated back-propagation or backward pass. Back-propagation starts by computing a scalar error of prediction, the loss function $\mathcal{L}$. After, the partial derivative of the loss function in relation to the output of the network, o, is computed as shown in Equation 3.14.

$$\nabla(\mathcal{L},\hat{y}) = \frac{\partial\mathcal{L}}{\partial\hat{y}} = \frac{\partial\mathcal{L}}{\partial o}\cdot\frac{\partial o}{\partial\hat{y}} \tag{3.14}$$

Then, it is possible to compute the derivatives of the loss function with respect to the weights going from the deepest layer to the shallower. By obtaining the contribution of each weight, a gradient, in relation to the loss function, it is possible to update the weights with certain magnitude and direction that minimizes the error of the loss function (Rummelhart and McClelland, 1986; Goodfellow et al., 2016).

Let us consider the previous example of an ANN with three layers. In this nested computational graph function f_3 follows f_2, which in turn follows f_1. The chain rule for this composition function, or nested function, is shown in Equation 3.15 (Goodfellow et al., 2016):

$$\frac{\partial f_3(f_2(f_1(x)))}{\partial x} = \frac{\partial f_3(f_2(f_1(x)))}{\partial f_2(f_1(x))}\cdot\frac{\partial f_2(f_1(x))}{\partial f_1(x)}\cdot\frac{\partial f_1(x)}{\partial x} \tag{3.15}$$

Equation 3.15 is designated the univariate chain rule. However, in a set of input nodes, U, the function f can also be applied by relating each input node-specific partial derivatives with its output, being designated the multivariate chain rule shown in Equation 3.16 (Goodfellow et al., 2016).

$$\frac{\partial f(f_1(x), \ldots, f_U(x))}{\partial x} = \sum_{i=1}^{U} \frac{\partial f(f_1(x), \ldots, f_U(x))}{\partial f_i(x)} \cdot \frac{\partial f_i(x)}{\partial x} \tag{3.16}$$

Reminding that ANNs are a composition of nested functions, the backpropagation consists of successive applications of the multivariate chain rule of differential calculus. Therefore, the gradient of an ANN, with respect to its trainable weights, is given by Equation 3.17 (Goodfellow et al., 2016).

$$\nabla(h_l, o) = \frac{\partial \mathcal{L}}{\partial h_l} = \sum_{h_{l+1}:h_l \Rightarrow h_{l+1}} \frac{\partial \mathcal{L}}{\partial h_{l+1}} \frac{\partial h_{l+1}}{\partial h_l} = \sum_{h_{l+1}:h_l \Rightarrow h_{l+1}} \frac{\partial h_{l+1}}{\partial h_l} \nabla(h_{l+1}, o) \tag{3.17}$$

Since the back-propagation occurs in the opposite direction of the forward pass, in Equation 3.17, h_{l+1} denotes the output of a hidden layer in a posterior layer, while h_l the output of a hidden layer at the current layer l, and o designates the final output of the network. To encompass all the possible connections between the node at layer l and the nodes at layer $l + 1$, the multivariate chain rule uses the notation $\sum_{h_{l+1}:h_l \Rightarrow h_{l+1}}$. Since h_{l+1} is the output of a posterior layer, the term $\nabla(h, o)$ was computed in a previous iteration. The term $\frac{\partial h_{l+1}}{\partial h_l}$ is obtained by applying the chain rule to Equation 3.17, which relates the output of a layer with respect to its previous nodes, as shown in Equation 3.18 (Goodfellow et al., 2016).

$$\frac{\partial h_{l+1}}{\partial h_l} = \frac{\partial h_{l+1}}{\partial z_{h_{l+1}}} \cdot \frac{\partial z_{h_{l+1}}}{\partial h_l} = \frac{\partial \phi(z_{h_{l+1}})}{\partial z_{h_{l+1}}} \cdot \mathcal{W}_{(h_l, h_{l+1})} = \phi'(z_{h_{l+1}}) \cdot \mathcal{W}_{(h_l, h_{l+1})} \tag{3.18}$$

The weight of the connection between the nodes h_{l+1} and h_l, is denoted by $\mathcal{W}_{(h_{l+1}, h_l)}$. This weight matrix is the target of the gradient descent algorithms. Hence, combining Equation 3.18 and Equation 3.17 we obtain the following:

$$\nabla(h_l, o) = \sum_{h_{l+1}:h_l \Rightarrow h_{l+1}} \phi'(z_{h_{l+1}}) \cdot \mathcal{W}_{(h_l, h_{l+1})} \cdot \nabla(h_{l+1}, o) \tag{3.19}$$

Hence, the back-propagation algorithm can be characterised by four major steps:

- From the forward pass, compute the output o, and the loss $\mathcal{L}$ in regard to y;
- Compute $\frac{\partial \mathcal{L}}{\partial o}$
- Use Equation 3.18 to compute each $\nabla(h_l, o)$ and obtain a gradient with respect to each weighted connection.
- From the computed gradients perform updates of the parameters that optimizes the loss function.

The latter step will be discussed in the following section.

3.2.3.3 Gradient descent optimization

Optimization of a loss function in a neural network differs from classical optimization problems. In classical optimization problems the goal is to achieve the best possible value for a given problem, depending on its objective. However, in a neural network, the loss function generally does not represent directly the goal of optimization. Particularly, in classification problems, the update of the weights are performed to achieve higher performance metrics on the validation set, without losing the capacity of generalization, observed on the test data (Goodfellow et al., 2016).

ANNs fall in the group of machine learning algorithms, where the learning occurs via an optimization process over a loss function. Gradient descent algorithms aim to minimize the loss function magnitude by updating the parameters in agreement with the negative direction of the gradient. There are several gradient descent strategies that can be employed to train a neural network (Goodfellow et al., 2016). The following sections will address the stochastic gradient descent, the momentum-based learning and the Adam algorithm.

Stochastic gradient descent Gradient descent is performed in regard to the weights of an ANN. However, updating all the weights of a given network on the totality of input records is an impractical task. Computing the gradient in a single step, that considers all data, has huge memory requirements since the backward pass needs to store intermediate and final outputs for each input to compute the gradient descent. Moreover, as the complexity and deepness of the network increases the demand of computational resources is higher (Goodfellow et al., 2016).

To better understand how to surpass such problem, let us first consider the following hypothesis. After computing the first forward pass, the following backward pass will often have a high magnitude of partial derivatives, which means that the weights are incorrect to a level that even a small sample of data points can be enough to provide a good estimation of the gradient, $\hat{g}$, and its direction. The concept of having a small data sample led to the appearance of the designation mini-batch, and the respective optimization method called mini-batch stochastic gradient descent method. Mini-batch stochastic gradient descent method performs an iterative process where small subsets of samples are used to perform an estimation of the gradient, as shown in Equation 3.20 (Goodfellow et al., 2016).

$$\hat{g}^u = \frac{1}{m}\nabla_\theta \sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i, \theta), \tag{3.20}$$

where m denotes the cardinality of the mini-batch, u indexes the update number and θ represents the whole set of parameters of the network. When all the the mini-batches of the training set are covered, without repetition, the gradient descent algorithm finishes an iteration, designated as epoch. During each update number u, the parameters of the network are updated as shown in Equation 3.21 (Goodfellow et al., 2016):

$$\theta^{u+1} = \theta^u - \epsilon\hat{g}^u \tag{3.21}$$

In Equation 3.21, ϵ designates the learning rate. This hyper-parameter is one of the most important hyper-parameters when training ANN. Ultimately, it controls the magnitude of the updates on the weights of the network. If set too high, the learning might be unstable, and in some situations the iterations may overshoot the loss function and lose the minima. On the contrary, when ϵ is low, the learning process is slow, and might be stalled at a minimum. Usually, the best compromise to avoid both behaviours is achieved by scheduling the learning rate, which means it is set to high values at initial epochs and decreases as the epochs increase (Goodfellow et al., 2016). Fig. 3.3 illustrates the described behaviour, which can occur in any gradient descent optimization.

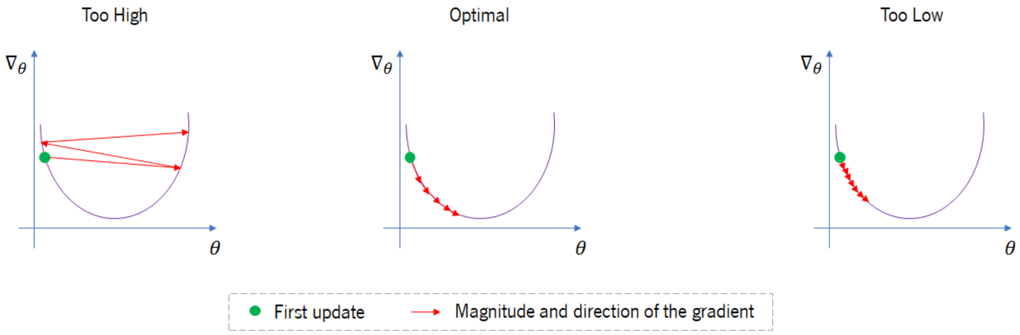


Figure 3.3: Three case examples of a gradient descent optimization, with different learning rate parameters.

Momentum Momentum based algorithms focus on speeding the convergence of the updates by performing consistent gradient optimizations in the same direction (Polyak, 1964). These optimization algorithms avoid contradictory updates that cancel one another, and can consequently lead to a gradient descent trapped on a local minimum, or slow down the effective size of an update when the loss landscape has low curvature. To accomplish this, a velocity term, v, stores the gradients of the previous iterations to drive the training towards a desirable direction. Such parameter is controlled by the momentum parameter $\beta \in [0, 1]$ as shown in Equation 3.22 (Goodfellow et al., 2016).

$$v^{u+1} = \beta v^u - \epsilon \hat{g}^u \tag{3.22}$$

$$\theta^{u+1} = \theta^u + v^{u+1} \tag{3.23}$$

The momentum parameter, β, when settled to 0 “brakes” the velocity term, and the gradient descent algorithm presents its normal behaviour. High values of β help the model to learn faster, since the velocity terms allows the optimization to occur without great oscillations. To avoid overshooting in the initial iterations, the value β may be schedule to be higher as the iterations increase and the optimal solution is closer. Once again the parameter, ϵ designates the learning rate (Goodfellow et al., 2016).

A traditional modification of Momentum was proposed by Sutskever et al. (2013), called the Nesterov Momentum. Nesterov Momentum allows a faster and less oscillatory convergence since it incorporates the Momentum when computing the gradient $\hat{g}^u$. This is achieved by computing how the momentum influences the gradient before an update, as shown in Equation 3.24.

$$\hat{g}^u = \frac{1}{m}\nabla_\theta \sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i, \theta + \beta v^u) \tag{3.24}$$

Adam Adam optimization algorithm belongs to a class of adaptive learning rate algorithms, which means that the learning rate is adapted for the different parameters of the network (Goodfellow et al., 2016). Adam algorithm was designed specifically for neural networks, being extremely popular since it incorporates most of the advantages of other algorithms, such as momentum characteristics and exponential gradient smoothing to avoid overshooting. The first step of the Adam algorithm starts by computing moving averages, μ and υ, at two different scales, both conditioned by a hyper-parameter γ, as shown in Equation 3.25 and Equation 3.26 (Kingma and Ba, 2014).

$$\mu^u = \gamma_1 \mu^{u-1} + (1-\gamma_1)\hat{g}^u = (1-\gamma_1)\sum_{i=0}^{u} \gamma_1^{u-i}\hat{g}^u \tag{3.25}$$

$$\upsilon^u = \gamma_2 \upsilon^{u-1} + (1-\gamma_2)(\hat{g}^u)^2 = (1-\gamma_2)\sum_{i=0}^{u} \gamma_2^{u-i}(\hat{g}^u)^2 \tag{3.26}$$

Note however, that by using moving averages, during learning the first moment estimates have a bias. To cope with this problem, after computing μ and υ one must perform a bias correction, given as follows (Goodfellow et al., 2016):

$$\hat{\mu}^u = \frac{\mu^u}{1-\gamma_1^u} \tag{3.27}$$

$$\hat{\upsilon}^u = \frac{\upsilon^u}{1-\gamma_2^u} \tag{3.28}$$

The last step of the Adam optimization considers μ and υ, constrained to the learning rate of each parameter, as shown in Equation 3.29 (Goodfellow et al., 2016).

$$\mathcal{W}_j^u = \mathcal{W}_j^{u-1} - \epsilon \frac{\hat{\mu}^u}{\sqrt{\hat{\upsilon}^u} + \delta} \tag{3.29}$$

The term δ, in the previous equation, is a fixed constant for stability purposes, and the term $\mathcal{W}_j^u$ designates the weights of node j of the network at an update step u (Goodfellow et al., 2016).

The main purpose of Adam algorithm is that the parameters with large partial derivatives are more prone to oscillations. On the contrary, parameters with smaller partial derivatives tend to have consistent updates that point slightly in the same direction (Goodfellow et al., 2016; Kingma and Ba, 2014).

3.2.4 Deep Learning

ANNs started to be developed in 1958 (Rosenblatt, 1958) and were explored during the 1980s (Rummelhart and McClelland, 1986). However, training an ANN with a high number of hidden layers was an impractical task due to two major problems. The exploding and vanishing gradient that can occur during the optimization process. Once identified, the exploding gradient problem was mitigated by restricting the magnitude of the backpropagation through gradient clipping, or by employing regularization such as $L1$ and $L2$ norm. As for the vanishing gradient problem, it remained unsolved for several years (Goodfellow et al., 2016).

The vanishing gradient problem arises when the gradient computed by the algorithm responsible for updating the weights of each layer has a magnitude tending to zero (vanishing). This leads to a network where some learnable weights are not optimized or, in the worst-case scenario, the whole neural network stops learning. Consider, for example, an ANN with l layers all using a TanH activation function. In these conditions, the gradients have a magnitude in the range of $[0, 1]$. When performing the backpropagation algorithm by the chain rule, the effect of multiplying each small ranged value l times, decreases the gradient magnitude exponentially. Thus, if the signal gradient that reaches the top layers is near zero, the weights practically do not change (Glorot and Bengio, 2010; Goodfellow et al., 2016).

Hinton et al. (2006) proposed the first approach that successfully mitigated both vanishing and exploding gradients. A deep model was achieved by stacking Restricted Boltzmann Machines, which were trained in an unsupervised way, followed by a fine-tuning step that employed supervised training backpropagation. The first learning phase was designated as pre-training. Pre-training allowed a first initialization of weight parameters among the layers of RBMs, before being conditioned to the supervised learning step. Therefore, the proper weight initialization became a gold standard for training deep networks and overcome vanishing gradients. These findings turned the focus of machine learning research to study effective ways to train deep networks through proper weight initialization. The proposals, first of Glorot and Bengio (2010) and then of He et al. (2015), allowed the training of deep networks from scratch, avoiding a previous time-consuming step of pre-training.

Alongside the referred improvements, other factors played an important role in improving the train of deep ANNs. One of them was the activation function, such as the ReLU and Leaky ReLU. These activation functions overcame the saturation problem that can occur with the TanH and the sigmoid activation functions, hindering the training (Glorot et al., 2011; Maas et al., 2013). Also, techniques such as skip connections in residual neural networks (He et al., 2016), allowed the training of even deeper networks since it propagates the gradient from deeper layers to earlier ones. Other factors were the advances on the gradient descent algorithms (Kingma and Ba, 2014) and more efficient regularization procedures (Srivastava et al., 2014; Tompson et al., 2015).

Krizhevsky et al. (2012) was able to take advantage of previous advances and propose a deep neural network with convolutional layers that won the ImageNet competition in 2012. In a competition where most of the machine learning algorithms were based on SVMs, this was an important mark in the history of Machine Learning.

In this thesis work, for stroke tissue outcome prediction, three representation learning methods are employed: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Restricted Boltzmann Machine (RBM). All of them will be described in the following sections.

3.3 Convolutional Neural Networks

Convolutional Neural Networks allow a hierarchical representation of the data as the depth increases, granting higher levels of abstraction (LeCun et al., 1998). This property potentiated the development of very deep neural networks (LeCun et al., 2015). CNNs aim to intuitively summarize spatial relationships, through convolution operations, making them inherently designed for grid-structured data that contains strong spatial correlations in local regions of the grid. Hence, CNNs have as the most obvious application images. In fact, the majority of applications of CNNs is for image data, but one can still find applications of these networks in temporal and spatio-temporal data (Grefenstette and Blunsom, 2014; Karpathy et al., 2014). However, image data shows an important characteristic, that to a certain degree does not hold for other kinds of grid-structured data – translation invariance; meaning that similar patterns can be present in different locations of an image but still being identified as equal (LeCun et al., 1998).

The convolution layer present in the CNN consists of a dot-product between a small grid of the input data and a set of learnable parameters with the same dimensions. This allows the capturing of information from a local neighbourhood of a pixel and search for correlations in regard to it. The output of the convolution layer is generated when the set of learnable parameters goes through all possible spatial overlaps of the grid-structure input data (Goodfellow et al., 2016).

3.3.1 Fundamental structure

The CNN description given in this section will be based on its most widely known application: 2D images. In an image the elements at a given spatial location are designated as picture elements, or pixels. However, one can apply the same knowledge for many other grid-structured data.

The vanilla CNN receives as input a 3-dimensional array of data, characterising the height, H, width W, and channels, ch, of the input data. The terms H and W index all the available grid points of the input, corresponding to spatial information, whereas ch characterises independent properties along the spatial location. Given an input of a layer l, from the output of a previous layer $l-1$, of size $H_{l-1} \times W_{l-1} \times ch_{l-1}$, the parameters of the layer l are organized in a 3-dimensional array, $k_H \times k_W \times k_{ch}$, known as filter or kernel, k. Commonly, the kernel has a squared configuration $k_H = k_W$ in the spatial dimensions, with a much smaller size when compared to the input data. To perform the convolution operation, the third dimension of the kernel, k_{ch}, must be equal to the input channels dimension (Goodfellow et al., 2016).

The kernel is placed at each possible spatial position, that has a full overlap with the input data, and the output results by performing the dot product between the kernel parameters and the matching grid of the input with the same size. The total number of possible operations over the input, defines the height and width of the output layer, and consequently the input size of the following layer. Formally, at a layer l the output height and width are given respectively as follows (Goodfellow et al., 2016):

$$H_l = H_{l-1} - k_H + 1 \tag{3.30}$$

$$W_l = W_{l-1} - k_W + 1 \tag{3.31}$$

Hence, the dimensions H_l and W_l will be smaller, when compared to the input. This behaviour is explained by the fact that only full overlaps between the kernel and the data are considered. Since in the borders the convolution operation does not take place, it leads to a dimension reduction of the image. This characteristic is known as receptive field or field of view. In order to deal with this loss of information at the borders, two different strategies can be employed. One that simply discards the information at the borders. The other that performs a padding in accordance to the kernel size, granting the same output spatial dimensions. In the former, the number of convolutional layers may be limited due to a continuous reduction of the image size. Whereas for the latter, due to the image size being maintained, the number of convolutions is not limited, but due to artificial patterns placed near the borders it can influence the learning process (Goodfellow et al., 2016).

Each kernel is responsible for outputting a spatially arranged output called feature map. Consequently, higher the number of kernels will result in a higher number of feature maps, and higher parameter footprint given by: $k_{H_l} \times k_{W_l} \times ch_l \times ch_{l-1} + b_l$. The term b_l designates the bias of each new output feature map. Each filter in the convolutional neural network aims to identify patterns within the kernel size, thus with a higher number of features, the capacity of the model increases as it is possible to characterise many kinds of patterns. However, increasing the number of kernels can lead to redundant features and over-fitting of the model to the training data (Goodfellow et al., 2016).

After describing the convolution operation and its different aspects, we can now define formally a convolutional layer, as shown in Equation 3.32.

$$M_k = \sum_{ch=1}^{n_{ch}} I_{ch} * \mathcal{W}_{k,ch} + b_k \tag{3.32}$$

In Equation 3.32, I denotes the set of input channels indexed by ch, $\mathcal{W}_{k,ch}$ the weight matrix of the kernel, and b_k a bias which is summed element-wise. Therefore, M_k results from the convolution operation, $*$, of all input maps by the kernel k. In a $2D$ grid-structure the output feature map M of a kernel k is shown in Equation 3.33.

$$M_{k,x,y} = \sum_{ch=1}^{n_{ch}} \sum_{m=1}^{k_H} \sum_{n=1}^{k_W} I_{ch,(x-1)\times s+m,(y-1)\times s+n} \mathcal{W}_{k,ch,m,n}, \tag{3.33}$$

where $s \in \mathbb{N}^+$ designates the stride, which can be distinct for each spatial dimension, but in the above equation was the same for both. When $s = 1$ the convolution is performed on all possible inputs, whereas for $s > 1$ the convolution jumps s pixels from the previous location. By performing a smaller number of convolutions, the output feature map is down-sampled by the factor s (Goodfellow et al., 2016). Fig. 3.4 illustrates a practical example of a CNN with two kernels of size 3×3 applied with stride of 1.

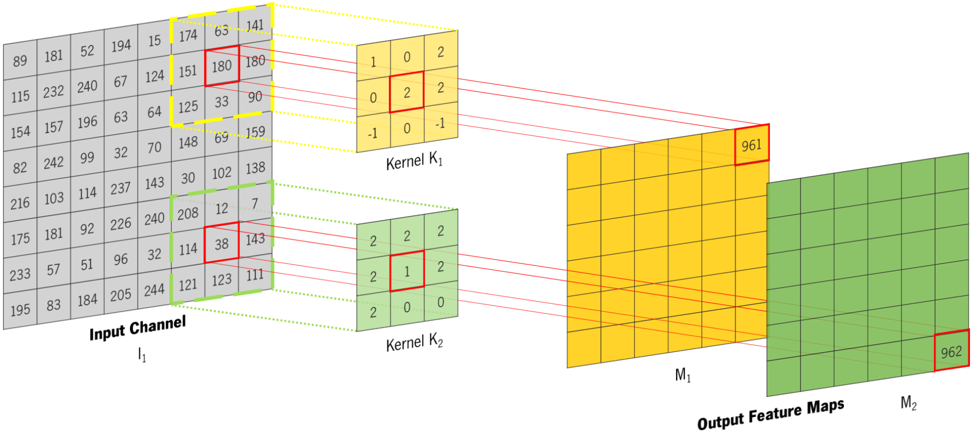


Figure 3.4: Convolution example applied for a single input channel, outputting two feature maps from two kernels k_1 and k_2. The bias was not considered for sake of simplicity.

The final step of a CNN layer encompasses an element-wise non-linear activation function applied to each output feature map.

Advantages of CNNs

CNNs have three major properties that made them widely popular in the computer vision field: translation invariance, parameter sharing and sparse connectivity.

In the convolution layer, the same kernel is convolved over the input. Therefore, since the kernel parameters are the same for a given output feature map, the detected patterns will be the same regardless of their spatial location in the image. This property, designated by translation invariance, is of great importance when we are dealing with grid-structured data such as images where the same patterns might appear in different locations. Nonetheless, affine transformations such as flipping, rotations or scaling, that alter the pattern in the grid-structured data, might produce different outputs (Goodfellow et al., 2016).

Parameter sharing was intuitively described in the translation invariance property. To generate the output feature map, the parameters of a kernel are applied to the whole input data, i.e. each point of the output feature map is then computed with the same parameters. Thus, it avoids the need for many parameters to extract high levels of abstraction (Goodfellow et al., 2016).

Finally, the property of convolutional layers that avoids an escalation of the number of parameters is the sparse connectivity. CNNs do not possess full connectivity with all the input data, but only with a small region of it, controlled by the kernel size. Such property is designated sparse connectivity, being the output node dependent on a small spatial grid of the input nodes. On the contrary, in fully connected layers, all the output nodes are connected to all the input nodes, leading to a higher parameter footprint (Goodfellow et al., 2016).

3.3.2 Pooling

The pooling layer is often seen in architectures alongside convolutional layers, being performed independently for each feature map, therefore maintaining the total number of feature maps. The pooling operation does not increase the number of trainable parameters and has low computational costs. In a pooling layer, the first step starts by defining a grid of size $(p_h \times p_w)$, which will then return the values that fulfil an operation performed all over the input image (Goodfellow et al., 2016).

Similarly, to convolutional layers, pooling layers can be performed with a given stride $s \geq 1$. Normally, the default value of the stride is equal to $(p_h \times p_w)$. Conditioned to the defined grid where the operation function is applied, the spatial dimension of the input can be reduced drastically. Reducing the spatial dimension of the input, pooling layers allow the following convolutional layers to consider patterns from distant regions of an image, therefore leading to an increase of the receptive field. Similarly, to the convolutional operation, the pooling operation grants a certain amount of translation invariance, since by shifting slightly the input, the output will barely change (Goodfellow et al., 2016).

One commonly used pooling layer is the max-pooling, which retrieves the maximum values within a grid $(p_h \times p_w)$. Fig. 3.5 illustrates this particular pooling layer.

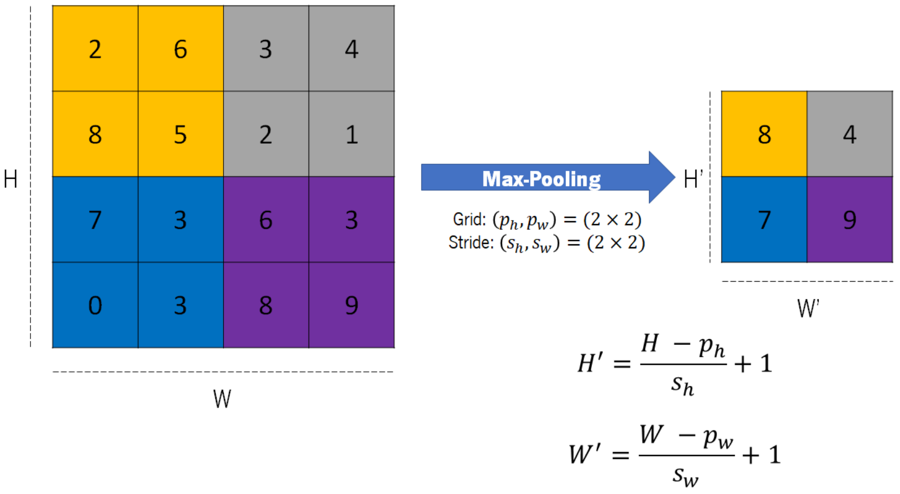

H
2 6 3 4
8 5 2 1
7 3 6 3
0 3 8 9
W
Max-Pooling
Grid: $(p_h, p_w) = (2 \times 2)$
Stride: $(s_h, s_w) = (2 \times 2)$
H'
8 4
7 9
W'
$H' = \frac{H - p_h}{s_h} + 1$
$W' = \frac{W - p_w}{s_w} + 1$

Figure 3.5: Max-pooling example applied to an image, with a grid size of 2×2, and with a stride of 2×2.

Max-pooling is frequently used alongside convolutional layers, but there are other operations of aggregation that can be applied to the spatial domain such as: average, and global pooling (Lin et al., 2013). Another approach to increase the receptive field without pooling operations is achieved by applying convolutions with a stride larger than 1 (He et al., 2016). However, these proposals are not largely employed since max-pooling still provides higher non-linearity and capacity of being invariant to translation, without increasing the network parameters.

3.4 Recurrent Neural Networks

Recurrent Neural Networks are naturally designed for data structures with sequential dependencies such as time-series and text. Nonetheless, RNNs can also be applied to grid-structured data, such as images (Visin et al., 2015, 2016) and also to video data (Xingjian et al., 2015).

In feed-forward neural networks, due to their connections, the flow of information occurs in one direction. However, for recurrent neural networks the presence of cyclical connections allows the flow of information in different directions. Furthermore, cyclical connections lead to the consecutive application of the function computed at a previous state. This property, despite requiring high computational resources, allows RNNs to be independent of the dimensions of the input, such as the case of sequential data. When dealing with sequential data, feed-forward neural networks are limited to a fixed size of the input data, since the input nodes of the network are fixed and, consequently, considering longer input sizes is not computable. In RNNs for each position of the input data there is a correspondent state of the network, identified by a time-step. Since it is possible to compute a state for each position of the input data (i.e. time-step), RNNs can process sequential data regardless of their size. It is thus verified that RNNs need higher computational resources than feed-forward networks, to consider their different instances. Additionally, one can raise the question on how the parameters of the network will increase as the sequences grow longer. RNNs take advantage of a property called parameter sharing, different as described for CNNs. The RNN shares the weights across the time-steps to allow a similar modelling for all the sequential data. This results in a deep computational graph with a fixed number of parameters. In RNNs, parameter sharing has a higher impact, when compared to CNNs, as it allows the network to generalize across inputs of different configuration, and at the same time respect the input data ordering when processing it, and assuring that each time-step is viewed as equally in regard to the previous steps (Graves, 2012).

RNNs are a good resource in the machine learning field. However, in the early beginnings, their learning process was difficult due to the vanishing and exploding gradient phenomena. Recent proposals based on gated recurrent neural networks, such as Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) were proposed to cope with the learning challenge of RNNs.

3.4.1 Fundamental structure

A fundamental characteristic of the feed-forward neural networks is that all nodes are connected acyclically. In the case of recurrent neural networks this is not verified due to the presence of cyclical connections forming a recurrent neural network. Common examples of cyclical connections in recurrent neural networks emerge when the recurrence occurs from hidden-to-input nodes, from hidden-to-hidden nodes or from output-to-hidden nodes (Elman, 1990; Lang et al., 1990; Jaeger, 2001).

This section focus on a simple one node RNN, where the recurrence occurs from hidden-to-hidden, i.e. in a self-loop configuration, as illustrated in Fig. 3.6 (Goodfellow et al., 2016).

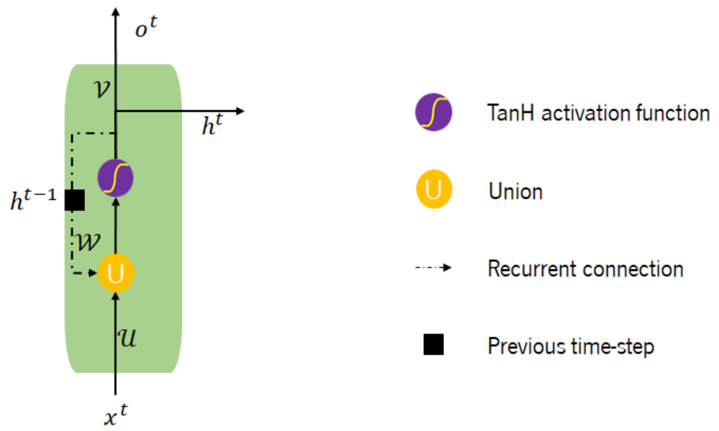


Figure 3.6: Recurrent neural network, with the recurrence applied among the hidden node. For sake of simplicity the bias was not considered.

The RNN depicted in Fig. 3.6 characterises the input-to-hidden connection with a weight matrix $\mathcal{U}$, the hidden-to-hidden recurrent connection with a weight matrix $\mathcal{W}$ and the hidden-to-output connection with a weight matrix V. The input, x contains a finite number of time-steps τ, with each time-step indexed by t (Goodfellow et al., 2016). Hence, the forward pass, at time-step t, of a RNN is formally given as follows (Goodfellow et al., 2016):

$$a^t = \mathcal{U}x^t + \mathcal{W}h^{t-1} + b \tag{3.34}$$

$$h^t = \text{TanH}(a^t) \tag{3.35}$$

$$o^t = \mathcal{V}h^t + c \tag{3.36}$$

In Equation 3.34, a^t denotes the first operation performed at time-step t, whereas in Equation 3.35 h^t denotes the output of the hidden node at time-step t that results from applying the TanH activation function to a^t. Equation 3.36 denotes the output o of the network at time-step t. The terms b and c designate bias vectors defined for each input-to-hidden, and for hidden-to-output layers, respectively.

In Fig. 3.6, the RNN maps the input sequence to an output sequence both with the same length. After applying the forward step, and computing the loss value, one can compute the backward pass. From a first perspective, the presence of the recurrent connection makes the backward pass a complex and demanding task. However, by turning a recurrent graph into an unrolled computational graph that encompasses a series of repetitive structures, it is possible to directly apply the back-propagation algorithm described in Section 3.2.3.2. The back-propagation algorithm applied to an unrolled computational graph is designated Back-Propagation Through Time (BPTT). Note however that, due to parameter sharing across time-steps, the backward pass has no chances of parallelization, being performed sequentially. Thus, RNNs are very powerful networks but also resource and time demanding when training (Goodfellow et al., 2016).

3.4.2 Unfolding and Back-propagation

Let us consider the RNN of Fig. 3.6. From the theory of classic dynamic systems, in the presence of an external signal, the hidden node of a RNN can be described as (Goodfellow et al., 2016):

$$h^t = f(h^{t-1}, x^t, \theta) \tag{3.37}$$

Equation 3.37 is recurrent since the state of the hidden node at a time t, is dependent on its state at time $t-1$. Nonetheless, for a finite number of steps, τ, these dependencies can be unfolded resulting in a non-recurrent graph. For demonstration purposes, when considering a sequential data with 3 time-steps, the unfolded computation of such steps is given as follows (Goodfellow et al., 2016):

$$h^3 = f(h^2; x^2; \theta) = f(f(h^1; x^1; \theta); x^2; \theta) \tag{3.38}$$

Equation 3.38 is no longer recurrent and can be represented by an acyclic computational graph, by repeatedly applying the operation function ϕ across all time-steps in τ, with the same parameters. Unfolding consists of transforming a recurrent graph into a computational graph that does not contain recurrence. Consequently, unfolding allows the learned model to have a fixed input size, since the length of the sequential data is encoded in number of time-steps. These two properties are the major advantages of RNNs (Goodfellow et al., 2016; Graves, 2012). Fig. 3.7 illustrates the unfolded version of the RNN in Fig. 3.6, for a sequence data with three time-steps, $t-2$, $t-1$, t.

