Universidade do Minho
Escola de Engenharia

João Vieira da Silva Fernandes

A deep learning approach for quantification of brain characteristics using MRI

Dissertação de Mestrado
Mestrado integrado em Engenharia Biomédica
Ramo de Informática Médica

Trabalho efetuado sob a orientação de
Victor Alves
Pim Moeskops

Novembro de 2018

DECLARAÇÃO

Nome: João Vieira da Silva Fernandes

Endereço eletrónico: joaovieirafernandes@hotmail.com

Título da dissertação: *A deep learning approach for quantification of brain characteristics using MRI*

Orientadores: Victor Alves, Pim Moeskops

Ano de Conclusão: 2018

Designação do Mestrado: Mestrado Integrado em Engenharia Biomédica

Área de Especialização: Informática Médica

Escola de Engenharia

Universidade do Minho, _____/_____/_____

Assinatura: _____

ACKNOWLEDGEMENTS

The Netherlands

In first place I would like to thank my supervisor at Eindhoven University of Technology (TU/e), Pim Moeskops for his guidance and help during this project. His knowledge and disposition were vital in several steps of the dissertation.

During my project at TU/e, I was immersed in a helpful and kind environment. To every person belonging to the Medical Image Analysis group at TU/e (Maxime, Koen, Mitko, Friso, Josien, Marcel) for all the availability, counselling and nice conversations during group meetings.

A special thanks to my fellow "co-workers" Cristiana Tiago and Leroy Volmer, who were also doing their own dissertations, for the tips but, mainly, for the friendship we created for life.

To all the friends from different countries that I met, for all the many adventures we experienced both inside and outside of Eindhoven. Javier, Nikos, Giannis, Filippo, Mixalis, Enrico, I will see you soon...

Portugal

I would like to express my gratitude to my supervisor at University of Minho, Victor Alves for his knowledge and availability. His support was always necessary in the last stages of the dissertation.

During my road until I graduate, I will never forget the *SpamCrew* members, which were my friends since the beginning and accompanied me until the end. To all the café meetings, dinners, conversations, headaches and adventures we shared.

Finally, to my parents who supported me since day one, in every single matter possible. They had been with me during this road and helped me climb many obstacles during my life.

ABSTRACT

Preterm birth is connected to impairments, altered brain growth and underdevelopment. Preterm infants have a high risk of problems related to cognition, socialization and behavior later in life since most part of their brain development is in extra-uterine conditions. By studying the anatomy of the brain, many disabilities can be found, and MRI brain scans are important in the identification of these details.

Accordingly, the goal of this dissertation is the development of different Deep Learning (DL) approaches with the objective of quantifying the volumes of 8 brain tissues and other 5 descriptors, necessary for the prediction of social and cognitive disabilities. Without DL, the first step for this quantification is a previous segmentation of the studies. With DL, it is possible to quantify directly from the brain MRI scans, skipping the segmentation step.

As data, two datasets were used: one with 86 studies of preterm patients around 30 weeks PMA and the other with 153 preterm patients around 40 weeks PMA. Non-image data was also needed for some ground truth descriptors. Three main procedures are performed: 2D approaches, 3D approaches using chunks of the brain and a 3D approach with full brains. Each approach consists of image processing techniques such as image cropping and downscale, a DL training and validation steps, prediction of unseen studies and, finally, statistical metrics combined with prediction vs ground truth graphs to assess the results.

The 2D approaches allowed high correlation (a mean of 0.95 for the 8 tissues for the dataset at 30 weeks PMA and a mean of 0.98 for the 40 weeks one) and low normalized RMSE (respectively, a mean of 6.26% and 5.98% for the 8 tissues). The 3D approaches using (small) chunks allowed a slightly improvement on the prediction vs ground truth graphs and metrics for both datasets (respectively, a mean correlation of 0.97 and 0.98 and a mean normalized RMSE of 5.69% and 5.38%, for all tissues). Using full 3D brains only allowed good predictions for half of the descriptors.

A second case study, using a dataset of MRI brain scans on rats (139 studies), was also executed in order to predict the volumes of 3 tissues. A 2D approach and a 3D approach were developed and the 2D one revealed to be the most accurate. These approaches are very similar to the ones developed for the neonatal brain.

RESUMO

Os nascimentos prematuros estão relacionados com deficiências, crescimento cerebral alterado e subdesenvolvimento. Os bebés prematuros têm um risco maior de ter problemas relacionados com cognição, socialização e comportamento mais tarde na vida, dado que a maior parte do seu desenvolvimento cerebral ocorre em condições extrauterinas. Ao estudar a anatomia do cérebro, muitas deficiências podem ser detetadas, e os exames cerebrais de ressonância magnética são importantes na identificação desses detalhes.

Nesse sentido, o objetivo desta tese é o desenvolvimento de diferentes abordagens de *Deep Learning* (DL) com o objetivo de quantificar os volumes de 8 tecidos cerebrais e de outros 5 descritores, necessários para a previsão de deficiências cognitivas e sociais. Sem DL, o primeiro passo para esta quantificação e uma segmentação prévia dos *studies*. Com DL, é possível quantificar diretamente a partir dos exames de ressonância magnética do cérebro, ignorando o passo da segmentação.

Como material, foram utilizados dois conjuntos de dados: um com 86 *studies* de pacientes pré-termo por volta das 30 semanas de idade pós-menstrual (PMA) e outro com 153 pacientes pré-termo por volta de 40 semanas IPM. Material não-imagem também foi necessário como *ground truth* para alguns descritores. Três abordagens principais são executadas: abordagens 2D, abordagens 3D usando pedaços de cérebro e uma abordagem 3D com cérebros inteiros. Cada método consiste em técnicas de processamento de imagem tais como cortes de imagem e redução de tamanho da mesma, passos de treino e validação em DL, previsão de *studies* nunca vistos e, finalmente, métricas de estatística combinadas com gráficos previsão vs *ground truth* para avaliar os resultados.

As abordagens 2D permitiram uma correlação alta (média de 0.95 para os 8 tecidos para o *dataset* de 30 semanas PMA e uma média de 0.98 para o de 40 semanas PMA) e um baixo RMSE normalizado (respetivamente, uma média de 6.26% e 5.98% para os 8 tecidos). As abordagens 3D usando pedaços (pequenos) permitiram uma ligeira melhoria nas métricas para os dois conjuntos de dados (respetivamente, uma correlação de 0.97 e 0.98 e um RMSE normalizado médio de 5.69% e 5.38% para todos os tecidos). Usando cérebros 3D inteiros permitiu boas previsões para metade dos descritores.

Um segundo caso de estudo, usando um *dataset* com exames de ressonância em cérebros de ratos (139 *studies*), foi também executados no sentido de prever os volumes de 3 tecidos. Um método 2D e um 3D foram desenvolvidos e o 2D revelou ser mais preciso. Estes métodos são muito semelhantes aos que foram desenvolvidos para o cérebro neonatal.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Abbreviations

A

ACD Automated Cardiac Diagnosis

Adam Adaptive Moment Estimation

ANN Artificial Neural Network

B

BET Brain Extraction Tool

BGT Basal Ganglia and Thalami

BS Brain Stem

C

CB Cerebellum

CNN Convolutional Neural Network

CPU Central Processing Unit

CSF Cerebrospinal Fluid

eCSF extracerebral Cerebrospinal Fluid

vCSF ventricular Cerebrospinal Fluid

CT Computer Tomography

D

DL Deep Learning

DNN Deep Neural Network

E

EPVS Enlarged Perivascular Spaces

F

FA Fractional Anisotropy

FC Fully Connected

FCN Fully Convolutional Network

FCRN Fully Convolutional Regression Network

FSL FMRIB Software Library

G

GA Gestational Age

GB Gigabyte

GI Gyrification Index

GM cortical Gray Matter

GPU Graphics Processing Unit

I

IS Inner cortical Surface area

M

MC global Mean Curvature

MICCAI Medical Image Computing and Computer-Assisted Intervention

ML Machine Learning

MR Magnetic Resonance

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

MT Median cortical Thickness

N

kNN k-Nearest Neighbor

NPC1 Niemann Pick Type C1

O

OCT Optical Coherence Tomography

OS Outer cortical Surface area

P

PET Positron Emission Tomography

PMA Postmenstrual Age

PNBI Premature Neonatal Brain Injuries

PReLU Parametric Rectified Linear Unit

R

R pearson correlation

ReLU Rectified Linear Unit

RF Radio Frequency

RMSprop Root Mean Square Propagation

RNN Recurrent Neural Network

ROI Region of Interest

RPN Region Proposal Network

S

SGD Stochastic Gradient Descent

SVM Support Vector Machine

U

UMCU University Medical Center Utrecht

cUS cranial Ultrasounds

W

WM White Matter

mWM myelinated White Matter

uWM unmyelinated White Matter

GLOSSARY

Artificial Intelligence Area that focuses on the creation of computer systems, based in the human intelligence, that solve difficult tasks that are usually solved by humans (or animals).

Artificial Neural Network Systems, whose architecture is inspired in the human brain, that replicate the way that humans learn. In a simple, way, they are composed of layers with units and connections between them. Convolutions neural networks are one example of a complex/deep artificial neural network, focused on convolutions.

Automatic Feature Extraction Determination of relevant characteristics in a set of descriptive attributes from raw input data, performed automatically by the system.

Basal ganglia and Thalami Basal ganglia are a group of nuclei located in the base of the brain. These structures are connected to involuntary movements such as tremors. Thalami are all the thalamus' parts. The thalamus is responsible, for example, for consciousness and sleep.

Batch Sample containing some training examples (and corresponding labels), in order to train the neural network.

Brain Stem A simple region that connects the spinal cord to the brain. It is related to reflexes' coordination and control of some organs like the heart and lungs.

Cerebellum Very easy to identify, it is responsible for coordination and balance of the muscles of the body.

Cerebrospinal Fluid Formed in the ventricles, it is a colorless liquid that surrounds the brain (and the spinal cord). Helps in maintaining pressure in the cranium and serves as a lubricator between brain and bones.

Deep Learning System System that was trained and validated with data and is capable of making a prediction with relative accuracy for a new example of that data.

Grey Matter The grey tissue of the brain (and spinal cord). It consists of nerve cell bodies, dendrites, bare axons and synapses. It includes regions involved in sensory perception and muscle control.

Gyrification Index Measures the magnitude of cortical convolutions (i.e. cortical fording) of the brain.

Medium Thickness Helps in tracking a quantitative cortical development in preterm infants and can be helpful in diagnosing and study neurodegenerative and psychiatric disorders.

Mean Curvature Important is assessing cortical development, helping in the detection of pathologies.

MRI study Also called MRI scan, it uses a magnet and radio waves connected to a computer with the objective of creating detailed cross-sectional images of the body.

NIfTI File format widely used in Medical Image Analysis, more specifically in neuroimaging.

Overfitting When the model performs very well on training data, but very poorly on new samples of data.

Preterm Infants All the babies that were born alive before 37 weeks of pregnancy. They can be divided in 3 subgroups: moderate to mild preterm, very preterm and extremely preterm.

Preterm Outcome Associated to the risks that preterm babies suffer due to their underdevelopment. A preterm infant can be healthy or suffer cognitive, social and/or behavioral impairments due to those risks.

Surface of the brain Other quantitative measure to assess cortical growth. It can be divided in inner surface and outer surface.

Test Process with the sole objective of assessing the model's performance.

Training Process that uses samples, so the model can learn the task.

Underfitting It happens when the model performs poorly on training data due to its inability in learning a relationship between the data and a target value.

Validation Process to check if the training is going well and to adjust hyperparameters.

Voxel Basic unit of information in a three-dimensional space.

White matter This deeper tissue is considered the biggest one in the brain. It contains the myelin-covered axons and it improves the speed of transmission of electrical nerve signals. Preterm infants have small myelinated structures, but most of their white matter is still unmyelinated.

Wistar rats Breed of albino rats created in 1906 at the Wistar institute. This rat is currently one of the most popular for laboratory research. Some other laboratory breeds of rats were developed from this one.

1 Introduction

1.1 CONTEXT

Medical imaging deals with imaging data that need to be effectively and accurately processed in order to contribute to high-quality health care and progresses in science. This contribution leads to an improvement in the quality of life not only of patients, but also of physicians, by simplifying their tasks [1]. With the recent boost in technology, image processing applications have become a central part in medical imaging systems. As new different, robust and accurate methods are being created, this field is under a quick development allowing, in the current research, more insight into the human anatomy and functions [2]. Magnetic resonance imaging (MRI) of the brain is a painless and safe technique of medical imaging that produces detailed images by using magnetic field and radio waves, instead of radiation. MRI scans, specifically in brain, are superb on distinguishing between masses and lesions, adjacent soft tissues and blood vessels in three orthogonal planes without the need to reposition the patient [3].

The brain is considered the most complex organ in the human body. Although there are many studies related to its investigation and a lot of information about it, the brain is still one of the most mysterious and undiscovered parts of the human body. Brain anatomy is well recorded and understood, however, in the brain function field, there is still a lot to be uncovered.

Preterm children have now a higher survival rate due to the recent progresses in neonatal care and obstetrics. However, preterm birth is related to brain altered growth, underdevelopment and disabilities [4]. In the last third of pregnancy, the cerebral cortex starts developing in a rapid rate, from a simple and smooth structure to a complex, folded and more adult-like one [4], [5]. Preterm infants have most part of this important brain development in extra-uterine conditions, and that is the main reason why they have a higher risk of problems related to cognition, socialization and behavior later in life. In these conditions, several abnormalities can be found, and MRI brain scans play a vital role in this identification. Cerebral white matter (WM) and cortical grey matter (GM) are some of the structures that need to be evaluated in terms of tissue growth and injuries to determine if there are problems [6].

With the recent technology evolution in the Medical Image Analysis field, it is possible to retrieve more information about the brain functions and problems related to those functions. With brain MRI, it is possible to determine the brain tissues' regions and potential lesions, in order to access the brain functionalities and its relation to the classification of a subject as having/not having future

impairments. In other words, by assessing the different tissues that are shown in brain MRI, it is possible to quantify certain brain descriptors that are related to motor and cognitive outcomes in preterm infants. Then, MRI is a valuable tool to determine neurodevelopment disabilities by evaluating cortical development, more specifically volumes, surface areas, thickness and folding [5]. This prediction of disabilities can be very important in a real clinical environment to inform both physicians and parents if a preterm infant is likely or not to have the same cognitive and motor capacities as their normal counterparts.

Rats have a close resemblance to humans in terms of biological, genetic and behavior characteristics. In addition, there are many human processes and symptoms that are very similar in both humans and rats. Specifically, the brain is one of the organs that shares several conditions between the two mammals so, the rat brain can be used to answer many research questions about humans. In fact, several projects focus of the study of rat brain disorders that are shared by humans, like Parkinson [7] and Alzheimer [8] diseases.

1.2 MOTIVATION

Machine Learning (ML) (and more specifically Deep Learning (DL)) techniques are slowly replacing some of the tasks humans usually were doing and even some that they were not doing at all, since they were so time-consuming. In the medical field, these techniques have the capacity to perform the tasks in a much quicker way and, in some cases, with more accuracy and specificity than physicians [9]. Deep Learning is considered a very promising field inside Machine Learning, with no theoretical limitations and has already achieved great results in many different fields: from Medical Diagnosis [9] to Finance [10]. This way, a Deep Learning approach on image recognition tasks, like the quantification of brain MRI descriptors would be a step further in science by potentially discovering new information about this very interesting organ.