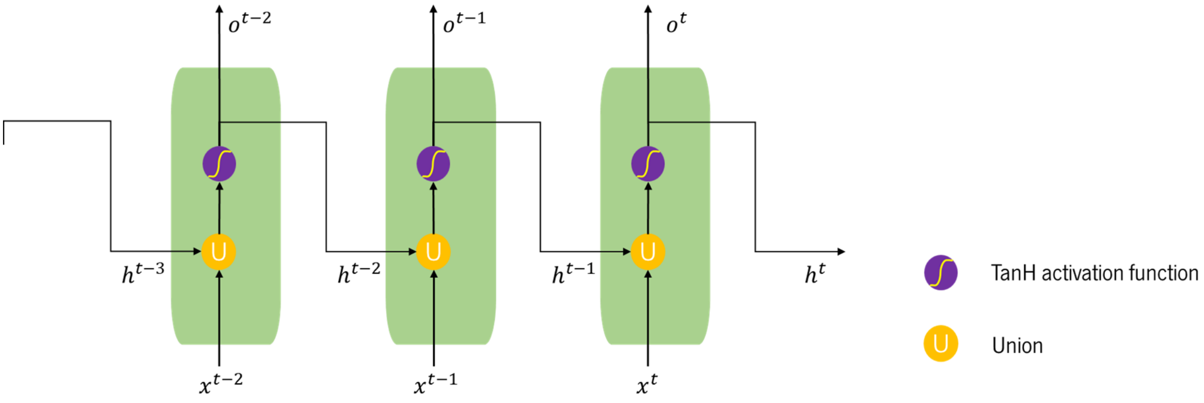


Figure 3.7: Unfolded RNN of Fig. 3.6 for three time-steps. For sake of simplicity the bias was not considered.

For RNNs, unfolding helps to understand how the forward pass, and specially the backward pass occurs. At the same time, unfolding helps to comprehend the influence of a previous output, at $t-1$, in the current time-step output t. By knowing the flow of information in a recurrent neural network, the back-propagation algorithm becomes a simpler task to perform. As a matter of fact, the backward pass of an unfolded recurrent neural network is performed accordingly to the general back-propagation algorithm. Once again, backward pass starts by computing the value of the loss function. For RNNs where the recurrence occurs as illustrated in Fig. 3.7, where there is an output at each time-step, the total value of the loss is computed as shown in Equation 3.39.

$$\mathcal{L} = \sum_{t=1}^{\tau} \mathcal{L}^t(y^t, \hat{y}^t) \tag{3.39}$$

Afterwards, the BPTT algorithm aims to compute the partial derivatives (Graves, 2012):

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathcal{V}} \\ \frac{\partial \mathcal{L}}{\partial \mathcal{W}}, \frac{\partial \mathcal{L}}{\partial b} \\ \frac{\partial \mathcal{L}}{\partial \mathcal{U}}, \frac{\partial \mathcal{L}}{\partial c} \end{cases} \tag{3.40}$$

Hence, the first computational step is given as follows (Goodfellow et al., 2016):

$$\nabla(\mathcal{V}, \mathcal{L}) = \frac{\partial \mathcal{L}}{\partial \mathcal{V}} = \sum_{t=0}^{\tau} \frac{\partial \mathcal{L}^t}{\partial \mathcal{V}} \tag{3.41}$$

In Equation 3.41 there is no recurrence, and $\frac{\partial \mathcal{L}^t}{\partial \mathcal{V}}$ only depends on the output at time-step t. Therefore, by applying the chain rule, computing $\frac{\partial \mathcal{L}^t}{\partial \mathcal{V}}$ is shown as follows (Goodfellow et al., 2016):

$$\frac{\partial \mathcal{L}^t}{\partial \mathcal{V}} = \frac{\partial \mathcal{L}^t}{\partial \hat{y}^t} \frac{\partial \hat{y}^t}{\partial \mathcal{V}} \tag{3.42}$$

The next step focus on $\nabla(\mathcal{W}, \mathcal{L})$. However, the chain rule cannot be applied, due to the fact that $\nabla(\mathcal{W}, \mathcal{L})$ is recurrent, which means that the weight matrix, $\mathcal{W}$, depends on the previous time-steps. Hence:

$$\nabla(\mathcal{W}, \mathcal{L}) = \sum_{t=0}^{\tau} \frac{\partial \mathcal{L}^t}{\partial \mathcal{W}}, \quad \text{where} \quad \frac{\partial \mathcal{L}^t}{\partial \mathcal{W}} = \frac{\partial \mathcal{L}^t}{\partial \hat{y}^t} \frac{\partial \hat{y}^t}{\partial h^t} \frac{\partial h^t}{\partial \mathcal{W}} \tag{3.43}$$

In Equation 3.43, the rightmost term is not feasible and the chain rule is not valid in the presence of recurrence. To compute $\nabla(\mathcal{W}, \mathcal{L})$ we need to expand this equation to all time-steps and use the formula of the total derivative.

$$\frac{\partial \mathcal{L}^t}{\partial \mathcal{W}} = \frac{\partial \mathcal{L}^t}{\partial \hat{y}^t} \frac{\partial \hat{y}^t}{\partial h^t} \left(\frac{\partial h^t}{\partial \mathcal{W}} + \frac{\partial h^t}{\partial h^{t-1}} \frac{\partial h^{t-1}}{\partial \mathcal{W}} + \cdots \right) \tag{3.44}$$

The last term in Equation 3.44 consists of the sum of the contribution of all the previous time-steps until time-step t. Therefore, we roll backwards in time from t to prime time-step t' as shown in Equation 3.45.

$$\sum_{t'=0}^{t} \left(\prod_{i=t'+1}^{t} \frac{\partial h^i}{\partial h^{i-1}} \right) \frac{\partial h^t}{\partial \mathcal{W}} \tag{3.45}$$

Equation 3.45 characterises the dependences between the hidden nodes and the weight matrix $\mathcal{W}$. Furthermore, for the specific case of a TanH as activation function can be further simplified as follows:

$$\prod_{i=t'+1}^{t} diag(\phi'(\cdots)\mathcal{W}^i) \tag{3.46}$$

Lastly, computing $\frac{\partial \mathcal{L}^t}{\partial u}$ falls under the same process as described for $\frac{\partial \mathcal{L}^t}{\partial w}$, since there is also recurrence, given by Equation 3.47

$$\frac{\partial \mathcal{L}^t}{\partial \mathcal{U}} = \frac{\partial \mathcal{L}^t}{\partial \hat{y}^t}\frac{\partial \hat{y}^t}{\partial h^t}\left(\sum_{t'=0}^{t}\prod_{i=t'+1}^{t}\frac{\partial h^i}{\partial h^{i-1}}\right)\frac{\partial x^t}{\partial \mathcal{U}} \tag{3.47}$$

3.4.3 Gated Recurrent Neural Networks

The earlier implementations of RNNs where known for the difficulties in learning and updating its parameters. In theory the recurrence property provides a high level context, but in practice it can influence a given input on the hidden layer, consequently making the output negligible. Thus, when performing the backward pass, the gradient tends to either vanish or explode. While the exploding gradient was mitigated by performing clipping, mitigating the vanishing gradient was challenging. Various attempts were proposed to deal with the vanishing gradient (Schmidhuber, 1992; Bengio et al., 1994; Lin et al., 1996). A widely adopted solution in deep learning is the Long-Short Term Memory (LSTM) proposed by Hochreiter and Schmidhuber (1997), which falls under the category of gated recurrent neural networks. Gated recurrent neural networks contain paths to ensure that when performing the BPTT, the partial derivatives neither vanish nor explode. Gated RNNs reveal a memory effect, which allows the network to retain information from the previous iteration, when performing a new iteration (Goodfellow et al., 2016).

The LSTM is a recurrent neural network constituted by four elements: one cell, and three gates; as shown in Fig. 3.8

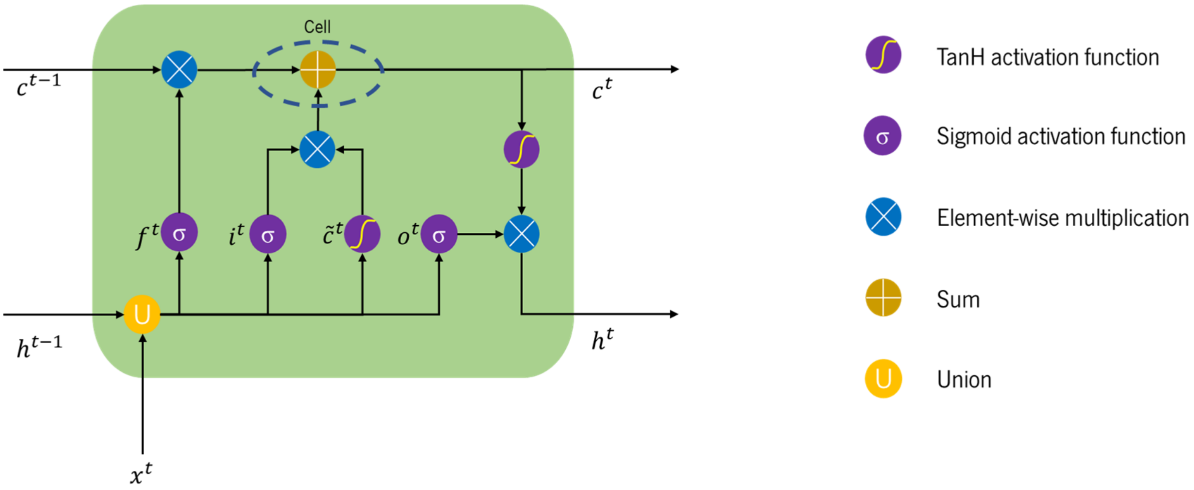


Figure 3.8: Long-Short Term Memory network at time-step t. For sake of simplicity the bias was not considered.

In Fig. 3.8, the term f^t denotes the forget gate, i^t the input modulation gate, $\tilde{c}^t$ the cell state estimation, and o^t the output gate. The cell, also called state node, is the element connected to the previous time-steps, therefore controlling the recurrence, replacing the hidden node in the vanilla configuration of the RNN. In the LSTM, the recurrent connection is controlled by the forget gate, which outputs a value between 0 and 1 (via σ function) further multiplied by the recurrent connection. Thus, the forget gate receives as input the data at time-step t, alongside the hidden state from the previous time-step $h^{(t-1)}$ as shown in Equation 3.48 (Goodfellow et al., 2016).

$$f^t = \sigma\Bigg(\mathcal{U}_f x^t + \mathcal{W}_f h^{t-1} + b_f\Bigg), \tag{3.48}$$

where b_f denotes to the biases, $\mathcal{U}_f$ input weights, and $\mathcal{W}_f$ recurrent weight of the forget gates, respectively. After computing the value of the forget gate it is possible to update the state unit, restricted to the weight of the forget gate f, as follows (Goodfellow et al., 2016):

$$c^t = f^t c^{t-1} + i^t TanH\Bigg(\mathcal{U}_c x^t + \mathcal{W}_c h^{t-1} + b_c\Bigg), \tag{3.49}$$

b_c, $\mathcal{U}_c$, and $\mathcal{W}_c$ designates biases, input weights and recurrent weight of the LSTM cell, respectively. The input gate unit is designated by i^t, and since it is a gate unit, its computation is similar to Equation 3.48 (Goodfellow et al., 2016):

$$i^t = \sigma\Bigg(\mathcal{U}_i x^t + \mathcal{W}_i h^{t-1} + b_i\Bigg) \tag{3.50}$$

The last gate of the LSTM is the output gate o^t, which follows the same principles as the other two already described gates (Goodfellow et al., 2016):

$$o^t = \sigma\Bigg(\mathcal{U}_o x^t + \mathcal{W}_o h^{t-1} + b_o\Bigg) \tag{3.51}$$

The output gate is responsible for controlling the output of the LSTM, given by:

$$h^t = TanH(c^t)o^t \tag{3.52}$$

Besides LSTM other gated recurrent variants have been proposed, such as the Gated Recurrent Unit (GRU). The GRU proposed by Cho et al. (2014a) was developed as an alternative over the Long-Short Term Memory network by reducing the number of gates and connections, which leads to a decreasing of the parameter footprint, consequently less computationally demanding. Fig. 3.9 characterises the graph of a GRU.

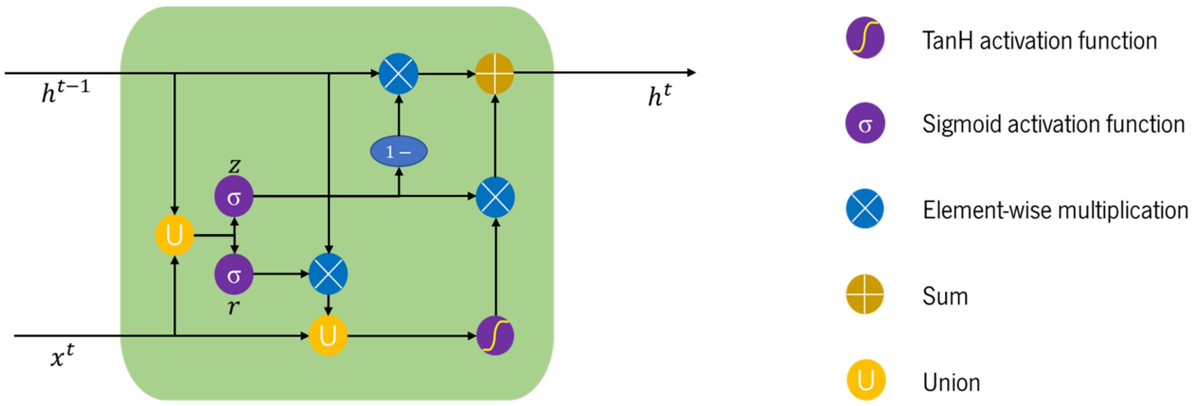


Figure 3.9: Gated Recurrent Unit Network at time-step t. For sake of simplicity the bias was not considered.

The output h^t depends on a given input x at the time-step t and the output from the previous time-step h^{t-1}. Here $\cup$ denotes the union of two weight matrices (concatenation), σ is the sigmoid activation function, $\times$ is the element-wise multiplication and, $+$ the sum. Hence, the forward pass of such recurrent layer is given in Equation 3.53.

$$\begin{cases} h^t & = (1 - z^t) \odot h^{t-1} + z^t \odot TanH(\mathcal{W}_x x^t + W_h(r^t \odot h^{t-1}) + b_h), \\ z^t & = \sigma(\mathcal{W}_z + b_z), \\ r^t & = \sigma(\mathcal{W}_r + b_r), \end{cases} \tag{3.53}$$

where $\mathcal{W}_{z,r} \in \mathbb{R}^{(d_h+d_x)\times d_h}$, $\mathcal{W}_x \in \mathbb{R}^{(d_x \times d_h)}$, $\mathcal{W}_h \in \mathbb{R}^{(d_h \times d_h)}$, and $b_{z,r,h} \in \mathbb{R}^{d_h}$ are model parameters. The mathematical operator $\odot$ designates the element wise multiplication (Cho et al., 2014a). Whereas the LSTM has three gates, the GRU contains only two gates, the reset gate r and the update gate z.

The BPTT of both gated recurrent neural networks is computed based on the same algorithm as described for the vanilla RNN, by expanding Equation 3.44 to each of the weight matrices present in each gate.

3.4.4 Gated Recurrent Networks in Computer Vision

The Gated Recurrent Network was developed for processing one-dimensional temporal data, e.g. time series. However, it has been extended to grid-structured data to provide a notion of spatial context (Stollenga et al., 2015; Tseng et al., 2017) or spatio-temporal context (Wu et al., 2016). When considering the current pixel or group of pixels, gated recurrent networks have the capability to correlate previous observations with the current one.

Wu et al. (2016) presented a video segmentation framework for person re-identification, which consisted in identifying individuals over disjoint camera views. The algorithm combined CNNs with LSTMs, where the latter considers both temporal and spatial data in an encoder-decoder network. The encoder path of this network focused on capturing the movement of a person, being the hidden representations fed to the decoder path to perform the video segmentation. To ensure that the LSTM network considers both spatial and time-dependant information, all gates were replaced by convolutional operators, designating these layers of Convolutional LSTM. However, to provide both properties the computational resources demanded are high, forcing the authors to use a reduced number of feature maps in the convolutional operations.

In the field of Biomedical Computer Vision, Stollenga et al. (2015) presented a method for several biomedical applications, such as brain segmentation from MRI images and segmentation of neuronal structures from electron microscopy, using Multi-Dimensional Recurrent ANNs. The authors employed LSTMs to take into consideration the notion of spatial context by sweeping all pixels several times. The model connects LSTMs in a grid like manner receiving information from the pixels under analysis and also from neighbouring LSTM nodes. Instead of considering the standard directions along the grid axes, the proposed method scans for spatial relationships within a pyramid structure, hence designating the method by PyraMiD-LSTM. When the context considers the standard directions along the grid axes, depending on the number of dimensions of the input, d, the number of LSTMs required is 2^d. Contrarily, the PyraMiD-LSTM due to its pyramidal context requires less LSTMs to process the data, $2 \times d$, therefore being computationally faster. In addition, the authors also demonstrated that in these biomedical applications, combining RNN blocks provides higher levels of performance instead of using isolated ones. Similarly, Kalchbrenner et al. (2015) presented the Grid-LSTM, which is a network of LSTMs that expands the views from one or more dimensions. The proposed model achieved success both in synthetic and real data. In addition, the architecture ensures no dependency in what concerns the short term memory size and its parameters.

Spatio-temporal context was also explored in the biomedical field, through the usage of Convolutional-LSTMs. Proposed for precipitation nowcasting, the Convolutional-LSTM replaces the operations of the LSTM computed at each gate by convolutions (Xingjian et al., 2015). Tseng et al. (2017) explored these properties in brain tumour segmentation, encoding the transverse plane (z) as time-steps to ensure a 3D correlation. First, the authors aim to characterise correlations among MRI sequences within a slice, followed by the spatial correlation ensured by the Convolutional-LSTM. Although with increased computational costs, the proposed method allowed a better characterisation among different types of brain tumour tissue.

In this thesis, we employed Gated Recurrent Networks similarly to Visin et al. (2016). A native 1D network is applied to 2D data, by performing an online 2D partition block. The partition block is responsible for defining a neighbourhood of $n \times n$ to be considered at each time-step. Therefore, each neighbourhood is characterised by a feature space of n^2 elements. Fig. 3.10 depicts the bidirectional Gated-Recurrent networks employed in the horizontal and vertical dimension to capture the local and global context of the input grid-patch.

In the schematic illustrated in Fig. 3.10, green arrows comprehend the vertical bidirectional gated-recurrent layer, while blue arrows the horizontal bidirectional gated-recurrent layer. The yellow circles denote the neighbourhood $n \times n$ grid structure.

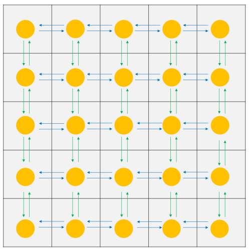


Figure 3.10: Proposed method of global and local context using Gated-Recurrent layers in grid-structured data.

3.5 Restricted Boltzmann Machines

Restricted Boltzmann Machines are undirected graphs that aim to learn joint probabilistic distributions. This joint probability distribution is modelled by the observed input data alongside hidden states of the RBM by a stochastic hidden representation of each data point (Goodfellow et al., 2016).

RBMs are responsible for one of the major contributions in deep learning, which consisted in a stack of RBMs trained in two phases, first with unsupervised learning and then with supervised learning. This training procedure was later on designated as pre-training (Hinton et al., 2006). Intrinsically, RBMs are unsupervised networks used for different learning tasks, such as to generate latent feature representations of the data, reduction of dimensionality, and matrix factorization (Goodfellow et al., 2016). By learning latent feature representations, one can combine the generated data into feed-forward networks (Goodfellow et al., 2016). Latent representations of data can also be achieved by autoencoders, with the slight difference that most of them generate deterministic hidden representations of each data point (Goodfellow et al., 2016).

3.5.1 Fundamental Structure

The fundamental structure of the RBM encompasses two layers of nodes: the visible layer and the hidden layer (Rummelhart and McClelland, 1986). RBMs arose as an evolution of the Boltzmann machines by having only connections among hidden nodes and visible nodes, which allows a more efficient learning process (Goodfellow et al., 2016). Each visible and hidden layer is characterised by a set of states represented by the vectors $v = [v_i : i = 1, \dots, m]$ and $h = \{h_j : j = 1, \dots, n\}$ for the visible and hidden layers, respectively. Each node has a weighted connection to all nodes in the other layer, represented by $\mathcal{W} = [w_{ij}]$, having no connections among nodes of the same layer (Rummelhart and McClelland, 1986). Moreover, the flow of information is non-directional meaning that the values can go from the visible node to the hidden node, or vice-versa (Goodfellow et al., 2016; Hinton, 2012), as illustrated in Fig. 3.11.

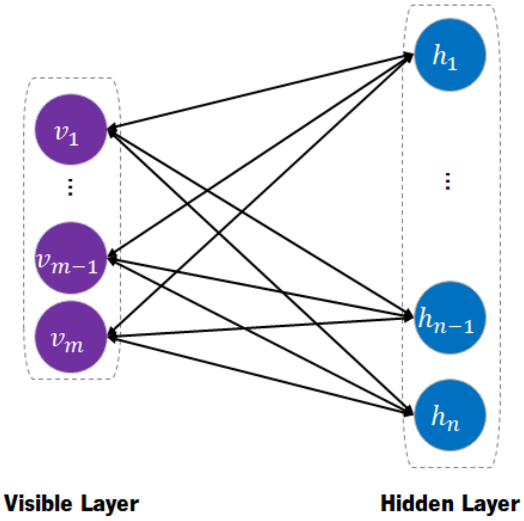


Figure 3.11: Restricted Boltzmann Machine layer. Each connection between a visible node v_i and a hidden node h_j has no implicit direction, hence it can occur on both possible ways. Once more, the bias was not considered for the sake of simplicity.

The RBM is an intrinsic binary state network, although it can be used for real-valued data (Goodfellow et al., 2016). When applied to binary data, with $v_i \in \{0,1\} \, \forall \, i = 1,\ldots,m$, and $h_j \in \{0,1\} \, \forall \, j = 1,\ldots,n$, RBMs compute the joint probability, p, of a hidden state and visible node as shown in Equation 3.54.

$$p(v,h) = \frac{e^{-E(v,h)}}{Z} \tag{3.54}$$

$$E(v,h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} v_i h_j \mathcal{W}_{i,j}, \tag{3.55}$$

$$Z = \sum_{i,j} e^{-E(v_i,h_j)} \tag{3.56}$$

The term E designates the energy term, which was initially defined for the Hopfield network as a definition of the objective function in an unsupervised learning environment. Hopfield network uses the sign of δE, denoting the difference between a pair (v_i, h_i) and predefined values, to set the connection of v_i to h_j to 1. In RBMs δE computes a conditioned probability. Hence, E is defined as shown in Equation 3.55 (Hinton, 2012; Bengio et al., 2013), where a_i denotes the bias of each visible node i, and b_j the bias of the hidden node j, and $\mathcal{W}_{i,j}$ the weighted connection between the i^{th} visible node and the j^{th} hidden node. Last, as shown in Equation 3.56, the term Z, designated normalization factor or partition function, is computed to ensure that the probabilities of all possible connections sum to 1 (Hinton, 2012). Computing the term Z is intractable, since the set v and h at a computation time i and j, respectively, need to be known to obtain Z. Therefore, the exact computation of Equation 3.54 is impossible since it is undefined. However, in most case scenarios of joint probability computation, the values of the conditional probabilities are ratios, cancelling the Z normalization factor (Hinton, 2012). For the particular case of the

RBMs, in the absence of intra-layer connections, the conditional probability distribution is given as (Bengio et al., 2013, 2009):

$$p(v|h) = \prod_i p(v_i|h) \tag{3.57}$$

$$p(h|v) = \prod_j p(h_j|v) \tag{3.58}$$

Due to this intrinsic properties, one can obtain the output of the hidden layer, given a binary input, by setting a node of such layer h_j to 1. This output is characterised by a probability shown in Equation 3.59.

$$p(h_j = 1|v) = \sigma\left(b_j + \sum_i v_i \mathcal{W}_{ij}\right) \tag{3.59}$$

In the inverse direction, often called as signal reconstruction, the same approach can be employed as shown in Equation 3.60 (Hinton, 2012; Bengio et al., 2013, 2009).

$$p(v_i = 1|h) = \sigma\left(a_i + \sum_j h_j \mathcal{W}_{ij}\right) \tag{3.60}$$

Originally, Rummelhart and McClelland (1986) proposed RBMs to model binary data, having the sigmoid as activation function for both the visible and the hidden nodes. Nonetheless, real-valued data can also be modelled by RBMs by changing the activation function. One of the approaches for real-valued data is the Gaussian-Bernoulli RBM. Here, the hidden nodes remain binary, but the visible nodes are instead linear nodes with independent Gaussian noise, which allows the modelling of continuous inputs (e.g. MRI image patches). The energy function is shown in Equation 3.61 (Hinton and Salakhutdinov, 2006).

$$E(v,h) = \sum_i \frac{(v_i - a_i)^2}{2\phi_i^2} - \sum_j b_j h_j - \sum_{i,j} \frac{v_i}{\phi_i} h_j \mathcal{W}_{ij} \tag{3.61}$$

The term ϕ_i designates the standard deviation of the Gaussian noise of v_i (Hinton, 2012). The first term in Equation 3.61 is of positive magnitude to assure a containment of E, more specifically to restrict the value of each v_i close to a_i, and mitigate an exponential growth of the probabilities, when computing Equation 3.54. Nonetheless, it cannot ensure boundaries, useful for the reconstruction phase, where v_i is computed. When compared to the binary RBM, the Gaussian-Bernoulli RBM has as major disadvantage its learning process, that is unstable. This is due to the absence of bounding restrictions for the computed values. To diminish this disadvantage, one solution is modifying the Gaussian-Bernoulli RBM so that it includes the independent Gaussian noise in the hidden nodes as well as shown in Equation 3.62.

$$E(v,h) = \sum_i \frac{(v_i - a_i)^2}{2\phi_i^2} - \sum_j \frac{(h_j - b_j)^2}{2\phi_j^2} - \sum_{i,j} \frac{v_i}{\phi_i} \frac{h_j}{\phi_j} \mathcal{W}_{ij} \tag{3.62}$$

However, by adding the independent Gaussian noise term, ϕ, the learning becomes even more unstable (Goodfellow et al., 2016). Furthermore, defining the appropriate value of ϕ is feasible but impractical, since the magnitude of the updates computed in the visible nodes tend to be small, whereas the magnitude of the updates of the hidden nodes tend to be large. Thus, a simpler solution for this problem is achieved by normalizing the input data to zero mean and unit variance, allowing setting $\phi = 1$ (Nair and Hinton, 2010; van Tulder and de Bruijne, 2016).

In this setting, based on the Equation 3.59 and Equation 3.60 sampling the state of the hidden and visible layers occurs from a Gaussian distribution $\mathcal{N}$, being respectively given as follows:

$$p(h_j|v) = \mathcal{N}\left(b_j + \sum_i v_i \mathcal{W}_{ij}, \phi_j\right) \tag{3.63}$$

$$p(v_i|h) = \mathcal{N}\left(a_i + \sum_j h_j \mathcal{W}_{ij}, \phi_i\right) \tag{3.64}$$

When dealing with real-valued data there are different approaches for the activation function. Equations 3.63 and 3.64 employed linear nodes of independent Gaussian noise. However, non-linear activation functions such as ReLU or Noisy REctifier Linear Units (NReLU) can also be applied in various settings.

In this thesis, when RBMs were employed, the hidden nodes of the RBM were set to be the NReLU, since they proved to be suitable for feature extraction (Hinton, 2012). Hence, the NReLU is only applied in the hidden layer (Nair and Hinton, 2010). The sampling equation for the hidden node is shown in Equation 3.65.

$$p(h_j|v) = \text{max}\left(0, \sum_i \mathcal{W}_{ij} v_i + b_j + \mathcal{N}\left(0, \sigma\left(\sum_i \mathcal{W}_{ij} v_i + b_j\right)\right)\right) \tag{3.65}$$

Since RBMs map the input data into a feature vector through the interaction of states between the visible and hidden nodes, equation 3.65 is only valid during sampling. Consequently, in the process of feature extraction where the parameters are learned and static, computation of p for the hidden node, h_j, uses as activation function the noise-free variant, the ReLU. The ReLU grants intensity equivariance, allowing that when the inputs of the visible layer are scaled by a positive value κ, the correspondent outputs are scaled by the same factor κ (Nair and Hinton, 2010).

3.5.2 Optimization

The learning process of RBMs occurs by minimizing the negative log-likelihood of the training data, given by:

$$\nabla \text{log} p(v) = \nabla \tilde{p}(v) - \nabla \text{log}(Z) \tag{3.66}$$

In Equation 3.66, the left term characterises the positive phase, while the right term characterises the negative phase. The negative phase of the gradient depends on the partition function, Z shown in Equation 3.56, which increases the complexity of the learning process. Recalling that Z depends on the parameters of the RBM, so the gradient computation still remains intractable. However, after some derivation ∇ log(Z) can be defined as shown in Equation 3.67.

$$\nabla \log Z = E_{v\,p(x)} \nabla \tilde{p}(v) \tag{3.67}$$

To cope with this problem ∇ log(Z) is approximated with Markov Chain Monte Carlo, alongside Gibbs sampling to provide an estimation of the model. At a starting state v^0, the Gibbs sampling generates $h^0 \sim p(h|v^0)$, then $v^1 \sim p(v|h^0)$ and so forth until the chain reaches the convergence. However, one clear disadvantage of computing the Gibbs Markov Chain Monte Carlo is the need to reach the convergence of the chain, which is a time demanding task. A faster and more common alternative to train RBMs is achieved by the Contrastive Divergence algorithm. The chain is initiated with a training example, and it is restricted to a small amount of Gibbs steps, k. Contrastive Divergence performs a coarse estimation of the gradient, but with the right direction, allowing the parameters to decrease the objective function properly. Generally, fixing $k = 1$ suffices to train RBMs (Hinton, 2012; Bengio et al., 2013, 2009).

3.6 Summary

Machine learning comprehends a vast area of methods, with the goal of performing a task given numeric data. The learning process can either be supervised or unsupervised. Classical approaches achieve this at the cost of a previous feature engineering step, which requires expertise knowledge of the field. On the other hand, Representation Learning focus on learning how to better extract features from the data, changing the learning paradigm from feature engineering to architecture design.

Currently, Representation Learning, namely Deep Learning-based methods are the most commonly used approach in machine learning. Both Convolutional Neural Networks, mainly applied for image data, and Recurrent Neural Networks, mainly used in language processing, have unlocked ground-breaking results. On one hand, Convolutional Neural Networks with learnable filters, allow the generation of higher levels of abstraction from the data, considering at the same time a neighbouring context. On the other, Recurrent Neural Networks, namely Gated Recurrent Networks, delve even further in the notion of context, by allowing that previous observations from data can influence the current observation. Regardless of being recurrent or convolutional, in its earlier beginnings, when stacked, these methods were difficult to optimize. Due to initialization, non-linear activation functions and efficient techniques of parameter sharing it became possible to efficiently train deep networks. Initialization algorithms began with the proposal of a pre-training technique applied to a stack of Restricted Boltzmann Machines. This unsupervised learning method, initially proposed for binary data, focus on learning the distribution of the data. However, this training is intricate since RBMs characterise the connections between nodes based on conditioned probabilities, with normalization factors. Nonetheless, RBMs can be used for real-valued data allowing its application in a vast number of machine learning problems.

Chapter 4

Capturing cerebral blood flow from temporal DSC-MRI acquisitions

Perfusion DSC-MRI is the spatio-temporal data, which is used to generate the standard parametric perfusion maps (MTT, TTP, T_{max}, rCBV and rCBF) used in clinical context for stroke assessment alongside diffusion parametric maps. These standard parametric perfusion maps are viewed as surrogate maps of the DSC-MRI. However, the mathematical foundations responsible for generating these parametric maps have already been recognized as an ill-posed problem, since they can generate numerical non-physiological solutions when processing the time-concentration curve (Fieselmann et al., 2011). Additionally, to characterize ischaemic stroke lesions, clinicians apply thresholds to parametric maps based on the deconvolution method employed, potentiating the loss of relevant information.

This chapter describes a fully automatic method to predict ischaemic stroke tissue outcome, employing deep learning-based algorithms. We propose an end-to-end two-pathway multi-data deep neural network that extracts features from the DSC-MRI perfusion scans, and the standard parametric perfusion/diffusion maps.

The first section provides the motivation behind this work, followed by a description of the proposed method. Afterwards, we detail the experimental set-up for final infarct tissue prediction. The final section contains the discussion of the results.

4.1 Motivation

In an ischaemic stroke context, time is critical, since stroke is a dynamic process where, in the absence of clinical intervention, the hypo-perfused region becomes irreversibly damaged, translating to a growth of the infarct region (Gonzalez et al., 2007). Hence, by characterizing the underlying phenomena that occur in the presence of an ischaemic stroke one can estimate the final infarct stroke lesion. Predicting the final stroke lesion, at a follow-up time, provides useful information to physicians when pondering on the benefits and risks of clinical intervention.

In this work, we present an end-to-end two-pathway multi-data deep neural network that incorporates information from the DSC-MRI perfusion sequence alongside its standard parametric maps and a diffusion parametric map, i.e. the ADC map. We hypothesize that features extracted from the perfusion DSC-MRI might complement the standard parametric maps, and surpass the possible loss of relevant information from the standard deconvolution methods. From the temporal acquisitions of the DSC-MRI, containing time-attenuation curves for each voxel, we propose to characterize blood flow circulation of the brain with data-driven maps extracted automatically in a dedicated deep neural network path, having as ultimate goal the identification of tissue at risk of infarction that will evolve to infarct tissue.

Another contribution of this work is a fully automatic pipeline that avoids the definition of an AIF for each patient, when dealing with DSC-MRI acquisitions. We encode the temporal information as channels, where an end-to-end deep neural network approach is responsible to extract spatio-temporal information of relevance avoiding the variability associated to spatially defining a reference input function.

The work contained in this chapter extends and improves the preliminary work accepted at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (Pinto et al., 2018b), with further validation and a more detailed description and discussion of the method. To the best of our knowledge this was the first time that a solution to deal with spatio-temporal images and combine them with standard parametric maps for stroke tissue outcome prediction was proposed. Currently, all the state-of-the-art methods only consider the information depicted by the standard parametric perfusion and diffusion maps.

4.2 Methods

This section starts by describing the perfusion dynamics associated with the DSC-MRI for ischaemic stroke. Afterwards, the details on the proposed two-pathway deep neural network architecture are given, being responsible for the combination of features extracted from DSC-MRI scans with the features extracted from the standard parametric perfusion and diffusion maps.

Based on neuroimaging scans acquired at the acute phase, the main goal of the research presented in this chapter is to assign one of two classes to each MRI voxel: healthy tissue or stroke tissue that will appear as infarct at a 90-day follow-up.

4.2.1 Perfusion DSC-MRI in Ischaemic Stroke

When performing the acquisition of perfusion DSC-MRI, the passage of contrasting agent through the brain, the bolus, is responsible for a drop on the MRI signal, and consequently for attenuating the intensity values. In the course of time, the contrasting agent is diluted by the renal function and consequently the intensity values recover to their basal value. Temporally, this behaviour is characterized by a time-attenuation curve (Song et al., 2017). However, in the presence of an ischaemic stroke, in the infarct core region the concentration of contrasting agent is mostly absent, and the intensity values of such region barely change across time. As for the hypo-perfused tissue, due to residual blood flow circulation or other factors, there might be a slight decrease in the intensities of such region (Song et al., 2017). For demonstration purposes, Fig. 4.1 depicts the signal intensity behaviour in a patient with an acute ischaemic stroke, during contrasting agent injection.

Fig. 4.1 illustrates a coarse perspective behind the perfusion spatio-temporal MRI acquisition. As explained previously, on average, after injection of the contrasting agent, the MRI signal of the healthy

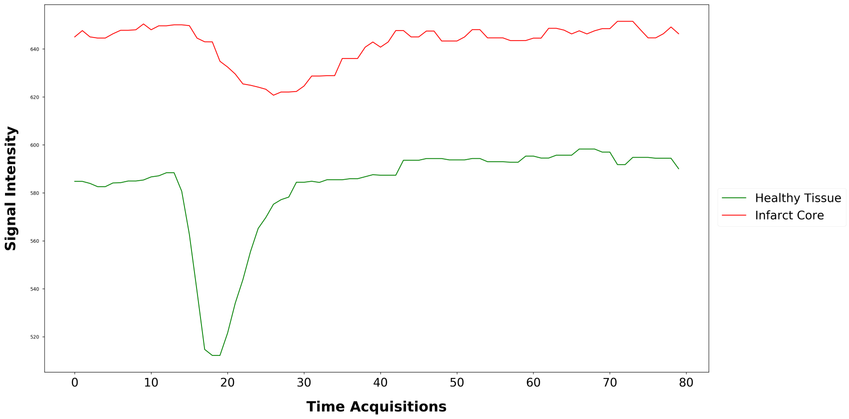


Figure 4.1: Whole brain perfusion DSC-MRI time-attenuation curve of a patient with acute ischaemic stroke, in the healthy tissue (green line) and in the final infarct core (red line).

tissue gradually drops, and recovers after the contrasting agent is fluxed out from the brain.

At a voxel level the time-attenuation curve directly translates to a time-concentration curve, since a higher intensity attenuation corresponds to higher concentration of contrasting agent. Based on the latter curve, through deconvolution in the time space, and clinical thresholding, it is possible to obtain 3D MRI perfusion maps that characterize different cerebral perfusion properties, as addressed in Chapter 2. However, the perfusion DSC-MRI is highly complex to characterize. As the contrasting agent enters the brain through principal feeding vessels, at a given point in time and space, healthy brain tissue can be under the effect of contrasting agent. However, other portions of healthy tissue, distal from the feeding arteries, maintain the basal MRI signal magnitude. This behaviour is illustrated in Fig. 4.2.

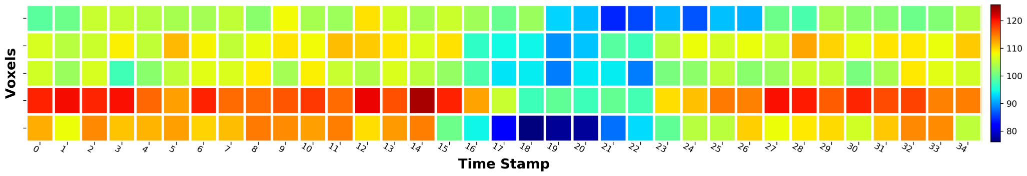


Figure 4.2: MRI signal intensity across time of healthy tissue voxels.

The intensity values of the voxels displayed in Fig. 4.2 are ordered by decreasing distance to the main feeding arteries, from top to bottom. Warmer colour correspond to higher intensity values. To capture this perfusion cerebral blood flow dynamics, machine learning methods need to be aware of two key-aspects of brain physiology. The first is that the arrival of contrasting agent to a specific point in space occurs constrained to its distance to main feeding arteries. The second aspect is that even when the contrasting agent arrives, its concentration might be influenced by the presence of a nearby occlusion or due to a patient clinical conditioning. Additionally, since the main goal is to estimate the final infarct core lesion at a 90-day follow-up, the model also needs to be aware that tissue surrounding the onset infarct core, might become irreversibly damaged at a posterior time-point, even after performing clinical intervention. For demonstration purposes, Fig. 4.3 illustrates a small region of a brain in the acute phase, characterized by a voxel, that has the intensity signal variation similar to healthy tissue, but was classified as lesion at the 90-day follow-up.

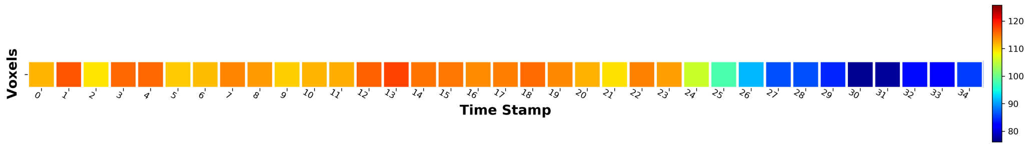


Figure 4.3: MRI signal intensity across time of an infarct core tissue. Warmer colours correspond to higher intensity values.

As can be observed, at the onset time, the voxel classified as lesion has intensity variations of healthy perfused tissue. Analysing Fig. 4.2 and Fig. 4.3, DSC-MRI poses as data difficult to handle in order to characterize and extract information. Furthermore, note that until now all the addressed phenomena concern only patient-specific conditions. The DSC-MRI acquisition and bolus injection protocols may vary across patients, hindering the learning process and increasing its complexity, as illustrated in Fig. 4.4.

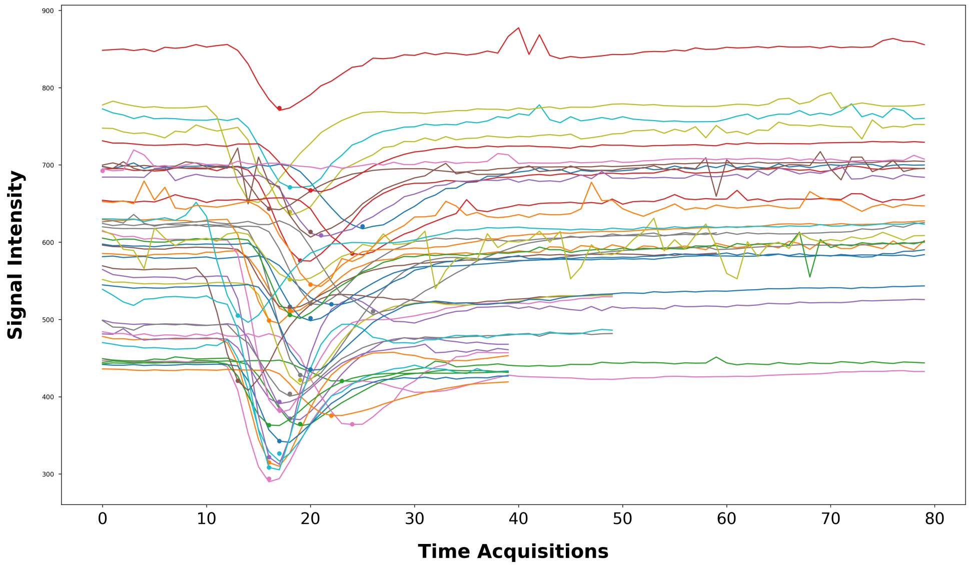


Figure 4.4: Perfusion DSC-MRI acquisitions from different patients with acute ischaemic stroke designed form ISLES 2017 dataset.

To surpass this variability, this work developed an automatic approach capable of retrieving a set of contiguous temporal acquisitions from a pre-defined temporal window, as illustrated in Fig. 4.5. Our approach focus on the time-stamp where the whole brain has the highest attenuation of intensities, and therefore the highest concentration of contrasting agent. This time-stamp characterizes the point in time when the differences of perfusion between healthy tissue, ill-perfused tissue and infarct core tissues are higher (Hosseini and Liebeskind, 2018). The detection of this time-point is obtained automatically with a k-means algorithm, with a number of classes set to two. One class contains the time-acquisitions with considerable variations in the intensity value and standard deviation, and the other has the time-acquisitions with despicable variation. From the first class, a contiguous set of temporal acquisitions is retrieved, containing the time-point of interest. Afterwards, from the temporal acquisitions contained in this class, a temporal window is extracted, as illustrated in Fig. 4.5. In this way, it is possible to reduce the number of temporal slices needed to characterize the blood flow dynamics, and perform an estimation of the tissue at risk of infarction. In addition, we also enforce the same spatial-temporal space across patients.

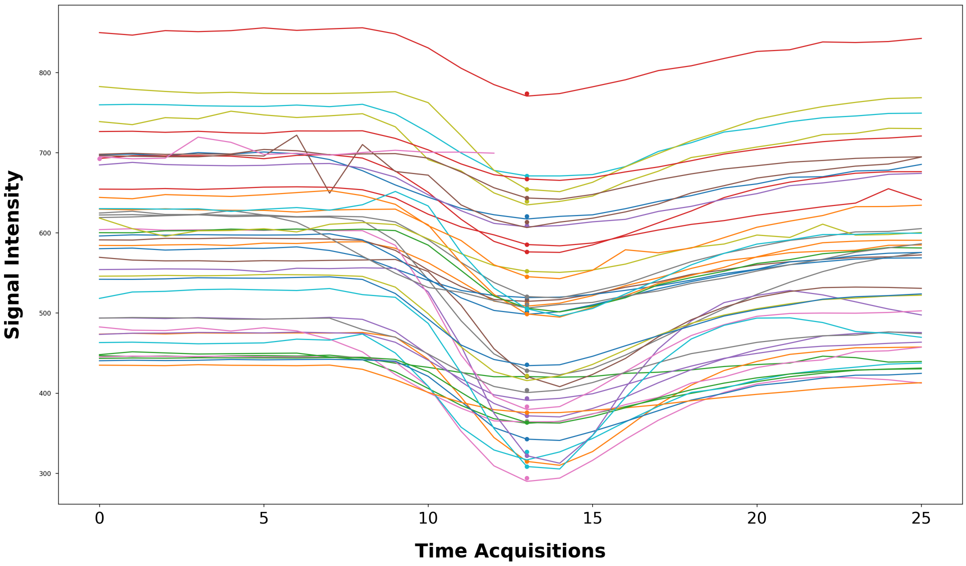


Figure 4.5: Perfusion DSC-MRI acquisitions from different acute ischaemic stroke patients in ISLES 2017 dataset, after retrieving a predefined set of time acquisitions.

To better understand, the method developed for the temporal alignment across patient cases, Fig. 4.6 illustrates a practical example of the automatic k-means window selection.

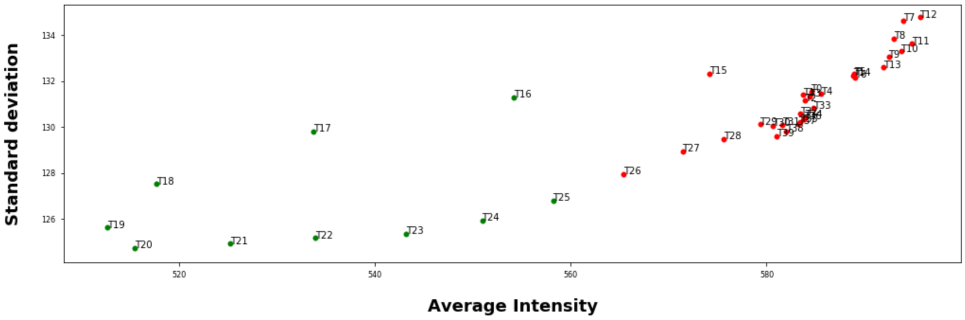


Figure 4.6: Average and standard deviation of the intensities in case 14 of ISLES 2017 training set, alongside the respective two groups selected by k-means.

After defining the group that encompasses higher intensity variation among the temporal MRI acquisitions, the algorithm is responsible to identify the temporal point, where the concentration of the bolus is higher, and consequently the intensity signal is lower.

In the light of the DSC-MRI image properties depicted previously, due to the complexity of the data and the underlying principles, we proposed to consider the spatio-temporal acquisitions of the perfusion DSC-MRI in a dedicated deep neural network, encoding the temporal relationships as channels.

4.2.2 Deep neural network architecture

To predict the final infarct core volume, we propose a deep neural network architecture divided in two functional blocks: Data-driven DSC-MRI block and Standard Diffusion/Perfusion block. The first block has as input the perfusion DSC-MRI. After temporally processing the spatio-temporal data, as described in the previous section, 2D patches, along the pre-defined time-acquisitions, are extracted as inputs to the Data-driven DSC-MRI block. Considering the DSC-MRI aims to capture the information regarding the blood dynamics needed to estimate the tissue at risk of infarction. The second block, the Standard Diffusion/Perfusion block, encompasses the diffusion ADC map alongside the perfusion T_{max}, TTP, MTT, rCBV and rCBF maps, computed from the perfusion DSC-MRI with standard deconvolution methods. This block is functionally equivalent and competitive to other state-of-the-art approaches (Winzeck et al., 2018). Afterwards, the feature sets extracted from both blocks are combined into a single pathway block. Merging the output of the two paths in the Fusion block, takes advantage of the information captured in both functional blocks. The deep neural network architecture is illustrated in Fig. 4.7.

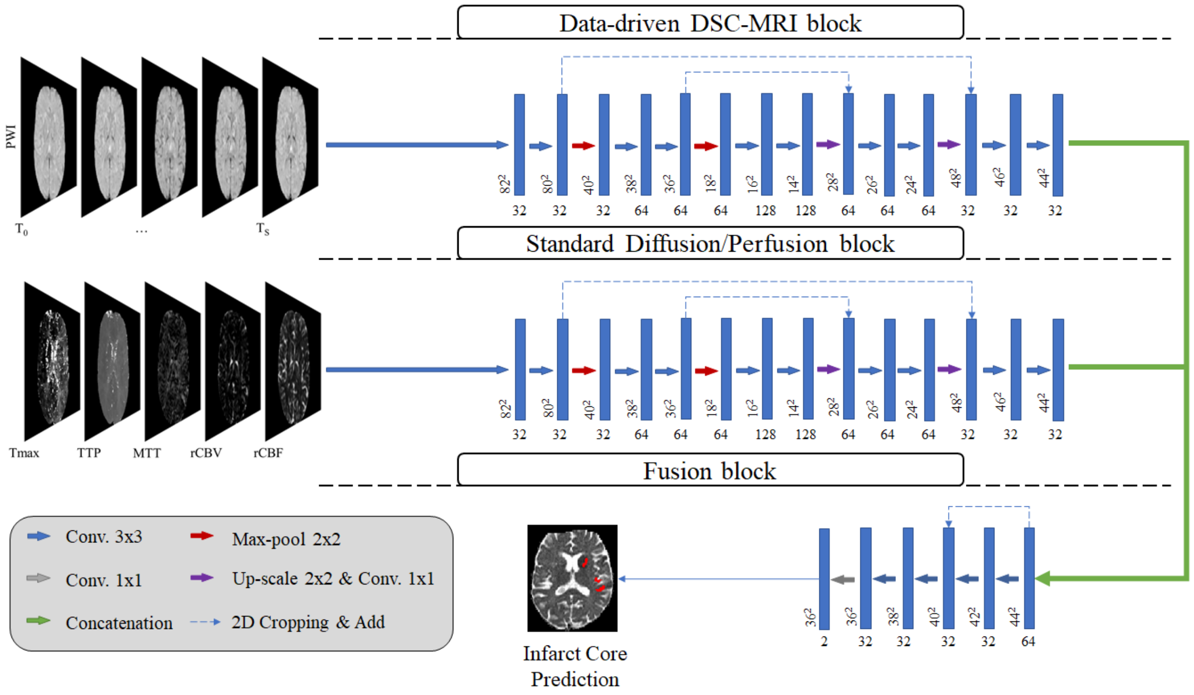


Figure 4.7: Overview of the proposed architecture for stroke lesion tissue prediction.

For the Data-driven DSC-MRI block a 2D U-Net based scheme (Ronneberger et al., 2015) is employed, where the temporal information was coded as channels. Hence, we use 3×3 convolutional kernels to simultaneously correlate the information among temporal slices and the local context.

In the Standard Diffusion/Perfusion block, we also employed an U-Net based scheme. Similarly, to the Data-driven DSC-MRI block, convolutional layers belonging to the same level of the network were defined with 32, 64, and 128 channels, for the first, second, and third level, respectively.

Finally, the Fusion block is responsible for combining the outputs from both functional blocks and elaborate on the most suitable features to predict the final stroke infarct. We hypothesize that the first two blocks contribute with different specific features, requiring a smaller block to take advantage of complementary information. Therefore, in the Fusion block we focus on combining the information across channels. Additionally, to ensure the most relevant extracted features from the first two blocks, an attention mechanism was also employed in the Fusion block.

4.2.3 Attention based network

Attention based networks are able to focus on regions (channel-wise and/or spatial-wise) of their input space, and specifically attend to relevant information. Therefore, one can view attention mechanisms as active filters on the input. These mechanisms can provide higher performance besides an increased interpretation of the network behaviour and decision-making (Xu et al., 2015).

For classification, one of the commonly used attention mechanisms is based on squeeze-and-excitation (SE) operations proposed by Hu et al. (2018). The Squeeze-and-excitation method, proposed for classification, consists on a gating mechanism that enhances the representational capability of the network. It employs feature recalibration by modelling channel-wise relationships, which gives higher relevance to certain features to the detriment of others.

Consider a bi-dimensional feature set $U \in R^{H \times W \times C}$, where H and W are the height and width of the input and C the cardinality of such feature set. The first operation of the SE block is a squeeze across the spatial domain by average global pooling. The output of this operation contains channel-wise statistics that globally highlight feature maps. On the second operation, the excitation is responsible for modelling channel-wise dependencies having as input the result of the squeeze operation. A set of weights, defined channel-wise, is then multiplied by the input feature maps, highlighting the features with higher representational power. The excitation operation learns complex interactions across channels, therefore being flexible on learning non-linear interactions and emphasizing various channels simultaneously. The selection of the feature maps with higher discriminative power is controlled by a hyper-parameter designated reduction ratio, r, which also correlates to the computational complexity and performance of a SE block. Therefore, the SE block can be achieved by having a layer that reduces the dimensionality of the feature space by a reduction ratio, r. After, a non-linear activation (e.g. ReLU) is employed, followed by a second layer that ensures the same dimensionality through a rescaling operation.

When dealing with semantic segmentation the same principles do not hold true. To cope with this limitation Pereira et al. (2018a) proposed a segmentation SE block, the SegSE. The SegSE linearly expands and compresses the feature space to enrich it, followed by a recalibration phase, with dilated convolutions, to increase the contextual information. The work of Roy et al. (2018) presents a concurrent spatial and channel SE layer, designated the scSE block. Similarly to Pereira et al. (2018a), this block can focus both on the channel-wise and spatial-wise spaces. It combines both the vanilla SE block and a SE layer to squeeze channel-wise and recalibrate spatially. When comparing both proposals of SE blocks applied to segmentation (Roy et al., 2018; Pereira et al., 2018a), there are differences. On one hand, the SegSE block employs an adaptive spatial recalibration with a reduction ratio of 10, while the scSE only focus on recalibrating whole feature maps. On the other, the scSE has the capability to select between channel-wise or spatial recalibration, through a voxel-wise max-out layer, with a reduction ratio of 2. Fig. 4.8 both attention blocks employed.

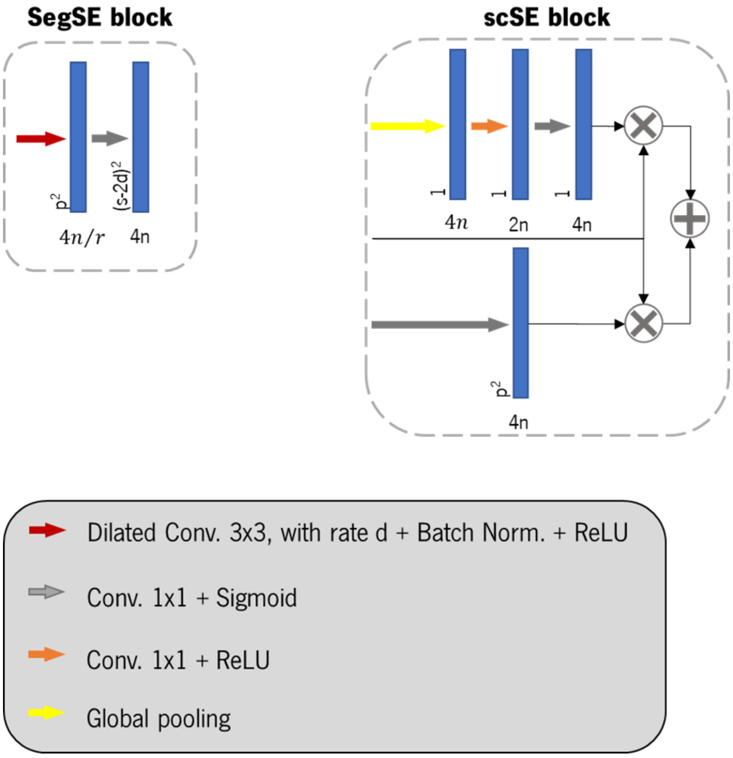


Figure 4.8: Attention blocks employed in the Fusion block.

In this work, to further improve the capability of the data fusion block by increasing the representational power of the learnable feature maps, we employ the SegSE of (Pereira et al., 2018a) and compare it with the proposal of Roy et al. (2018).

4.2.4 Feature Analysis

Machine learning methods, regardless of being supervised or unsupervised, are associated with high dimensional data, which has been further increasing due to the availability of more data and more information to characterize this data (Janecek et al., 2008). However, this leads to the presence of redundant information, raising the need for feature selection algorithms that decrease the training computational costs and might lead to a performance increase of the methods (Zhang et al., 2015). The feature selection step can be employed during the learning stage, or as a step independent of the classifier being used. The latter approach are designated as classifier-independent methods or filter methods. They focus on ranking features with respect to its importance in characterizing the label, or between features belonging to an extracted feature set. This ranking can be obtained from several criteria,e.g. distance measures, information, correlation and consistency measures. Filter methods have as major advantages their computational efficiency and scalability in what concerns the dimension of the dataset, and their independence from the classifier (Cover and Thomas, 2006).

Mutual Information (MI) is a measure of statistical independence, belonging to filter class. Henceforth, it can be used to quantify feature relevance. However, MI methods do not make an assumption on the linearity between features, having the capability to measure any kind of related information among all of them (Battiti, 1994). Furthermore, MI is invariant to transformations of the feature space that are invertible and differentiable, such as translations and rotations.

In the scope of this chapter, we followed the procedure of Battiti (1994), to quantify the relationship among non-linear variables. Therefore, after training the deep neural network, we extract n features from a specific layer for each training sample S, that characterizes the brain of a patient. Afterwards, for each feature, $k \in \{1, \cdots, n\}$, a feature vector $f_k = [f_r : r = 1, \cdots, S]$ characterizes the feature k for S samples. Then, the Mutual Information between each feature f_k and the values of the intensities from each MRI image, c ($i_c = [i_r]$), quantifies the statistical dependence between the feature k and the MRI c as shown in Equation 4.1.

$$MI_k(f_k, i_c) = \sum_c H(f_k) + H(i_c) - H(f_k, i_c) \tag{4.1}$$

The term H denotes an entropy computation commonly used in information retrieval, more specifically the Shannon entropy (Battiti, 1994).

4.3 Experimental Set-up

Our method was evaluated on the publicly available ISLES 2017 Challenge dataset (SMIR, 2017), which has an online benchmark platform responsible for the evaluation of the testing set data. In this section, we detail the dataset used and the metrics employed for evaluation. Lastly, we describe set-up of our deep neural network architecture.

4.3.1 Data

ISLES 2017 dataset encompasses 75 ischaemic stroke patients, which are separated into two sets: training ($n = 43$) and testing ($n = 32$). Both sets are constituted by patients who underwent mechanical thrombectomy, and are characterized by perfusion DSC-MRI and the standard 3D perfusion parametric maps rCBV, rCBF, MTT, TTP and T$_{\text{max}}$, alongside a 3D diffusion map, the ADC. However, in four training patients (cases 31, 42, 43, 45) the DSC-MRI images were corrupted and the spatio-temporal analysis is not feasible. One of these corrupted patients is shown in Fig. 4.9.

In addition to the MRI images, the dataset contains the manual delineation of the final infarct lesion from a 90-day follow-up T2-weighted MRI. However, the ground truth is only disclosed for public access in the training set. The testing set can only be evaluated by the online platform (SMIR, 2017), each Monday of the week.

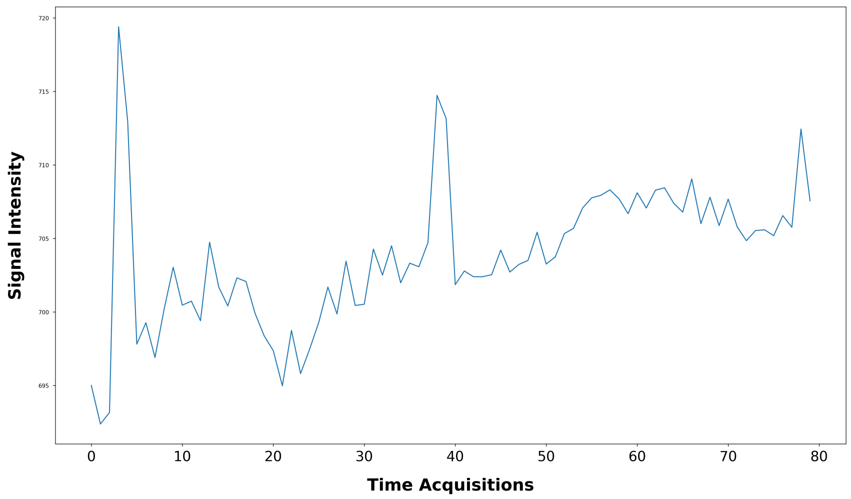


Figure 4.9: Whole brain perfusion DSC-MRI time-attenuation curve of the acute ischaemic stroke patient 42 form ISLES 2017 training set.

4.3.2 Evaluation

For evaluation purposes, the metrics used were the Dice score (DSC), Hausdorff Distance (HD), Average Symmetric Surface Distance (ASSD), Precision, and Recall. Each metric is mathematically defined, respectively, as follows:

$$DSC = \frac{2TP}{FP + 2TP + FN} \tag{4.2}$$

$$HD(A,B) = \max\{\max_{a \in A} \min_{b \in B} d(a,b), \max_{b \in B} \min_{a \in A} d(b,a)\} \tag{4.3}$$

$$ASSD(A,B) = \frac{1}{2}\left(\frac{\sum_{a \in A} \min_{b \in B} d(a,b)}{|A|} + \frac{\sum_{b \in B} \min_{a \in A} d(b,a)}{|B|}\right) \tag{4.4}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.6}$$

In the Dice score, Precision and Recall scores, TP denotes the true positives, which in the context of predicting the stroke tissue outcome comprehends voxels correctly assigned to the lesion class, while FP and FN are the false positives and the false negatives, respectively. The false positives correspond to voxels that do not belong to the lesion class but were classified as such, while false negatives encompasses samples that belong to the lesion class but were not identified as such. Correlating these principles, the DSC measures the spatial overlap between two volumes. Precision characterizes the probability of assigning correctly voxels to the lesion class, while Recall consists of a probability in identifying positive cases as such.

As for the distance metrics, HD allows the identification of the farthest spatial outlier present in the prediction, where $d(\dot{;})$ denotes the Euclidean distance between two voxels. The remaining distance metric, ASSD, computes the average distances between volumes surface points, namely the ground truth and the prediction, $|\,|$ denoting the cardinality of elements present in each surface volume.

4.3.3 Pre- and post-processing

In ISLES 2017, all MRI data is already co-registered and skull-stripped (Winzeck et al., 2018). However, as preprocessing, we first resize all image volumes to the same volume space of $256 \times 256 \times 32$, since the dataset contains acquisitions from different centers. Bias field correction was performed to the perfusion DSC-MRI, using the N4ITK method (Tustison et al., 2010), followed by the temporal processing that extracted a fixed temporal window size of 26 acquisitions. The choice of this temporal window was based on the sampling rate of the MRI acquisition. Finally, a linear scaling was applied between $[0, 255]$ to all maps. Before linear scaling, the T$_{\text{max}}$ was clipped to $[0, 20s]$, and the ADC was clipped to be within the range $[0, 2600] \times 10^{-6} mm^2/s$, as values out of these ranges are known to be biologically meaningless (McKinley et al., 2016). When uploading the predictions to the online platform, each volume is resized back to its original dimensions.

Fig. 4.10 illustrates a training case example with the standard parametric maps alongside some temporal acquisitions of perfusion DSC-MRI and the manual segmentation of the tissue lesion, the Ground Truth (GT).

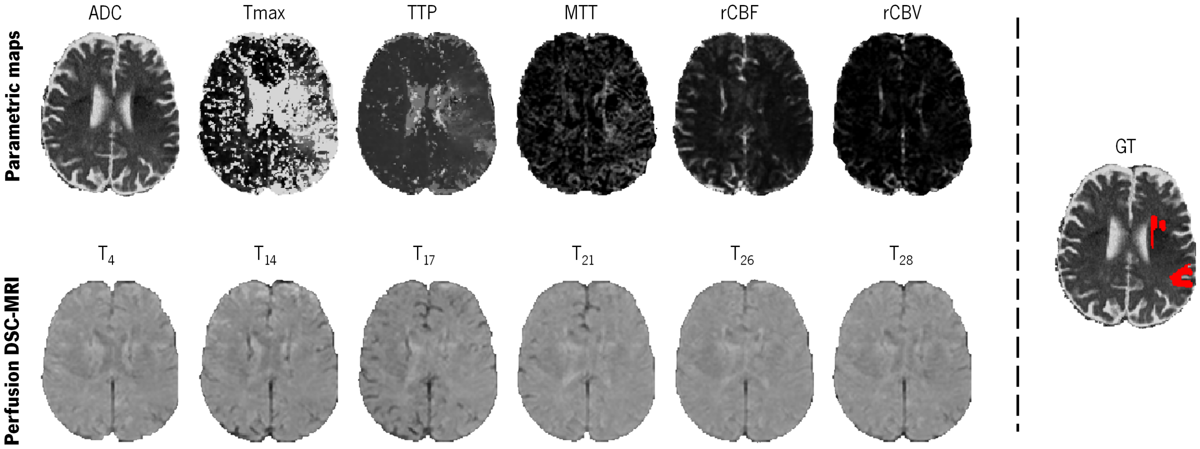


Figure 4.10: Standard parametric maps, DSC-MRI acquisitions and the respective manual segmentation of validation case 5.

4.3.4 Model training & parameters

The overall deep neural network architecture, including both blocks as shown in Fig. 4.7, was trained with 35 cases, alongside a validation set that encompassed 4 cases. In each case, 1000 patches of dimensions 84×84 were extracted with a random sampling scheme.

The network was trained with ADAM optimizer (lr=1×10^{-5}) (Kingma and Ba, 2014) and a mini-batch of size 4. For regularization, we employed a spatial drop-out (Tompson et al., 2015) of 0.25 at each two convolutions. As for the loss function, we used the soft-dice loss (Milletari et al., 2016), where the gradient of the Dice score for the j^{th} voxel of prediction is given by:

$$\frac{\delta Dice}{\delta p_j} = \frac{g_j(\sum_i^N p_i^2 + \sum_i^N g_i^2) - 2p_j \sum_i^N p_i g_i}{(\sum_i^N p_i^2 + \sum_i^N g_i^2)^2} \tag{4.7}$$

In Equation 4.7, the sum is performed for the N voxels of the patch both in the binary prediction $p_i \in P$ and the ground truth $g_i \in G$.

All the models were developed using Keras (Chollet, 2015) with Tensorflow, and trained on an Nvidia GeForce GTX 1070 8 GB, with a prediction time around 15 seconds per patient.

4.4 Results and Discussion

This section starts by discussing the ablative study, which measures the importance of the main contributions of our proposal. In this ablative study, we first evaluate the importance of including spatio-temporal imaging data with the standard parametric maps. Second, we measure the importance of the two-pathway architecture and key components of the method, namely the temporal processing of the perfusion DSC-MRI, the data-fusion block and the inclusion of attention mechanisms. After, we delve in the information extracted from the DSC-MRI with deep neural network. Finally, we compare our proposal with state-of-the-art methods in ISLES 2017 Challenge.

4.4.1 Ablative Study

To measure the importance of key components of our deep neural network, we start by evaluating the impact of considering the DSC-MRI data alongside the standard parametric maps. Then, we compare our two-pathway architecture with the one-pathway U-Net based scheme. Finally, we measure the impact of performing the temporal processing of the DSC-MRI and of the Fusion block. The obtained results are presented in Table 4.1.

Table 4.1: Results obtained in the ablative study, considering different source data and its separation in the network, alongside key components of the architecture in ISLES 2017 testing dataset. Each metric contains the mean ± standard deviation. Underline values correspond to the highest score of the respective performance metric (column-wise).