However, as previously mentioned, brain descriptors may include, for example, volumetric measurements. This way, to obtain accurate measurements, it is necessary to label the regions of interest in an image in the first place, i.e. segmentation. Segmentation also allows the computation of other descriptors so, it needs to be considered as an important step before doing the quantitative evaluation of cortical characteristics for brain development assessment in preterm newborns. Based

on those segmentations, many quantifications can be made by using some mathematical formulas [11].

Although the segmentation method has improved quality in the health environment, a deep learning system which directly predicts those quantifications is even more optimal and automatic. The biggest advantage of this system is that there is no need to perform an automatic segmentation, so we can skip this first step and still obtain the quantifications. This is a step further in the possibility of building a complete system that, given an MRI image as an input, returns a single number (for example, like age or motor and cognitive outcome) as an output.

Over the last few years, many deep learning approaches are being used to study the brain. These approaches focus on different aspects like, for example, segmentations, anatomical/functionalities identifications or detection of certain diseases. In this line of sight, an implementation of a system which quantifies several brain descriptors would be extremely useful for clinical practice. These quantifications are extremely useful when predicting if a certain neonatal subject is likely to have cognitive and/or motor impairments 2-3 years after birth. In other words, building a DL system for quantification of brain descriptors helps in obtaining insight for a future development of an ideal DL system that receives neonatal brain MRI scans as inputs and predicts the motor and cognitive outcomes for those subjects [12].

An analogy of a quantification system for a case study concerning rats is also important to evaluate in what extent can this quantification method be generalized to differently shaped brains. In addition, the study of rats is also relevant for human research. This way, the construction of a system that receives rat brain MRI scans to predict volumes can be seen as a first step for future studies concerning both human and rat brain.

This work presents an automatic method for quantification of neonatal and rat brain descriptors based on MRI scans, using Deep Learning, more specifically, Convolutional Neural Networks (CNNs). For the preterm infants, it is a step further in the construction of a complete system which investigates motor and cognitive impairments on preterm newborns, in the first years after birth. These descriptors are computed at 30 and 40 weeks postmenstrual age (PMA) images, with the purpose of evaluating the quantification in these two specific points. The developed techniques for the preterm brain are also adapted to a rat brain dataset in order to quantify 3 volumes of the brain of these animals.

1.3 OBJECTIVES

The dissertation main objective consists on the development of deep learning models to construct a system for brain descriptors' quantification of unusual MRI studies like the preterm and rat brain. For the preterm infants, this quantification can be studied in order to predict cognitive and motor impairments in these infants at 2-3 years' chronological age. The extension of the project to the quantification of the rat brain was due to the resemblance between human and rat brain. This way, instead of calculating descriptors based on segmentations, it is possible to directly predict a brain quantification from an MRI scan.

In a more detailed way, the following objectives are required to fulfil the overall purpose of the dissertation:

- Review related work in both neonatal and rat brain studies, and deep learning techniques.
- Implementation of different approaches/models, in terms of pre-processing techniques.
- Implementation and optimization of Deep Learning algorithms, more specifically CNNs to achieve a better performance of the system, using algorithm tuning.
- Tracking and analysis of loss curves and statistical results.
- Evaluation of the models with test data and comparison with reality.
- Evaluate and comparison of the different models/techniques constructed.

1.4 STRUCTURE OF THE DISSERTATION

This dissertation was incorporated in the fifth year of the Integrated Master's Degree in Biomedical Engineering, more specifically in the field of Medical Informatics. It is structured in 5 chapters, including this first chapter which gives an overview of the context of the project and clearly defines the problem as well as its motivation, objectives and methodology.

An introduction and description of some general theoretical concepts and technologies related to brain MRI and deep learning was necessary to provide an important background in these fields. This second chapter also focuses in the literature review which is essential to describe in which state of development these two main topics are nowadays. With the objective of being up-to-date with the

development in these areas, several papers had to be consulted for an accurate review. These articles were obtained by using queries with specific key words in Web of Science and Google Scholar.

The third chapter is dedicated to the quantification of descriptors using the neonatal brain datasets. It is divided in 3 main subsections. The first one exposes important considerations on the two neonatal datasets used for the study. The second one describes all the approaches developed to predict the descriptors. The third one is related to the results of the developed approaches, along with their discussion in order to identify some patterns and get some new knowledge concerning this area.

The topics related to the case study on the dataset of the rats is reported in a fourth chapter: data description, developed approaches and results with their assessment and discussion.

The final chapter serves to expose the conclusions and contributions of this research and ends with some ideas for future work.

2 Concepts and Related Work

This chapter focuses on the main topics of the dissertation. It describes important clinical considerations about neonatal care and covers brief deep learning concepts, as well as some useful information about rats. It has the objective to provide basic information for readers without any background on these areas. In addition, the chapter also tracks some important works developed in all these fields.

2.1 Neonatal Infants and the Brain

2.1.1 General considerations on neonatology

In a global scale, an estimated 15 million babies are born prematurely every year. This corresponds to more than 1 in 10 babies born before 37 completed weeks of gestation. Prematurity can be divided in extremely preterm (less than 28 weeks PMA), very preterm (28 to 32 weeks PMA) and moderate to late preterm (32 to 37 weeks PMA). Preterm birth has increased over the last decades, including both very preterm and moderate to late preterm births. Preterm birth complications are the main reason for neonatal deaths (around 1 million children), even surpassing intrapartum-related events or sepsis and meningitis [13], [14].

Preterm babies are often more susceptible to complications than normal babies. Although the increasing number of moderate to late preterm births (30%) is higher than very preterm ones (4%), in these last ones the mortality rate is much higher [15]. Despite some of the infants facing death, many survivors will have impairments, including learning, behavioral and social problems. Therefore, prediction of outcome after preterm birth is vital (but very difficult), mainly in the postnatal period.

For these children, many of those impairments are due to their immature brain development. For example, the brain weight at 34 weeks PMA is 35% lower than a term-born peer. As mentioned before, in the last weeks of a normal gestation, the brain tissues still have much to develop. A comparison of the brain at 4 different stages of development is performed in figure 2.1. Brain development can also be in jeopardy due to the fact that some organs, such as heart and lungs will have to adapt to extrauterine conditions earlier than normal births [4], [5], [15].

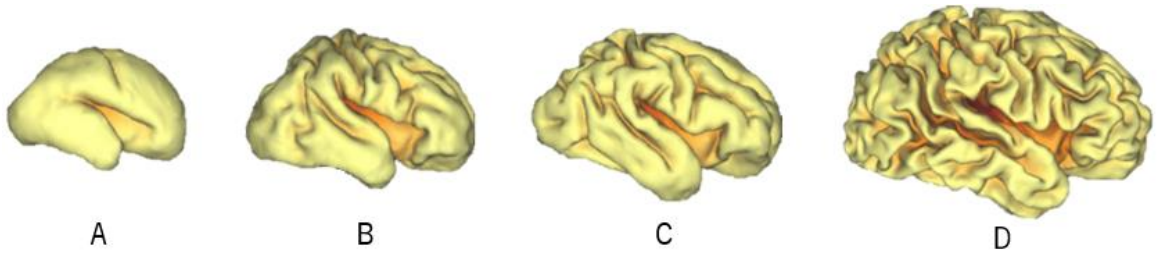


Figure 2.1 - The different stages of development of the brain: A- at 25 weeks PMA. B- at 30 weeks PMA. C- at 33 weeks PMA. D- term.

In a more personal way, the children and parents involved must endure financial, emotional and physical problems. In an economical way, these infants represent an additional burden for healthcare institutions. An improvement in the prediction of outcome for the neonatal period would increase the prevention of cerebral injuries, by providing a better selection of children for therapeutic treatment and a more accurate parental counseling [16].

Very preterm infants are more susceptible to severe disabilities, such as cerebral palsy, mental retardation or visual and hearing problems. These disabilities can be detected by the age of 2 in contrast to other less problematic impairments which are only detected until school by lower school/academic performance [16].

2.1.2 Brain MRI and neonatal care

2.1.2.1 The Basics of MRI

As previously remarked, MRI is considered a great tool in medical imaging of soft tissues (e.g. the brain) by not only assessing diseases or treatment progresses in a very effective way, but also because it is safer than other imaging modalities as it does not use radiation [3].

The concepts of MRI may sometimes be a little confusing due to the physics behind them. In a simple way, MRI focuses on the nucleus proton present in each hydrogen atom that composes water molecules all over the human body. Therefore, MRI takes pictures of the body parts that contain water. These hydrogen atoms are usually arranged in a random order, yet they produce a magnetic field (figure 2.2 A). When these atoms are exposed to a MRI scanner, it releases a first constant magnetic field that aligns the hydrogens in the same direction as this field (figure 2.2 B). A second magnetic field (Radio Frequency (RF) pulses) disrupts the protons as they absorb the energy and become misaligned to the first magnetic field, i.e. they are in a different direction (figure 2.2 C). When these pulses are switched off, the excited protons tend to realign to the first constant field,

emitting energy that is captured by MRI detectors (figure 2.2 D) [17]. These detectors have the capacity of identifying the shape and location of different tissues. The steps of this process are all illustrated in figure 2.2.

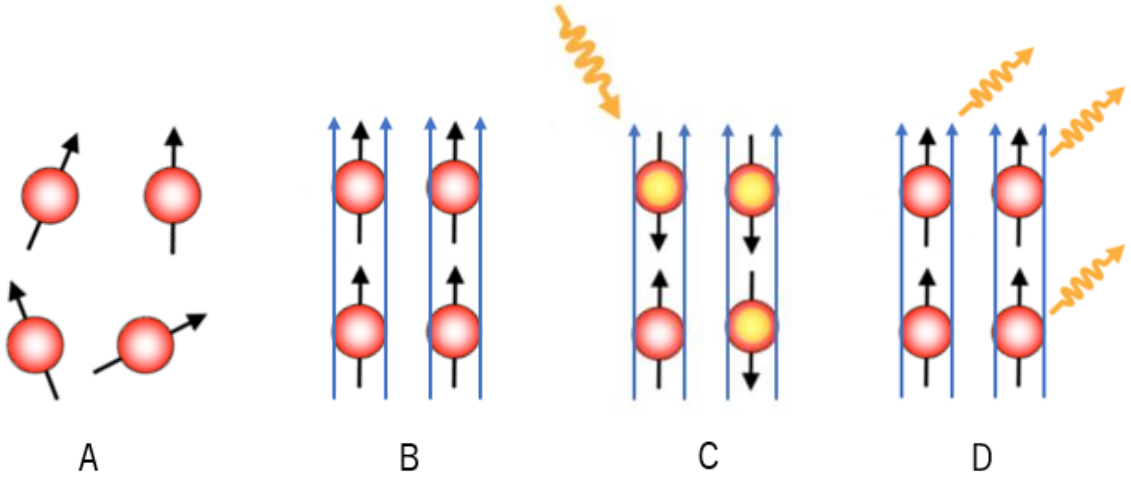


Figure 2.2 - Hydrogen protons' behavior in the stages of the MRI process: A- randomly arranged. B- arranged in the direction of the magnetic field (in blue). C- change of direction due to RF pulses. D- emission of energy by the excited protons.

2.1.2.2 MRI AND NEONATAL BRAIN

Some of the benefits of MRI described in the sections before underline the its importance in neonatology. Although it may be more challenging and expensive, this technique is proving to be more useful compared to head cranial ultrasounds (cUS).

In order to assess premature neonatal brain injuries (PNBI), white matter is a relevant tissue to examine. When imaging this tissue, cUS have limited temporal and spatial resolution. While cUS is reliable for the diagnosis of some severe WM lesions, MRI is more sensitive in the identification of smaller cerebellar hemorrhages, diffuse white matter injury and punctate white matter lesions, contributing even more for an improvement in the outcome prediction [18], [19]. This difference in accuracy in the detection of WM injuries is crucial in the choice of MRI over cUS for the correct assessment of PNBI.

In sum, in a research environment, MRI not only helps in outlining the neonatal brain development and its injuries, but also in deciding if a certain patient would heal from a new therapeutic treatment, by precisely evaluating the outcome of the strategies. In a clinical environment, MRI is a strong predictor of long-term prognosis and gives a perception of the extent, time and cause or a certain injury [20].

In newborns, especially preterm newborns, the brain is under constant development. As previously stated, these special infants have the disadvantage of having part of its brain development outside uterine conditions. For example, the brain of a patient at 30 weeks PMA will be less complex than the brain at 40 weeks PMA, as depicted in figure 2.3. As MRI can detect accurately WM and other tissues of the brain, it is the method of choice that can show, differentiate and clearly track all the tissues in the small neonate brain, which has its peculiarities depending on the stage, i.e. at 30 or 40 weeks PMA or in between.

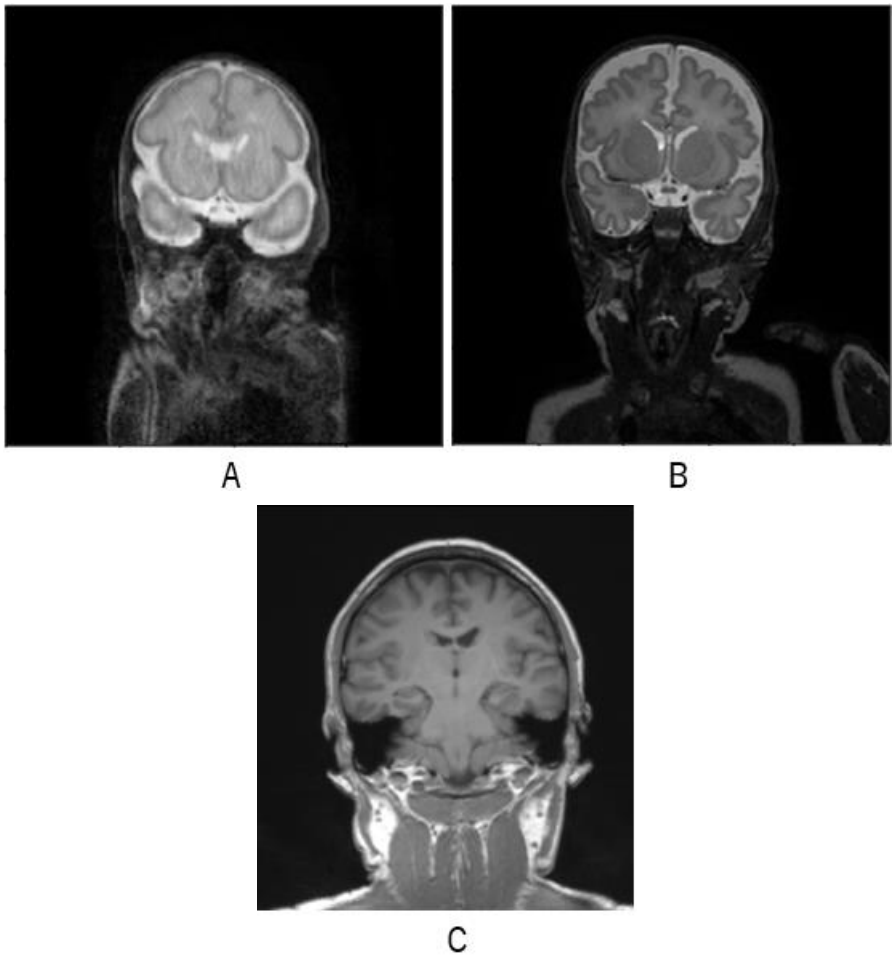


Figure 2.3 – The brain of a patient at 30 weeks PMA (A) and at 40 weeks PMA (B), in coronal T_2-weighted images. A coronal T_1-weighted example of an adult brain (C) is represented for comparison.