	Params.	Dice	HD	ASSD	Precision	Recall
Standard Maps	382 154	0.30 ± 0.21	38.83 ± 21.10	7.08 ± 5.15	0.26 ± 0.23	$\underline{0.64}$ ± 0.30
One-pathway	816 916	0.28 ± 0.21	37.47 ± 16.05	6.90 ± 4.43	$\underline{0.32}$ ± 0.28	0.54 ± 0.30
Two-pathway no temporal proc.	838 594	0.28 ± 0.21	43.66 ± 23.57	7.89 ± 6.31	0.25 ± 0.23	0.66 ± 0.33
Two-pathway no fusion block	787 778	0.27 ± 0.21	40.89 ± 18.68	8.21 ± 6.68	0.28 ± 0.26	0.53 ± 0.34
Two-pathway	836 002	$\underline{0.31}$ ± 0.21	$\underline{33.94}$ ± 17.43	$\underline{5.99}$ ± 4.58	0.29 ± 0.23	0.63 ± 0.30

On the inclusion of spatio-temporal imaging data

Considering our hypothesis and contributions, first it is evaluated the importance of the DSC-MRI alongside the standard perfusion and diffusion maps, the Two-pathway result presented on Table 4.1. This method is compared with the Standard Maps method, which only receives as input the standard parametric maps, in a single U-Net based scheme equal to the Standard Diffusion/Perfusion block. Comparing the Standard Maps method with the Two-pathway, we demonstrate the benefit of considering both input data sources. The two-pathway achieved a higher average Dice score, alongside lower average distance metrics and higher average Precision score. Despite our proposal not being independent of the mathematically ill-posed problem of the generated perfusion maps, considering the source data (i.e. DSC-MRI) responsible for generating them, it allowed the extraction of discriminative information. Thus, we conclude that direct extraction of features from the DSC-MRI imaging data is of importance to predict the final infarct stroke lesion.

One-pathway vs. Two-pathway

The combination of DSC-MRI with the standard parametric maps can be performed directly by aggregating both input data into a single U-Net based scheme. However, by combining them in a single U-Net based scheme, it may lack the capacity in extracting features that characterise the bolus passage from the spatio-temporal DSC-MRI. Since the standard parametric characterise the different physiological processes at a different level from the DSC-MRI, we hypothesize that it is more effective to elaborate specific features before combining them. Hence, the Two-pathway is compared with a One-pathway network that combines all the input data into a single U-Net based scheme. For sake of a fair comparison, we ensured that the learnable parameters of both methods are similar, leading to an increase of the number of channels when using the one-pathway block. Additionally, both methods employed temporal slicing and alignment. From the obtained results presented on Table 4.1, comparing with the Two-pathway method, it is possible to verify that aggregating the DSC-MRI with the standard parametric maps into a single network, achieved lower performance in all metrics. The Two-pathway method achieved 10.71% higher Dice score on average, and

16.67% higher average Precision, while the Recall lowered 10.34%. So, in the context of predicting stroke lesion evolution, we conclude that the use of dedicated paths to extract information from distinct physiological processes is more effective than using a single network to process all of them simultaneously. Since each network will elaborate the features along its convolutions, it seems to be better to correlate elaborated features from distinct sources, than to correlate simpler features from the beginning.

On the importance of temporal alignment and the data Fusion block

In the last study, we evaluate the presence of two key components of our proposal, namely the Fusion block and the temporal slicing and alignment. First, with the removal of the Fusion block, we demonstrate that elaborating over the extracted features aggregated from both paths, allows an overall increase on the performance of our proposal. Comparing our proposal with the Two-pathway no fusion block, we obtained an increase in the average Dice of 14.8%, and an increase of 3.6% and 18.9% in the average Precision and Recall, respectively. Afterwards, analysing our proposal against the two-pathway with no temporal alignment, demonstrates an overall increase of 10.71% in the average Dice score, alongside lower average distance metrics. Hence, we conclude that performing temporal slicing and alignment allows a spatio-temporal standardization useful for the extraction of higher discriminative features when predicting the final infarct stroke lesion. Furthermore, the spatio-temporal standardization allows a reduction on the temporal acquisitions without impairing the performance of the method.

To better assess the importance of the ablative studies conducted so far, Fig. 4.11 depicts an example case of the validation set when considering the one-pathway multi-data network, the two-pathway multi-data with no temporal slicing, the two-pathway multi-data without the Fusion block and our proposal.

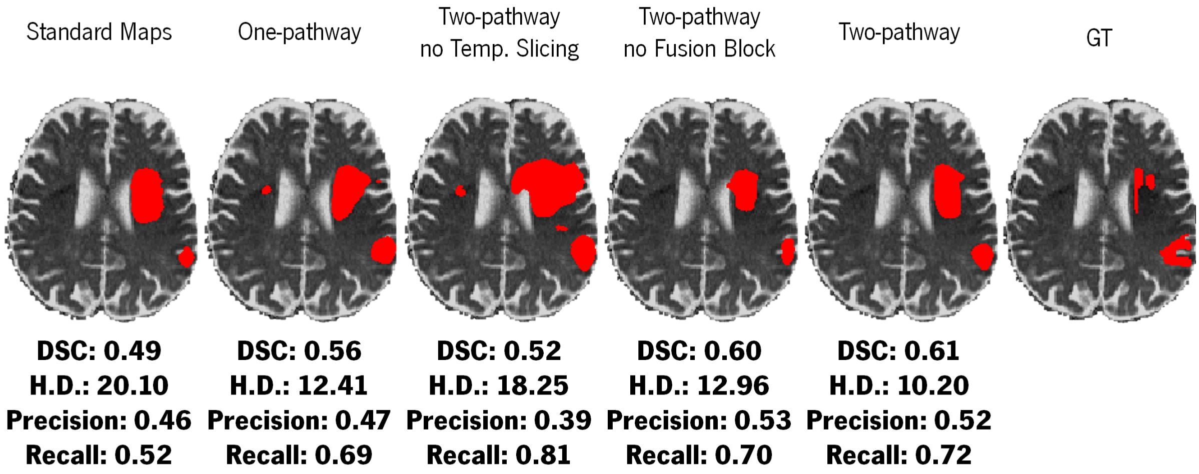


Figure 4.11: Predictions obtained for validation case 5 of ISLES 2017 training set, in the ablative studies conducted on Table 4.1.

Analysing Fig. 4.11, the Standard Maps method achieved the lowest Dice score and highest Hausdorff distance, when compared with the other methods. The Two-pathway no temporal slicing method obtained the worst prediction scenario, sustained by the high Recall metric. The Two-pathway method yields the best prediction of the final stroke lesion, achieving the highest Dice score alongside a good balance between Precision and Recall. Additionally, the Hausdorff distance achieved by our proposal was the lowest.

Attention-based mechanisms

Due to the distinct nature behind the standard parametric maps and the perfusion DSC-MRI, when combining automatically the extracted features from both paths, the network needs to be capable of extracting the most relevant ones for the task of stroke tissue prediction outcome. Hence, having a mechanism intrinsically developed to capture relevant information among a given feature space, would ultimately provide a higher capacity in predicting the final stroke lesion. Therefore, Table 4.2 presents the results obtained with both attention mechanisms referred in Section 4.2.3. In order to keep an equal output patch size across all approaches, when employing the SegSE block, due to its receptive field, we increased the input patch to 92×92.

Table 4.2: Results obtained in ISLES 2017 testing dataset. Each metric contains the mean ± standard deviation. Underline values correspond to the highest score of the respective performance metric (column-wise).

	Params.	Dice	HD	ASSD	Precision	Recall
Two-pathway	836 002	0.31 ± 0.21	33.94 ± 17.43	5.99 ± 4.58	0.29 ± 0.23	0.63 ± 0.30
Two-pathway + SegSE	946 430	0.29 ± 0.22	38.83 ± 21.30	8.81 ± 13.59	0.27 ± 0.25	0.58 ± 0.30
Two-pathway + scSE	926 787	0.29 ± 0.23	40.82 ± 16.87	7.88 ± 4.92	0.28 ± 0.27	0.61 ± 0.31

When comparing the obtained results from the attention mechanisms, both achieve the same Dice score, being the scSE block capable of reaching higher average Precision and Recall metrics. Nonetheless, neither approaches are capable of surpassing the performance of our proposal. Our proposal achieves higher average score in all metrics apart from the average Recall. From these results, incorporating attention based mechanisms lacks in the capacity of selecting properly relevant features that allow a better prediction of the final stroke lesion. Furthermore, we also conclude that the complexity of predicting the final stroke lesion requires all the discriminative information present in the feature space.

Fig. 4.12 illustrates the results obtained in the same validation patient described by the imaging data of Fig. 4.10.

In this validation case, when using the scSE attention block the obtained prediction was the first to consider a single connected region of final stroke lesion. This behaviour indicates that the scSE block might be selecting simpler and coarse features over elaborated ones. In the presence of another attention block, the SegSE block, the prediction result is closer to the Two-pathway, but with a lower capacity in predicting the final infarct core lesion, as can be observed by the Dice score of 0.61 vs. 0.58 of the Two-pathway compared with the Two-pathway with SegSE.

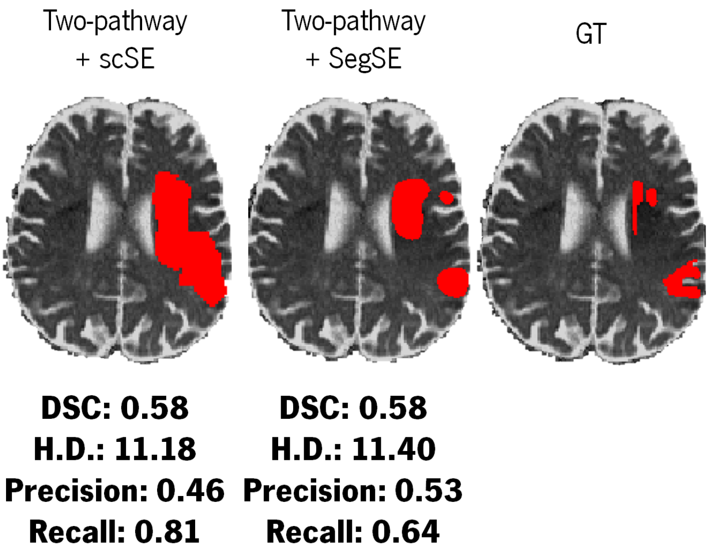


Figure 4.12: Slice results of validation case 5 from ISLES 2017 training set, in the Two-pathway method with the employed attention mechanisms.

4.4.2 Feature analysis

In our proposal, with the extraction of features from the DSC-MRI data in a dedicated deep neural network, we hypothesize that this information grants additional perfusion dynamics information, useful for predicting the final infarct core lesion. To assess the added value of the feature space extracted from the Data-driven DSC-MRI block, we study the correlation level of its extracted features with the input of the Standard Diffusion/Perfusion block, the standard parametric maps. The correlation was computed through the normalized MI (Estévez et al., 2009), whose results are illustrated in the graph bar of Fig. 4.13. Note that values closer to 0 mean low mutual information and closer to 1 represent a high association.

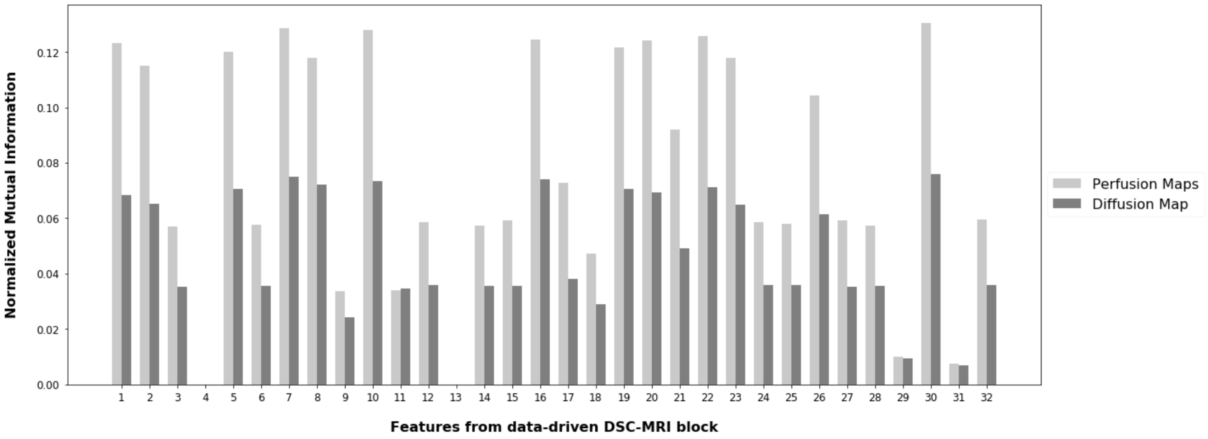
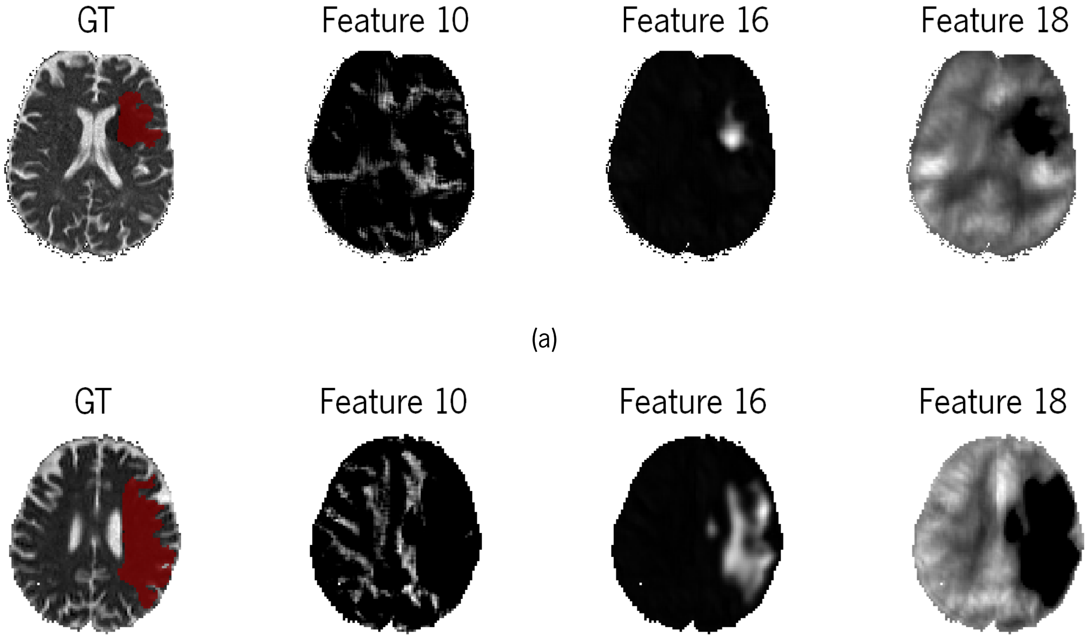


Figure 4.13: Normalized Mutual Information between the standard perfusion/diffusion maps and the feature maps from the data-driven block for the training and testing sets.

Analysing Fig. 4.13, the normalized mutual information achieved low association values (less than

0.2) among the extracted feature maps and the standard parametric maps. Regardless of the fact that the DSC-MRI acquisition only characterizes perfusion properties, we extended our analysis to both major vascular properties, diffusion and perfusion. Therefore, in the light of the performance results obtained in the testing set, we hypothesize that both functional blocks introduce distinct and complementary features, useful when predicting the final infarct core volume.

Fig. 4.14 illustrates examples of the extracted features from the Data-driven DSC-MRI block, with the highest MI values.



(a)

(b)

Figure 4.14: Example of extracted features alongside the corresponding GT over the ADC map for case 33 (Fig. 4.14a) and 36 (Fig. 4.14b) of ISLES 2017 training set.

Delving in the information illustrated in Fig. 4.14a and Fig. 4.14b, feature 10 can reflect some descriptions of collateral blood flow, where features 18 focus on the surrounding area of the stroke lesion and feature 16 retrieves areas that have a high overlap with the final infarct. This analysis poses as a crucial factor, since it is possible to observe that our proposal was capable of extracting important and interpretable information, which might provide the physicians a better understanding of the lesion growth through time. Simultaneously, it provides an understanding on how the method performed the prediction of the final infarct lesion at a 90-day follow-up. However, one disadvantage is the absence of clear clinical interpretation from such learned feature maps, as opposed to the standard parametric maps.

4.4.3 State-of-the-art: ISLES 2017 Challenge

On Table 4.3, we compare our proposal with state-of-the-art methods from ISLES 2017 Challenge (Winzeck et al., 2018), ranked accordingly to the average Dice score. Regardless of the model topology, predicting final infarct core is still a challenging and intricate task, that needs to consider scenarios of successful and unsuccessful reperfusion. Furthermore, in each reperfusion scenario, predicting the infarct growth, and consequently the final stroke lesion, needs to be aware of various haemodynamic factors (e.g. location or collateral circulation) which hinders the learning process.

Table 4.3: Published methods in ISLES 2017 Challenge testing dataset and our proposal. Each metric is represented by the mean ± standard deviation. Underlined values correspond to the highest mean.

		Dice	HD	ASSD	Precision	Recall
Ensemble	Mok et al. *	0.32 ± 0.23	40.74 ± 27.23	8.97 ± 9.52	0.34 ± 0.27	0.39 ± 0.27
	Kwon et al. *	0.31 ± 0.23	45.26 ± 21.04	7.91 ± 7.31	0.36 ± 0.27	0.45 ± 0.30
	Robben et al. *	0.27 ± 0.22	37.84 ± 17.75	6.72 ± 4.10	0.44 ± 0.32	0.39 ± 0.31
	Pisov et al. *	0.27 ± 0.20	49.24 ± 32.15	9.49 ± 10.56	0.31 ± 0.27	0.39 ± 029
Single Model	Monteiro et al. *	0.30 ± 0.22	46.60 ± 17.50	6.31 ± 4.05	0.34 ± 0.27	0.51 ± 0.30
	Lucas et al. *	0.29 ± 0.21	33.85 ± 16.82	6.81 ± 7.18	0.34 ± 0.26	0.51 ± 0.32
	Choi et al. *	0.28 ± 0.22	43.89 ± 20.70	8.88± 8.19	0.36 ± 0.31	0.41 ± 0.31
	Niu et al. *	0.26 ± 0.20	48.88 ± 11.20	6.26 ± 3.02	0.28 ± 0.25	0.56 ± 0.26
	Sedlar et al. *	0.20 ± 0.19	58.30 ± 20.02	11.19 ± 9.10	0.23 ± 0.24	0.40 ± 0.29
	Rivera et al. *	0.19 ± 0.16	63.58 ± 18.58	11.13 ± 7.89	0.27 ± 0.25	0.21 ± 0.17
	Islam et al. *	0.19 ± 0.18	64.15 ± 28.51	14.17 ± 15.80	0.29 ± 0.28	0.25 ± 0.25
	Chengwei et al. *	0.18 ± 0.17	65.95 ± 25.94	9.22 ± 6.99	0.37 ± 0.30	0.21 ± 0.23
	Yoon et al. *	0.17 ± 0.16	45.23 ± 19.14	12.43 ± 11.01	0.23 ± 0.27	0.36 ± 0.32
	Baseline	0.30 ± 0.21	38.83 ± 21.10	7.08 ± 5.15	0.26 ± 0.23	0.64 ± 0.30
	Two-pathway	0.31 ± 0.21	33.94 ± 17.43	5.99 ± 4.58	0.29 ± 0.23	0.63 ± 0.30

* Results retrieved from (Winzeck et al., 2018).

In ISLES 2017 Challenge testing set, our single model was capable of achieving competitive results, with a Dice score among the top two ranked methods, tied with Kwon et al. 2018, and with the second lowest average Hausdorff distance and first ASSD, in such ranking.

Encompassing ensemble strategies, we obtained only 3.2% bellow the top performing method, Mok and Chung (2017). However, for a single model approach, we remark our consistency by the distance metrics obtained in the testing set, being lower than both ensemble methods. As for the precision and recall metrics, we observe a slight trade-off. Considering the top two ensemble approaches, our model achieved higher average Recall, alongside the average Precision of 0.29, which was the lowest. From our perspective, employing several adversarial deep neural network approaches, as proposed by Mok and Chung (2017), granted the authors a higher capability to distinguish slight intensity variations present in the standard parametric maps, which can be sustained by the balance between the average precision and average recall. As for the proposal of Kwon et al. 2018, we hypothesize that the presence of deep neural networks capable of predicting if patches either contain or not final infarct lesion voxels explains the higher average precision, when compared with Mok and Chung (2017).

Considering only single model strategies, we observe that our proposal achieved the highest Dice average in the testing set, followed by Monteiro and Oliveira (2017). The proposal of the authors considers a weighted scheme between cross-entropy and soft dice losses, using a V-Net based scheme, which might sustain the higher average Precision, when compared to our proposal. Nonetheless, in this group, the Precision and Recall metrics ranked as 5th and 1st place, respectively. We remark the robustness of our proposal in predicting stroke tissue outcome, observed by the low standard deviation values. Furthermore, we note the benefits of the proposed approach to extract and model information that might not be fully characterized by the standard perfusion and diffusion maps.

4.5 Summary

Clinical intervention aims to restore the perfusion deficits by chemical or mechanical approaches. Regardless of the reperfusion procedure, the clinicians need to ponder on the risks and benefits based on multi-modal neuroimaging acquisitions, such as MRI, and clinical experience. Hence, automatic prediction of final ischaemic infarct lesion would help the physician in such intricate decision-making process, providing information about tissue that will probably infarct.

Parametric perfusion maps can be affected by intrinsic patient physiology (Song et al., 2017). To cope with this effect, mathematical models are applied to standardize the behaviour of the contrasting agent. Nonetheless, it cannot be independent of patient specific blood flow haemodynamic, which can highly affect the perfusion parametric maps by adding a wide variability in the penumbra delineation (Song et al., 2017).

In this chapter, we propose a deep neural network architecture, that can process the information from perfusion DSC-MRI data and generate complementary information to the perfusion parametric sequences.

Chapter 5

Incorporating clinical meta-data alongside MRI acquisitions

Predicting the final infarct lesion in ischaemic stroke, at a 90-day follow-up, needs to be aware of different infarct evolution scenarios, to provide the clinician information useful for the decision-making process that ponders on the therapeutic intervention. In this chapter, we aim to combine imaging information, of the standard parametric MRI maps, with non-imaging information, namely clinical meta-data. The clinical meta-data considered was the Thrombolysis in Cerebral Infarction (TICI) score (Higashida et al., 2003), which characterizes the success of reperfusion by thrombectomy.

The first section contains the motivation behind this study, enumerating its contributions. Afterwards, Section 5.2 describes the method proposed, which is then followed by the experimental set-up. Finally, the results and respective discussion are presented.

5.1 Motivation

Predicting stroke lesion outcome (i.e. 90-day follow-up), and the potential efficacy of the treatment according to the nature of the lesion, has a great potential to guide the decision-making by physicians. Furthermore, automatic methods of stroke tissue outcome prediction would help the physician in such time-critical decision-making process (Maier et al., 2015).

In this chapter, our main contribution is the proposal of an end-to-end deep neural network architecture that combines imaging information with clinical meta-data, namely the TICI score. The deep neural network incorporates clinical meta-data at two levels. First, at the population level, which implicitly considers correlations between tissue loss and the TICI score, through a custom loss function. Second, at a patient level, which explicitly encodes the TICI score of each patient as an extra input channel, thus allowing it to be considered during training and prediction.

The second contribution is about how the clinical information is instated in the loss function. This proposal considers a customized loss function to learn the relationships between imaging and non-imaging information at a population level.

The final contribution comprehends the inclusion of clinical information during the prediction phase at a patient-specific level, allowing the prediction of different lesion outcome scenarios in clinical environment.

Our proposal was evaluated using the publicly available ISLES 2017 dataset, where we demonstrate the potential value of incorporating imaging and clinical meta-data for stroke tissue outcome prediction at a 90-day follow-up.

5.2 Methods

Prediction of the final stroke lesion consists in characterizing changes in location and extension of lesions over time from standard parametric MRI maps and non-imaging clinical information gathered at onset time. Hence, automatic methods assign to each voxel of the MRI volume one out of two classes, healthy tissue or stroke lesion tissue. However, when re-establishing the brain blood flow, depending on the success of clinical reperfusion, the onset stroke lesion can grow or shrink over time. In order to evaluate the level of reperfusion achieved, the clinical intervention is evaluated through the TICI score (Higashida et al., 2003). Fig. 5.1 and Fig. 5.2 illustrates two cases of MRI maps with different TICI scores, alongside the final stroke lesion (ground truth – GT), manually delineated from a 90-day follow-up T2 sequence.

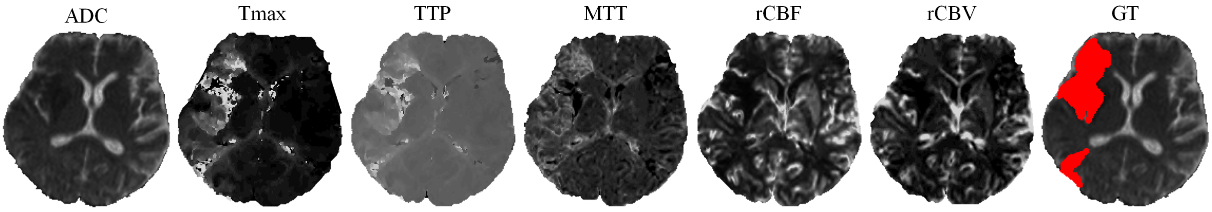


Figure 5.1: MRI parametric maps of a stroke patient with TICI score 0, and the respective manual segmentation. Only one class is defined, describing simultaneously the infarct core and the penumbra regions.

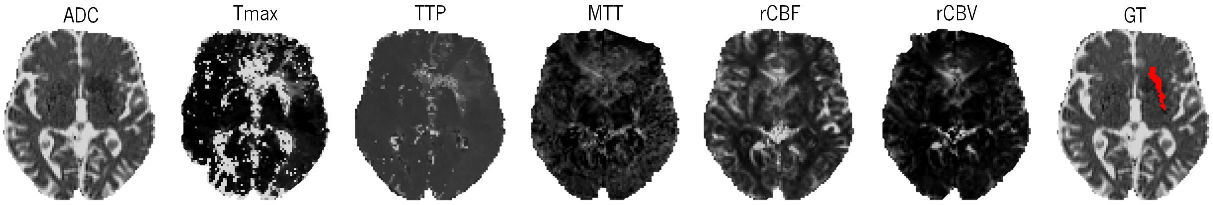


Figure 5.2: MRI parametric maps of a stroke patient with TICI score 3, and the respective manual segmentation.

5.2.1 Pre-processing

Our proposal uses diffusion and perfusion maps, adding up to six MRI parametric maps: diffusion ADC map, perfusion T$_{\text{max}}$, TTP, MTT, rCBF, rCBV, maps, as illustrated in Fig. 5.1 and Fig. 5.2.

ISLES 2017 dataset provides MRI scans acquired from different centers (Winzeck et al., 2018). So, the perfusion and diffusion maps result from different configuration conditions. Therefore, for each patient we first resized all maps to a common dimension of $256 \times 256 \times 32$. Afterwards, the ADC maps were clipped between $[0, 2600] \times 10^{-6} mm^2/s$, and the T$_{\text{max}}$ maps were clipped between $[0, 20s]$, since values beyond these ranges are known to be biologically meaningless (McKinley et al., 2016). As a final step of pre-processing, we applied a linear scaling across all maps transforming them to the range $[0, 255]$.

5.2.2 Deep neural network architecture

Our proposal is inspired by the fully convolutional U-Net architecture (Ronneberger et al., 2015), which due to its success was rapidly known and employed in several biomedical imaging problems, specially for segmentation (Isensee et al., 2019). The encoder-decoder architecture encompasses three levels of encoding and decoding. At each level we employed two 2D convolutional layers, with 32, 64 and 128 channels per layer. In the encoding path, from one level to the following we apply 2D max-pooling operators to increase the translation invariance and to extract features of higher complexity and detail. As for the decoder path, to allow the sum of its extracted features with features from the correspondent encoder level, it employs 2D up-sampling layers followed by convolutions (Fig. 5.3). Thus, the 2D up-sampling layers provide the same spatial dimensions, while the 2D convolutions, with a kernel of size 1×1, provide the same feature size as in the extracted features on the same level of the encoder. In addition, the proposed U-Net based scheme is combined with a 2D-dimensional GRU layer (Cho et al., 2014b) to obtain smoother and structured predictions. The motivation behind the Gated-RNN resides in its capacity of global and local context. While convolutional layers only can relate the information depicted within the kernel dimensions, the recurrence property of Gated-RNNs allows the correlation of previously observed voxels with the current one, which translates into a higher notion of context, ultimately outputting a better stroke lesion prediction. Similarly to Visin et al. (2016), we employ the GRU layer into two directions, vertical and horizontal, in a bi-directional approach. Note however, that RNN were intrinsically developed for sequences. To ensure a correct reconstruction of the image to its original dimensions, a partition layer was developed, being responsible for transforming the grid-structured input into a one-dimensional sequence capable of being applied to the GRU layer, and back to its original dimension.

The details of the proposed architecture are illustrated in Fig. 5.3. The convolutional layers are responsible for the generation of discriminative feature vectors. Afterwards, the feature maps are fed into the GRU layer to enforce the spatial context of the network. The last convolutional layer comprehend a kernel size of 1×1, to simultaneously reduce the feature space and combine the imaging information with the non-imaging clinical information.

In Fig. 5.3, blue rectangle shapes represent the computed feature maps, where the first dimension corresponds to the number of feature maps and the second dimension to the input patch size. The green rectangle shapes represent the output of the 2D-dimensional GRU layer, and the dashed line consists of a 2D cropping layer, applied before connecting the output of the U-Net into the GRU layer. Finally, the prediction is provided by the last layer, corresponding to the softmax activation.

5.2.3 Combining imaging with non-imaging data

Besides MRI data, non-imaging clinical information is also gathered during the acute phase of stroke, such as the Time Since Stroke (TSS), Time to Treatment (TTT), mRS score, and TICI score. TSS and TTT are time measures that mark the time-points when the stroke incident was diagnosed and when clinical intervention was performed. Although, only the TTT was available for all patients in the used dataset,

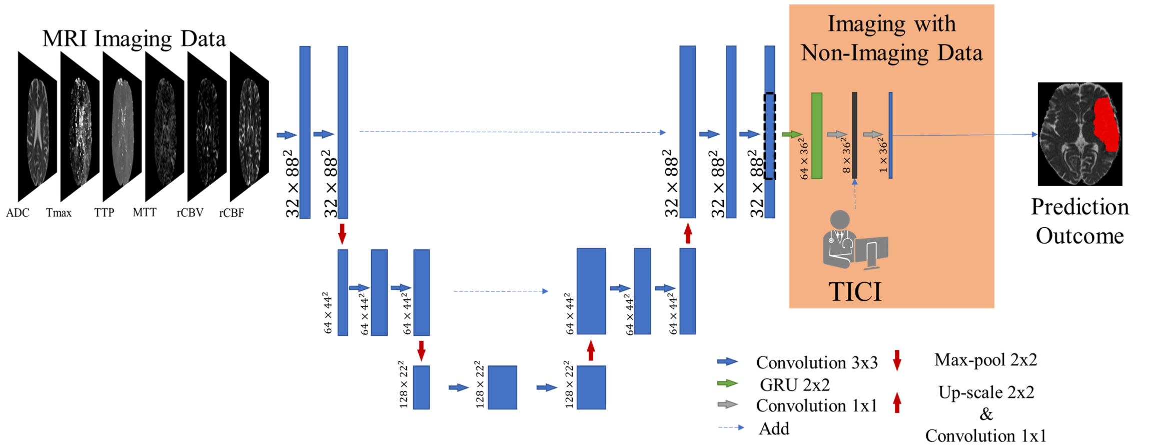


Figure 5.3: Overview of the proposed architecture. Blue feature maps result from 2D-dimensional convolutions.

since this variable is continuous and presents high variability, we refrained from using it in this work. Additionally, the number of patients available is insufficient to properly learn the temporal relationships between the TTT and the final infarct lesion. The mRS score characterizes the degree of disability 90-days after a stroke incidence. However, the most relevant factor is the TICI score (Higashida et al., 2003), which indicates the degree of success of the mechanical thrombectomy, based on cerebral angiography. Low TICI scores (TICI $\in \{0, 1\}$) describe cases with minimal perfusion or no perfusion at all. Mid-range TICI scores (TICI $\in \{2a, 2b\}$) characterize cases with progressively better partial perfusion. The highest TICI score (TICI $= 3$) characterizes a complete flow-restoration (Higashida et al., 2003). Consequently, it is expected that higher TICI scores naturally lead to increased levels of tissue being salvaged, and conversely, lower TICI scores might indicate increased levels of tissue loss. In our proposal, we aim to integrate this information in a deep neural network architecture, to relate imaging (e.g. stroke location, extension) with clinical information. We accomplish this integration by including the TICI information during the learning and testing phases of the method, being the TICI scale considered at a population-level and patient-level.

5.2.3.1 Population-level

In the presence or absence of perfusion beyond the location of the occlusion, stroke lesion extension can present changes between the TSS and the follow-up acquisitions. For cases with no perfusion, it is expected an infarct growth between the two time-points (the onset and the follow-up), while cases with existent perfusion should present a lower infarct growth rate, or even a stall in the infarct growth, leading to smaller lesion volumes. When the lesion shrinks, our method must learn that even though the lesion presents a larger extension at the onset MRI maps, it should produce a smaller segmentation, and when the lesion grows, it should learn to predict a larger segmentation, although the provided imaging information might indicate otherwise. Modelling these ischaemic stroke dynamics when predicting the final infarct lesion from the MRI parametric maps at the first exam to a future time is the focus of our work. Interpreting the lesion growth as oversegmentation, and the lesion shrinkage as undersegmentation in relation to the information provided by the MRI maps in the present time, we may interpret the oversegmentation as an increase in false positives (FP) and the shrinkage as an increase in false negatives (FN). Note however that, this information is not present in the MRI maps acquired at the first medical exam. Incorporating the clinical knowledge behind lesion growth/shrinkage is achieved, at a population-level, through a custom loss function, which drives the learning process to gradient optimizations conditioned to the clinical TICI score. This dynamic in our proposal is modelled by the F_β score that combines the Precision and Recall scores as follows:

$$F_\beta = (1+\beta^2)\frac{precision \times recall}{(\beta^2 \times precision) + recall}. \tag{5.1}$$

The Precision score, defined as $Precision = \frac{TP}{TP+FP}$, measures the presence of FP, while the Recall, given by $Recall = \frac{TP}{TP+FN}$, considers the presence of FN (TP corresponds to the number of true positives). Hence, Equation 5.1 can be rewritten as follows:

$$F_\beta = \frac{(1+\beta^2)\cdot TP}{(1+\beta^2)\cdot TP + \beta^2 \cdot FN + FP}. \tag{5.2}$$

Equation 5.2 provides an easier relationship between the ground-truth and the prediction. This relation is controlled by β, which in our proposal encodes the TICI score. To be applicable to a supervised learning approach, F_β needs to relate the predictions with the ground truth, which is defined in the following way:

$$F_\beta = (1+\beta^2)\frac{\sum_i^N p_i g_i}{\sum_i^N \beta^2 p_i^2 + \sum_i^N g_i^2}. \tag{5.3}$$

The sum is performed for the N voxels of the patch in the prediction, $p_i \in P$, and the ground truth, $g_i \in G$. The gradient of the F_β score for the j^{th} voxel prediction is computed as:

$$\frac{\delta F_\beta}{\delta p_j} = (1+\beta^2)\left(\frac{g_j(\sum_i^N \beta^2 p_i^2 + \sum_i^N g_i^2) - (2\beta^2 p_j)\sum_i^N p_i g_i}{(\sum_i^N \beta^2 p_i^2 + \sum_i^N g_i^2)^2}\right). \tag{5.4}$$

5.2.3.2 Patient-level

The inclusion of the TICI at a patient-level aims to drive the learning process to search for correlation between the imaging features extracted and the success of the clinical intervention. With this approach we hypothesize that the model should be aware that different TICI scores should predict different lesion outcomes, during the estimation phase. Therefore, our proposal would be capable of predicting the amount of salvageable tissue loss in the presence and absence of successful reperfusion. This property allows the clinician to explore different scenarios and study patients that can actually benefit from clinical intervention. The inclusion of the TICI score at a patient level is achieved by an extra channel before the final layer of the architecture (Fig. 5.3).

5.2.4 Post-processing

As post-processing step, we performed simple morphological filtering. Stroke lesions vary significantly in size. The post-processing should take this variation into account to avoid the complete removal of stroke lesions; therefore, a threshold to remove only connected components with less than 25 voxels was defined using cross-validation.

5.3 Experimental Set-up

We evaluated our proposal on the ISLES 2017 training and testing datasets, where the online platform also includes an automated evaluation of prediction results submitted to the online benchmark tool. In this work, we compared the performance of our proposal with and without using clinical meta-data.

5.3.1 Dataset

To evaluate our proposal, we used ISLES 2017 dataset, where Section 4.3 already provided detail on the MRI imaging information. Note however, that alongside the diffusion and perfusion parametric MRI maps and the perfusion DSC-MRI, each patient is also characterized by the TICI score, TSS, TTT, and mRS Score. Although other clinical information is available, only the TICI scores were used in this study. Table 5.1 describes the distribution of TICI score for each available dataset.

Table 5.1: TICI distribution for ISLES 2017 dataset.

	TICI 0	TICI 1	TICI 2a	TICI 2b	TICI 3
Training	6 (14%)	3 (7%)	3 (7%)	11 (26%)	20 (46%)
Testing	3 (9%)	2 (6%)	4 (13%)	6 (19%)	17 (53%)

As presented in Table 5.1, we observe that there is no equal representation for all the TICI scores. Furthermore, it allows us to conclude that the majority of the clinical interventions performed in patients of the training and testing sets were successful. To gain further acquaintance on the dataset, and the variability present on each TICI scale, we performed an estimation of the lesion variation at the onset time compared with the ground-truth delineated at the follow-up time-point (90-days after). The estimate of the ischaemic stroke lesion at the onset time was accomplished by applying thresholds to the ADC and T_{max} maps, alongside morphological filtering to ensure a final delineation without holes or disconnected elements. The choice of the thresholds was based on clinical knowledge, commonly employed, with the purpose of attaining a rough lesion location and delineation (Austein et al., 2016; Straka et al., 2010). Afterwards, we computed the lesion variation from both time-points and grouped each case per TICI score, obtaining a final average lesion variation for each scale. Fig. 5.4 illustrates the obtained results.

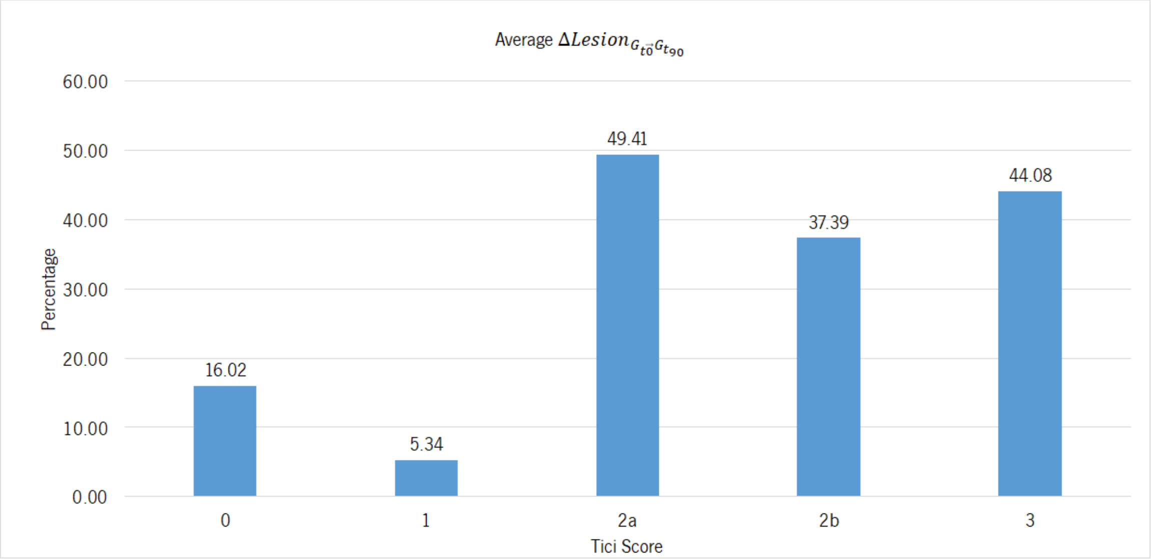


Figure 5.4: Average percentage lesion variation across the onset time and the follow-up time for each TICI score class.

As can be seen, successfully revascularized patients, which have a TICI score higher than $2a$, demonstrate higher average lesion variation, when compared to the non-reperfused patients (TICI bellow 1). This behaviour is the motivation for considering non-imaging information with imaging information in a deep learning-based method. Note however, that due to the representativeness of some TICI classes, we decided to merge TICI scores into three different ranges, following the clinical reasoning behind the reperfusion.

5.3.2 Evaluation

The performance of each method was evaluated using five metrics already described in Chapter 4, Section 4.3.2.

5.3.3 Set-up

The validation set comprised 7 cases, while the training set encompassed the remaining 36 cases from ISLES 2017 training set. As for the testing set, it comprehends 32 cases. To assess the added value of our contributions, we perform a 7-fold-cross-validation scheme within the training set. We compare our proposal with a baseline architecture, which does not encompass any clinical meta-data. In addition, we changed the loss function to the soft dice (Milletari et al., 2016), which is a standard loss function for segmentation tasks. Furthermore, we also report that the training of our deep neural network architecture with categorical cross-entropy was not possible. This finding goes with the encounter of Choi et al. (2016).

5.3.4 Model training & parameters

For each subject, 500 patches of size 88×88 were extracted, using a uniform random sampling scheme. We also employed a data augmentation scheme that encompasses rotations of $90°$, $180°$, $270°$.

The network was trained with Adam optimizer (Kingma and Ba, 2014) (learning rate of 1×10^{-5}) using a mini-batch size of 4 during 160 epochs. We employed spatial drop-out (Tompson et al., 2015) with a probability of 0.25, at each two convolutional layers. The work was implemented on Keras (Chollet, 2015), with Theano backend. All tests were conducted on a workstation equipped with a GeForce GTX 1070 with 8 GB. For each patient, prediction took around 15 seconds.

5.3.4.1 Inclusion of clinical information

When considering cases where the TICI score is low, and the onset infarct core lesion will evolve over time, having the capability to predict the maximal extent of the infarct core tissue, will provide the clinicians the worst clinical scenario, and consequently clinicians may ponder to perform clinical intervention, to decrease the chances of increasing the tissue death by hypo-perfusion. In such circumstances, with the inclusion of the TICI score we aim to drive the model to predict the worst-case scenario of stroke lesion outcome. Conversely, in a case with a high TICI score we would prefer a prediction where the recovered hypo-perfused tissue due to reperfusion is achieved with success, holding on the same principles as before. It is worth mentioning that such relationship is further affected by several other clinical and patient-specific pathophysiological aspects, such as collateral flood, onset time of the stroke, cardiovascular conditions and others.

Giving the available number of cases per TICI in ISLES 2017 dataset, we merged TICI scores, increasing the number of cases per score. Therefore, at a population level, β in Equation 5.5 encodes the TICI score as follows:

$$\beta = \begin{cases} 2, & \text{if } TICI \in \{0, 1\} \\ 1, & \text{if } TICI \in \{2, 2a, 2b\} \\ 0.5, & \text{if } TICI = 3 \end{cases} \tag{5.5}$$

In this way, for TICI$= 3$ (i.e. complete perfusion) we defined $\beta = 0.5$, so recall is weighted four times less than precision. Hence, we drive the model to give higher importance to the expression of false positives rather than false negatives, preferring scenarios with low tissue loss. Conversely, for TICI $\in \{0, 1\}$ (i.e. poor recanalisation), we defined a $\beta = 2$, where recall is weighted four times higher than precision. For such cases, the motivation is to give preference to high tissue loss. Finally, for TICI $\in \{2a, 2b\}$ the value of $\beta = 1$, obtaining the F1-score commonly known as the Dice Score, where precision and recall are equally taken into consideration. This scale of β was defined through cross-validation.

5.4 Results and Discussion

In this section, we first evaluate the main contribution of our proposal in the training set. Using cross-validation we compare the performance of the baseline method without non-imaging clinical information against our proposal. Afterwards, we present the results obtained in ISLES 2017 testing dataset, performing a comparison with state-of-the-art methods.

5.4.1 Incorporation of non-imaging clinical information

Due to the large diversity of appearance, size and shape, the tissue outcome prediction presents as a challenging task (Maier et al., 2015). In this study, we show the importance of having non-imaging clinical information in a neural network, to characterize principal and collateral blood flow haemodynamic and obtain better prediction outcomes. The results for the training set are presented in Table 5.2.

Table 5.2: Results obtained through cross-validation in ISLES 2017 training dataset for the baseline method and our proposal. Each metric contains the average ± standard deviation.

	Dice	Hausdorff Distance	ASSD	Precision	Recall
Baseline	0.34 ± 0.22	35.09 ± 17.27	6.08 ± 5.27	0.37 ± 0.29	0.54 ± 0.26
Proposal	0.35 ± 0.22	31.38 ± 15.81	5.55 ± 5.00	0.41 ± 0.30	0.47 ± 0.24

In the cross-validation study, when comparing with the baseline, our proposal is capable of achieving higher average Dice and lower Hausdorff Distance and ASSD. However, in this study, the gain is not considerably high, not allowing us to demonstrate the added value of incorporating the TICI score into the neural network. Considering the average precision and recall metrics, our proposal achieved higher precision but lower recall. This suggests a higher capability to perform stroke lesion outcome prediction, by depicting gradual changes in the hypo-perfused tissue. We hypothesize that making the model aware to intrinsic biological phenomena of lesion growth or shrinkage (TICI dependent) lead to more precise predictions, which is sustained by the lower average values of distance metrics and higher average Dice score.

However, in clinical practice the TICI score is only obtained after recanalisation. Being so, predicting the stroke lesion at a 90-day follow-up, during the sub-acute phase, needs to consider different reperfusion scenarios. In our proposal, we grant such property at patient-level domain. By adding an extra input channel that contains the TICI score, we aim to obtain tissue outcome predictions with successful and unsuccessful reperfusion scenarios. When accessing both case scenarios, during the decision-making process, our method could provide to clinicians additional information on the salvaged tissue if mechanical thrombectomy was performed with success or not. In Fig. 5.5 and Fig. 5.6, we show the added value of incorporating clinical information on two patients with different TICI scores: one with an unsuccessful reperfusion (TICI=0), and one with a successful reperfusion (TICI=3).

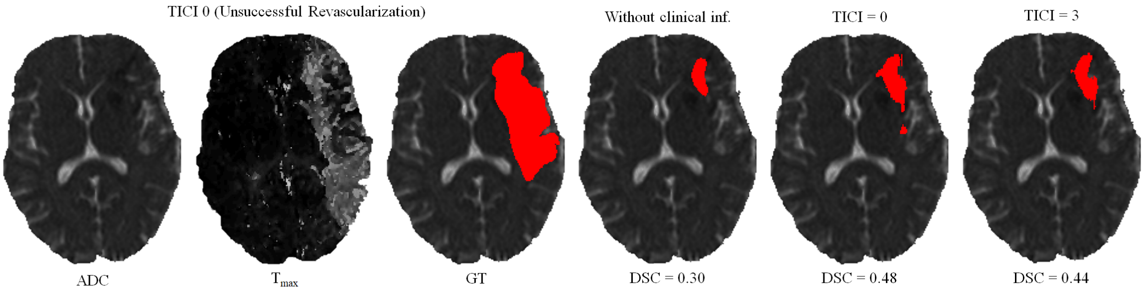


Figure 5.5: Example case of stroke lesion outcome prediction, with and without non-imaging clinical information in a patient with unsuccessful reperfusion. For sake of description we present the ADC and T_{max} maps and the GT. In the presence of clinical information, we show the two possible outcomes: unsuccessful (TICI=0) and successful reperfusion (TICI=3), respectively.

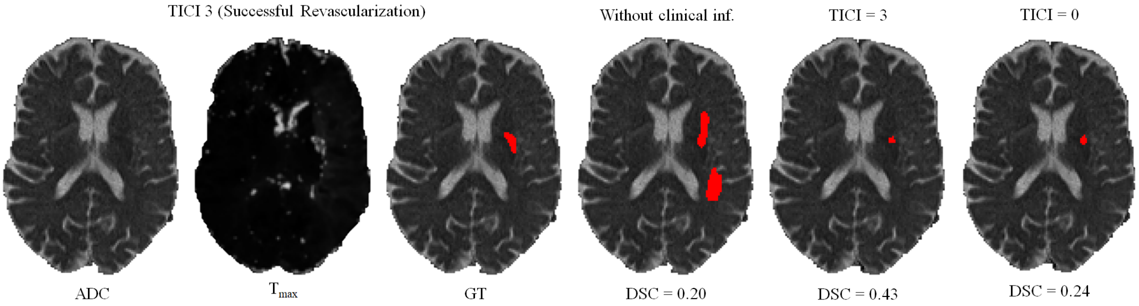


Figure 5.6: Example case of stroke lesion outcome prediction, with and without non-imaging clinical information in a patient with successful reperfusion. We also present the ADC and T_{max} maps and the GT. In the presence of clinical information, we show the two possible outcomes: successful (TICI=3) and unsuccessful reperfusion (TICI=0), respectively.

For each case, we present the tissue outcome predictions with and without non-imaging clinical information. In the absence of the TICI score, the tissue outcome prediction performs worse than our proposal, for both cases. Our proposal is capable of employing the TICI score to yield better predictions, which are corroborated by higher Dice scores, but also provides a result that is physiologically more plausible. Observing the stroke lesion outcome predictions of our proposal against the baseline, it is noticeable the presence of physiologically infeasible isolated regions in the latter. Additionally, we also tested if our method was capable of predicting different lesion outcomes by changing the TICI score. When changing the TICI score, we obtained different lesion outcomes for each patient. Furthermore, such scenarios agreed with the expected outcome describe for each TICI score (e.g. by changing from a TICI score of 3 to 0 it was observed a larger lesion outcome volume). From the latter study, we show that our proposal gained awareness to scenarios of no-perfusion and complete perfusion. This capability could provide the clinicians useful insight on the benefits and risks associated to the mechanical thrombectomy. Moreover, it can also be used to forecast recovery, which is important for patient treatment and the complete standard care associated to patient recovery. To corroborate our qualitative analysis, Table 5.3 contains the ground-truth lesion volume for each case, alongside the predicted volume outcome for the original TICI score and for the opposite case scenario, respectively.

On Table 5.3 we demonstrate the effect of the TICI score in our proposal. When changing the TICI score we observe different stroke lesion outcome predictions, in agreement to the reperfusion success. When increasing the TICI score the volume of salvaged hypo-perfused tissue becomes higher, which corresponds to a stroke lesion shrinkage. Case 24, with TICI$= 0$, illustrates this behaviour. After increasing the TICI score to TICI$= 3$, we obtain a smaller stroke lesion volume. As for case 42 with TICI$= 3$, when we decrease the TICI score from TICI$= 3$ to TICI$= 0$ the prediction volume characterized the opposite phenomena. With TICI$= 0$ there is higher hypo-perfused tissue loss, and the final infarct volume predicted is larger. From both case scenarios, the observed changes in the final infarct volumes predicted shows that the TICI score was capable of driving the tissue outcome prediction scenario, and simultaneously grant a lesion growth or shrinkage in accordance with the physiological dynamics of each TICI score and without infeasible isolated regions.

Table 5.3: Results obtained by our proposal on two patient cases with different TICI scores, alongside the obtained result after changing the original TICI score to its opposite (marked with a *).

Case	GT volume (voxels)	TICI	Dice	Precision	Recall	Predicted volume (voxels)
24	21310	0	0.48	0.87	0.33	8170
		3*	0.44	0.90	0.29	6840
42	288	3	0.43	0.59	0.33	163
		0*	0.24	0.17	0.39	651

5.4.2 State-of-the-art: ISLES 2017 Challenge

In Table 5.4 we compare our proposal with methods from ISLES 2017 testing dataset, evaluated by the online platform (SMIR, 2017) grouped by ensemble and non-ensemble methods, and ordered decreasingly by the average Dice score. To reinforce our analysis, we also included the baseline method.

Table 5.4: Recently published methods in ISLES 2017 testing dataset and our proposal. Each metric is represented by the mean ± standard deviation. Underlined values correspond to the highest mean.

		Dice	HD	ASSD	Precision	Recall
Ensemble	Mok et al. *	0.32 ± 0.23	40.74 ± 27.23	8.97 ± 9.52	0.34 ± 0.27	0.39 ± 0.27
	Kwon et al. *	0.31 ± 0.23	45.26 ± 21.04	7.91 ± 7.31	0.36 ± 0.27	0.45 ± 0.30
	Robben et al. *	0.27 ± 0.22	37.84 ± 17.75	6.72 ± 4.10	0.44 ± 0.32	0.39 ± 0.31
	Pisov et al. *	0.27 ± 0.20	49.24 ± 32.15	9.49 ± 10.56	0.31 ± 0.27	0.39 ± 029
Single Model	Monteiro et al. *	0.30 ± 0.22	46.60 ± 17.50	6.31 ± 4.05	0.34 ± 0.27	0.51 ± 0.30
	Lucas et al. *	0.29 ± 0.21	33.85 ± 16.82	6.81 ± 7.18	0.34 ± 0.26	0.51 ± 0.32
	Choi et al. *	0.28 ± 0.22	43.89 ± 20.70	8.88± 8.19	0.36 ± 0.31	0.41 ± 0.31
	Niu et al. *	0.26 ± 0.20	48.88 ± 11.20	6.26 ± 3.02	0.28 ± 0.25	0.56 ± 0.26
	Sedlar et al. *	0.20 ± 0.19	58.30 ± 20.02	11.19 ± 9.10	0.23 ± 0.24	0.40 ± 0.29
	Rivera et al. *	0.19 ± 0.16	63.58 ± 18.58	11.13 ± 7.89	0.27 ± 0.25	0.21 ± 0.17
	Islam et al. *	0.19 ± 0.18	64.15 ± 28.51	14.17 ± 15.80	0.29 ± 0.28	0.25 ± 0.25
	Chengwei et al. *	0.18 ± 0.17	65.95 ± 25.94	9.22 ± 6.99	0.37 ± 0.30	0.21 ± 0.23
	Yoon et al. *	0.17 ± 0.16	45.23 ± 19.14	12.43 ± 11.01	0.23 ± 0.27	0.36 ± 0.32
	Baseline	0.24 ± 0.20	53.29 ± 26.95	10.59 ± 4.98	0.27 ± 0.27	0.50 ± 0.35
	Proposal	0.29 ± 0.22	47.17 ± 22.13	7.20 ± 4.14	0.26 ± 0.23	0.61 ± 0.28

* Results retrieved from (Winzeck et al., 2018).

Incorporating clinical information through the proposed custom loss function and the extra TICI channel resulted in a higher performance, in comparison to the baseline. Our proposal was able extract information from non-imaging data and to drive its training and testing phases towards better predictions. Therefore, the simultaneous incorporation of the reperfusion status, as an additional feature and in the loss function, improved the performance of the classifier. In addition, we show the higher generalization capability of our proposal, since the performance metrics or our proposal for both datasets present less variation.

Although a previous work (McKinley et al., 2016) had investigated the use of non-imaging clinical information to conduct the training of machine learning methods, such information has not been evaluated directly in the context of deep learning-based methods. The results on the ISLES 2017 indicate the benefits of incorporating non-imaging clinical information in a deep neural network architecture, implicitly during the training phase and explicitly by extra channels, incorporating patient-specific information.

When comparing to the state-of-the-art methods, our proposal can reach competitive results, being placed among top scoring methods. As a single method approach, our proposal yields results within the top five methods, alongside ensemble approaches (e.g. Choi et al. (2016)). In the same group, our method achieved the highest average recall metric, with lower average precision. As for the distance metrics, our proposal can provide competitive ASSD score, with low standard deviation, and a Hausdorff Distance among of top methods. We emphasize that, as post-processing step, our method only applies a simple morphological removal of small connected components. Therefore, elaborate schemes of post-processing such as Conditional Random Fields or even weighted schemes of ensemble can boost the performance of such approaches. Even in such cases, our approach provides a good and precise estimation of the final stroke lesion. To enforce such analysis in Fig. 5.7, we show the average DSC score and the Hausdorff Distance obtained by each state-of-the-art method in ISLES 2017 testing dataset. Besides our proposal, we included the baseline method.

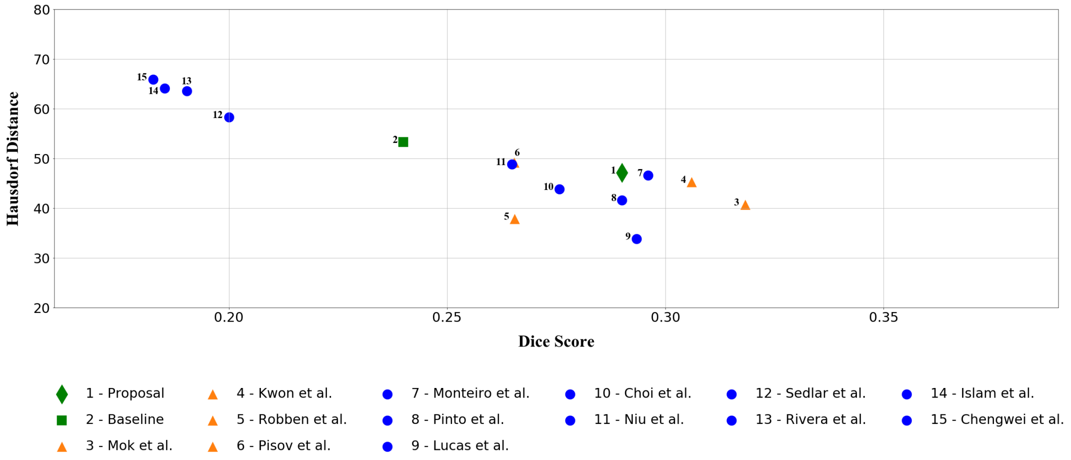


Figure 5.7: Hausdorff Distance versus Dice score from methods of ISLES 2017 in the testing set.

In Fig. 5.7, we can observe the overall increase in robustness of our proposal over the baseline, when considering the average Dice versus the average HD. Moreover, our proposal was capable of reaching average DSC and HD metrics similar of top scoring methods. Note that closer to the horizontal axis and further away from the origin is better (i.e. high Dice and low Hausdorff). Ensemble methods are marked with a triangle shape.

However, there is still room from improvement since none of the current state-of-the-art methods, provides the robustness and accuracy needed for clinical practice, and are currently bellow the inter-rater performance of expert radiologists (DSC$= 0.58$) (Winzeck et al., 2018). Furthermore, we argue the need for more imaging and non-imaging data in order to characterize the underlying dynamics in predicting the final ischaemic stroke lesion, and at the same time to decrease the variability and susceptibility of the TICI score present in the evaluated dataset (presented in Table 5.1). Jung et al. (2013) demonstrated that the great majority of the cases of ISLES 2017 dataset had a high inter-rater variability for reperfusion (kappa 0.81). Recently, Robben et al. (2018) showed evidence on the importance of non-imaging information in a larger dataset, only using CTP perfusion images. In the future, we would like to investigate on adding other clinical information, such as TTT and TSS. We esteem that the proposed approach can be further applied to other diseases where clinical information complements imaging information.

5.5 Summary

Prediction of the final infarct lesion in ischaemic stroke patients has the potential to assist physicians when assessing the risks and benefits associated to mechanical thrombectomy. Therefore, having a model that can provide useful information during the clinical decision process.

In this chapter, we propose a novel deep neural network architecture that beyond previously proposed architectures incorporates clinical information in a principled way. Our proposal integrates clinical information at two different levels of the architecture. The first level considers the population domain-knowledge, achieved through the development of a custom loss function, to depict relationships between the TICI score and the tissue outcome prediction. The second level considers the patient-specific domain, where the TICI is encoded into an input channel of the architecture. From the latter level, we demonstrated that our proposal was able to characterize different outcome scenarios of successful and unsuccessful reperfusion. This method presents itself as a tool with potential to assess the risks and benefits associated to the mechanical thrombectomy. The evaluation of our proposal was conducted on the publicly available ISLES 2017 online benchmark tool. We observe that the proposed method has benefited from the combination of imaging and non-imaging information. In addition, when comparing to the state-of-the-art methods, we observed that a single architecture with fewer parameters, such as ours, yields competitive performance results similar to more elaborate and/or ensemble methods.

Chapter 6

Combining unsupervised and supervised learning for stroke tissue outcome prediction

In the context of predicting the final infarct stroke lesion from onset MRI acquisitions, principal and collateral blood flow has been either considered directly by modeling the temporal perfusion imaging (shown in Chapter 4), indirectly by perfusion and diffusion parametric maps (Choi et al., 2016; Maier et al., 2017; Scalzo et al., 2012), or through clinical information that characterises the success of the revascularization, by dichotomizing the training set (McKinley et al., 2016) or guiding the learning process of a Machine Learning method (Chapter 5). We hypothesize that modeling the hemodynamics of the brain, when artery occlusion occurs, can be beneficial in predicting the final stroke lesion. So, in this work, we propose to model such hemodynamics with an unsupervised learning model. Contrary to previous approaches, we propose that modeling different input groups of the time-resolved perfusion maps (i.e. T_{max}, TTP, MTT), and of the blood-flow-dynamic related maps (i.e. rCBF, rCBV) can lead to a better stroke lesion outcome prediction.

The chapter is organized as follows. First, we address the motivation and the main contributions. Section 6.2 describes the fundamental components of the proposed method. Section 6.3 describes the database used, the evaluation performed and the set-up. Results and the discussion are shown in Section 6.4. Finally, in Section 6.5 we present the main conclusions of this chapter.

6.1 Motivation

This chapter presents an automatic method based on unsupervised and supervised deep methods. As motivation, we know that unsupervised methods learn structural features when encoding and decoding the original image, while the supervised methods learn features conditioned on the label, so there is potential for obtaining richer and more discriminative features by joining both types of methods. Thus, the research conducted in this chapter employs RBMs in a two-pathway approach, to extract structural features from time-resolved parametric maps and blood-flow-dynamics of parametric maps. One subset encompasses the TTP, MTT, T_{max}, and ADC. The second set contains the ADC, the rCBV and rCBF. In a second stage, the extracted structural features are combined with the standard parametric maps to form the input of a supervised deep neural network architecture composed by Convolutional Neural Networks and Recurrent Neural Networks.

One contribution of the work presented in this chapter is the use of unsupervised methods for extracting structural features of time-resolved perfusion and blood-flow-dynamic related MRI maps for predicting stroke lesion. Additionally, we can identify two other contributions. First, the use of long spatial context provided by gated recurrent neural networks for relating structural features and image information, when learning features conditioned on the label in a supervised method. Second, the proposal of a competitive system which outperforms state-of-the-art methods to predict the final infarct stroke lesion, in ISLES Challenge testing set.

6.2 Methods

Once again, predicting the final infarct lesion consists of delineating the lesion's spatial extension at a 90-day follow-up time-point, using the multi-parametric MRI imaging ADC, MTT, TTP, T_{max}, rCBF, and rCBV, which are acquired at the onset time-point. The architecture of the proposed system and its main components are described in the following subsections.

6.2.1 Deep neural network architecture

The overall architecture of the proposed method can be divided into two functional blocks illustrated in Fig. 6.1.

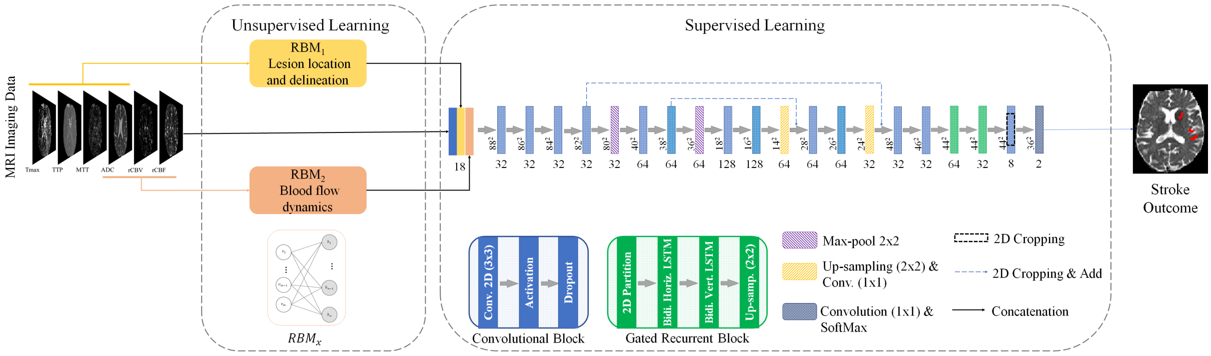


Figure 6.1: Overview of the proposed method for stroke lesion outcome prediction.

In the proposed architecture, the first functional block performs unsupervised representation learning using two unsupervised models, namely RBMs. This unsupervised block provides new features that represent structural information that complements the standard parametric MRI maps, enhancing the capacity of our model to predict the final infarct lesion volume. In our approach, we aim to model the clinical procedure, which first locates and delineates the lesion at current time, and then considers the blood flow haemodynamic that might influence the final stroke lesion prediction. This procedure is encoded in our two-path RBM. The first RBM is responsible for capturing information on lesion location and extension, referred to as the $\text{RBM}_{\text{Lesion}}$. The second RBM, $\text{RBM}_{\text{Haemo}}$, aims to capture blood flow haemodynamics information (e.g. collateral circulation), which has been identified as a key factor by physicians when assessing stroke final infarct lesion in clinical reports (Berkhemer et al., 2016; Menon et al., 2015). On one hand, to locate the onset ischaemic stroke lesion, the RBM$_{\text{Lesion}}$ considers standard parametric maps that characterise the brain perfusion rates. In the presence of an ischaemic lesion, the occluded vessel can decrease or interrupt the normal brain perfusion, translating into hyperintense regions on time-related parametric maps (Butcher and Emery, 2010b). On the other hand, the RBM$_{\text{Haemo}}$ considers standard parametric maps that characterise the amount of blood being delivered in unit of time, which correlates to the cerebral blood flow haemodynamics (Butcher and Emery, 2010b). Thus, the RBM$_{\text{Lesion}}$ considers the MTT, TTP and T$_{\text{max}}$ perfusion maps, while the RBM$_{\text{Haemo}}$ the rCBV and rCBF perfusion maps. Regarding the ADC standard diffusion map, it is present in both RBM$_{\text{Lesion}}$ and RBM$_{\text{Haemo}}$, since it provides higher brain structural information and allows the identification of tissue that is already infarcted. This separation of the input imaging allows the RBM to learn specific feature sets, which may enable the method to analyse difficult cases where information concerning the blood flow can have a favourable impact on the lesion outcome.

The second functional block consists of a deep neural network architecture that comprehends 2D convolutional blocks in a U-net structure, alongside recurrent blocks. As imaging input data, we combine the standard parametric maps with feature maps from each RBM, totalling 18 input feature maps.

6.2.2 Restricted Boltzmann Machines

The RBM is an undirected graphical model constituted by two layers of nodes: a visible layer and a hidden layer (Rumelhart and McClelland, 1986). Each node has a weighted connection to all nodes in the other layer (Rumelhart and McClelland, 1986). However, there are no connections among nodes of the same layer. Originally, Rumelhart and McClelland (1986) proposed RBMs to learn from binary data on both layers. However, this does not represent well continuous real-valued input data, which is the case of MRI data. Therefore, we model the visible nodes as linear units with independent Gaussian noise. The hidden nodes are modelled as Noisy Rectifier Linear Units (NReLU), since they have been reported to be suitable for feature extraction (Hinton, 2012). This kind of RBM was previously used in segmentation tasks, such as in Pereira et al. (2018a). Mapping the input data into a feature vector is performed through the interaction of states between the visible and hidden units, which is learned by minimizing an energy function.

The RBM$_{\text{Lesion}}$ and RBM$_{\text{Haemo}}$ function as feature generators that output two complementary sets of feature maps $\mathcal{N}_1$ and $\mathcal{N}_2$. These features characterise the structure of the images; however, we are interested only on the most distinctive details. So, after training the RBMs, we perform feature selection to reduce the generated feature space, obtaining smaller but representative feature sets $\mathcal{M}_1$ and $\mathcal{M}_2$, such that $|\mathcal{M}_i| \ll |\mathcal{N}_i|$, for $i \in [0, 1]$, where the operator $|.|$ denotes the cardinality of a set. In the literature there are several methods for feature selection (Chandrashekar and Sahin, 2014). In this work, the feature selection step was inspired on the method proposed by Pereira et al. (2018b). Hence, we start by computing the Normalized Mutual Information between each feature map from the two generated feature sets and each MRI map of the respective input, to quantify the statistical dependence between the generated features and each MRI sequence. Afterwards, each feature is ranked decreasingly according to the Normalized Mutual Information, allowing the selection of a subset $\mathcal{M}_i$ of relevant features. We also use a supervised RF classifier trained with mean decrease impurity to assess the selection of features $\mathcal{M}_i$ for each set, and to verify the representation capacity of the selected features (Pereira et al., 2018b).