2.1.3 Brain Segmentation and Quantification

Although this trend is slowly changing, analysis of MRI brain scans to evaluate the brain condition is usually performed visually. This analysis includes brain quantification, which is critical when assessing injuries and pathologies and has a relation with the functional development of this organ,

and can comprise both 2D and 3D measurements [21], [22]. Doing quantification with MRI images is considered a credible task due to the high sensibility of this imaging modality.

However, manual segmentation performed by doctors is tiring, boring, time-consuming and sensitive to changes between observers. For few images, a manual segmentation and descriptor quantification would be feasible, although it would take a lot of time. However, in order to have a clear line of knowledge and establish patterns related peculiar details and impairments of preterm newborns, it is necessary to analyze and process many more images. Automatic methods should lead to a great performance for the system in classifying, segmenting or quantifying images, and many times even surpassing physicians' own accuracy and efficiency. In the segmentation step, which is represented in figure 2.4, the input is an image and the output is its segmentation. The quantification step, based in the segmentations, is depicted in figure 2.5.

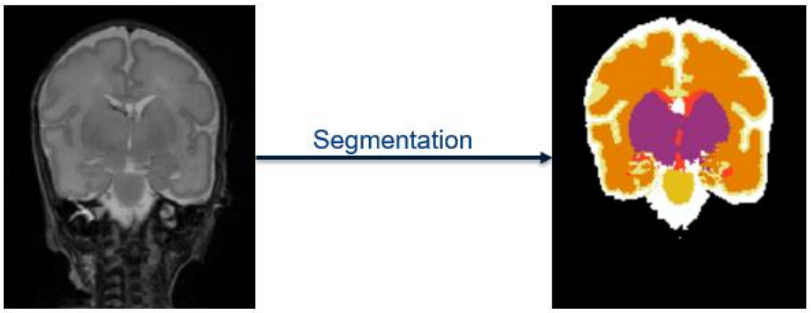


Figure 2.4 - The segmentation step of an image.

Figure 2.5 – The quantification step, based on a previous segmentation.

It is observed in figure 2.5 that the input is a segmentation of an image and the output is a series of quantified descriptors.

An automatic segmentation is a more feasible step in order to obtain the following quantification due to the fact that it allows a faster analysis of more scans and the outputs can easily be used to compute

volumes and other morphological descriptors such as, in case of the neonatal brain, cortical folding and cortical thickness. Naturally, the path for an automatic quantification directly from a scan is more complicated than the one for an automated segmentation owing to the fact that it combines two steps (i.e. automatic segmentation and posterior quantification) in one.

2.2 The study of rats in research

The study of the human brain is not easy. Many important discoveries about the human brain and its functionalities are connected to analogous studies on model organisms. Rats are inserted in these model organisms and they facilitate the study of the human brain due to the similar characteristics between these mammals' brains.

Even though many animals such as worms or zebra-fish are used in biomedical research, rats (and mice) have been the leading organisms to perform studies for over a century. In fact, nowadays, these rodents make nearly all laboratory animals. Consequently, the publications regarding research on these animals are constantly increasing [23].

The benefits provided by rat MRI on human research range from the solid foundation of molecules and cells to the creation of new therapies and medical procedures. Due to its significant sensitivity and reliability, MRI allows a viable connection between laboratory discoveries and the potential treatment of several diseases (translational research) [24].

As for the rat brain automatic segmentation, similar to any other segmentation, would allow a faster and less toilsome process as many examples have to be analyzed. The automatic quantification of volumes using DL techniques also permits skipping the segmentation step, making the process even more optimal.

2.3 Deep Learning

Learning is defined as being better at a certain task over the course of time. In the case of identifying a certain disease, humans need to learn how to identify this disease, i.e., they need to collect many cases where the patients had that disease and not and identify the exact features (for example, age, gender, physical condition) that would help them make the final prediction. Machine Learning tries

to mimic the human learning process. Essentially, it focuses on developing algorithms (i.e. complex mathematical functions) that can determine a correct prediction, given some features (like a age, gender, physical condition) [25]. In a simple way, ML algorithms combine these features given in the best manner with the objective of getting the correct conclusion. However, the main limitation of ML is that most features need to be identified by an expert and then hand-coded to serve as inputs for ML algorithms.

Deep Neural Networks (DNNs) represent the latest trend associated to Machine Learning. They can surpass other machine learning algorithms by having a better performance with a large amount of data. With DL, the explicit definition of hand-crafted features is not needed. Instead, depending on the whole task at hand, the network learns a set of kernels (i.e. features) which are trained to solve the initial problem [26]. In other words, with DL, it is possible to generate features from raw data.

A perfect analogy for DL is when someone must learn to identify, for example, a tree, provided by some tree examples. That person learns, from raw data, to predict/identify a tree, without someone providing features to him/her. If someone identified the exact features for the learner (like size, trunk, branches), it would be much more closer to a ML analogy [25].

2.3.1 From biology to Deep Learning

Artificial Neural Networks (ANN) are a popular classification and regression algorithms being increasingly used in the ML field in recent years. They have shown remarkable performances in finding solutions for certain problems. These networks were inspired in a similar way that signal transmission is performed in the human brain.

In brain signal transmission, a neuron receives various input signals from other neurons through the dendrites, processes and then, transmits them to other neurons through the axon terminals. ANNs are the translation of this biological fact to a computational model and are made up of many interconnected artificial neurons. Each of these neurons is associated to weights and outputs a response [26]. The mimic is represented in figure 2.6, which compares brain neurons (figure 2.6 A) with an artificial neuron (figure 2.6 B).

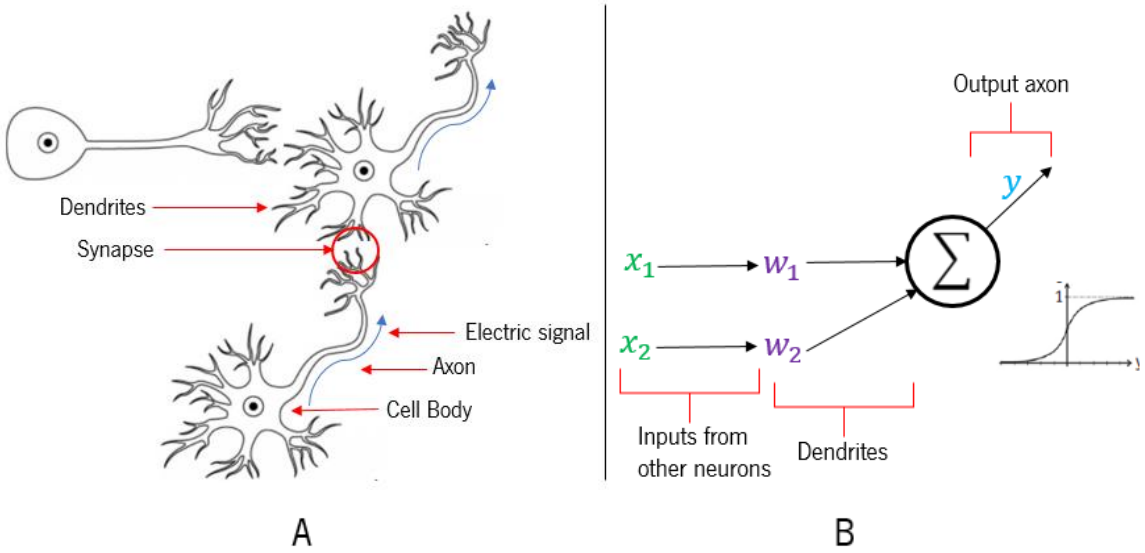


Figure 2.6 - Analogy from biology to the artificial neuron: A- brain neurons. B- artificial neuron.

These artificial neurons, when combined, they compose the ANNs. Although ANNs achieved excellent results in many fields, there were some limitations in these models like, for instance, overfitting [26] i.e. overtraining for the data. To surpass these limitations, the next step was to try deeper networks.

DNNs, where an example is represented in figure 2.7, are a specific type of ANN. The first layer, also called input layer, contains the observable examples in which the prediction focuses. The last layer, or output layer, presents that prediction, which could be a number, a series of numbers or classes. The hidden layers are between the first and last layers, and their neurons are connected to the inputs of other neurons so, their output is not visible. Their inputs are not also observable as they receive them from other layers [26] (hence the term hidden layer). This more complex structure in layers, allows to perform more simpler decisions each step and, produce, in total, a more complex decision.

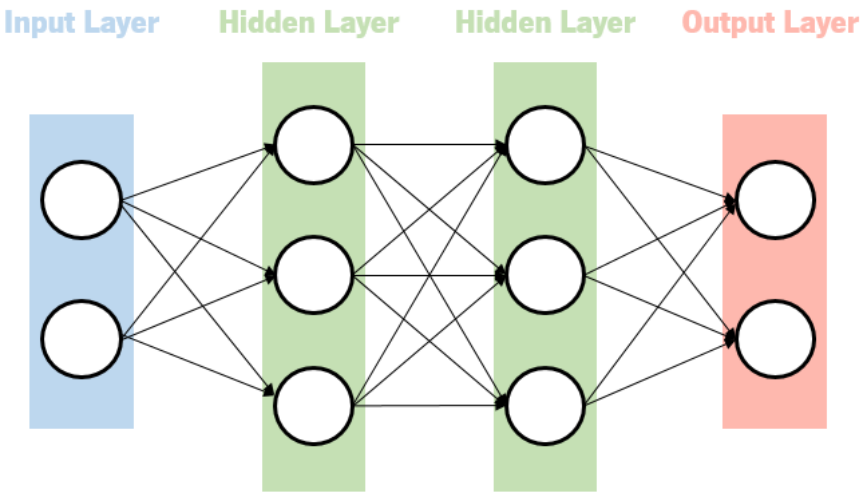


Figure 2.7 - An example of a DNN. This network is composed of an input layer, output layer and 2 hidden layers.

2.3.2 THE BASICS OF NEURAL NETWORKS

If a Neural Network was just input and output layers, it was just a linear approximation. Hidden layers are the ones that introduce non-linearities to the network. When certain inputs enter a determined neuron, that input data is weighted and after, a certain bias can be added. Then, the neuron has an **activation function** to make sure the output is not only an influence of the sum of weights. Considering a simple artificial neuron, figure 2.8 describes the structure of an artificial neuron and its standard output formula.

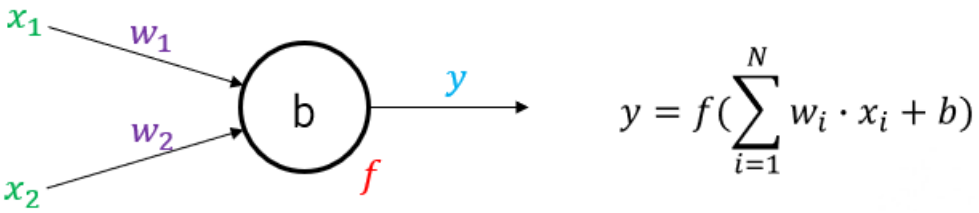
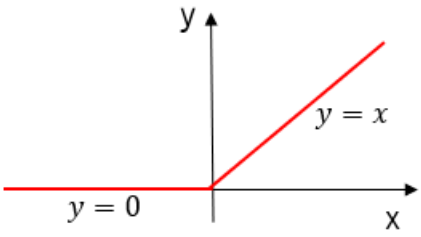


Figure 2.8 – The standard structure (left) and output formula (right) for each neuron.

Regarding the standard structure in figure 2.8, x_1 and x_2 represent the inputs of the neuron, while y represents the output; w_1 and w_2 represent the weights; b is the bias and f is the activation function of the neuron. In the formula on the right, w_i and x_i denote, respectively, all the weights and inputs connected to the neuron, while N denotes the number of connections.

Activation functions have the particularity of adding non-linearities to the neural network. Without these functions, the output of a network would be a simple linear function (i.e. a one-degree polynomial), which are limited in complexity and easy to solve. A Neural Network would be a linear regression model, which would be impossible to deal with complicated data like images, audio or video and would have less power to learn and have a good performance [25].

There are many examples of activation functions like sigmoid unit, linear unit, softmax layer, and others that are not in the scope of this work. In recent times, the most widely used [25] is the ReLU (Rectified Linear Unit), which graph and function are depicted in figure 2.9.



$$y = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

Figure 2.9 - ReLU graph (left) and formula (right).

ReLU is an easy computational step: as represented on the graph in figure 2.9, positive values stay the same and negative ones are set to 0. There are no complex operations as divisions or exponentials. In addition, ReLUs improve the networks by speeding up in training.

One of the main steps in building a DL model is to evaluate the **loss values** as the training occurs. When the network is being trained, each observable example will have a prediction associated. This prediction result will be compared to the labels that are also provided to the network. The further this prediction is from the ground truth labels, the highest the loss. Therefore, loss values suggest how the model behaves after each iteration of training [25]. There are some ways of evaluating the loss. For instance, it can be used the Squared Error loss function that plots a squared error of the differences between each estimated result and corresponding label.

The biggest objective when learning is to minimize the loss (or error) by changing the weights (and biases), after each training iteration, through **optimization methods**, as represented in Figure 2.10 A. There are many optimizers [27], such as RMSprop (Root Mean Square Propagation) or SGD (Stochastic Gradient Descent), which differ in the way that they update the weights in the network.

Related to optimization, the **learning rate** is the size of the step that is taken to optimize (i.e. update the weights). An adequate learning rate should be chosen as a higher learning rate would make the solution to bounce around too much (figure 2.10 B), and a lower one would have a slower convergence to the minimum (figure 2.10 C). In both cases, there would not be a good optimization of weights to reduce the error to the minimum.

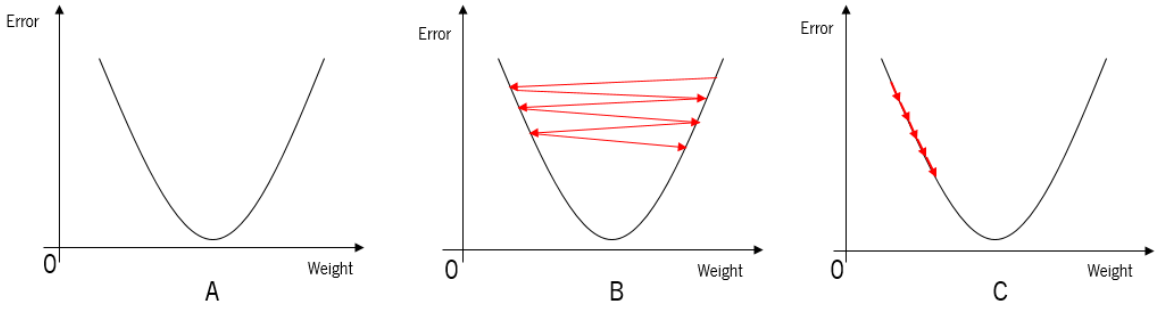


Figure 2.10 - Optimization and learning rate: A- a relation between error and weights (the objective is to minimize the error, by changing weights). B- a case of a high learning rate. C- a case of a low learning rate.