6.2.3 Convolutional and Recurrent Neural Networks

Our supervised functional block is based on the U-Net architecture as proposed by Ronneberger et al. (2015). The input of the U-Net considers the concatenation of standard parametric maps with the sets of feature maps extracted from the unsupervised block. In the first level of our encoder architecture we use four 2D convolutional blocks with kernel size of 3×3 and 32 channels. Afterwards, the output of the final convolutional block is down-sampled by a factor of 2, starting the second encoding level, formed by two convolutional blocks with equal kernel size but doubling the number of feature maps. The third level of encoding follows the same pattern. The decoder level mimics the encoder counterpart. As in Ronneberger et al. (2015) we only used long skip connections among encoder and decoder levels. These encoder-decoder deep CNNs provide high levels of abstraction from the input data, increasing the global notion of context as the network grows deeper. However, it comes at a cost of a high receptive field (Zeiler and Fergus, 2014). Thus, we used a 2D architecture in the plane with the highest resolution, since the acquisition resolution is anisotropic in the dataset. Also, in the end of the decoding path we expanded our learning block with Gated Recurrent Neural Networks (Gated RNNs). Due to their nature, Gated RNNs can capture short- and long-term spatial relations, by retaining information from previous nodes encoded in the time-steps. Hence, Gated-RNNs consider information from all previous nodes when analysing the current one. This property, when applied to imaging data, allows considering intra-slice contextual dependencies needed for the prediction of stroke lesions. In our work, we used a particular Gated-RNN, namely the LSTM (Hochreiter and Schmidhuber, 1997). However, the LSTM was intrinsically developed to process 1D data (Hochreiter and Schmidhuber, 1997) (e.g. time-series). To be applicable to 2D data, we developed an online 2D Partition layer that transforms a grid-structure input (e.g. an image) into an one-dimensional sequence. Inspired by Visin et al. (2016), the 2D Partition layer was predefined with a neighbourhood of 2×2, where each time-step is characterised by the feature space of four voxels. After, two Bidirectional LSTM layers are employed along the vertical and horizontal directions followed by an Up-sampling layer. These four layers are referred as the Gated Recurrent Block are depicted in Fig. 6.1. In our supervised functional block, two Gated Recurrent Blocks were used, where the Bidirectional LSTMs have 64 and 32 hidden layers, respectively. The impact of the main components is evaluated in an ablation study in the experiments.

6.3 Experimental Setup

We evaluated the proposed approach on the publicly available ISLES 2017 dataset. ISLES dataset has an online benchmark platform (Kistler et al., 2013) that performs automatic evaluation (SMIR, 2017). In this section we describe the dataset, the training and evaluation, and the main hyper-parameters of our method.

6.3.1 Data & Evaluation Metrics

To evaluate the value of our proposals, we used ISLES 2017 dataset, where Section 4.3 already provided the complete details of this dataset. Fig. 6.2 (top row) illustrates an example patient characterised by MRI maps, alongside the manual lesion outcome, the GT.

As for the evaluation procedure, it was kept the same as discussed in Section 4.3.2 using five metrics as in the online ISLES benchmark platform, allowing a fair comparison with other competitors.

6.3.2 Image pre- and post-processing

Since MRI acquisitions were acquired from different centers and configurations (Winzeck et al., 2018), for each patient we resized all maps to a common volume of dimension of $256 \times 256 \times 32$. Afterwards, the ADC maps were clipped between $[0, 2600] \times 10^{-6} mm^2/s$ and the T$_{\text{max}}$ maps were clipped to $[0, 20s]$, since values beyond these ranges are known to be biologically meaningless (Rose et al., 2001). Finally, a linear scaling was applied across all maps, to the range $[0, 255]$. The images are resized to its original size, after we perform the prediction.

As for the postprocessing, we applied a morphological filtering. Since stroke lesion outcome presents a wide variety of lesion volumes (McKinley et al., 2016), we focused on removing only small connected components with less than 25 voxels. This step was kept fixed for all the evaluated models.

6.3.2.1 Data Augmentation

Data augmentation can be used to increase the number of training samples and reduce over-fitting (Krizhevsky et al., 2012). Due to the relatively small training dataset of stroke lesion outcome prediction, we employed artificial data augmentation in the supervised portion of our proposal. For each sample, we employed rotations of $90°$, $180°$, $270°$.

6.3.3 Model training & parameters

The unsupervised functional block was trained by optimizing the negative log-likelihood of the data. However, since computing such gradient is generally intractable, we performed the training by approximating the gradient with Contrastive Divergence with one step of alternating Gibbs sampling (Hinton, 2012). The training process of an RBM can be difficult if one tries to learn the parameter σ. According to Hinton (2012), we normalize each component of the data with zero mean and unit variance, and define $\sigma_i = 1$. In Table 6.1, we present the settings used for the training of the unsupervised model. In each RBM, 3D image patches of shape $7 \times 7 \times 3$ are extracted from a set of MRI maps, $\mathcal{C}$. Then, the 3D patches are reshaped into a 1D vector and fed into the visible layer, having an input of size $m = 7 \cdot 7 \cdot 3 \cdot |\mathcal{C}|$.

After training, we extract features from the NReLU units noise-free activations. Such units exhibit intensity equivariance when the bias has zero value, and they are noise free units (Nair and Hinton, 2010). Due to the large number of feature maps extracted ($|\mathcal{N}_1| = |\mathcal{N}_2| = 600$), we perform a feature selection to reduce the feature space. The most appropriate number of features will be discussed in the following section. In this work, the unsupervised block encompasses RBMs with different sets of MRI images.

As for the supervised functional block, the settings of the training are also given in Table 6.1.

Table 6.1: Model training parameters for the unsupervised and supervised functional blocks.

Functional Block	Parameter	Description
Unsupervised	Optimizer	SGD with momentum($lr = 1 \times 10^{-5}$)
	Weight Decay	$L_1 = 2 \times 10^{-6}$, $L_2 = 2 \times 10^{-4}$
	Patch size	$7 \cdot 7 \cdot 3$
	Batch size	32
Supervised	Optimizer	ADAM($lr = 1 \times 10^{-5}$)
	Patch size	88×88
	Batch size	4

The loss function used was the soft-dice loss defined as Milletari et al. (2016).

$$\text{Soft Dice loss} = \frac{\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}, \tag{6.1}$$

In the soft dice loss, the sum occurs over the set N of voxels belonging to the predicted output patch, where $p_i \in \mathcal{P}$ denotes the probability of a voxel i in the output patch and $g_i \in \mathcal{G}$ corresponds to the respective ground-truth label voxel.

The method was implemented using Keras with Tensorflow backend, in a workstation equipped with GTX 1080 Ti 11 GB. Prediction time takes around 20 seconds per patient.

6.4 Results and Discussion

In this section, we discuss the impact of the main contributions, namely the incorporation of unsupervised learning with supervised learning and the importance of the Gated Recurrent blocks. Then, we compare our method with state-of-the-art in ISLES 2017 Challenge. Finally, we delve on the difficulty of predicting the final infarct stroke lesion.

6.4.1 Ablation Study

The ablation study aims to gradually measure the importance of the main components and consequently assert on which components contributed to the overall performance. Thus, we start by evaluating the importance of the unsupervised feature generator and the proposed input grouping. After, we focus on the use of the Gated Recurrent Block and the choice of the dimensionality of the spatial context.

6.4.1.1 Unsupervised feature generation

We hypothesize that grouping the parametric MRI maps according to their physical meaning and encoding each group with a RBM has the potential to extract better features to characterise the stroke lesion and the blood haemodynamics. We perform several experiments to corroborate this working hypothesis. In all experiments, the parametric MRI maps are also used as input to the supervised block. Fig. 6.2 illustrates the feature maps encoded by the RBMs and the respective MRI maps. The results are presented in Table 6.2.

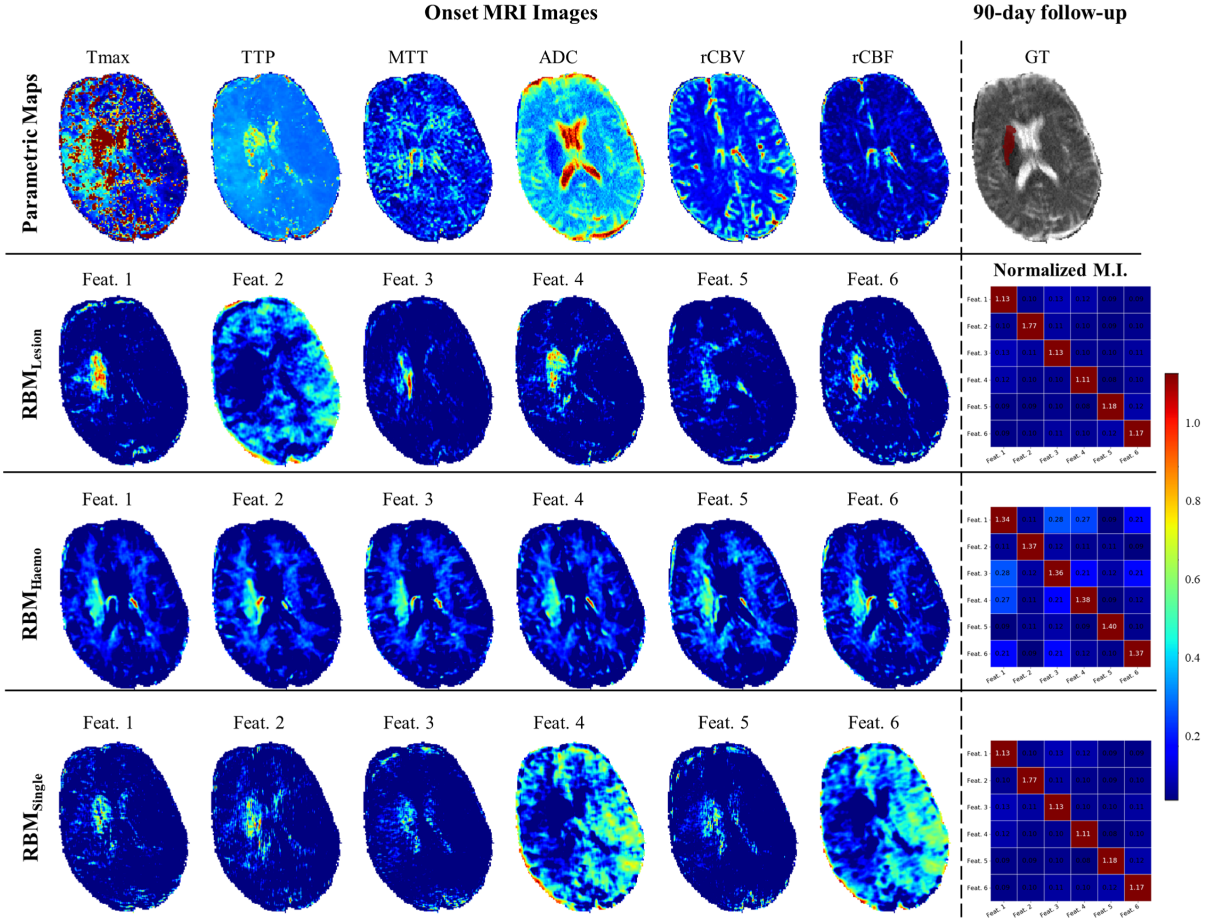


Figure 6.2: Onset parametric maps of Training case 11, alongside the final stroke lesion, at a 90-day follow-up, over the onset ADC map. The subsequent rows show the RBM features selected from the RBM_{Lesion}, RBM_{Haemo} and RBM_{Single}, respectively. The last column corresponds to normalized mutual information, across whole dataset, among features of the same RBM.

Table 6.2: Results obtained with different configurations of the unsupervised feature generator block in ISLES 2017 testing set. Each metric represents the mean ± standard deviation. Underlined values correspond to the highest mean.

Unsupervised Block		Supervised Block			Dice	HD	ASSD	Precision	Recall
Method	Params.	FCNN	G-RNN	Params.					
–		U-Net	LSTM	519 034	0.30 ± 0.21	36.58 ± 16.62	6.96 ± 5.08	0.30 ± 0.26	0.55 ± 0.31
$\text{RBM}_{\text{Single}}$ (3 Feat.) [3D]	519 898	U-Net	LSTM	519 034	0.30 ± 0.21	38.93 ± 18.80	6.55 ± 4.22	0.29 ± 0.24	0.61 ± 0.31
$\text{RBM}_{\text{Single}}$ (6 Feat.) [3D]	520 762	U-Net	LSTM	520 762	0.30 ± 0.21	36.94 ± 19.19	6.72 ± 4.43	0.29 ± 0.24	0.59 ± 0.31
$\text{RBM}_{\text{Single}}$ (12 Feat.) [3D]	522 400	U-Net	LSTM	519 034	0.28 ± 0.20	41.07 ± 18.67	6.81 ± 3.88	0.24 ± 0.21	0.65 ± 0.30
$\text{RBM}_{\text{Haemo}}$ [3D]	264 600	U-Net	LSTM	520 762	0.28 ± 0.24	38.50 ± 22.78	11.09 ± 14.79	0.35 ± 0.30	0.44 ± 0.34
$\text{RBM}_{\text{Lesion}}$ [3D]	520 762	U-Net	LSTM	519 034	0.31 ± 0.21	35.38 ± 15.75	6.44 ± 4.43	0.30 ± 0.24	0.59 ± 0.30
$\text{RBM}_{\text{Lesion}} + \text{RBM}_{\text{Haemo}}$ [3D]	617 400	U-Net	LSTM	522 490	$\underline{0.38} \pm 0.22$	$\underline{29.21} \pm 15.04$	$\underline{5.52} \pm 5.06$	$\underline{0.41} \pm 0.26$	0.53 ± 0.29
$\text{RBM}_{\text{Random}_1} + \text{RBM}_{\text{Random}_2}$ [3D]	522 490	U-Net	LSTM	519 034	0.27 ± 0.21	40.89 ± 14.63	6.92 ± 3.64	0.25 ± 0.23	$\underline{0.68} \pm 0.28$

Grouping all parametric MRI maps in a single group

We considered, first, the effect of encoding all parametric maps using a single RBM. We varied the number of selected features from the RBM, observing that in all cases, the average Dice score is equal or lower than using only the parametric maps as input to the supervised block. Also, using 12 features presented the lowest average Dice score. The use of 3 or 6 obtained the same average Dice score, having the second, a lower average Hausdorff distance. Since, the selection of 6 features includes the best 3, we also compared the normalized mutual information between them. As illustrated in Fig. 6.2, the mutual information has lower values, which indicates that second set of 3 features had additional information. For this reason, we chose 6 as the number of features in the subsequent experiments. So, based on the metrics, we may conclude that there is no clear gain in using the features generated by the RBM, at least, when we encode all the parametric maps with a single RBM.

Grouping parametric MRI maps according to the subjacent physical meaning

In this experiment, we grouped the parametric maps according to their underlining physical meaning together with ADC map in each group. Each group was encoded with a RBM. Comparing isolatedly the use of each group of features, we verify that $\text{RBM}_{\text{Lesion}}$ had a higher average Dice score compared to using only the parametric maps as input to the supervised block. The increase in the average Dice score was obtained by a higher average Recall. Also, we observe an improvement in all distance metrics. The experiment of using $\text{RBM}_{\text{Haemo}}$ presented the lowest average Dice and Recall, as well as higher average distance metrics. However, $\text{RBM}_{\text{Haemo}}$ presented higher average Precision, contrary to $\text{RBM}_{\text{Lesion}}$, which motivated the study on the combination of features from $\text{RBM}_{\text{Lesion}}$ with $\text{RBM}_{\text{Haemo}}$ besides the parametric maps. We may observe that this combination obtained the highest average Dice and Precision, as well as the lowest average distance metrics. However, this improvement could have been originated from the combination of maps according to a specific common property, subjacent physical meaning of the parametric maps, in each group, or because we reduced the number of maps from 6 to 3 in each group. And this reduction could have allowed a better training of the RBM. So, we performed a complementary experiment. In this experiment, we formed two groups with similar size, but we randomly chose the parametric maps to include in each group. As presented in Table 6.2, this experiment obtained the lowest average Dice score and higher average distance metrics.

Considering these experiments together, we may draw some conclusions. First, although CNNs are very effective in generating features from raw data, they can generate even better features if rich and complementary information is provided. A similar conclusion was inferred by Oliveira et al. (2018) that observed improvement when the coefficients of the Wavelet were added as input in the problem of retinal vessel segmentation. Here, we observe a similar effect, but using the encoding provided by a RBM trained unsupervisedly. Second, at least to the problem of stroke lesion prediction, when we have data with different latent factors and we are able to group it, according to those factors, then there is potential to extract complementary information from each group, but to mix them all together can be detrimental.

The importance of each generated feature

To assess the contribution of each feature extracted from the $\text{RBM}_{\text{Lesion}} + \text{RBM}_{\text{Haemo}}$ in predicting the final infarct stroke lesion, an additional test was performed. Each feature map of the $\text{RBM}_{\text{Lesion}} + \text{RBM}_{\text{Haemo}}$ was individually perturbed with noise, namely the Gaussian noise. This perturbation was performed after batch normalization, with zero mean and unitary variance, being the probability density function of the statistical noise characterised by a zero mean and unitary variance. The obtained results obtained on the Validation set are presented in Table 6.3. Additionally, Fig. 6.3 illustrates the Dice score gain when comparing our proposal with the Gaussian noise perturbation of each $\text{RBM}_{\text{Lesion}} + \text{RBM}_{\text{Haemo}}$ feature, individually.

Table 6.3: Results obtained in the Validation set with added Gaussian Noise to each individual feature map of our proposal. Each metric represents the mean ± standard deviation.

Features	Gaussian Noise	Dice	H. D.	ASSD	Precision	Recall
$\text{RBM}_{\text{Lesion}} + \text{RBM}_{\text{Haemo}}$	–	0.3212 ± 0.2756	24.7485 ± 18.8072	11.5441 ± 12.8435	0.2846 ± 0.2829	0.4570 ± 0.3857
$\text{RBM}_{\text{Lesion}}$	Feature 1	0.3062 ± 0.2796	26.3204 ± 18.0065	11.7388 ± 12.6907	0.2611 ± 0.2764	0.4944 ± 0.4162
	Feature 2	0.3115 ± 0.2790	25.4371 ± 18.4487	11.6741 ± 12.7404	0.2659 ± 0.2757	0.4889 ± 0.4110
	Feature 3	0.3075 ± 0.2780	25.7735 ± 18.1804	11.6933 ± 12.7255	0.2657 ± 0.2798	0.4855 ± 0.4095
	Feature 4	0.3026 ± 0.2775	26.4023 ± 17.9309	11.7658 ± 12.6698	0.2592 ± 0.2757	0.4864 ± 0.4088
	Feature 5	0.3082 ± 0.2785	26.2934 ± 18.0293	11.7296 ± 12.6973	0.2619 ± 0.2738	0.4923 ± 0.4142
	Feature 6	0.3165 ± 0.2798	24.8879 ± 18.7294	11.6049 ± 12.7949	0.2699 ± 0.2760	0.4913 ± 0.4138
$\text{RBM}_{\text{Haemo}}$	Feature 1	0.3049 ± 0.2745	26.7738 ± 17.7922	11.7667 ± 12.6686	0.2637 ± 0.2772	0.4814 ± 0.4068
	Feature 2	0.3071 ± 0.2771	25.6165 ± 18.2810	11.6863 ± 12.7306	0.2645 ± 0.2774	0.4798 ± 0.4031
	Feature 3	0.3110 ± 0.2770	25.7595 ± 18.2277	11.6780 ± 12.7372	0.2679 ± 0.2769	0.4832 ± 0.4076
	Feature 4	0.3181 ± 0.2765	24.7689 ± 18.8122	11.5746 ± 12.8191	0.2748 ± 0.2779	0.4835 ± 0.4095
	Feature 5	0.3048 ± 0.2771	26.6852 ± 17.8857	11.7600 ± 12.6742	0.2611 ± 0.2764	0.4904 ± 0.4136
	Feature 6	0.3025 ± 0.2794	29.4114 ± 17.9100	12.0080 ± 12.4937	0.2572 ± 0.2755	0.4986 ± 0.4200

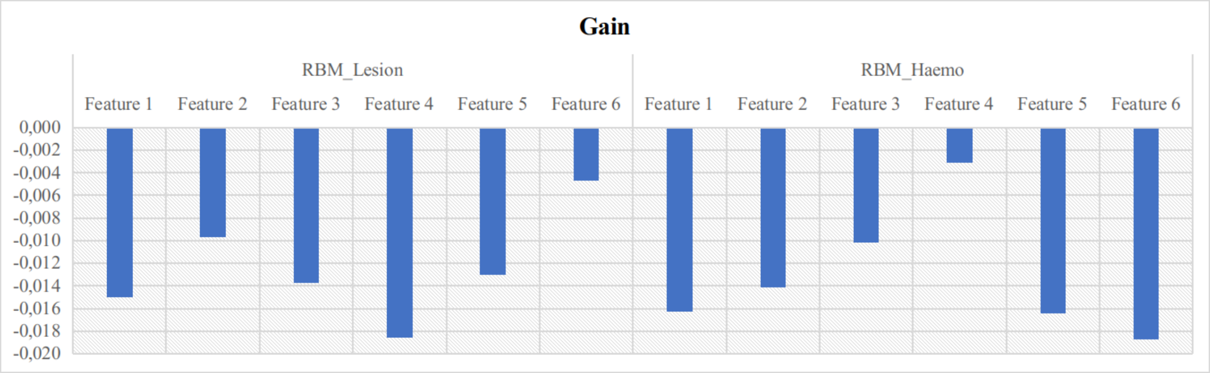


Figure 6.3: Results in the Validation set of the Dice score gain, emerged by comparing our proposal against its variations with Gaussian Noise added to each feature map, individually.

Analysing Table 6.3 and Fig. 6.3, when perturbing each feature extracted from the $\text{RBM}_{\text{Lesion}}$ and from the $\text{RBM}_{\text{Haemo}}$ the overall performance of our proposal decreases. Hence, each feature map is of relevance for predicting the final infarct stroke lesion. Furthermore, this study also demonstrates the adequacy of having 6 feature maps for each type of RBM. For the $\text{RBM}_{\text{Lesion}}$, perturbing feature map 4 lead to the highest score gain drops in our proposal, while for the $\text{RBM}_{\text{Haemo}}$, perturbing feature 6 resulted on the lowest average Dice score.

6.4.1.2 Context aggregation through gated recurrent blocks

In medical imaging segmentation, which is similar to our problem of inferring the extension of the lesion 90 days ahead, the use of a cascade of convolutional layers to elaborate the features is the prevalent practice. However, as discussed previously, Gated-RNN layers are able to capture long distance spatial relations among input voxels, so we performed some experiments to evaluate its contribution. The results are presented in Table 6.4.

Table 6.4: Results obtained when considering the Gated Recurrent block with and without the unsupervised learning block with ISLES 2017 testing set. Each metric represents the mean ± standard deviation. Underlined values correspond to the highest mean.

Unsupervised Block		Supervised Block			Dice	HD	ASSD	Precision	Recall
Method	Params.	FCNN	G-RNN	Params.					
–		U-Net	–	411 770	0.30 ± 0.21	38.83 ± 21.10	7.08 ± 5.15	0.26 ± 0.23	$\underline{0.64} \pm 0.30$
		U-Net	LSTM	519 034	0.30 ± 0.21	36.58 ± 16.62	6.96 ± 5.08	0.30 ± 0.26	0.55 ± 0.31
$\text{RBM}_{\text{Lesion}} + \text{RBM}_{\text{Haemo}}$ [3D]	617 400	U-Net	–	415 226	0.32 ± 0.23	34.09 ± 16.51	7.60 ± 7.14	0.35 ± 0.27	0.48 ± 0.32
		U-Net	LSTM	522 490	$\underline{0.38} \pm 0.22$	$\underline{29.21} \pm 15.04$	$\underline{5.52} \pm 5.06$	$\underline{0.41} \pm 0.26$	0.53 ± 0.29

Analysing Table 6.4, we verify that just having parametric maps as input to the supervised block, adding a LSTM layer increased the average Precision, but the average Recall decreased, resulting in the same average Dice score. But, when we added RBM features as input, we verify that using just CNN layers improved over having just parametric maps. This improvement came by a higher average Precision.

However, when we add the LSTM, we observe that the improvement is even higher, having originated from a larger increase in the average Precision, and a decrease in the average distance metrics.

Based on these experiments, we may conclude that the aggregation of CNN layers was able to extract additional information from the RBM features; however, at least to the problem of inferring the extension of the lesion days ahead, long and local distance spatial relations among input voxels introduced by gated RNN was critical to reduce the detection of false positive cases, increasing substantially the average Dice score by 6%.

6.4.1.3 Spatial context: 2D or 3D?

MRI images are 3D by nature, so the use of 3D filters would allow capturing more context, which has the potential to provide better prediction. Since 2D filters are confined to a plane, unnatural discontinuous contour may occur in the perpendicular axis. However, as presented previously, the resolution of MRI images in ISLES dataset is not equal in all axis, being coarser along the axial axis. So, we studied the effect of the spatial context in our architecture. As we have two blocks, unsupervised and supervised blocks, the effect on each one was evaluated separately. The results are presented in Table 6.5. Considering the results, we observe that using 2D patches in both blocks has lower average Dice score, than using only the parametric maps as input (baseline), because the increase in the average Precision was not enough to compensate the drop in the average Recall. Using, 3D patches for both blocks had the same performance as our baseline. However, when we used 3D patches for the RBM but 2D blocks for the U-Net block, we improved over our baseline. This is the model with the highest average Dice score without LSTM. So, we may conclude that for our architecture, larger context using 3D patches was more effective for encoding features (unsupervised block), while 2D patches are better suited for the U-Net.

Table 6.5: Evaluation metrics obtained with different spatial context configurations in the unsupervised and supervised learning blocks in ISLES 2017 testing set. Each metric represents the mean ± standard deviation. Underlined values correspond to the highest mean.

Unsupervised Block		Supervised Block			Dice	HD	ASSD	Precision	Recall
Method	Params.	FCNN	G-RNN	Params.					
$\text{RBM}_{\text{Lesion}} + \text{RBM}_{\text{Haemo}}$ [2D]	205 800	U-Net [2D]	–	415 226	0.27 ± 0.23	36.35 ± 14.89	9.14 ± 12.35	0.31 ± 0.28	0.53 ± 0.34
$\text{RBM}_{\text{Lesion}} + \text{RBM}_{\text{Haemo}}$ [3D]	617 400	U-Net [2D]	–	415 226	0.32 ± 0.23	34.09 ± 16.51	7.60 ± 7.14	0.35 ± 0.27	0.48 ± 0.32
		U-Net [3D]	–	720 122	0.30 ± 0.21	34.17 ± 14.86	6.16 ± 3.82	0.32 ± 0.27	0.54 ± 0.30

6.4.1.4 Incorporating clinical information

In Chapter 5 incorporating clinical information improved the overall performance of the supervised deep neural network. Thus, the importance of considering clinical information is once again measured, evaluating at the same time the presence of Gated-RNNs. The results obtained are presented in Table 6.6.

Clinical information plays a peculiar interaction, in an architecture that combines unsupervised and supervised methods. When the TICI score is considered in the supervised block based on the U-Net,

Table 6.6: Evaluation metrics obtained when incorporating non-imaging clinical information alongside Gated-RNNs in the presence of the unsupervised and supervised blocks. Each metric represents the mean ± standard deviation. Underlined values correspond to the highest mean.

Unsupervised Block		Supervised Block				Dice	HD	ASSD	Precision	Recall
Method	Params.	FCNN	G-RNN	Clin. Info.	Params.					
$\text{RBM}_{\text{Lesion}} + \text{RBM}_{\text{Haemo}}$	617 400	U-Net	–		415 226	0.32 ± 0.23	34.09 ± 16.51	7.60 ± 7.14	0.35 ± 0.27	0.48 ± 0.32
		U-Net	–	✓	415 228	0.32 ± 0.21	34.92 ± 13.90	5.41 ± 3.11	0.30 ± 0.23	0.66 ± 0.27
		U-Net	GRU		495 674	0.33 ± 0.21	34.08 ± 13.47	5.69 ± 3.67	0.29 ± 0.22	$\underline{0.67} \pm 0.26$
		U-Net	GRU	✓	495 676	0.36 ± 0.22	30.57 ± 13.98	$\underline{5.36} \pm 3.64$	0.38 ± 0.27	0.55 ± 0.29
		U-Net	LSTM		522 490	$\underline{0.38} \pm 0.22$	$\underline{29.21} \pm 15.04$	5.52 ± 5.06	$\underline{0.41} \pm 0.26$	0.53 ± 0.29
		U-Net	LSTM	✓	522 492	0.34 ± 0.22	31.79 ± 16.48	6.75 ± 7.28	0.36 ± 0.26	0.54 ± 0.31

the average Dice score is the same, while the average ASSD improves. However, the TICI score obtains a higher imbalance between Precision and Recall, due to a decrease of the average Precision with an increase in the average Recall. Nonetheless, this method provides a better delineation of the final stroke lesion. After, the importance of considering the TICI score is studied in the presence of Gated-RNNs, namely the LSTM and the GRU. Having as Gated-RNN the GRU the average Dice score increases by 9.1%, the average Precision increases by 35.7% and the average Recall decreases by 21.8% (U-Net + GRU vs. U-Net + GRU + TICI). However, a different behaviour occurs when using as Gated-RNN the LSTM. For this supervised block, considering the TICI score decreased the average Dice score by 11.8%, the average Precision also decreased by 13.9% and the average Recall increased 1.9% (U-Net + LSTM vs. U-Net + LSTM + TICI). From the latter study, we hypothesize that the LSTM retains longer dependencies, due to the presence of a gate that controls its memory state and is less influenced by updates in the gradient, which in turn are dependent on the TICI through the custom loss function. However, in the presence of an unsupervised feature generator, which increases the complexity of the input data, we hypothesize that the LSTM was not capable of interacting properly with the loss function, whereas the GRU layer, since it exposes its hidden state, grants a tighter interaction with the customized loss function. However, we recognized that this hypothesis needs further elaboration to understand this interaction.

6.4.1.5 Post-processing

The post-processing is an offline procedure, hence not optimized nor updated, aiming for the removal of small objects with less than 25 voxels. Table 6.7 presents the results of our proposal with and without this post-processing technique.

From the results, we show that this step impacts positively the performance of a model by increasing the average precision and both distance metrics. Furthermore, we observe that there is no loss of relevant information, since the average recall and its standard deviation was kept unchanged.

Table 6.7: Results in ISLES 2017 testing set evaluating the impact of the post-processing. Each metric represents the mean ± standard deviation. Underlined values correspond to the highest mean.

Unsupervised Block		Supervised Block			Dice	HD	ASSD	Precision	Recall
Method	Params.	FCNN	G-RNN	Params.					
$\text{RBM}_{\text{Lesion}} + \text{RBM}_{\text{Haemo}}$	617 400	U-Net*	LSTM	522 490	0.37 ± 0.20	30.61 ± 15.97	5.56 ± 4.08	0.40 ± 0.25	0.53 ± 0.29
		U-Net	LSTM	522 490	0.38 ± 0.22	29.21 ± 15.04	5.52 ± 5.06	0.41 ± 0.26	0.53 ± 0.29

* Without post-processing.

To conclude the ablative study, Fig. 6.4 illustrates a validation case, and the predicted final infarct stroke lesion, on the most important experiments of the ablation study. We note that our proposal ($\text{RBM}_{\text{Lesion}}$ + $\text{RBM}_{\text{Haemo}}$ with U-Net + LSTM) achieved the highest Dice score.

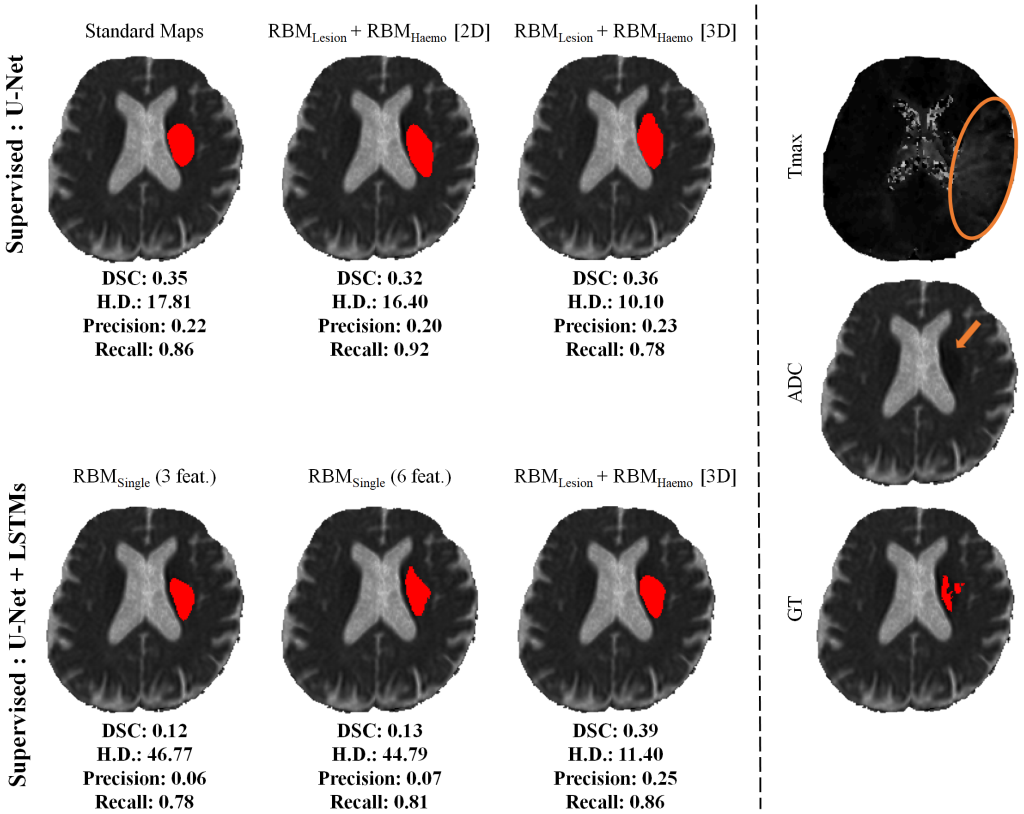


Figure 6.4: Example of the final infarct predicted for a validation case (0006) with different methods, alongside the ground-truth delineated at a 90-day follow-up. The ADC map depicts areas of diffusion restriction (arrow), whereas the T_{max} shows perfusion prolongation of the temporal parameters (delineation).