In many cases in ML (and DL), algorithms perform very well in training data, but when new test cases are presented, the performance will be worse (i.e. overfitting to the training data). There are many approaches that have the objective to reduce the test error, possibly at the expense of an increased training error. This definition is related to a large research field in DL called **regularization**. One of those regularization techniques is **dropout**. Dropout consists in randomly dropping units (and correspondent incoming and outgoing connections) during the training of the neural network. This reduces overfitting problems while training due to the fact that units will not adapt that much to the training dataset. This way, in training, layers will have less neurons [28]. Figure 2.11 compares a standard neural network with one with dropout.

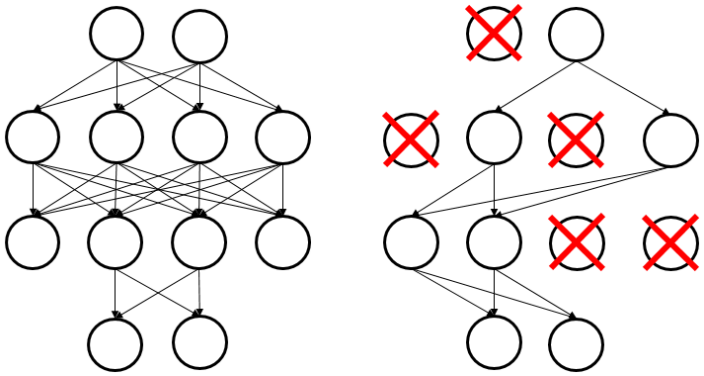


Figure 2.11 - A neural network without (left) and with dropout (right).

2.3.3 The Specific case of Convolution Neural Networks

Deep learning models are constituted by different types of DNN architectures. However, CNNs have become the most used architecture and the leading one for most detection tasks, image recognition and classification tasks. Their popularity is parallel to the recent success in the DL field as many advances occurred in Graphics Processing Unit (GPU) implementations and better algorithms and larger datasets emerged [29].

CNNs have biological inspiration just like the ANNs. The architecture was based on the fact that the animal visual brain cortex consists on alternating layers of complex and simple cells [26], [29]. They have many variations on account of people making changes with the objective of reducing computational costs or improve accuracy. Basically, CNNs consist of convolutional layers, activation functions (most used are ReLUs), pooling layers and fully connected layers. An example of this architecture of a CNN is shown in Figure 2.12.

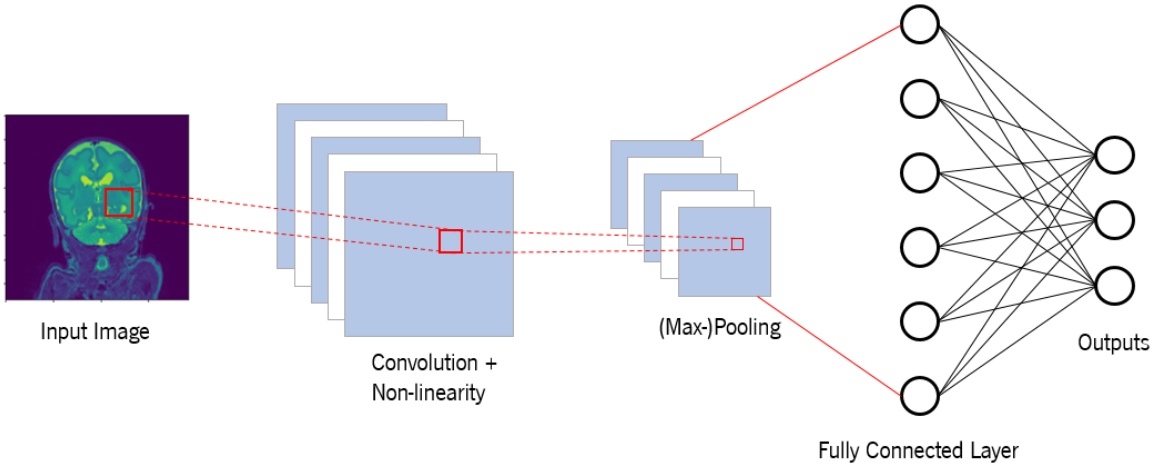


Figure 2.12 - The basic architecture of a CNN.

Convolutional layers are the feature extractors in the CNNs, serving mainly to detect lines, edges and other type of visual elements. To accomplish that, CNNs extract features from the input image by performing **convolutions**. This process applies several filters to an image in order to extract a "feature map" and is associated to non-linear functions (like ReLUs). Figure 2.13 describes in a simple way how this process is executed.

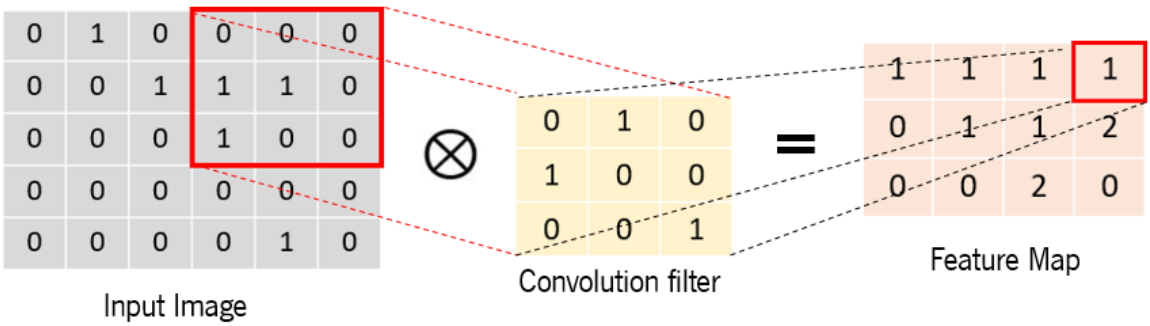


Figure 2.13 - The convolution process.

In the example of figure 2.13, a 6x5 image is convolved by a 3x3 convolutional filter (stride of 1), resulting in a 4x3 feature map.

After the convolution layers, the **pooling** layer reduces the dimension of the feature maps, retaining the most important information and being spatially invariant to translations and distortions in inputs. This reduction of the features allows a less computation complexity in the network. Although there are many ways of performing pooling (max-pooling or average-pooling, for example), max-pooling is the most used. Figure 2.14 describes how of the max and average pooling processes are performed.

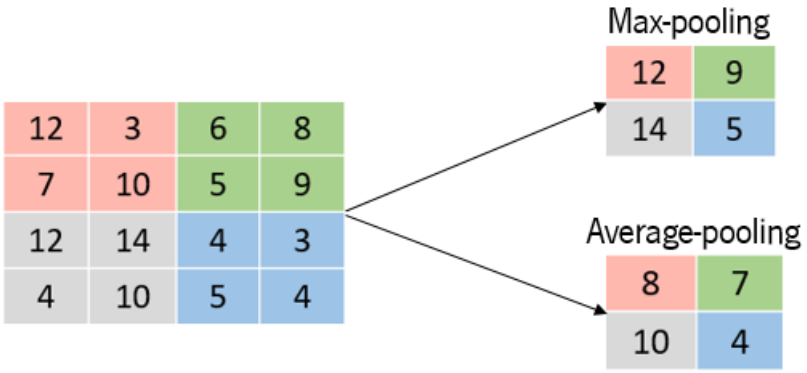


Figure 2.14 - Comparison between average and max-pooling processes. In these 2 specific cases, a pooling window of size 2x2 is used.

In figure 2.14, a pooling window of size 2x2 is used. For max-pooling, for each window, it is determined the maximum of all input values, while for average-pooling it is computed the average of those values.

After many convolution and pooling layers, the network outputs high-level features obtained. These features can now be used for classification or regression, depending on the final objective at hand. This is the purpose of **Fully Connected layers**. The term "Fully Connected" suggests that every neuron of the previous layer connects to each neuron of the next layer. The final output of the CNN will be a classification in classes based in a threshold (classification task) or a prediction of continuous numbers (regression task).

2.4 Related Work

The following literature review approaches the current state of brain neonatology, deep learning and rat MRI investigations. Section 2.4.1 and section 2.4.2 describe, respectively, the recent studies in neonatology and deep learning fields. Section 2.4.3 reviews the insertion of deep learning in

healthcare and, particularly, in the brain neonatology field. Section 2.4.4 approaches the rat brain MRI studies.

2.4.1 Brain Neonatology

In their beginnings, neonatal brain studies only had the objective of describing the development of the preterm infants' brains and its differences compared to the term-born ones. However, during the last decade, several investigations started by, not only diagnosing pathologies, but also focusing on the effects in childhood of these neurodevelopment alterations [30].

The recent progresses in automation and the numerous relevant studies performed in the field of neonatology in the past years led to the introduction of automatic methods in order to carry out an even better evaluation in this neonatal field. As previously remarked, segmentation is a must-do process to assess neonatal brain development and study possible impairments.

Recently, although some studies are connected to brain extraction or skull stripping methods, most techniques used for neonatal brain MRI segmentation are centered in the tissues [31]. Inside brain tissue assessment, many studies focus on classification techniques, where a classifier is trained using image features and a corresponding label. *Wang et al.* developed an effective approach, using random forest techniques to integrate features of images from different sources, including FA, T_1 and T_2 images for neonatal brain tissue segmentation. Their method allows the automated classification of three brain tissues, more specifically, it estimates the probability maps of WM, GM and cerebrospinal fluid (CSF). In the MICCAI grand challenge, this work got the best score comparing to the other methods in the competition [32]. *Moeskops et al.* presented a segmentation algorithm of uWM, GM, and extracerebral cerebrospinal fluid (eCSF), using a sequential supervised voxel classification on T_1 and T_2-weighted images of neonatal infants at 30 and 40 weeks PMA. While in the first and third stages k-nearest neighbor (kNN) classifiers were used, in the second one, the classifier was a support vector machine (SVM). Although this method only achieved accurate segmentations on the T_2-weighted images, it is robust to different acquisition protocols and age [33].

Although the segmentation task has its utilities in a clinical environment, other studies went further and started studying the patterns that neonatal babies can develop and their relation in a future outcome. Using MRI scans at 30 weeks PMA of extremely preterm infants (<28 weeks), *Kersbergen et al.* investigated not only the volumetric brain development on these babies, but also, the clinical

risk factors. An automatic segmentation of the given data had to be performed. The results provided evaluated the growth of 50 specific brain structures and showed that, while brain injury has a localized effect on the brain, other risk factors such as mechanical ventilation, surgery and birth weight have a more global effect. These factors were evaluated in their relation to brain growth, effect at 30 weeks and effect seen at Term-Equivalent Age [34].

In another perspective, *Keunen et al.* studies the relation between structural organization of WM in preterm infants at a term-equivalent age and their future performance in early school. In this study, a cohort of MRI scans of 30 preterm infants (<31 weeks) was used. The study also performed diffusion-weighted imaging at term-equivalent age and built structural connectivity maps. It was concluded that WM connectivity strength has a significant impact in relation to Performance IQ and that the information that the neonatal brain gives in at a term-equivalent age is important in the assessment of future cognitive capacities [35].

Kono et al. performed a study with the objective of assessing the changes in the outcomes (death and neurodevelopmental impairments) at the third year of age in extremely preterm babies (<25 weeks PMA) for the past decade. For that, they used two datasets from 52 tertiary centers in Japan from 2 different periods (2003-2007 and 2008-2012) and used logistic regression models, considering only pregnancy-related and maternal factors. The authors concluded significant improvements in survival and in some impairments such as cerebral palsy and visual problems but not in the cognitive ones [36]. This study recognizes that a further research on the factors influencing neurodevelopmental impairments could be an interesting future work.

2.4.2 Non-Medical applications of Deep Learning

Nowadays, although deep learning is a successful and popular approach inside machine learning with many applications, it doesn't mean that it only appeared in the last years. In fact, the term "Deep Learning" started being used in the beginning of this century as neural networks became more "deep", with more layers, but its successes were not that well spread in industry and research [26]. The main ideas behind deep learning date to the decades even before the 2000s, however, it was just when computers became powerful and fast machines and large datasets were created that it was possible to show its true capabilities.

The biggest boom is this field was the ImageNet open challenge. The objective of this competition was to create the best computer vision algorithm to classify images into labels for tasks such as classification, detection, localization, among others. While there were around 1.2/1.3 million images for training, the labels consisted in 1000 different categories. Different pictures were used for the validation and testing [37].The eventual winner of the competition, *Krizhevsky et al.* not only created a deep network and showed the importance of the "deep architecture", a CNN with 5 convolutional layers and 3 fully connected ones, but also used some new clever techniques at the time such as ReLUs to introduce non-linearities, data augmentation to generate more data, enlarge the dataset and minimize overfitting problems and dropout to reduce the network complexity and also to reduce overfitting [38]. This paper and some others involved in the ImageNet challenge achieved record-breaking results and deep learning started being considered as one of the most promising for the future that even big companies such as Google and Microsoft started to invest in research in this field.

From that point, one of the applications of deep learning is object classification. For example, *Ren et al.* created a Region Proposal Network (RPN) which shares convolutional layers with object detection networks, such as Fast R-CNN. In other words, trained RPNs generate high-end proposals of regions in an almost cost-free step, which will be utilized for detection tasks by a Fast R-CNN [39].

Another deep learning applications with experienced remarkable progresses in the past years were connected to speech recognition. *Amodei et al.* successfully built an end-to-end approach based on Recurrent Neural Networks (RNNs) to recognize two languages: Mandarin Chinese and English. As training data, they utilized 11940 and 9400 hours of labeled speech with several different utterances for both English and Mandarin Chinese, respectively [40].

2.4.3 Deep Learning in Healthcare and brain neonatology

Medical imaging was born from the moment it was possible to deal with images in a computer (scan and load). By then, although the objective was constructing automated systems for medical image, this field was still quite simple, as only mathematical modelling for rule-based systems and low-level pixel techniques were used [41].

Around the beginning of the 21st century, machine learning became very popular in the medical imaging field. It was the first time when the idea of using training data to build a system showed up

and, due to its good results, machine learning techniques are still used nowadays. As the time progressed, more and more systems were being more computerized than being designed by humans. As previously remarked, machine learning approaches use handcrafted features. One more step to automatization was the ability of computers to learn by themselves the features that represent the information for the task at hand. This is one of the reasons for the now extended use of deep learning in medical imaging.

Deep learning has been used in many imaging modalities such as X-ray, MRI, Computer Tomography (CT) and Positron Emission Tomography (PET), enhancing the potential benefits of these scans for discoveries and breakthroughs in different organs.

Deep learning can contribute to increase of the study in several different organs. Besides the brain, for instance by studying tumors [42], it can be applied to other organs, for example, in distinguishing the density categorization of the mammographic breast ("heterogeneously dense" vs "scattered density") [43]. Another example is the retinal eye, by creating and evaluating the effectiveness of a deep learning approach based on fully convolutional networks (FCN) combined with a post-processing regression technique to segment optical coherence tomography (OCT) scans of the retina [44], for instance. Cardiac image analysis had also been influenced by deep learning and is the subject of many studies. For example, *Baumgartner et al.* developed a new approach to segment the myocardium and both left and right ventricles, using cardiac MR scans from the Automated Cardiac Diagnosis (ACD) challenge. For this method, they investigated four different neural network architectures based on state-of-the-art approaches: a 2D FCN based on the VGG-16 architecture, a 2D U-Net, a modified 2D U-Net and a modified 3D U-Net [45]. Besides all these organs, musculoskeletal, abdomen, chest and microscopy images are also object for studies using neural networks.

As previously remarked, in the field of MRI, deep learning has several applications. It can be applied to image registration [46], i.e. the matching of images taken with different equipment, from different angles and positions and at different times. Image classification [47] is also another application that can acknowledge many brain activities or classify different diseases. Classification is well connected with the study of the brain and its disorders. In the field of image detection [48], the final goal is to discover certain tissues of interest and evaluate if they are connected to lesions.

However, in the neonatology field, the brain is by far the most focused organ of study by deep learners, being the segmentation of the many tissues of the brain the most useful application for further quantifications and evaluations in the preterm brain. In that way, *Ghafoorian et al.* built an accurate and efficient deep neural network approach with the objective of segmenting the lateral ventricles of the brain (right and left). Using a dataset of 397 patients and noisy labels, they also show that their system achieves a better performance when compared to a similar one which uses accurate labels. Although these segmentations help in the study of brain ventricles' diseases such as schizophrenia and certain forms of dementia, this use of less accurate labels is of importance in the medical field where many times it is expensive to build large manually datasets [49].

At first, deep learning was focusing on brain segmentation. Recently, more studies regarding neonatal brain are being performed due to the immense potential benefits in healthcare they could lead to, both in terms of the patients and hospital management. *Moeskops et al.* created an approach for the automatic segmentation of 6-8 different tissue classes using a CNN. As for material, they use 5 data sets: 3 corresponding to T_2-weighted scans of neonatal infants (coronal at 30 weeks PMA, coronal at 40 weeks PMA, axial at 40 weeks PMA) and 2 corresponding to T_1-weighted images of young and senior adults, respectively. This specific method is based in a voxel classification problem and uses multiple patch and kernel sizes combined without the need of defining explicit features as the network learns them, estimating both intensity and spatial characteristics [50].

In the line of segmentation, *Xu et al.* produced a faster and versatile method, based on a FCN and using transfer learning, to segment brain MR scans in 3D that works on both adult and neonatal patients. The method achieved considerable results using 2 datasets, one concerning to preterm infants and the other related to adults (respectively, NeoBrainS12 and MRBrainS13). The use of transfer learning showed to be of extreme usefulness: it allowed to obtain promising results even for just one training image. This is important due to the difficulty in generating datasets with accurate annotations in the medical imaging field [51].

In many segmentation tasks in medical image analysis, the final objective is to use the results to quantify some characteristics of the segmented organ. This way, some studies like the following are also focusing on the quantification part of the organs, which would lead to different kind of problem in neural networks: a regression one.

Xie et al. built a method based on Fully Convolutional Regression Networks (FCRNs) to predict the number of cells in a microscopy image. They generate density maps from dot-annotations (each dot corresponds to one cell) and then regress those maps. The two FCRNs built are inspired in the deep VGG-net. They perform experiments in both synthetic/generated and real data. For the latter case, they use two datasets, one of retinal pigment epithelial (RPE) scans and other of T-Cell lymphoblastic lymphoma ones [52]. Based on a dataset with systole and diastole volumes as labels provided by the 2015 "*Kaggle.com"* competition, *Liao et al.* created a fully CNN-based method to both segment cardiac MRI scans and quantify the left ventricular volume with the objective of diagnosing cardiac diseases [53].

Dubost et al. also performed a quantification study, but this time on enlarged perivascular spaces (EPVS) in the basal ganglia (BG) in the brain. After segmentation, a 3D CNN, based on object detection inside a region of interest (ROI), is implemented to output a quantification score of EPVS. The modified CNN architecture is inspired in the VGG of Simonyan and Zisserman [54]. The dataset used consisted of 2000 brain MRI scans, which were annotated manually. Successful tests on EPVS focusing were carried out, along with a comparison with other methods of EPVS quantification. Hyperparameter tuning and a correlation of results with age were also performed [55].

2.4.4 Brain MRI in rodents

Around 80 % of the human genome is shared with rats and mice. As previously remarked, these animals are used for some preclinical studies of the human brain, being their brain one of the focused organs.

For example, *Scheenstra et al.* focused on constructing and an automated segmentation method concerning mice MRI for registration of the brains in other modalities. They used two different types of mice MRI datasets: a T_2-weighted in vivo dataset and T_1-weighted dataset from the same mice. They successfully reached their objective of increasing the speed of the time-consuming atlas-based segmentations, while the accuracy of the segmentation of the main structures was sufficient [56].

Another work on mice MRI was performed by *Totenhagen et al*. This work consisted in assessing atrophy in specific regions of the mice brain with a neurodegenerative disease named Niemann Pick Type C1 (NPC1). As data, it was used in vivo 3D mice MRI from 3-9 weeks of age. They concluded

that, at the late stages of the disease, some brain regions were unequal when compared to the early stages, because of genotype and because of both genotype and age [57].

3 Quantification of brain characteristics using neonatal MRI

As stated before, for the neonatal part, this dissertation is inserted in a larger and more complex project, which can be divided in 3 steps.

When evaluating or predicting neonatal brain impairments, physicians need to analyze the constitution of the brain of those infants in order to find different patterns that can lead to the discovery of those problems. As stated before, MRI is a medical imaging study that can expose with detail the anatomical parts in soft tissues, so it is clear why MRI scans are performed on these peculiar babies. On other hand, deep learning is a very good technique in finding those patterns in an automatic way, sometimes even better than doctors, which is essential for the classification of a preterm infant as impaired or not.

However, predicting impairments only analyzing an MR scan takes a very complex process in between. The study performed in [12] describes 3 steps in order to identity infants with motor and cognitive impairments at 2-3 years after birth. These impairments are determined based on their cognitive and motor outcome.

The first step consists on the **segmentation** of the MRI images. This step used the deep learning approach described in [50], applied for the neonatal data. The result is the segmentation of the neonatal images in 8 different brain tissues. Figure 2.4 describes this step: the input is a neonatal MRI scan and the output is its segmentation.

The second step consists on the **quantification** of the segmented images, obtained on the first step. For each image, it will be produced a series of quantitative descriptors. Eight of those descriptors correspond to the volumes of the 8 tissues segmented in the previous step. The other computed descriptors are, for example, gyrification index (GI), Median Cortical Thickness (MT), among others, corresponding to measures for cortical morphology [12]. Figure 2.5 describes the quantification step: input is the segmented scan and output is a series of descriptors (volumes and measures of cortical morphology).

Finally, the last step is **classification**. In this phase, patients are classified in two classes depending on their outcome. SVMs classifiers, associated with a threshold, are used to distinguish patients with a low cognitive and motor outcome, i.e. with impairments, from the ones with high outcomes, without impairments. This classification was performed using the descriptors quantified in the second step in with 3 different kind of experiments: using descriptors from scans acquired at 30 weeks PMA;

using descriptors from scans acquired at 40 weeks PMA; differences in descriptors between 30 and 40 weeks PMA [12]. Figure 3.1 shows clearly the inputs (a series of descriptors) and outputs (a classification between "Healthy" or "Risk of impairment", according to a threshold) for this step.

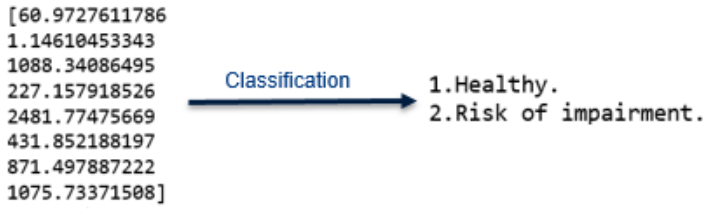


Figure 3.1 – The classification step.

As previously stated, the final goal of the whole project described in the section before is to use deep learning to classify a MRI scan as with impairments or not. Although the segmentation step is performed and optimized using deep learning, the other two steps (i.e. quantification and classification) are not. With the objective of building this whole system, the next step is to be able to quantify descriptors using deep learning.

This dissertation is the continuation of this whole task. As already stated, the dissertation's objective is to directly predict quantifications for several descriptors of the brain of neonatal MRI scans using deep learning techniques. By directly quantifying the descriptors, it is possible to skip the segmentation task.

For a future work, it will be necessary to use deep learning in the classification step to get the fully functioning deep learning system. The MRI scans would be inputted in a DNN and they would be relevant in the prediction of preterm impairments. This way, all the 3 steps of the whole project would be automated using DL techniques so, the development of a DL system which combined those 3 steps would be possible. Figure 3.2 depicts the stages of the whole process, the step in which this project centers on, its importance and future work.

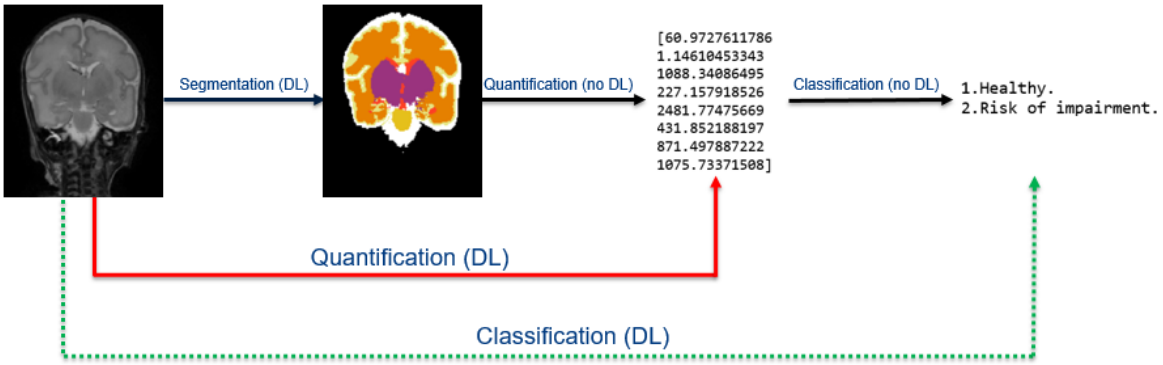


Figure 3.2 - The whole process and the objective of the dissertation: black arrows represent the steps already done; the red arrow shows the focus of this dissertation and its importance and the green dashed arrow indicates the future work.

3.1 Materials

In this subsection, all the data used in this project is described. The neonatal brain images are considered the main source of data and are used for the deep learning tasks. More specifically, image origin, characteristics and objective, for example, are some of the topics addressed. The corresponding masks and segmentations are used for pre-processing steps and the descriptors serve as some of the labels.

3.1.1 Neonatal brain images

A cohort of MRI scans of preterm infants born before 28 weeks gestational age (GA) was acquired on a *Phillips* (Best, The Netherlands) *Achieva 3T* scanner. Both parental consent and permission for the use of the clinically acquired data from the institutional review board of University Medical Center Utrecht (UMCU) were obtained.

Around 30 weeks PMA is when these infants are still developing their brain and have a higher risk of brain injury. The infants that were considered clinically stable by the neonatologist were scanned using a MR compatible incubator (LMT Medical Systems, Lübeck, Germany), creating a dataset of 86 coronal T_2-weighted studies at an average PMA of 30.8 ± 0.8 weeks. At 40 weeks PMA, the infants were scanned using a *Phillips SENSE* head coil. The dataset consisted on 153 coronal T_2-weighted studies at an average PMA of 41.2 ± 0.7 weeks. Images with motion artifacts and with brain abnormalities were excluded if that resulted in inaccurate automatic analysis. A summary and

more additional details concerning the data described are presented in table 3.1. An example of a neonatal data used, a brain slice at 30 weeks PMA, is represented in figure 3.3.

Table 3.1 - Acquisition parameters for the images used in this study

	Data at 30 weeks	Data at 40 weeks
Age (weeks PMA)	30 ± 0.8	41.2 ± 0.7
Number of scans	86	153
Acquisition protocol	Coronal T_2-weighted	Coronal T_2-weighted
Reconstruction matrix	384 × 384 × 54	512 × 512 × 110
Reconstruction voxel size (mm^3)	0.34 × 0.34 × 2.0	0.35 × 0.35 × 1.2

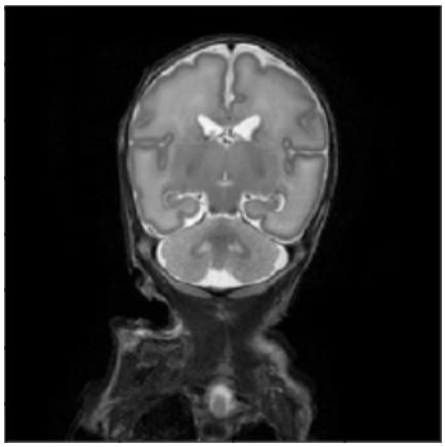


Figure 3.3 - A neonatal brain slice at 30 weeks PMA.

3.1.2 Brain segmentations and masks

To implement preprocessing techniques, brain segmentations and masks were used. The segmented images were a result of the first stage of the work in [12], i.e. they are the result of running the neonatal data described in 3.1.1 with a pre-trained model for segmentation. The images were automatically segmented in 8 different tissues, using the system constructed in [50]: basal ganglia and thalami (BGT), brain stem (BS), cerebellum (CB), extracerebral cerebrospinal fluid (eCSF), ventricular cerebrospinal fluid (vCSF), cortical gray matter (GM), myelinated white matter (mWM) and unmyelinated white matter (uWM). Figure 3.4 depicts an example of segmentation of a determined brain slice. The tissues for prediction are also labeled in figure 3.4.

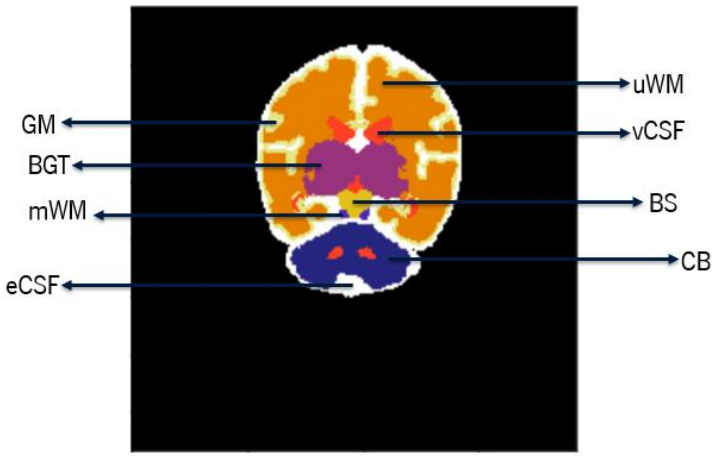


Figure 3.4 - A segmented slice of a brain at 30 weeks PMA.