6.4.2 State-of-the-art: ISLES 2017 Challenge

The results of published methods for final infarct stroke lesion prediction using ISLES 2017 dataset Winzeck et al. (2018), together with our baseline and proposal methods are presented in Table 6.8. The metrics were computed by an online platform, so the ground-truth data, which is manually delineated based on a follow-up T2 MRI acquisitions, are not disclosed for public access.

Table 6.8: Published methods in ISLES 2017 testing dataset and our proposal. Each metric is represented by the mean ± standard deviation. Underlined values correspond to the highest mean.

		Dice	HD	ASSD	Precision	Recall
Ensemble	Mok et al. *	0.32 ± 0.23	40.74 ± 27.23	8.97 ± 9.52	0.34 ± 0.27	0.39 ± 0.27
	Kwon et al. *	0.31 ± 0.23	45.26 ± 21.04	7.91 ± 7.31	0.36 ± 0.27	0.45 ± 0.30
	Robben et al. *	0.27 ± 0.22	37.84 ± 17.75	6.72 ± 4.10	0.44 ± 0.32	0.39 ± 0.31
	Pisov et al. *	0.27 ± 0.20	49.24 ± 32.15	9.49 ± 10.56	0.31 ± 0.27	0.39 ± 029
Single Model	Monteiro et al. *	0.30 ± 0.22	46.60 ± 17.50	6.31 ± 4.05	0.34 ± 0.27	0.51 ± 0.30
	Pinto et al. (2018b)	0.29 ± 0.21	41.58 ± 22.04	7.69 ± 5.71	0.21 ± 0.21	0.66 ± 0.29
	Lucas et al. *	0.29 ± 0.21	33.85 ± 16.82	6.81 ± 7.18	0.34 ± 0.26	0.51 ± 0.32
	Choi et al. *	0.28 ± 0.22	43.89 ± 20.70	8.88 ± 8.19	0.36 ± 0.31	0.41 ± 0.31
	Niu et al. *	0.26 ± 0.20	48.88 ± 11.20	6.26 ± 3.02	0.28 ± 0.25	0.56 ± 0.26
	Sedlar et al. *	0.20 ± 0.19	58.30 ± 20.02	11.19 ± 9.10	0.23 ± 0.24	0.40 ± 0.29
	Rivera et al. *	0.19 ± 0.16	63.58 ± 18.58	11.13 ± 7.89	0.27 ± 0.25	0.21 ± 0.17
	Islam et al. *	0.19 ± 0.18	64.15 ± 28.51	14.17 ± 15.80	0.29 ± 0.28	0.25 ± 0.25
	Chengwei et al. *	0.18 ± 0.17	65.95 ± 25.94	9.22 ± 6.99	0.37 ± 0.30	0.21 ± 0.23
	Yoon et al. *	0.17 ± 0.16	45.23 ± 19.14	12.43 ± 11.01	0.23 ± 0.27	0.36 ± 0.32
	Baseline	0.30 ± 0.21	36.58 ± 16.62	6.96 ± 5.08	0.30 ± 0.26	0.55 ± 0.31
	Proposal	0.38 ± 0.22	29.21 ± 15.04	5.52 ± 5.06	0.41 ± 0.26	0.53 ± 0.29

* Methods presented in Winzeck et al. (2018).

Considering the results, we observe that our baseline is competitive with an average Dice, being among the top 3 methods together with Monteiro and Oliveira (2017), and surpassing the ensemble methods of Pisov et al. (2017) and Robben and Suetens (2017). Our method presented the lowest distance metrics among all methods, specially for the Hausdorff distance. It obtained the second best average Precision score, being surpassed by Robben and Suetens (2017). Robben and Suetens (2017) proposed the integration of meta-data information, using a two-pathway 3D network in an ensemble; however, our experiments did not indicate any improvement using 3D patches for the U-Net, at least for our architecture. So, this improvement could have come from a combination of the effect of the ensemble and the meta-data. But, we note that their method presented a much lower average Recall, which explain their lower average Dice score. Regarding the average Recall score, our method was fourth, but when we consider those methods, specially Pinto et al. (2018a), we conclude that it was obtained with a much lower average Precision, which means that to increase the true positive detections, they had to increase substantially the false positives. So, comparing with the state of the art, our method presented a better balance between Precision and Recall, which reflected into a higher average Dice score.

Based on the results, we may conclude that the use of complementary features provided by the RBMs and the use of LSTM for a larger context allowed our baseline to surpass current state-of-the-art methods on the average metrics.

Results from ChallengeR Benchmark

The SMIR platform of ISLES 2017 provides a weekly benchmark report of the current top-10 methods in the testing set, according to the average Dice score. So, some methods may not be published, lacking a description on their implementation, and, for this reason, were not included in the previous discussion. The boxplots of each method is illustrated in Fig. 6.5.

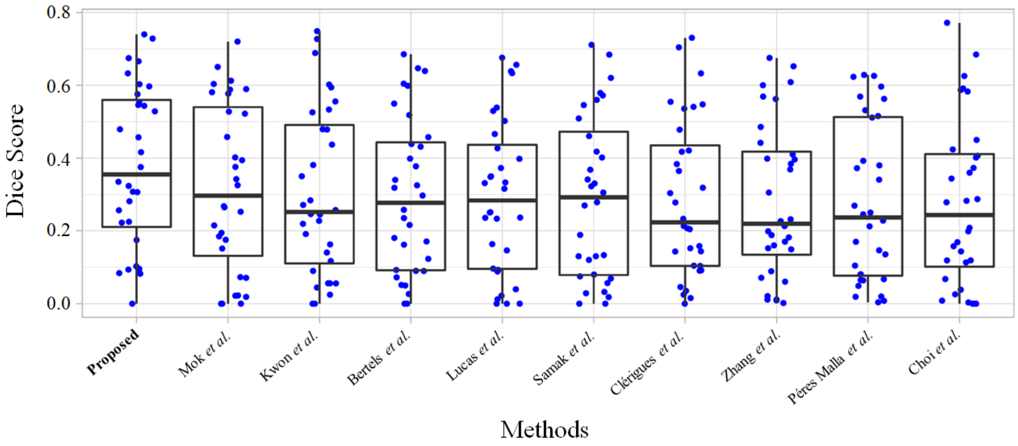


Figure 6.5: Boxplot of the top-10 ranking methods ordered by average Dice score in ISLES 2017 testing set.

We observe that the top-10 methods failed to predict the lesion of one or more cases (lowest outliers), which may indicate the degree of complexity of predicting infarct stroke lesion 90 days ahead in ISLES 2017 Challenge dataset. But, we verify that our method is the only one to have the first quartile above 0.20 in the Dice score. Fig. 6.6 illustrates the podium plot of each method for each case in the testing set, and its ranking. We observe that our proposal is the method, which ranked first most of the times, as well as second and third. Also, when we consider the methods ranked bellow fourth, our method is in general among those with the lowest counts. Analysing the cases individually, we note two trends, for some cases all methods presented similar performance, while for others, we find a large variation from the first to the other methods. The first trend may be found in the most difficult case, where all methods had zero or a close value for the Dice score. In the second trend, we observe that our method is ranked as first most of the cases.

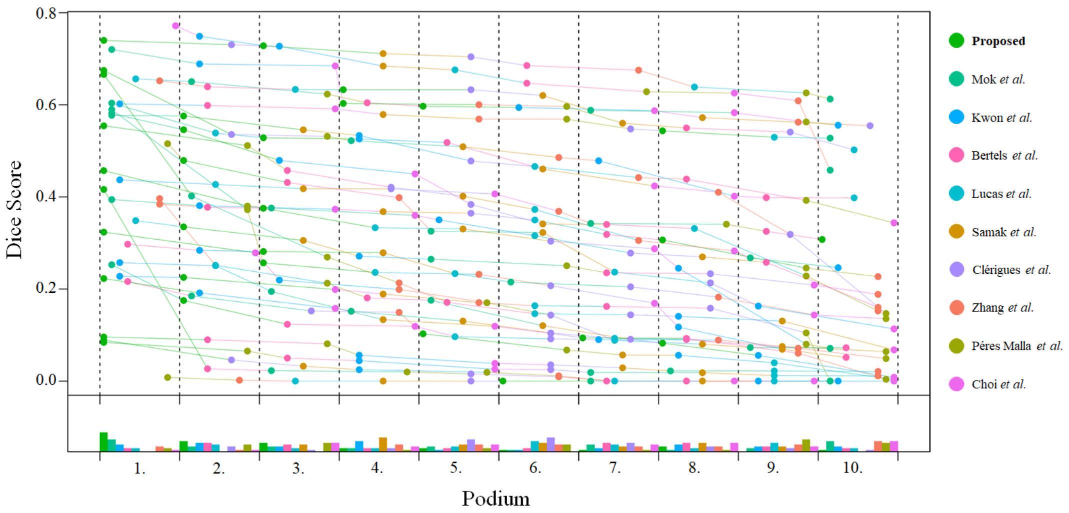


Figure 6.6: Podium plot of each testing case in ISLES 2017.

In Figure 6.7 we have the significance maps of the pairwise significant test with one-side Wilcoxon signed rank test (p-value $= 0.05$), showing that our method was statistically significant in Dice score against five of the top-10 ranked methods.

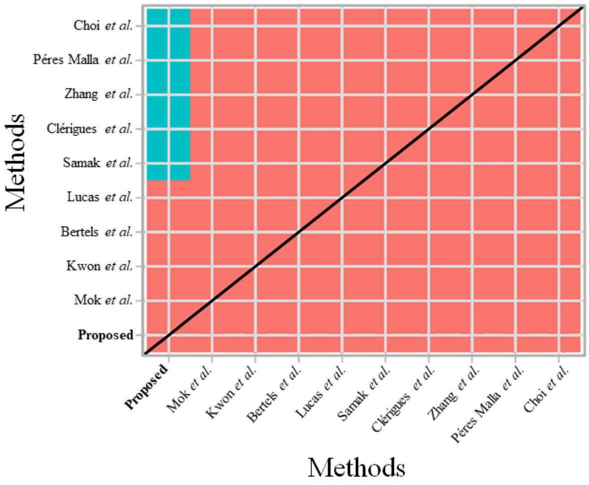


Figure 6.7: One-side Wilcoxon signed rank test in ISLES 2017 testing set.

Based on the results of the benchmark, we may infer that our method is competitive among current state of the art, presenting the highest average Dice score and lowest average distance score. Considering the ablation study, this performance was attained due to the combination of adding extra features obtained by encoding the parametric maps with RBMs, according to the underlining physical meaning, and the elaboration provided by the long context of the LSTM layers.

6.5 Summary

In this chapter, we present a deep learning-based method for stroke lesion outcome prediction, based on unsupervised and supervised learning. We proposed to group the input maps according to the underlying physical principle behind their creation, namely, the time-resolved perfusion maps (i.e. T_{max}, TTP, MTT), and the blood-flow-dynamic related maps (i.e. rCBF, rCBV). Each group was encoded using an unsupervised model to obtain structural features specific to its underlying physical principle. These structural features together with the standard parametric maps were fed to a supervised model to learn features conditioned on the label, which in our problem, means to condition on the results of the medical intervention – lesion at 90-days follow-up. We also investigated the use of Gated Recurrent Neural Networks to provide long spatial context, which were critical in relating the structural features to the information on input parametric maps. Our results showed that either the encoding or the long spatial context improved over our baseline. Also, these two together interacted positively increasing the performance when considering separately each one.

When evaluating our proposal on ISLES 2017 testing dataset, we observed a prediction improvement over current state-of-the-art methods. The proposed method obtained the first place in Dice and also in HD and ASSD.

Chapter 7

Conclusions

The main goal of this thesis was to be able to automatically predict the final infarct stroke lesion from onset neuroimaging acquisitions.

During the PhD., we conducted three different lines of research, which consist of the main contributions of this work. One of those lines focused on the combination of 4D spatio-temporal acquisitions with standard parametric maps. In another research line, we studied the incorporation of imaging and non-imaging information in a principled way. The final research line studied and proposed an unsupervised feature extractor block with a supervised functional block. Overall, we report that predicting the final infarct stroke lesion is an intricate task, but we were able to provide evidence of the importance of our proposals and its contribution to the medical imaging field. In this final chapter we sum up the main contributions and conclusions in ischaemic stroke prediction. The remainder of the chapter provides the perspectives on future lines of interest to research.

7.1 Overview and General Conclusions

Stroke still remains a clinical challenging task with a huge burden in society. Due to its rapidity and operation costs, CT is the most used neuroimaging technique for assessing and evaluating ischaemic stroke (González et al., 2011). Nonetheless, MRI is more sensitive in detecting early ischaemic stroke and its multi-parametric capability allows a robuster distinction between the hypoperfused tissue and permanently damaged tissue (González et al., 2011). In a context where elapsed time between stroke and treatment is related to the loss of brain tissue, assessment and treatment decision need to be performed in a short period, translating into a high cost of human resources. Clinicians often perform thresholding approaches and simplified measurements to obtain an overall notion of the perfusion and diffusion deficits (Austein et al., 2016; Straka et al., 2010), which might undermine the full potential of MRI maps, potentiating intra- and inter-rater variability. Thus, the motivation of this thesis focus on automatic methods for ischaemic stroke MRI imaging analysis, contributing in this way for the development of fully automated computerized systems.

Ischaemic stroke lesions emerge from occlusions in vessels of the brain. Besides being heterogeneous in location, shape and size, these lesions are restrained to underlying haemodynamic principles that vary across time. In a context where rapid intervention increases the chances of salvaging higher volumes of hypo-perfused tissue, MRI acquisitions are tuned for low resolution and fast acquisitions. Moreover, the standard parametric maps observed by clinicians emerge from deconvolution techniques applied to spatio-temporal acquisitions, which still are mathematically ill-posed approaches (Fieselmann et al., 2011). These factors combined with the intrinsic variability of the MRI equipment and the site, make the development of imaging analysis tools for predicting the stroke lesion outcome a very intricate task. In this thesis, the research focus resided on studying and developing Machine Learning methods, more specifically Deep and Representation Learning techniques capable of learning from MRI data.

In these last few years, Representation Learning has attracted a lot of attention in the scientific community. Representation Learning aims to learn how to extract the best set of features directly from the data. From the study and research conducted during this thesis, we confirmed that indeed it can extract powerful and discriminative features. We managed to achieve competitive results in predicting the final infarct lesion by using a FCNN-based approach. Furthermore, by incorporating non-imaging clinical information, we were able to increase the overall performance. This allowed us to surpass classical Machine Learning approaches based on hand-crafted and probabilistic methods, which require domain expert knowledge. Finally, when combining an unsupervised shallow model, using RBMs with a supervised deep neural network architecture, we achieved state-of-the-art results in ISLES 2017, demonstrating the capabilities of Representation Learning methods in learning to extract complex and discriminative features. Summarizing the marks achieved by the research conducted in this thesis, Table 7.1 presents its gradual progress.

Table 7.1: Summary of the results obtained in ISLES 2017 testing set using contributions from the three lines of research investigated in this thesis.

		Dice	HD	ASSD	Precision	Recall
Chapt. 4	Baseline	0.30 ± 0.21	38.83 ± 21.10	7.08 ± 5.15	0.26 ± 0.23	**0.64** ± 0.30
	Proposal	0.31 ± 0.21	33.94 ± 17.43	5.99 ± 4.58	0.29 ± 0.23	0.63 ± 0.30
Chapt. 5	Baseline	0.24 ± 0.20	53.29 ± 26.95	10.59 ± 4.98	0.27 ± 0.27	0.50 ± 0.35
	Proposal	0.29 ± 0.22	47.17 ± 22.13	7.20 ± 4.14	0.26 ± 0.23	0.61 ± 0.28
Chapt. 6	Baseline	0.30 ± 0.21	36.58 ± 16.62	6.96 ± 5.08	0.30 ± 0.26	0.55 ± 0.31
	Proposal	**0.38** ± 0.22	**29.21** ± 15.04	**5.52** ± 5.06	**0.41** ± 0.26	0.53 ± 0.29

From Table 7.1, we demonstrate that considering the DSC-MRI imaging data allowed the automatic extraction of complementary information from the source data responsible for generating the standard parametric perfusion maps. Afterwards, the incorporation of clinical information showed the importance of non-imaging information at a population-level and at a patient level, allowing better predictions. Lastly, with the proposal of an unsupervised learning block combined with a supervised block we conclude that taking into consideration clinical expertise and translating it into Machine Learning methods, allowed us to achieve the top score and current state-of-the-art performance in ISLES 2017.

Concluding, one can observe that predicting stroke is in fact a difficult task, even in an era dominated by Deep Learning, where features are extracted automatically from data and, at a first sight demand low domain knowledge expertise. During the course of this work, gaining expert knowledge on the ischaemic stroke imaging and how the infarct lesion progression is being characterized by such imaging data, was a key factor to gradually improve our methods and achieved state-of-the-art results. With the research developed in this thesis, we believe that is possible to develop Machine Learning models capable of predicting the final infarct ischaemic stroke lesions. However, there is still room for improvement, which we envision that will be a gradual and evolving process of research and development, allowing these methods to be applicable in clinical practice, easing the decision process of physicians and ultimately improving the quality of life of stroke patients.

7.2 Contributions

During the course of this thesis it is possible to identify several scientific contributions, which can be grouped accordingly to the lines of research conducted.

On deriving features maps of perfusion from DSC-MRI data using deep learning-based methods

Standard parametric maps of diffusion and perfusion are generated from post-processing techniques employed after MRI acquisitions. However, these techniques can lead to the loss of useful information, when assessing ischaemic stroke. Hence, the first line of research focused in combining spatio-temporal perfusion imaging with standard parametric maps of perfusion and diffusion. Due to the complexity of the data, we verified that combining spatio-temporal images directly with standard parametric maps was unable to retrieve complementary and discriminative information from both sources. Instead, we observed that having dedicated paths, to simultaneously extract features from different data sources, allows a better prediction of the final infarct lesion. Moreover, we were able to demonstrate that our proposed temporal preprocessing block, when applied to the DSC-MRI, allowed a reduction of time acquisitions without losing performance, which translates into a faster and lighter computerized method. Inspired by Milletari et al. (2016), we employed the soft dice loss function. Only by using this loss function, we were able to optimize our architecture and overcome the class imbalance, which was not possible with the categorical cross-entropy loss function, also reported in the work of Choi et al. (2016). Even by studying different sampling schemes that aim to overcome class imbalance, it was not possible to optimize successfully deep learning-based models with commonly used loss functions for segmentation tasks. Since in FCNNs each patch is mapped to a patch of labels that are spatially related, assuring a balanced training data scheme is not straightforward.

The contributions of this line of research are:

- A fully automatic algorithm to process spatio-temporal data, namely perfusion DSC-MRI.
- A study on the importance of the derived features from the DSC-MRI using mutual information analysis.

- The proposal of an independent learning block, to extract information from different data sources.

The preliminary studies that led to these contributions were published in an international peer-reviewed conference (MICCAI 2018) as means of sharing the proposed method and the results (Pinto et al., 2018b).

On incorporating non-imaging information with MRI imaging information in automatic deep learning-based methods

Prediction of ischaemic infarct growth can vary considerably across patients, which translates into a great lesion size variability. For sake of demonstration, in the ISLES 2017 training set we observed patients with lesions ranging from 23 voxels (0.036% of the whole brain) to 23961 voxels (10.71%). However, the success of the clinical intervention, when performed, impacts the final infarct core lesion. At ISLES 2017 Challenge, we demonstrated that considering non-imaging clinical information, that characterizes the success of the clinical intervention, at two different levels, allowed a better prediction of the final lesion. Indeed, other competitors at the Challenge, namely Robben and Suetens (2017) and Choi et al. (2016), also considered clinical information in their deep neural network. However, besides considering clinical information only as an extra channel, we proposed a custom loss function that encodes clinical information in the learning process. Later on, in a larger dataset, Robben et al. (2018) provided additional evidence that clinical information plays an important role as extra-input to the deep neural network architecture, combined with spatio-temporal CT imaging.

The contributions of this line of research are the:

- Proposal of a custom loss function guided by external non-imaging clinical information.
- Proposal of a deep neural network model that considers clinical information at a patient-specific level.
- Proposal of a deep neural network based on CNNs and Gated-RNNs.
- Proposal of an automatic machine learning method able to predict different reperfusion scenarios, to ease the decision-making process of physicians.

These contributions were published and detailed in two peer-reviewed journal papers (Winzeck et al., 2018; Pinto et al., 2018a).

On combining unsupervised feature generators with supervised learning

Capturing the underlying cerebral haemodynamic of the brain from static MRI imaging requires medical specialization. Furthermore, in the presence of an ischaemic stroke the cerebral haemodynamic processes are subdued to changes in order to restrain the loss of brain tissue. These changes are meant to be depicted and characterized by specific sets of perfusion parametric maps, which originated the last research line of this thesis. We tackled the characterization of cerebral blood flow dynamics and simultaneously the region(s) of the brain affected by perfusion and diffusion deficits. From the ablation study conducted, when discarding the aforementioned changes, we demonstrated that capturing the distribution of the data by employing a single RBM was not capable of proving discriminative features that could improve the prediction of the final infarct core. However, performing an input grouping in a two-pathway unsupervised learning block, based on the information depicted by each standard parametric map, achieved state-of-the-art results. Furthermore, we demonstrated that having only a two-pathway is not able to achieve a competitive method to predict the final infarct core lesion. Henceforth, we provide evidence that considering the medical expertise knowledge alongside the principles behind each standard parametric map, allowed the extraction of powerful and discriminative features. In the supervised learning block, we demonstrated the importance of considering Gated-RNNs, when having as input the standard maps alongside features extracted from the unsupervised learning block. Since Gated-RNNs consider the influence of the neighbourhood when analysing the current node, going through the input in a bidirectional approach, in both horizontal and vertical directions, allows a global and local notion of context. Employing Gated-RNNs allowed an overall increase in performance, specifically the average Dice, the average Precision and the distance metrics (HD and ASSD).

The contributions of this line of research are:

- Proposal of an unsupervised learning block, that takes into consideration expert domain knowledge on the standard parametric perfusion and diffusion maps to characterized different brain dynamics.
- A study on the importance of Gated-RNNs.
- A study on the impact of non-imaging clinical information in the presence of standard parametric maps alongside unsupervised deep generated features.

7.3 Dissemination of Scientific Research

During the course of this PhD., several manuscripts were published:

- Pinto, A., Mckinley, R., Alves, V., Wiest, R., Silva, C. A., & Reyes, M. (2018). Stroke lesion outcome prediction based on MRI imaging combined with clinical information. Frontiers in neurology, 9, 1060. – This manuscript resulted from the work developed for ISLES 2017 Challenge edition. The framework developed for this Challenge competition is also detailed in Chapter 5, hence contributing for the second line of research of this thesis.
- 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. – Preliminary study on the incorporation of DSC-MRI with the standard parametric maps of diffusion and perfusion, being part of the line of research detailed on Chapter 4.

- Amorim, J., Pinto, A., Pereira, S., & Silva, C. A. (2019, February). Segmentation Squeeze-and-Excitation Blocks in Stroke Lesion Outcome Prediction. In 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG) (pp. 1-4). IEEE. (Shared first authorship) **–** The work developed in this manuscript details further studies and research conducted on the first line of research of this thesis. Here, we studied the added value of attention mechanisms to select the best suited features for predicting the final infarct core lesion.

- Pinto, A., Pereira, S., Rasteiro, D., & Silva, C. A. Hierarchical Brain Tumour Segmentation using Extremely Randomized Trees. Pattern Recognition, 2018. **–** The research conducted in this manuscript reports a previous work conducted during the MsC. Thesis. The proposed method encompasses Extra-Trees to perform brain tumour segmentation in a hierarchical manner.

- 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. **–** This manuscript emerged from a collaborative work with Sérgio Pereira of brain tissue segmentation from MRI images, using a RF classifier.

- Winzeck, S., Hakim, A., McKinley, R., Pinto, A., Alves, V., Silva, C., ... & Oliveira, A. (2018). ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI. Frontiers in neurology, 9. **–** Benchmarking report of ISLES 2016 and ISLES 2017 Challenge competitions, where our method ranked 6^{th} among 15 competitors in the overall ranking performed by the online platform (SMIR, 2017). However, if we consider only single model approaches, we ranked 3^{rd} in this competition. This was an international challenge integrated in the MICCAI conference, where the goal of the Challenge was to predict the final infarct ischaemic stroke lesion from onset MRI images. For training purposes the competitors were provided with 43 cases, with access to the manual segmentation of the final stroke lesion performed in a follow-up acquisition. The testing phase, the Challenge, provided only the MRI images of 32 subjects, so that each participant was responsible to perform the predictions and submit them to an website platform which performed the evaluation.

- 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), where the preprocessing step employed was based on a previous work conducted during the MsC. degree, and combines it with a deep learning neural architecture, achieving top performance at brain tumour segmentation.

- 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. **–** Proposed method developed for the Brain Tumour Segmentation (BRATS) Challenge competition of 2015, held up at MICCAI 2015.

7.4 Opened Research Lines

Throughout the work developed and the research conducted during this thesis, it is possible to identify future lines of research.

Atlas constrained prediction. Understanding how a stroke lesion will evolve over time requires not only the knowledge about the volume of the lesion but also its location and the presence of secondary vessels providing blood flow to the hypo-perfused area. Therefore, this is a highly complex task where Machine Learning methods struggle to perform correct final infarct predictions from merely inputting the onset standard parametric maps. In this research, we provided evidence that cerebral blood flow dynamics and the success of the clinical intervention are key factors to consider in the development of methods for stroke tissue outcome prediction. The work of Habegger et al. (2018) demonstrated that lesion topography and lesion load (percentage that characterizes brain regions affected by an occlusion) correlates differently with the NIHSS clinical outcome in revascularized and non-revascularized patient cohorts. Furthermore, the authors observed that occlusions in cortical areas tend to have better clinical outcome due to the higher cerebral blood volume, when compared to sub-cortical areas. However, we recognize that this is still an open area of research. We envision that to further enforce brain vascular connectivity, when assessing ischaemic stroke lesions, vascular territory atlas maps, could allow the codification of proximal regions or connected regions affected by an ischaemic stroke. Recently, Schirmer et al. (2019) provided some research on this topic, by presenting an atlas capable of being applied to stroke.

The inclusion of more sequences. To perform stroke tissue prediction in this research we used standard parametric maps of diffusion and perfusion, and also perfusion DSC-MRI. Perfusion and diffusion maps and spatio-temporal sequences are the standard approaches used in clinical practice, regardless of the neuroimaging acquisition (González et al., 2011). Nevertheless, additional MRI acquisitions could be useful in providing more information. For example, conventional structural MRI, such as the FLAIR sequence, is useful in characterizing white matter hyperintensity, which in turn is correlated to the ischaemic stroke occlusion and outcome (Azizyan et al., 2011). However, it still needs to be investigated if additional sequences can boost the performance of Machine Learning based models, and on how to handle the new imaging information.

The interoperability. The beginning of this research overlapped with the release of the ISLES 2016 dataset, which was thought to be in the same line as its previous edition (ISLES 2015), where the purpose was the segmentation of ischaemic stroke lesions from MRI images. However, ISLES 2016 and ISLES 2017 editions provided an interesting and more clinically orientated direction for research. Having the notion about how the lesion will progress across time, can provide useful information to clinicians in an environment where ”time is brain”. Despite that the underlying objective of segmentation differs from the prediction objectives, we envision that a framework capable of performing segmentation of an onset stroke lesion could provide a starting point to predict how the infarct will evolve over time, in the presence of clinical intervention. Hence, a potential research direction would be a two-stage method were the first step focus on the definition of an onset ischaemic stroke lesion, and the second performs the final infarct prediction based on the cerebral flow dynamics occurring inside a ROI. The development of these hierarchical methods have already demonstrated its good results in segmentation problems in the imaging field. The works of Pinto et al. (2018c) and Pereira et al. (2017), developed for brain tumour segmentation, demonstrate the benefits of performing a first step responsible for identifying roughly the brain tumour location followed by a second step that performs the segmentation of the different types of brain tissue. Nonetheless, one viable solution would be performing these two steps end-to-end. Note however that, to do so one might need to guarantee that the segmentation and prediction tasks are performed over the same neuroimaging images.

Another direction of research could be the study of Machine Learning based methods which are agnostic to the neuroimaging acquisition, such as MRI and CT. We envision that methods based on transfer learning (Goodfellow et al., 2016) or Generative Adversarial Networks could be potential approaches. Xiang et al. (2018) proposed a method to generate CT images based on the T1 MRI sequence. Also, Nie et al. (2016) developed a 3D FCNN capable of generating CT images from MRI. However, for stroke tissue outcome prediction, research is still needed to identify which parametric maps should be translated into a standard pre-defined neuroimaging type. In addition, if the input data matches the standard pre-defined type, one might consider to investigate if the generator of the Generative Adversarial Network and the respective discriminator are in fact robust. In a similar line of thought, Song (2019) provided evidence on the benefits of employing a generative model to generate the DWI sequence from CTP imaging.

The search for reliability. Predicting the final stroke lesion has already attracted the attention of several research groups worldwide (Winzeck et al., 2018; Nielsen et al., 2018; Robben et al., 2018). However, we observe that as the time window of prediction increases from 24h to a 90-day, current methods still struggle to perform reliable and accurate predictions. From the Tables 4.3 to 6.8, we observe a large variabilities in metrics. Furthermore, other works report the same findings, such as in Fig. 3 of Robben et al. (2018) and in Fig. 3 of Nielsen et al. (2018), even when considering smaller time-windows.

One can hypothesize that this phenomena might be explained by ischaemic stroke lesions that present a large recovery of hypoperfused brain tissue, from the onset time to prediction time. This phenomena is mainly explained by a successful restoration of the perfusion deficits by clinical intervention. The location of the occlusion also influences the prediction scenario and, additionally the possibility to perform clinical intervention (González et al., 2011). Hence, in order to deal with this intrinsic phenomena, we consider that more data may be necessary. In addition, there is still room for improvement, and we envision that the future state-of-the-art results obtained in ischaemic stroke prediction will be the ones capable of predicting and understanding the cerebral vascular conditions reliably, which impact the prediction.

Another reliability challenge emerges with the intrinsic variability of MRI imaging. In ischaemic stroke, different acquisition protocols are implemented, varying across hospitals and vendors (e.g. different resolutions and slice thickness) (Winzeck et al., 2018). These factors combined with the intrinsic variability of MRI imaging impacts Machine Learning-based methods, due to the differences in training and data variability. A first line of thought would be the development of pre-processing approaches, such as histogram matching (Pinto et al., 2018c; Pereira et al., 2016), or even Machine Learning-based methods such as the works of Karani et al. (2018) and Kamnitsas et al. (2017a). While in the former group the mapping of a test image into a learned histogram does not require a re-training step, the method of Karani et al. (2018) requires annotated data to retrain the Batch Normalization blocks of a pre-trained CNN, and the method of Kamnitsas et al. (2017a) despite being an unsupervised method, based on adversarial training, still requires re-training. However, we note that in common practice, as observed in ISLES dataset, patients are characterized by functional MRI imaging, and not structural (e.g. T1 and FLAIR). Hence, performing non-linear transformations to parametric maps that characterize the metabolic functions is a non-interpretable task. Nonetheless, we recognize that dealing with data from different sources is a key factor that will improve the reliability of Machine Learning-based methods in performance but also in the testing and deployment phase.

References

Abulnaga, S.M., Rubin, J., 2018. Ischemic stroke lesion segmentation in ct perfusion scans using pyramid pooling and focal loss, in: International MICCAI Brainlesion Workshop, Springer. pp. 352–363.

Adams Jr, H.P., Bendixen, B.H., Kappelle, L.J., Biller, J., Love, B.B., Gordon, D.L., Marsh 3rd, E., 1993. Classification of subtype of acute ischemic stroke. definitions for use in a multicenter clinical trial. toast. trial of org 10172 in acute stroke treatment. Stroke 24, 35–41.

Albers, G.W., Caplan, L.R., Easton, J.D., Fayad, P.B., Mohr, J., Saver, J.L., Sherman, D.G., 2002. Transient ischemic attack—proposal for a new definition.

Allen, L.M., Hasso, A.N., Handwerker, J., Farid, H., 2012. Sequence-specific MR imaging findings that are useful in dating ischemic stroke. Radiographics 32, 1285–1297.

Astrup, J., Siesjö, B.K., Symon, L., 1981. Thresholds in cerebral ischemia-the ischemic penumbra. Stroke 12, 723–725.

Austein, F., Riedel, C., Kerby, T., Meyne, J., Binder, A., Lindner, T., Huhndorf, M., Wodarg, F., Jansen, O., 2016. Comparison of perfusion ct software to predict the final infarct volume after thrombectomy. Stroke 47, 2311–2317.

Azizyan, A., Sanossian, N., Mogensen, M., Liebeskind, D., 2011. Fluid-attenuated inversion recovery vascular hyperintensities: an important imaging marker for cerebrovascular disease. American journal of neuroradiology 32, 1771–1775.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2015. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561 .

Barber, P., Darby, D., Desmond, P., Yang, Q., Gerraty, R., Jolley, D., Donnan, G., Tress, B., Davis, S., 1998. Prediction of stroke outcome with echoplanar perfusion-and diffusion-weighted mri. Neurology 51, 418–426.

Barber, P.A., Demchuk, A.M., Zhang, J., Buchan, A.M., Group, A.S., et al., 2000. Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. The Lancet 355, 1670–1674.

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

Bauer, S., Gratz, P.P., Gralla, J., Reyes, M., Wiest, R., 2014. Towards automatic MRI volumetry for treatment selection in acute ischemic stroke patients, in: Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, IEEE. pp. 1521–1524.