The brain masks were automatically generated based on the T2-weighted images using *Brain Extraction Tool* (BET) from the *FMRIB Software Library* (FSL), using the same setting for all images. This work can be consulted in [33]. An example of a mask can be consulted in figure 3.5.

Figure 3.5 - A masked slice of a brain at 30 weeks PMA.

3.1.3 Other descriptors

As for the other descriptors, they were provided in an *Excel* file for each neonatal brain patient. They were quantified in [12] based on the neonatal segmentations. The computed descriptors are measures for cortical morphology: gyrification index (GI), inner (IS) and outer cortical surface area (OS), global mean curvature (MC) and median cortical thickness (MT).

3.2 METHODS

This section explains all the methods developed in the creation of the systems/models which predict the values of 13 descriptors from neonatal brain studies. In overall, four different methods were considered:

1. A **2D approach**, using 2D neonatal brain images.
2. A **2D approach with pre-processing**, with more pre-processing techniques.
3. A **3D approach using chunks**, i.e. 3D neonatal brain chunks of variable size.
4. A **3D approach for full brains**, using full 3D neonatal brains.

3.2.1 2D APPROACH

In general, this 2D approach has the objective of predicting the specific values of 8 of the 13 descriptors, i.e. the volumes for 8 brain tissues (described in 3.1.2), from 2D neonatal brain images. The other 5 descriptors (described in subsection 3.1.3) are 3D measurements so they can only be predicted from a full brain (3D approach for full brains).

3.2.1.1 LOADING PROCESS

Firstly, it is necessary to perform the loading of the data (studies, segmentations and masks) in the same way. These studies are first divided in training, validation and testing datasets. The training set is used to train the models, the validation set is the one used to adjust hyperparameters and optimize the model and the test set serves to perform the final test after the model is tuned. When performing the split, the same random seed is used, so that the 3 sets are always the same, between experiments. This way, reproducibility between experiments is guaranteed.

Since the given studies have different sizes, a function is created to check all the images in order to get their largest dimensions. Software package *SimpleITK* [58], [59] is used in order to get the width, height and depth of each study. When sequentially analyzing the studies, for each of these dimensions, if their value is bigger than the largest dimension until then, then this value is the new largest dimension. In the end, the function returns the largest dimension for each one of the 3 dimensions. As an example, for the brain images at 30 weeks PMA, the largest dimensions correspond to 384 × 384 × 54 (Width × Height × Depth).

The next step consists in the actual loading of the data. *SimpleITK* is used to get a *numpy* array from an image. Each study is compared towards the largest dimensions. If the study is smaller in relation to any largest dimension, it is padded with zeros in that dimension. This way, all the studies have the same dimensions, which correspond to the largest ones. Figure 3.6 represents this process in a case with 3 samples, while figure 3.7 presents a visual representation of this padding. The same logic is applied when loading the corresponding segmentations and masks. Then, the images (more specifically, all the slices for each study) are concatenated in a single variable.

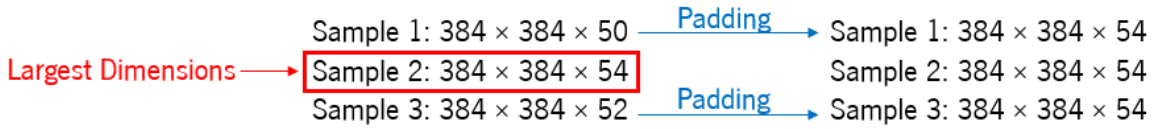


Figure 3.6 - Padding process, after the comparation towards the largest dimensions.

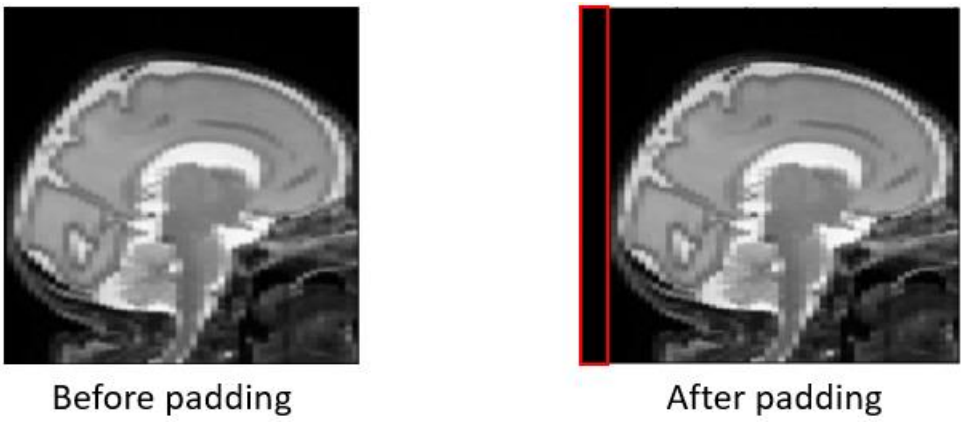


Figure 3.7 - Visual representation of padding.

The next step consists in determining the volumes for each one of the 8 brain tissues for each slice. The segmentations are used to determine those volumes in mm^3. These volumes consist in the ground truth labels.

In sum, the final result is a set of images (i.e. slices) (an example is represented in figure 3.3), the corresponding segmentations (figure 3.4, without the tissue labels) and masks (figure 3.5), the patient ID and the volumes per slice for the training, validation and test sets.

3.2.1.2 Pre-processing technique: normalization

The brain tissues can differ a lot in volumes, so, to prevent numerical instability and to help in the convergence in the training, those labels are normalized. All the 8 different tissues are normalized, according to the normalization formula, represented in equation 1.

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

In equation 1, x denotes the actual volume of a certain tissue in a certain slice; x_{min} and x_{max} correspond, respectively, to the minimum and maximum volume of the tissue in all slices and $x_{normalized}$ denotes the normalized value. The x_{min} and x_{max} computed for the training labels are the ones used to normalize the validation and test labels, i.e., the maximum and the minimum only correspond to the training set. All the values are normalized to the interval [0,1].

3.2.1.3 Data Augmentation

Data augmentation is performed in order to have more data samples for the posterior training. For this approach, this augmentation is performed in terms of rotations, scaling and translations, all 3 combined.

Rotated images are created between -10 and 10 degrees. The scaling is performed in terms of reduction (due to the fact that if size increase was performed, the brain could get out of the image borders). This reduction is performed until a factor of 0.8 i.e. 80% the size of the actual image. Since we are dealing with volumes, if a reduction is being performed in the image, the volumes must be reduced with the same reduction factor as the image. The translations are performed in both xx and yy axes in relation to the center of the image. For the xx axis, the translation is between -25 and 25 voxels. The translation in yy axis is set between -25 and 8 voxels. The positive and negative values have the center of the image as a reference. Figure 3.8 shows examples of this data augmentation.

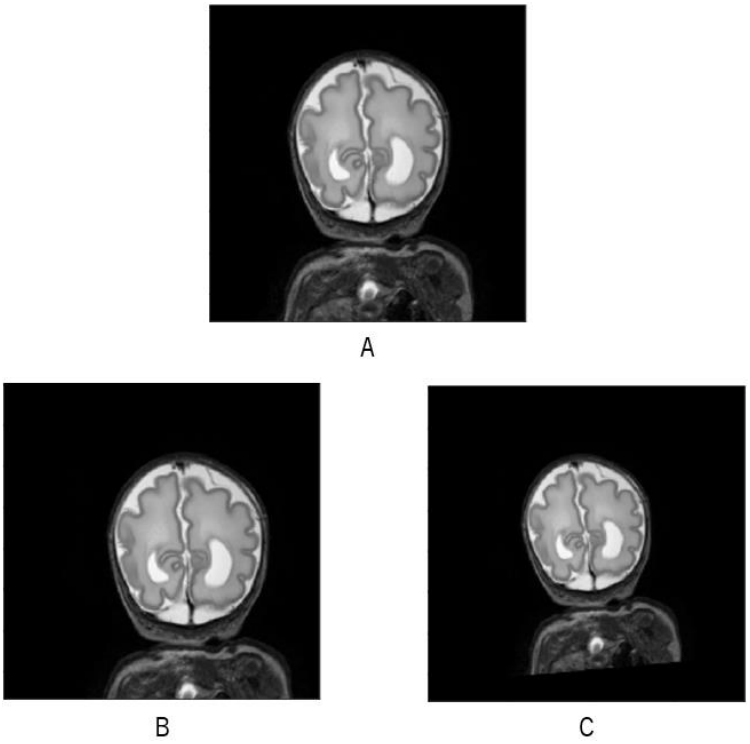


Figure 3.8 - Examples of data augmentation: A- original slice. B- augmentation in terms of translations. C- augmentation in terms of scaling and rotation.

Figure 3.8 shows 2 examples of data augmentation used in this approach. The original image (A) is transformed in image B by translating 25 voxels in both xx and yy axes. Image C shows the transformation of an original image if the rotation is equal to 5 degrees and the scaling is 0.8.

3.2.1.4 Neural network, training and validation

As previously remarked, the objective of this approach is to predict volumes of 2D neonatal brain slices. In this regression problem, for each 2D neonatal brain slice that goes into the neural network, it provides an estimation of 8 volumes, each corresponding to a brain tissue.

The neural network used in this approach is inspired and adapted from the successful VGG-Net [54]. As described in Figure 3.9 A, the network has an input of 2D slices with size 384 × 384 (width × height). Then, 3 convolutions are performed with 3 × 3 kernels, following by a 2 × 2 kernel max-pooling, as represented in figure 3.9 B. This step is executed 3 times. Each convolution produces 16 filters, using ReLU as the activation function. After the 3 steps of successive convolutions and pooling the network finishes with 2 fully connected (FC) layers. Both of those layers, use a dropout of 0.5 probability of setting a value to zero. The first fully connected layer has 256 units, using as

activation function ReLU, while the second one has only 8 units, corresponding to the same number of different tissues, whose volumes must be predicted, and a linear function.

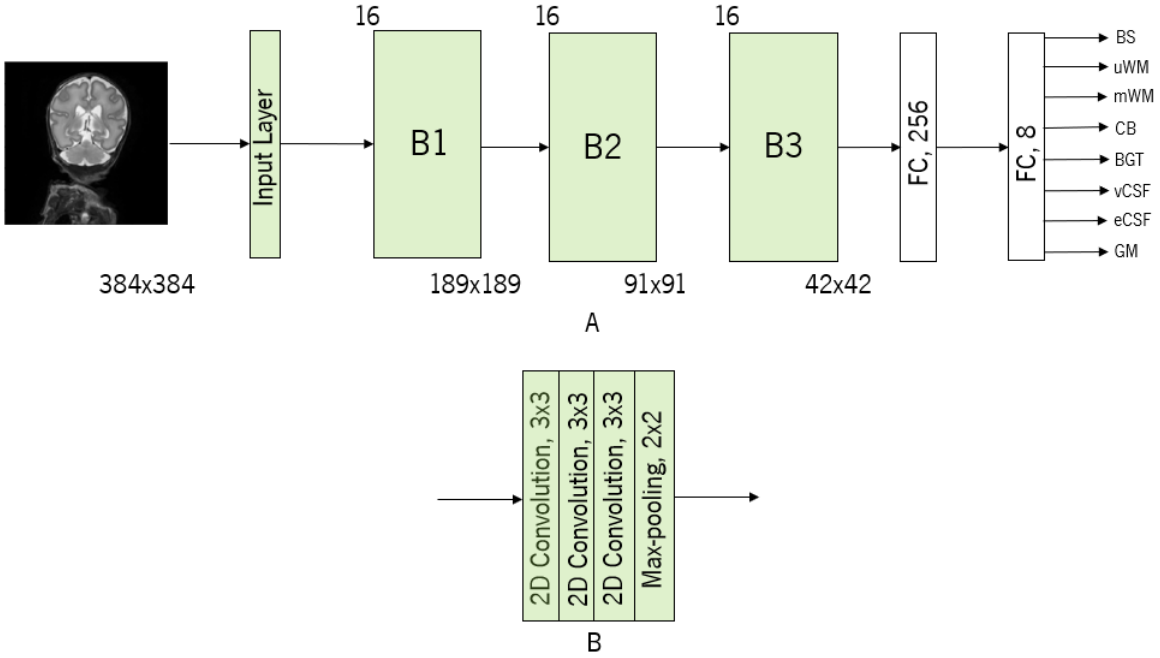


Figure 3.9 - Architecture of the neural network used in this approach: A- a general overview; B- convolutions and pooling used in each step (B1, B2 and B3). The number of filters for each step is represented on the top left of that step.

For the training part, a specific number of minibatches and minibatch size had to be chosen. The minibatches correspond to the number of iterations in training (and validation), while the minibatch size is related to the set of samples (i.e. images) that enter in the network for one iteration. For each iteration, a set of random images is selected from the training set, according to the minibatch size specified. The same is done when selecting the corresponding labels/volumes.

Due to memory constraints, the data augmentation step explained in 3.2.1.3 is performed on the images (and labels) of the random set at each iteration. It is more efficient to augment images for a small set each iteration, than to augment for the whole training set before the training. This way, the network gets as input a set with half “normal” and half “augmented” images. Consequently, the labels are half “normal” and half “augmented”. Figure 3.10 depicts this process.

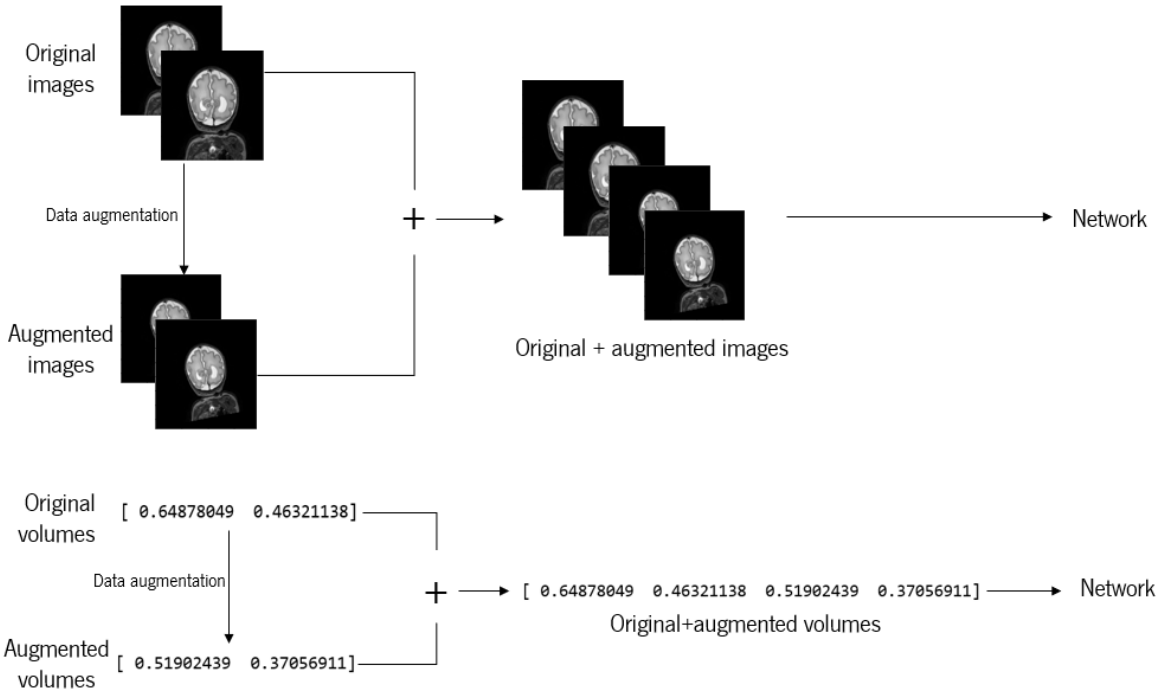


Figure 3.10 - Process of data augmentation: augmented images join original images, before they are fed to the network.

To build and train the neural network and define, optimize and evaluate during the training, validation and prediction processes, the DL library used is *Theano*, along with library *Lasagne*. When slices and ground truth volumes are fed for training, the network predicts the volumes from images and compares with the ground truth volumes of those images. The loss function used in the experiments is the squared error between predicted volumes and ground truth ones. Concerning the optimizers, two are tested [27]:

- RMSProp, with learning rates of 0.01, 0.001 and 0.0001.
- Adam (Adaptive Moment Estimation), with a learning rate of 0.001.

The steps described for training are also executed for the validation. In the end, the loss list presents the variation of the loss for both training and validation, according to the number of iterations.

Most of the models took a long time to train on a normal CPU. Therefore, a server with GPUs was necessary for a faster training. The following were the ones used:

- NVIDIA GeForce GTX TITAN X, with 12 GB (gigabyte) of Standard Memory Config.
- NVIDIA TITAN Xp, with 12 GB of Standard Memory Config.
- NVIDIA GeForce GTX 1080, with 8 GB of Standard Memory Config.

Due to the long time it took for the training to occur, a program that allows the use multiple windows, *screen*, was utilized. Besides other advantages, if the connection is lost or if the local computer crashes, the processes or login sessions established through *screen* don't go away.

After the training process, the trained models are saved in '.pkl' (*cPickle*) files.

3.2.1.5 PREDICTION AND MODEL EVALUATION

After training the network, it is necessary to use a different set of images to access the performance of the model. If the objective is to tune the model by changing and optimizing hyperparameters, validation images should be used to access the performance. Test images are only used for the final model, when it is completely tuned and will not be changed.

Either way, the prediction of these images originates a set of predicted values that must be compared with the ground truth ones. For each one of the 8 tissues, the predicted values are compared to the ground truth ones in graphs. Ideally, this comparison would be perfect in the graphs if the pairs (predicted values, ground truth values) followed the identity function y=x, which means that the predicted values are the same as ground truth ones. These graphs allow a clear visual representation of the results. *Bland-Altman* [60] graphs are also computed. This measure uses a difference plot to compare measurements for the same variable. In this case, these graphs can be an important tool when studying the differences between predicted and ground truth values.

Since it is a regression problem, statistical measures such as Pearson correlation (R), MSE (mean squared error) and RMSE (root mean squared error) are computed. Both RMSE and MSE formulas are represented by the equations 2 and 3, respectively.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \qquad (2)$$

$$\mathrm{MSE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2 \qquad (3)$$

In equation 3, $\mathrm{y_i}$ and $\hat{\mathrm{y}}_\mathrm{i}$ denote, respectively, the ground truth and predicted volumes and N correponds to the number of examples. The Pearson correlation evaluates the strength of the linear relationship between predicted and ground truth results. The RMSE is referred as the standard deviation of the prediction errors and gives a sense of how close or far the predicted values are from the ground truth. It is important for precision (accuracy) of the model. The normalized RMSE provides

a standard deviation of the prediction errors, with respect to the range of volumes predicted. The metric is presented in percentage (%) of the error in relation to the range of volumes.

3.2.2 2D APPROACH WITH PRE-PROCESSING

This approach is basically similar to the one described in subsection 3.2.1. with the only difference being in applying previous pre-processing techniques on the images. Here, this pre-processing consists on images that are cropped and downscaled in order to tackle memory issues and increase batch sizes and/or filters during training. This approach is used for both neonatal images at 30 and 40 weeks PMA.

3.2.2.1 IMAGE CROPPING

Even though the cropping process may seem simple in theory, the way it is performed in this step is so it can be applied to any stack of images with corresponding masks, not the only the ones used in this work.

The process consists of 3 main steps:

1. Determination of the minimum and maximum width and height for each slice, based on the slice's corresponding mask.
2. Comparison of the slice's width and height values with the values until that point. In the end, the minimum and maximum width and height of the whole stack will be computed.
3. Perform the cropping. Figure 3.11 shows an example of this step.

With this cropping, the translations in data augmentation used in the subsection before are not feasible anymore, due to the fact that if the same values were kept, the brain could get out of the borders of the cropped image. At the same time the cropping is performed, the center of the masks is also studied. This way, the cropping can be adapted so that the brain stays in the center of each slice. The reason for this method is so there can be translations in data augmentation, even though the values of these translations had to be adapted.

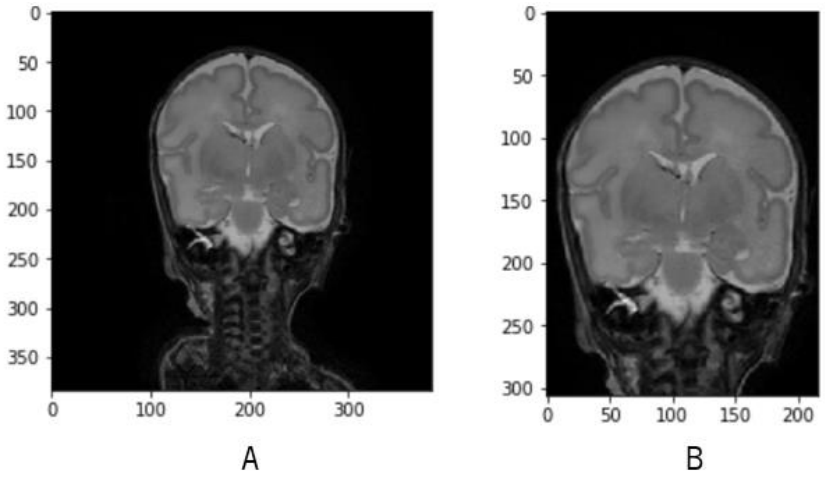


Figure 3.11 - Cropping step: A- the original image. B- cropped image.

3.2.2.2 *IMAGE DOWNSCALING*

Right after the cropping, a downscaling of all the cropped images is performed. The factor for downscale must be chosen so the image does not become much downgraded. The factor chosen for this approach is 2, i.e. the downscaled image has half the size of the cropped one, as depicted in figure 3.12.

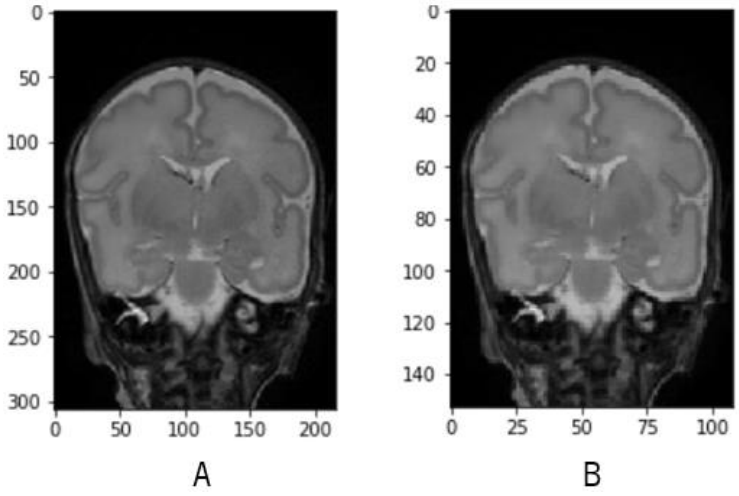


Figure 3.12 - Downscaling step: A- the original cropped image. B- downscaled image.

As previously remarked, the other steps already described, such as loading, normalization, training, validation, CNN, prediction and metrics are the same ones used in the 2D approach (subsection 3.2.1)

3.2.3 3D APPROACH USING CHUNKS

This approach uses chunks/parts of 3D images of the brain to predict the same 8 descriptors described in 3.1.2. Some of the processing is similar to the 2D approach with pre-processing techniques described in 3.2.2 with some additional characteristics owing to the fact that this method uses 3D chunks of brain instead of 2D slices.

3.2.3.1 3D TRANSFORMATION

After loading the data and performing the cropping and downscaling on those 2D images, the next step is to transform them in 3D chunks. This can be achieved by stacking the (downscaled) images according to a certain size of choice. The cropping explained in section 3.2.2.1 cannot be centered, for this approach, because that would change the 3D reconstruction to unrealistic examples. The results of this stacking in the 3 planes are depicted in figure 3.13.

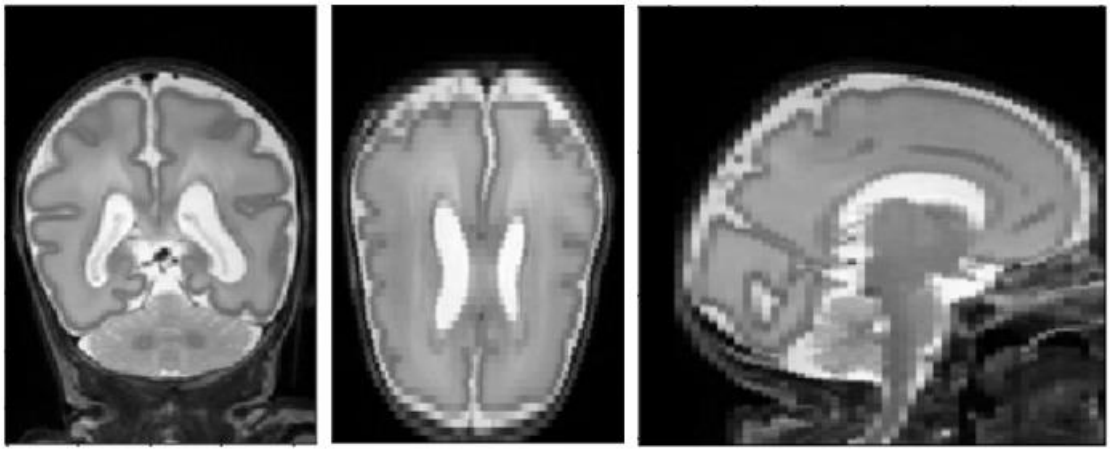


Figure 3.13 - Neonatal brain represented in 3 planes: A- coronal; B- axial; C- sagittal.

This "stacking" process could be performed in two different ways: with overlapping (figure 3.14 A) and without overlapping (figure 3.14 B). The overlapping process allows the use of more examples for both training, validation and test, but, for bigger chunks, it can sometimes happen the situation when the end of a brain could be in the same chunk as the beginning of the next one and this does not represent reality. Therefore, the method without overlapping was the most used.

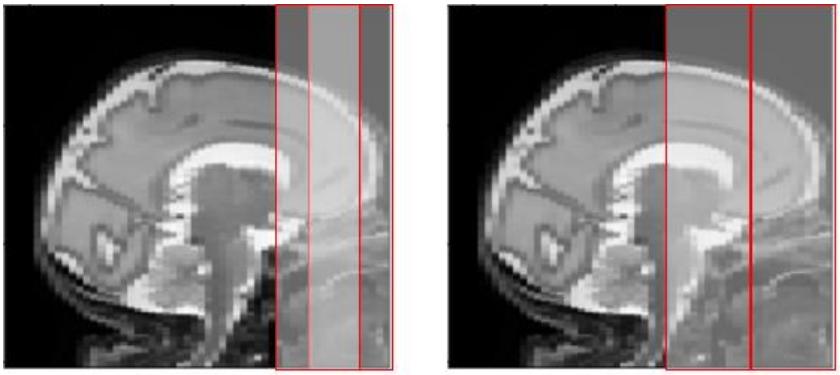


Figure 3.14 - Ways of stacking: A- with overlapping; B- without overlapping.

Since now chunks of brain are being used, the labels also must be changed. For each tissue, the volumes are summed according to the size of the chunk.

3.2.3.2 Data Augmentation

For this approach, since it's 3D, the augmentation is not of 2D slices but is of chunks. The biggest difference now is that all slices within a chunk need to have the exactly same transformation. Since the images are not centered after cropping, no translations are performed due to the fact that they can get out of the image borders. Intensity variation and flipping are added as new variables in the augmentation. Intensity variation in the augmented image was calculated using the following formula:

$$\text{New augmented image} = \text{image}^{\text{number}} \quad (4)$$

In formula 4, the variable "number" ranges from [0,95;1,05].

3.2.3.3 Neural network

One of the CNNs developed is an adaptation of the one presented in the 2D approaches (VGG-Net like), with the difference of now being 3D. As figure 3.15 depicts, a new dimension related to the size of the chunk is added to the network input (in the specific case of figure 3.15, the size of the chunk is 54 slices), as well as to the convolutions and pooling. Sometimes, the size of the chunks does not allow all the convolutions and pooling layers in the network to be 3D, so 3 different architectures in the CNN are tested:

- 3D on the first block (3 3D convolution layers and 1 3D pooling one). This architecture is represented in figure 3.15.
- Only 3D on pooling layers.
- Only 3D convolution layers.

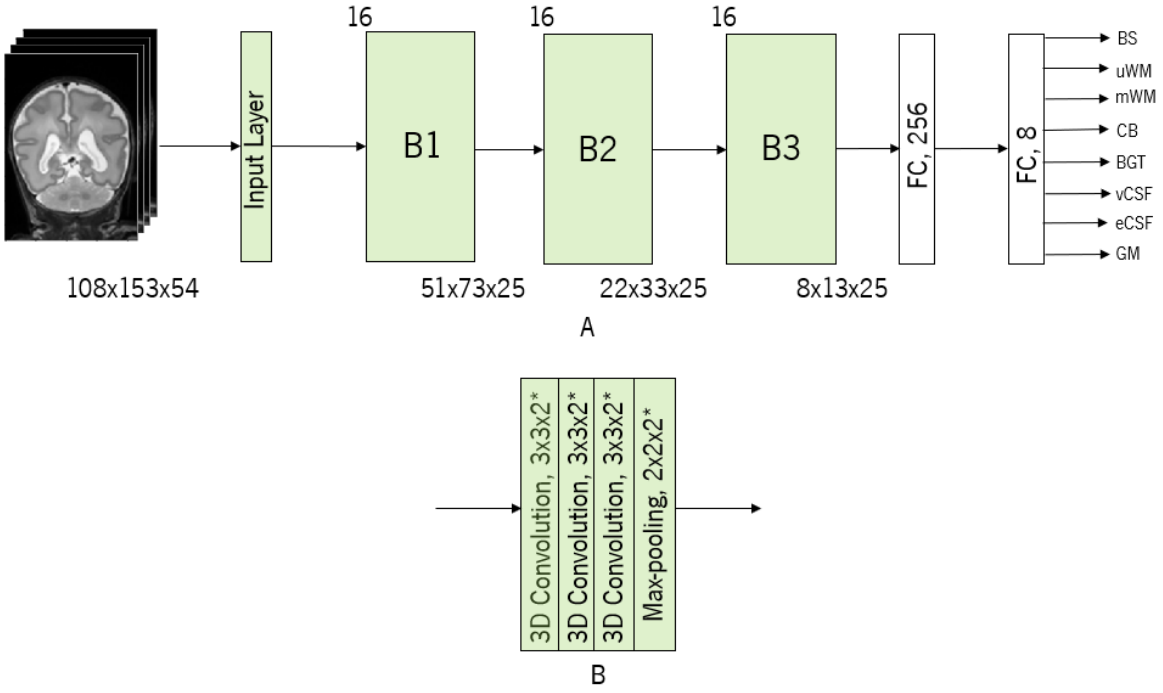


Figure 3.15 - Architecture of the adapted 3D neural network, based in VGG-Net: A- a general overview; B- convolutions and pooling used in each step (B1,B2 and B3). The * means that for steps B2 and B3 all 3D convolutions used are with 3x3x1 kernels and max-pooling is 2x2x1 due to the size of the chunks. The number of filters for each step is represented on the top left of that step.