Beloosesky, Y., Streifler, J., Burstin, A., Grinblat, J., 1995. The importance of brain infarct size and location in predicting outcome after stroke. Age and ageing 24, 515–518.

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., Simard, P., Frasconi, P., et al., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks 5, 157–166.

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

Berkhemer, O.A., Jansen, I.G., Beumer, D., Fransen, P.S., Van Den Berg, L.A., Yoo, A.J., Lingsma, H.F., Sprengers, M.E., Jenniskens, S.F., Lycklama à Nijeholt, G.J., et al., 2016. Collateral status on baseline computed tomographic angiography and intra-arterial treatment effect in patients with proximal anterior circulation stroke. Stroke 47, 768–776.

Brainin, M., Heiss, W.D., 2014. Textbook of stroke medicine. Cambridge University Press.

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

Brott, T., Adams Jr, H.P., Olinger, C.P., Marler, J.R., Barsan, W.G., Biller, J., Spilker, J., Holleran, R., Eberle, R., Hertzberg, V., 1989. Measurements of acute cerebral infarction: a clinical examination scale. Stroke 20, 864–870.

Brouns, R., De Deyn, P., 2009. The complexity of neurobiological processes in acute ischemic stroke. Clinical neurology and neurosurgery 111, 483–495.

Butcher, K., Emery, D., 2010a. Acute stroke imaging part i: Fundamentals. Canadian Journal of Neurological Sciences 37, 4–16.

Butcher, K., Emery, D., 2010b. Acute stroke imaging part ii: the ischemic penumbra. Canadian Journal of Neurological Sciences 37, 17–27.

Calamante, F., Gadian, D.G., Connelly, A., 2000. Delay and dispersion effects in dynamic susceptibility contrast mri: simulations using singular value decomposition. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 44, 466–473.

Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering 40, 16–28.

Chen, C.L., Tang, F.T., Chen, H.C., Chung, C.Y., Wong, M.K., 2000. Brain lesion size and location: effects on motor recovery and functional outcome in stroke patients. Archives of physical medicine and rehabilitation 81, 447–452.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014a. Learning phrase representations using rnn encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics. pp. 1724–1734. doi:10.3115/v1/D14-1179.

Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014b. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 .

Choi, Y., Kwon, Y., Lee, H., Kim, B.J., Paik, M.C., Won, J.H., 2016. Ensemble of deep convolutional neural networks for prognosis of ischemic stroke, in: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer. pp. 231–243.

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.

Clèrigues, A., Valverde, S., Bernal, J., Freixenet, J., Oliver, A., Lladó, X., 2018. Sunet: a deep learning architecture for acute stroke lesion segmentation and outcome prediction in multimodal mri. arXiv preprint arXiv:1810.13304 .

Cooper, R.L., Chang, D.B., Young, A.C., Martin, C.J., Ancker-Johnson, B., 1974. Restricted diffusion in biophysical systems: experiment. Biophysical Journal 14, 161–177.

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

Cover, T.M., Thomas, J.A., 2006. Elements of information theory 2nd edition (wiley series in telecommunications and signal processing) .

Dani, K.A., Thomas, R.G., Chappell, F.M., Shuler, K., MacLeod, M.J., Muir, K.W., Wardlaw, J.M., Study, T.M.R.C.M.A.S.I., 2011. Computed tomography and magnetic resonance perfusion imaging in ischemic stroke: definitions and thresholds. Annals of neurology 70, 384–401.

Deng, W., Teng, J., Liebeskind, D., Miao, W., Du, R., 2019. Predictors of infarct growth measured by apparent diffusion coefficient quantification in patients with acute ischemic stroke. World neurosurgery 123, e797–e802.

Dolz, J., Ayed, I.B., Desrosiers, C., 2018. Dense multi-path u-net for ischemic stroke lesion segmentation in multiple image modalities, in: International MICCAI Brainlesion Workshop, Springer. pp. 271–282.

Elman, J.L., 1990. Finding structure in time. Cognitive science 14, 179–211.

Emberson, J., Lees, K.R., Lyden, P., Blackwell, L., Albers, G., Bluhmki, E., Brott, T., Cohen, G., Davis, S., Donnan, G., et al., 2014. Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. The Lancet 384, 1929–1935.

Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M., 2009. Normalized mutual information feature selection. IEEE Transactions on Neural Networks 20, 189–201.

Feigin, V.L., Forouzanfar, M.H., Krishnamurthi, R., Mensah, G.A., Connor, M., Bennett, D.A., Moran, A.E., Sacco, R.L., Anderson, L., Truelsen, T., et al., 2014. Global and regional burden of stroke during 1990–2010: findings from the global burden of disease study 2010. The Lancet 383, 245–255.

Fiehler, J., Albers, G.W., Boulanger, J.M., Derex, L., Gass, A., Hjort, N., Kim, J.S., Liebeskind, D.S., Neumann-Haefelin, T., Pedraza, S., et al., 2007. Bleeding risk analysis in stroke imaging before thrombolysis (BRASIL) pooled analysis of t2*-weighted magnetic resonance imaging data from 570 patients. Stroke 38, 2738–2744.

Fieselmann, A., Kowarschik, M., Ganguly, A., Hornegger, J., Fahrig, R., 2011. Deconvolution-based ct and mr brain perfusion measurement: theoretical model revisited and practical implementation details. Journal of Biomedical Imaging 2011, 14.

Fisher, M., Garcia, J.H., 1996. Evolving stroke and the ischemic penumbra. Neurology 47, 884–888.

Ga, D., 2008. Fisher m, macleod m, davis sm. Stroke. Lancet 371, 1612–23.

Giesel, F.L., Mehndiratta, A., Risse, F., Rius, M., Zechmann, C.M., von Tengg-Kobligk, H., Gerigk, L., Kauczor, H.U., Politi, M., Essig, M., et al., 2009. Intraindividual comparison between gadopentetate dimeglumine and gadobutrol for magnetic resonance perfusion in normal brain and intracranial tumors at 3 tesla. Acta Radiologica 50, 521–530.

Gleason, S., Furie, K.L., Lev, M.H., O'Donnell, J., McMahon, P.M., Beinfeld, M.T., Halpern, E., Mullins, M., Harris, G., Koroshetz, W.J., et al., 2001. Potential influence of acute CT on inpatient costs in patients with ischemic stroke. Academic radiology 8, 955–964.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth 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.

Gonzalez, R., Hirsch, J., Koroshetz, W., Lev, M., Schaefer, P., 2007. Acute ischemic stroke: imaging and intervention. American Journal of Neuroradiology 28, 1622.

González, R.G., Hirsch, J.A., Koroshetz, W., Lev, M.H., Schaefer, P.W., 2011. Acute ischemic stroke. Springer.

Gonzalez, R.G., Schaefer, P.W., Buonanno, F.S., Schwamm, L.H., Budzik, R.F., Rordorf, G., Wang, B., Sorensen, A.G., Koroshetz, W.J., 1999. Diffusion-weighted mr imaging: diagnostic accuracy in patients imaged within 6 hours of stroke symptom onset. Radiology 210, 155–162.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Grant, P.E., He, J., Halpern, E.F., Wu, O., Schaefer, P.W., Schwamm, L.H., Budzik, R.F., Sorensen, A.G., Koroshetz, W.J., Gonzalez, R.G., 2001. Frequency and clinical context of decreased apparent diffusion coefficient reversal in the human brain. Radiology 221, 43–50.

Graves, A., 2012. Supervised sequence labelling, in: Supervised sequence labelling with recurrent neural networks. Springer, pp. 5–13.

Grefenstette, E., Blunsom, P., 2014. A convolutional neural network for modelling sentences, in: ACL.

Grysiewicz, R.A., Thomas, K., Pandey, D.K., 2008. Epidemiology of ischemic and hemorrhagic stroke: incidence, prevalence, mortality, and risk factors. Neurologic clinics 26, 871–895.

Habegger, S., Wiest, R., Haeni, L., Weder, B.J., Gralla, J., Mordasini, P., Jung, S., Reyes, M., McKinley, R., 2018. Lesion load in acute ischemic stroke prior to interventional treatment: Mismatch patterns vs. predictive models. Frontiers in Neurology 9, 600.

Hacke, W., Kaste, M., Bluhmki, E., Brozman, M., Dávalos, A., Guidetti, D., Larrue, V., Lees, K.R., Medeghri, Z., Machnig, T., et al., 2008. Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. New England Journal of Medicine 359, 1317–1329.

Harrison, J.K., McArthur, K.S., Quinn, T.J., 2013. Assessment scales in stroke: clinimetric and clinical considerations. Clinical interventions in aging 8, 201.

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, pp. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Henninger, N., Fisher, M., 2016. Extending the time window for endovascular and pharmacological reperfusion. Translational stroke research 7, 284–293.

Hess, A., Meier, R., Kaesmacher, J., Jung, S., Scalzo, F., Liebeskind, D., Wiest, R., McKinley, R., 2018. Synthetic perfusion maps: Imaging perfusion deficits in dsc-mri with deep learning. arXiv preprint arXiv:1806.03848 .

Higashida, R.T., Furlan, A.J., Roberts, H., Tomsick, T., Connors, B., Barr, J., Dillon, W., Warach, S., Broderick, J., Tilley, B., et al., 2003. Trial design and reporting standards for intraarterial cerebral thrombolysis for acute ischemic stroke. Journal of Vascular and Interventional Radiology 14, E1–E31.

Hinton, G.E., 2012. A practical guide to training restricted boltzmann machines, in: Neural networks: Tricks of the trade. Springer, pp. 599–619.

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., Zemel, R.S., 1994. Autoencoders, minimum description length and helmholtz free energy, in: Advances in neural information processing systems, pp. 3–10.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Hosseini, M.B., Liebeskind, D.S., 2018. The role of neuroimaging in elucidating the pathophysiology of cerebral ischemia. Neuropharmacology 134, 249–258.

Hossmann, K.A., 1994. Viability thresholds and the penumbra of focal ischemia. Annals of neurology 36, 557–565.

Hounsfield, G.N., 1973. Computerized transverse axial scanning (tomography): Part 1. description of system. The British journal of radiology 46, 1016–1022.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.

de Ipolyi, A., Wu, O., Schaefer, P., Macklin, E., Schwamm, L., Ackerman, R., Gonzalez, R., Copen, W., 2010. Cerebral blood volume measurements in acute ischemic stroke are technique-dependent and cannot substitute for dw imaging. Boston, Mass: American Society of Neuroradiology .

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2019. nnu-net: Self-adapting framework for u-net-based medical image segmentation, in: Bildverarbeitung für die Medizin 2019. Springer, pp. 22–22.

Islam, M., Vaidyanathan, N.R., Jose, V.J.M., Ren, H., 2018. Ischemic stroke lesion segmentation using adversarial learning, in: International MICCAI Brainlesion Workshop, Springer. pp. 292–300.

Jaeger, H., 2001. The “echo state” approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report 148, 13.

Janecek, A., Gansterer, W., Demel, M., Ecker, G., 2008. On the relationship between feature selection and classification accuracy, in: New challenges for feature selection in data mining and knowledge discovery, pp. 90–105.

Johnston, S.C., Gress, D.R., Browner, W.S., Sidney, S., 2000. Short-term prognosis after emergency department diagnosis of tia. Jama 284, 2901–2906.

Jolliffe, I., 2011. Principal component analysis. Springer.

Jung, S., Gilgen, M., Slotboom, J., El-Koussy, M., Zubler, C., Kiefer, C., Luedi, R., Mono, M.L., Heldner, M.R., Weck, A., et al., 2013. Factors that determine penumbral tissue loss in acute ischaemic stroke. Brain 136, 3554–3560.

Kalchbrenner, N., Danihelka, I., Graves, A., 2015. Grid long short-term memory. arXiv preprint arXiv:1507.01526 .

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, Springer. pp. 597–609.

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.

Kane, I., Sandercock, P., Wardlaw, J., 2007. Magnetic resonance perfusion diffusion mismatch and thrombolysis in acute ischaemic stroke: a systematic review of the evidence to date. Journal of Neurology, Neurosurgery & Psychiatry 78, 485–491.

Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E., 2018. A lifelong learning approach to brain mr segmentation across scanners and protocols, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 476–484.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732.

Kemmling, A., Flottmann, F., Forkert, N.D., Minnerup, J., Heindel, W., Thomalla, G., Eckert, B., Knauth, M., Psychogios, M., Langner, S., Fiehler, J., 2015. Multivariate dynamic prediction of ischemic infarction and tissue salvage as a function of time and degree of recanalization. Journal of Cerebral Blood Flow & Metabolism 35, 1397–1405.

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

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.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

von Kummer, R., Meyding-Lamade, U., Forsting, M., Rosin, L., Rieke, K., Hacke, W., Sartor, K., 1994. Sensitivity and prognostic value of early CT in occlusion of the middle cerebral artery trunk. American Journal of Neuroradiology 15, 9–15.

Labeyrie, M.A., Turc, G., Hess, A., Hervo, P., Mas, J.L., Meder, J.F., Baron, J.C., Touzé, E., Oppenheim, C., 2012. Diffusion lesion reversal after thrombolysis: a mr correlate of early neurological improvement. Stroke 43, 2986–2991.

Lang, K.J., Waibel, A.H., Hinton, G.E., 1990. A time-delay neural network architecture for isolated word recognition. Neural networks 3, 23–43.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Lev, M., Ackerman, R., Lustrin, E., Brown, J., Gahn, G., Chehade, R., 1995. Procedure for accurate spiral CT angiographic measurement of luminal diameter, in: Radiology, RADIOLOGICAL SOC NORTH AMER 20TH AND NORTHAMPTON STS, EASTON, PA 18042. pp. 133–133.

Lev, M., Gonzalez, R., 2002. CT angiography and CT perfusion imaging, in: Brain mapping: the methods. Elsevier, pp. 427–484.

Lev, M.H., Farkas, J., Rodriguez, V.R., Schwamm, L.H., Hunter, G.J., Putman, C.M., Rordorf, G.A., Buonanno, F.S., Budzik, R., Koroshetz, W.J., et al., 2001. CT angiography in the rapid triage of patients with hyperacute stroke to intraarterial thrombolysis: accuracy in the detection of large vessel thrombus. Journal of computer assisted tomography 25, 520–528.

Lev, M.H., Nichols, S.J., 2000. Computed tomographic angiography and computed tomographic perfusion imaging of hyperacute stroke. Topics in Magnetic Resonance Imaging 11, 273–287.

Liebeskind, D.S., 2003. Collateral circulation. Stroke 34, 2279–2284.

Lin, M., Chen, Q., Yan, S., 2013. Network in network, in: International Conference on Learning Representations (ICLR).

Lin, T., Horne, B.G., Tino, P., Giles, C.L., 1996. Learning long-term dependencies in narx recurrent neural networks. IEEE Transactions on Neural Networks 7, 1329–1338.

Lisboa, R.C., Jovanovic, B.D., Alberts, M.J., 2002. Analysis of the safety and efficacy of intra-arterial thrombolytic therapy in ischemic stroke. Stroke 33, 2866–2871.

Liu, P., 2018. Stroke lesion segmentation with 2d novel cnn pipeline and novel loss function, in: International MICCAI Brainlesion Workshop, Springer. pp. 253–262.

Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T., Murray, C.J., 2006. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. The lancet 367, 1747–1757.

Lou, M., 2019. Can imaging extend the thrombolytic time window after stroke? Nature Reviews Neurology , 1.

Lövblad, K.O., Altrichter, S., Pereira, V.M., Vargas, M., Gonzalez, A.M., Haller, S., Sztajzel, R., 2015. Imaging of acute stroke: Ct and/or mri. Journal of Neuroradiology 42, 55–64.

Lucas, C., Heinrich, M.P., 2017. 2d multi-scale res-net for stroke segmentation.

Lucas, C., Kemmling, A., Bouteldja, N., Aulmann, L.F., Mamlouk, A.M., Heinrich, M.P., 2018. Learning to predict ischemic stroke growth on acute ct perfusion data by interpolating low-dimensional shape representations. Frontiers in neurology 9.

Luengo-Fernandez, R., Leal, J., Gray, A., 2015. Uk research spend in 2008 and 2012: comparing stroke, cancer, coronary heart disease and dementia. BMJ open 5, e006648.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, p. 3.

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.

Maier, O., Wilms, M., von der Gablentz, J., Krämer, U.M., Münte, T.F., Handels, H., 2015. Extra tree forests for sub-acute ischemic stroke lesion segmentation in MR sequences. Journal of neuroscience methods 240, 89–100.

Markus, R., Reutens, D., Kazui, S., Read, S., Wright, P., Pearce, D., Tochon-Danguy, H., Sachinidis, J., Donnan, G., 2004. Hypoxic tissue in ischaemic stroke: persistence and clinical consequences of spontaneous survival. Brain 127, 1427–1436.

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 .

McKinley, R., Hung, F., Wiest, R., Liebeskind, D.S., Scalzo, F., 2018. A machine learning approach to perfusion imaging with dynamic susceptibility contrast mr. Frontiers in neurology 9, 717.

Menon, B.K., Qazi, E., Nambiar, V., Foster, L.D., Yeatts, S.D., Liebeskind, D., Jovin, T.G., Goyal, M., Hill, M.D., Tomsick, T.A., et al., 2015. Differential effect of baseline computed tomographic angiography collaterals on clinical outcome in patients enrolled in the interventional management of stroke iii trial. Stroke 46, 1239–1244.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE. pp. 565–571.

Mok, T.C., Chung, A.C., 2017. Deep adversarial networks for stroke lesion segmentation .

Monteiro, M., Oliveira, A.L., 2017. Fully convolutional neural network for 3d stroke lesion segmentation .

Muir, K.W., Weir, C.J., Murray, G.D., Povey, C., Lees, K.R., 1996. Comparison of neurological scales and scoring systems for acute stroke prognosis. Stroke 27, 1817–1820.

Mullins, M.E., Lev, M.H., Bove, P., O'Reilly, C.E., Saini, S., Rhea, J.T., Thrall, J.H., Hunter, G.J., Hamberg, L.M., Gonzalez, R.G., 2004. Comparison of image quality between conventional and low-dose nonenhanced head CT. American Journal of Neuroradiology 25, 533–538.

Mullins, M.E., Schaefer, P.W., Sorensen, A.G., Halpern, E.F., Ay, H., He, J., Koroshetz, W.J., Gonzalez, R.G., 2002. Ct and conventional and diffusion-weighted mr imaging in acute stroke: study in 691 patients at presentation to the emergency department. Radiology 224, 353–360.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. 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.

Nakamura, H., Yamada, K., Kizu, O., Ito, H., Yuen, S., Ito, T., Yoshikawa, K., Shiga, K., Nakagawa, M., Nishimura, T., 2005. Effect of thin-section diffusion-weighted mr imaging on stroke diagnosis. American journal of neuroradiology 26, 560–565.

National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995. Tissue plasminogen activator for acute ischemic stroke. New England Journal of Medicine 333, 1581–1588.

Nie, D., Cao, X., Gao, Y., Wang, L., Shen, D., 2016. Estimating ct image from mri data using 3d fully convolutional networks, in: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 170–178.

Nielsen, A., Hansen, M.B., Tietze, A., Mouridsen, K., 2018. Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. Stroke 49, 1394–1401.

Oliveira, A., Pereira, S., Silva, C.A., 2018. Retinal vessel segmentation based on fully convolutional neural networks. Expert Systems with Applications 112, 229–242.

Østergaard, L., 2005. Principles of cerebral perfusion imaging by bolus tracking. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 22, 710–717.

Østergaard, L., Weisskoff, R.M., Chesler, D.A., Gyldensted, C., Rosen, B.R., 1996. High resolution measurement of cerebral blood flow using intravascular tracer bolus passages. part i: Mathematical approach and statistical analysis. Magnetic resonance in medicine 36, 715–725.

rt PA Stroke Study Group, N., et al., 1998. Effect of rt-PA on ischemic stroke lesion size by computed tomography: preliminary results from the NINDS rt-PA stroke trial. Stroke 29, 287.

dela Peña, I., Borlongan, C., Shen, G., Davis, W., 2017. Strategies to extend thrombolytic time window for ischemic stroke treatment: an unmet clinical need. Journal of stroke 19, 50.

Pendlebury, S.T., 2007. Worldwide under-funding of stroke research. International Journal of Stroke 2, 80–84.

Pereira, S., Alves, V., Silva, C.A., 2018a. Adaptive feature recombination and recalibration for semantic segmentation: application to brain tumor segmentation in mri, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 706–714.

Pereira, S., Meier, R., McKinley, R., Wiest, R., Alves, V., Silva, C.A., Reyes, M., 2018b. 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: 2017 IEEE 5th Portuguese meeting on bioengineering (ENBENG), IEEE. pp. 1–4.

Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE transactions on medical imaging 35, 1240–1251.

Petty, M.A., Wettstein, J.G., 1999. White matter ischaemia. Brain Research Reviews 31, 58–64.

Pinheiro, G.R., Voltoline, R., Bento, M., Rittner, L., 2018. V-net and u-net for ischemic stroke lesion segmentation in a small dataset of perfusion data, in: International MICCAI Brainlesion Workshop, Springer. pp. 301–309.

Pinto, A., McKinley, R., Alves, V., Wiest, R., Silva, C.A., Reyes, M., et al., 2018a. Stroke lesion outcome prediction based on MRI imaging combined with clinical information. Frontiers in Neurology 9, 1060.

Pinto, A., Pereira, S., Meier, R., Alves, V., Wiest, R., Silva, C.A., Reyes, M., 2018b. 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 2018, pp. 107–115.

Pinto, A., Pereira, S., Rasteiro, D., Silva, C.A., 2018c. Hierarchical brain tumour segmentation using extremely randomized trees. Pattern Recognition 82, 105–117.

Pisov, M., Belyaev, M., Krivov, E., 2017. Neural networks ensembles for ischemic stroke lesion segmentation .

Polyak, B.T., 1964. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics 4, 1–17.

Powers, W.J., Rabinstein, A.A., Ackerson, T., Adeoye, O.M., Bambakidis, N.C., Becker, K., Biller, J., Brown, M., Demaerschalk, B.M., Hoh, B., et al., 2018. 2018 guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the american heart association/american stroke association. stroke 49, e46–e99.

Pressman, B.D., Tourje, E.J., Thompson, J.R., 1987. An early CT sign of ischemic infarction: increased density in a cerebral artery. American Journal of Roentgenology 149, 583–586.

Quinn, T., Dawson, J., Walters, M., 2008. Dr john rankin; his life, legacy and the 50th anniversary of the rankin stroke scale. Scottish medical journal 53, 44–47.

Rekik, I., Allassonnière, S., Carpenter, T.K., Wardlaw, J.M., 2012. Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. a critical appraisal. NeuroImage: Clinical 1, 164–178.

Rempp, K.A., Brix, G., Wenz, F., Becker, C.R., Gückel, F., Lorenz, W.J., 1994. Quantification of regional cerebral blood flow and volume with dynamic susceptibility contrast-enhanced mr imaging. Radiology 193, 637–641.

Rha, J.H., Saver, J.L., 2007. The impact of recanalization on ischemic stroke outcome: a meta-analysis. Stroke 38, 967–973.

Rimmele, D., Thomalla, G., 2014. Wake-up stroke: clinical characteristics, imaging findings, and treatment option–an update. Frontiers in neurology 5, 35.

Robben, D., Boers, A.M., Marquering, H.A., Langezaal, L.L., Roos, Y.B., van Oostenbrugge, R.J., van Zwam, W.H., Dippel, D.W., Majoie, C.B., van der Lugt, A., et al., 2018. Prediction of final infarct volume from native ct perfusion and treatment parameters using deep learning. arXiv preprint arXiv:1812.02496 .

Robben, D., Suetens, P., 2017. Dual-scale fully convolutional neural network for final infarct prediction, in: Ischemic stroke lesion segmentation-ISLES challenge 2017, held in conjunction with MICCAI 2017, Date: 2017/09/10-2017/09/10, Location: Quebec City, Quebec, Canada.

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, Springer. pp. 234–241.

Rose, S.E., Chalk, J.B., Griffin, M.P., Janke, A.L., Chen, F., McLachan, G.J., Peel, D., Zelaya, F.O., Markus, H.S., Jones, D.K., et al., 2001. Mri based diffusion and perfusion predictive model to estimate stroke evolution. Magnetic resonance imaging 19, 1043–1053.

Rosen, B.R., Belliveau, J.W., Vevea, J.M., Brady, T.J., 1990. Perfusion imaging with nmr contrast agents. Magnetic resonance in medicine 14, 249–265.

Rosenblatt, F., 1958. Two theorems of statistical separability in the perceptron. United States Department of Commerce.

Roy, A.G., Navab, N., Wachinger, C., 2018. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 421–429.

Rumelhart, D.E., McClelland, J.L., 1986. Parallel distributed processing: explorations in the microstructure of cognition. volume 1. foundations .

Rummelhart, D.E., McClelland, J.L., 1986. Parallel distributed processing: Explorations in the microstructure of cognition. Foundations 1.

Sandercock, P., Willems, H., 1992. Medical treatment of acute ischaemic stroke. The Lancet 339, 537–539.

Saver, J.L., Goyal, M., Van der Lugt, A., Menon, B.K., Majoie, C.B., Dippel, D.W., Campbell, B.C., Nogueira, R.G., Demchuk, A.M., Tomasello, A., et al., 2016. Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: a meta-analysis. Jama 316, 1279–1289.

Scalzo, F., Hao, Q., Alger, J.R., Hu, X., Liebeskind, D.S., 2012. Regional prediction of tissue fate in acute ischemic stroke. Annals of Biomedical Engineering 40, 2177–2187.

Schaefer, P.W., Barak, E.R., Kamalian, S., Gharai, L.R., Schwamm, L., Gonzalez, R.G., Lev, M.H., 2008. Quantitative assessment of core/penumbra mismatch in acute stroke: CT and MR perfusion imaging are strongly correlated when sufficient brain volume is imaged. Stroke 39, 2986–2992.

Schaefer, P.W., Hunter, G.J., He, J., Hamberg, L.M., Sorensen, A.G., Schwamm, L.H., Koroshetz, W.J., Gonzalez, R.G., 2002. Predicting cerebral ischemic infarct volume with diffusion and perfusion mr imaging. American Journal of Neuroradiology 23, 1785–1794.

Schellinger, P.D., Fiebach, J.B., Hacke, W., 2003. Imaging-based decision making in thrombolytic therapy for ischemic stroke: present status. Stroke 34, 575–583.

Schirmer, M.D., Giese, A.K., Fotiadis, P., Etherton, M.R., Cloonan, L., Viswanathan, A., Greenberg, S.M., Wu, O., Rost, N., 2019. Spatial signature of white matter hyperintensities in stroke patients. Frontiers in neurology 10, 208.

Schmidhuber, J., 1992. Learning complex, extended sequences using the principle of history compression. Neural Computation 4, 234–242.

Selim, M., Fink, J.N., Kumar, S., Caplan, L.R., Horkan, C., Chen, Y., Linfante, I., Schlaug, G., 2002. Predictors of hemorrhagic transformation after intravenous recombinant tissue plasminogen activator: prognostic value of the initial apparent diffusion coefficient and diffusion-weighted lesion volume. Stroke 33, 2047–2052.

Simonsen, C.Z., Madsen, M.H., Schmitz, M.L., Mikkelsen, I.K., Fisher, M., Andersen, G., 2015. Sensitivity of diffusion-and perfusion-weighted imaging for diagnosing acute ischemic stroke is 97.5%. Stroke 46, 98–101.

SMIR, 2017. Ischemic stroke lesion segmentation challenge. URL: `http://www.isles-challenge.org`.

Smith, W.S., Roberts, H.C., Chuang, N.A., Ong, K.C., Lee, T.J., Johnston, S.C., Dillon, W.P., 2003. Safety and feasibility of a CT protocol for acute stroke: combined CT, CT angiography, and CT perfusion imaging in 53 consecutive patients. American Journal of Neuroradiology 24, 688–690.

Smith, W.S., Sung, G., Saver, J., Budzik, R., Duckwiler, G., Liebeskind, D.S., Lutsep, H.L., Rymer, M.M., Higashida, R.T., Starkman, S., et al., 2008. Mechanical thrombectomy for acute ischemic stroke: final results of the multi merci trial. Stroke 39, 1205–1212.

Smolensky, P., 1986. Information processing in dynamical systems: Foundations of harmony theory. Technical Report. Colorado Univ at Boulder Dept of Computer Science.

Song, S., et al., 2017. Temporal similarity perfusion mapping: A standardized and model-free method for detecting perfusion deficits in stroke. PloS one 12.

Song, T., 2019. Generative model-based ischemic stroke lesion segmentation. ArXiv abs/1906.02392.

Sorensen, A.G., Buonanno, F.S., Gonzalez, R.G., Schwamm, L.H., Lev, M.H., Huang-Hellinger, F.R., Reese, T.G., Weisskoff, R.M., Davis, T.L., Suwanwela, N., et al., 1996. Hyperacute stroke: evaluation with combined multisection diffusion-weighted and hemodynamically weighted echo-planar mr imaging. Radiology 199, 391–401.

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.

Stollenga, M.F., Byeon, W., Liwicki, M., Schmidhuber, J., 2015. Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation, in: Advances in Neural Information Processing Systems, pp. 2998–3006.

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.

Stys, P.K., 1998. Anoxic and ischemic injury of myelinated axons in CNS white matter: from mechanistic concepts to therapeutics. Journal of Cerebral Blood Flow & Metabolism 18, 2–25.

Sutskever, I., Martens, J., Dahl, G.E., Hinton, G.E., 2013. On the importance of initialization and momentum in deep learning. ICML (3) 28, 5.

Taghanaki, S.A., Zheng, Y., Zhou, S.K., Georgescu, B., Sharma, P., Xu, D., Comaniciu, D., Hamarneh, G., 2019. Combo loss: Handling input and output imbalance in multi-organ segmentation. Computerized Medical Imaging and Graphics 75, 24–33.

Thijs, V.N., Lansberg, M.G., Beaulieu, C., Marks, M.P., Moseley, M.E., Albers, G.W., 2000. Is early ischemic lesion volume on diffusion-weighted imaging an independent predictor of stroke outcome? a multivariable analysis. Stroke 31, 2597–2602.

Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C., 2015. Efficient object localization using convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 648–656.

Tseng, K.L., Lin, Y.L., Hsu, W., Huang, C.Y., 2017. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6393–6400.

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., et al., 2010. N4itk: improved n3 bias correction. IEEE T MED IMAGING 29, 1310–1320.

Villringer, A., Rosen, B.R., Belliveau, J.W., Ackerman, J.L., Lauffer, R.B., Buxton, R.B., Chao, Y.S., Wedeenand, V.J., Brady, T.J., 1988. Dynamic imaging with lanthanide chelates in normal brain: contrast due to magnetic susceptibility effects. Magnetic resonance in medicine 6, 164–174.

Visin, F., Ciccone, M., Romero, A., Kastner, K., Cho, K., Bengio, Y., Matteucci, M., Courville, A., 2016. Reseg: A recurrent neural network-based model for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 41–48.

Visin, F., Kastner, K., Cho, K., Matteucci, M., Courville, A., Bengio, Y., 2015. Renet: A recurrent neural network based alternative to convolutional networks. arXiv preprint arXiv:1505.00393 .

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.

Warach, S., 2001. Tissue viability thresholds in acute stroke. Stroke 32, 2460–2461.

Wardlaw, J., 2010. Neuroimaging in acute ischaemic stroke: insights into unanswered questions of pathophysiology. Journal of internal medicine 267, 172–190.

Wardlaw, J., Brazzelli, M., Miranda, H., Chappell, F., McNamee, P., Scotland, G., Quayyum, Z., Martin, D., Shuler, K., Sandercock, P., et al., 2014. An assessment of the cost-effectiveness of magnetic resonance, including diffusion-weighted imaging, in patients with transient ischaemic attack and minor stroke: a systematic review, meta-analysis and economic evaluation .

Wardlaw, J., Dorman, P., Lewis, S., Sandercock, P., 1999. Can stroke physicians and neuroradiologists identify signs of early cerebral infarction on CT? Journal of Neurology, Neurosurgery & Psychiatry 67, 651–653.

Wardlaw, J.M., Murray, V., Berge, E., Del Zoppo, G., Sandercock, P., Lindley, R.L., Cohen, G., 2012. Recombinant tissue plasminogen activator for acute ischaemic stroke: an updated systematic review and meta-analysis. The Lancet 379, 2364–2372.

Weinman, J., Bissias, G., Horowitz, J., Riseman, E., Hanson, A., 2003. Nonlinear diffusion scale-space and fast marching level sets for segmentation of mr imagery and volume estimation of stroke lesions, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 496–504.

WHO MONICA Project, et al., 1990. Monica manual, part iv: Event registration, section 2: Stroke event registration data component.

Williams, D.S., Detre, J.A., Leigh, J.S., Koretsky, A.P., 1992. Magnetic resonance imaging of perfusion using spin inversion of arterial water. Proceedings of the National Academy of Sciences 89, 212–216.

Winzeck, S., Hakim, A., McKinley, R., Pinto, J.A., Alves, V., Silva, C., Pisov, M., Krivov, E., Belyaev, M., Monteiro, M., et al., 2018. Isles 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral mri. Frontiers in neurology 9.

World Health Organization, et al., 2007. The world health report 2007: a safer future: global public health security in the 21st century .

Wu, L., Shen, C., Hengel, A.v.d., 2016. Convolutional lstm networks for video-based person reidentification. arXiv preprint arXiv:1606.01609 .

Xiang, L., Wang, Q., Nie, D., Zhang, L., Jin, X., Qiao, Y., Shen, D., 2018. Deep embedding convolutional neural network for synthesizing ct image from t1-weighted mr image. Medical image analysis 47, 31–44.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c., 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Advances in neural information processing systems, pp. 802–810.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, pp. 2048–2057.

Young, F.B., Weir, C.J., Lees, K.R., 2005. Comparison of the national institutes of health stroke scale with disability outcome measures in acute stroke trials. Stroke 36, 2187–2192.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer. pp. 818–833.

Zhang, Y., Yang, C., Yang, A., Xiong, C., Zhou, X., Zhang, Z., 2015. Feature selection for classification with class-separability strategy and data envelopment analysis. Neurocomputing 166, 172–184.

Zivin, J.A., 1998. Factors determining the therapeutic window for stroke. Neurology 50, 599–603.