The other developed neural network is an adaptation of ResNet, more specifically, ResNet-34 [61], for 3D. This network is based on residual mapping and, in theory, is easier to optimize in relation to the unreferenced mapping. As shown in figure 3.16 A (in this specific case, the size of the chunk is 54 slices), the chunks suffer a first convolution and pooling after the input layer. Then, the architecture has 4 big steps (represented in green), each of them consisting in the repetition of several residual blocks. The residual learning blocks that create the steps mentioned before, are exemplified in figure 3.16 B. For each of the second, third and fourth steps, the first convolution has a stride of 2. All the convolutions use ReLUs as activation functions. The network ends with 2 fully connected layers, being the final output, the 8 descriptors related to the volumes.

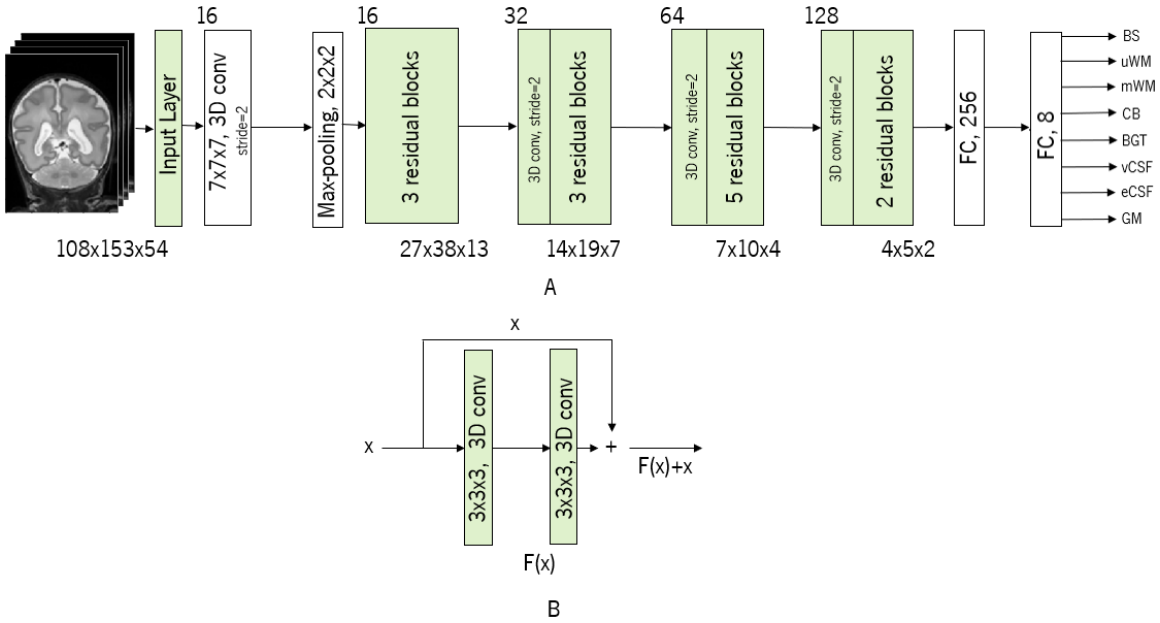


Figure 3.16 – Architecture of the adapted 3D neural network, based on Resnet-34: A- a general overview (the 4 big steps are represented in green); B- exemplification of a residual block. The number of filters associated to each step is located on the top left of that step.

In figure 3.16 B, the original input x is added to the "convoluted" one F(x), originating the output F(x)+x.

3.2.4 3D APPROACH USING FULL BRAINS

For this approach, full neonatal 3D brains are used to predict all 13 descriptors. This approach has the same logic as the one described in 3.2.3, apart from dealing with 5 new descriptors, the ones described in 3.1.3.

3.2.4.1 OTHER DESCRIPTORS AND NEW CONSIDERATIONS

The other descriptors apart from the 8 volumes are loaded from an *Excel* file, using *Pandas*, in the same order as the *random.seed* provided for the loading of the images, segmentations, masks and labels in order to match the corresponding patient. After the volumes are summed for the entire brain, these descriptors are concatenated with them in order to create 13 ground truth labels for each brain (8 volumes and 5 other descriptors). Then, all these labels will be normalized as explained before in sub-section 3.2.1.2.

The same neural networks, VGG-like and ResNet-like, are used for this approach with the sole difference of predicting 13 outputs instead of 8. It is used the same learning rate values and optimizers mentioned in the subsection 3.2.1.4.

In the data augmentation step, scaling must be removed due to the fact that these 5 new descriptors change in a non-linear manner in a relation to a decrease (or increase) in brain size. In sum, augmentation is performed in terms of rotation, intensity and flipping.

3.2.4.2 Other techniques

During the validation/pre-test phase, in order to obtain improved results, other techniques are used. **Batch normalization** [62] is tested in both neural networks (VGG-based and ResNet based). For each batch of brain images, this method consists on the normalization of a part of the model architecture, allowing, for example, a better regularization of the model, the use of higher learning rates and saturating non-linearities, in order to get a better model convergence. In this work, batch normalization is performed after each convolutional layer, after the input layer and after the first dense layer.

Other technique is the implementation of parametric rectified linear units (**PReLUs)** [63], which are a generalization of ReLUs. PReLUs have the ability of learning the parameters of the rectifiers and they can achieve a better accuracy with little computational cost. A comparison of both ReLU and PReLU graphs is shown of figure 3.17.

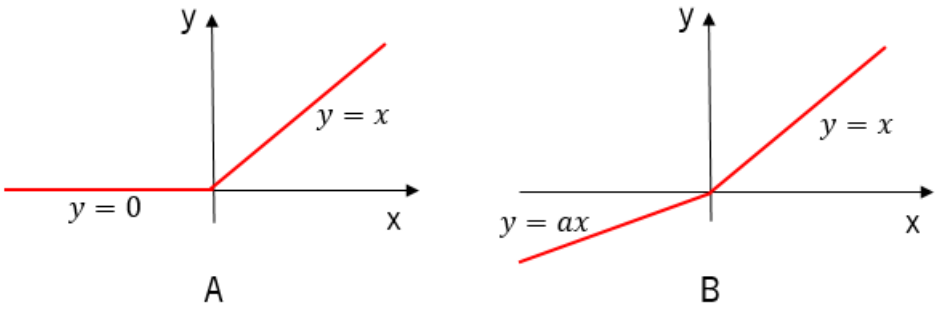


Figure 3.17 - ReLU and PReLU comparison: A- ReLU graph; B- PReLU graph.

In figure 3.17 B, in the negative part of the PReLU graph, the coefficient is not constant and learned iteratively in the network.

3.3 RESULTS AND DISCUSSION

Separate experiments are performed for each approach. Since the problem consists in a regression task, comparison between predicted and ground truth (real) values is needed in graphs. For each set of results, the graphs for each brain tissue are shown, along with correlation (R) and normalized RMSE, for comparison.

3.3.1 2D APPROACH

This approach consists on using 2D neonatal brain images/slices to predict the volumes of 8 different tissues. For the data at 30 weeks PMA, 86 studies are used: 60 for training, 13 for validation and 13 for testing. The dataset with 40 weeks PMA data was not used in this approach due to their size which would not fit in memory. The hyperparameters fixed for this test are specified in table 3.2. Figure 3.18 shows the prediction vs ground truth graphs for the 8 tissues, using unseen images (test set). Table 3.3 shows the correlation and normalized RMSE results for each graph presented.

Table 3.2 - Parameters used in the 2D approach

Minibatches	12000
Batch size	80
CNN	VGG-like
Filters per convolution	16
Learning rate	0,001
Optimizer	RMSProp
Filters' size	3 × 3
Max-pooling size	2 × 2

Table 3.3 - Values of correlation and normalized RMSE for the 2D approach, with the 30 weeks PMA test set

	CB	mWM	BGT	vCSF	uWM	BS	GM	eCSF
Correlation	0,946	0,799	0,963	0,921	0,981	0,953	0,968	0,917
Normalized RMSE (%)	8,5	8,4	9,3	7,4	6,9	6,5	7,5	10,2

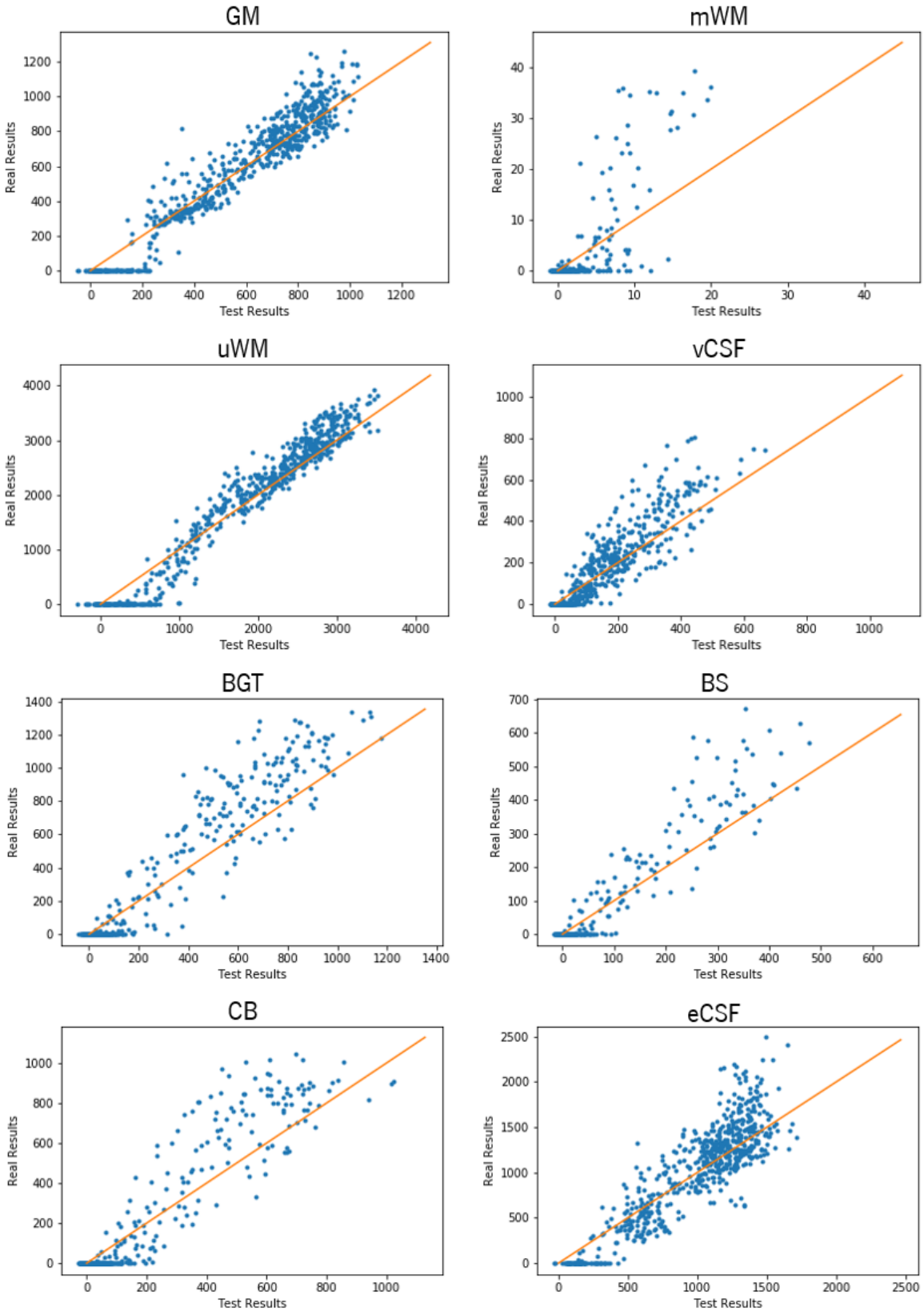


Figure 3.18 – Test (prediction) vs Real (ground truth) graphs of the 8 tissues for the 2D approach. The name of the tissue is on top of its corresponding graph. Values are in mm^3.

This approach predicts the 8 volumes for each slice. On a whole, by observing both graphs and metrics, GM and uWM seem to be the easiest tissues to predict. Even though, vCSF and BS look worse in the graphs, the normalized RMSE is low in those two tissues, when compared to the best-looking tissues. This is due to the fact that in many of these slices there is no presence of tissue (volume of 0 mm^3), and, the system predicts easily the non-existence of a tissue, leading to a better normalized RMSE. On the same line of sense, normalized RMSE for CB and BGT was expected to be higher, when looking to the graphs, but it is not due to the same reason: the number of slices without tissue. Also, for most tissues, the bigger the volume in the slice, the harder it is to get an accurate prediction. The ground truth given segmentations are produced automatically, and the system used to segment these brains could not do it accurately for mWM. In addition, the number of samples with positive volumes in training may not be sufficient for the system to "learn" how to predict these positive volumes. These 2 reasons can explain the bad results for this specific tissue.

3.3.2 2D APPROACH WITH PRE-PROCESSING

This next approach consists on using 2D neonatal brain slices/images with some pre-processing techniques (centered cropping and downscaling per 2) to predict the volumes of 8 different tissues.

3.3.2.1 30 WEEKS

The same split of 60 for training, 13 for validation and 13 for testing is used for the data at 30 weeks PMA. The hyperparameters fixed for this test are specified in table 3.4. Figure 3.19. shows the prediction vs ground truth graphs for the 8 tissues, using unseen images (test set). Table 3.5 shows the correlation and normalized RMSE results for each graph presented.

Table 3.4 - Parameters used in the 2D approach with pre-processing, using the 30 weeks PMA dataset

Minibatches	50000
Batch size	400
CNN	VGG-like
Filters per convolution	16
Learning rate	0,001
Optimizer	RMSProp
Filters' size	3 × 3
Max-pooling size	2 × 2

Table 3.5 - Values of correlation and normalized RMSE for the 2D approach with pre-processing, using the 30 weeks PMA test set

	CB	mWM	BGT	vCSF	uWM	BS	GM	eCSF
Correlation	0,972	0,823	0,983	0,967	0,988	0,953	0,974	0,961
Normalized RMSE (%)	5,9	7,4	5,8	4,9	5,9	5,9	6,7	7,6

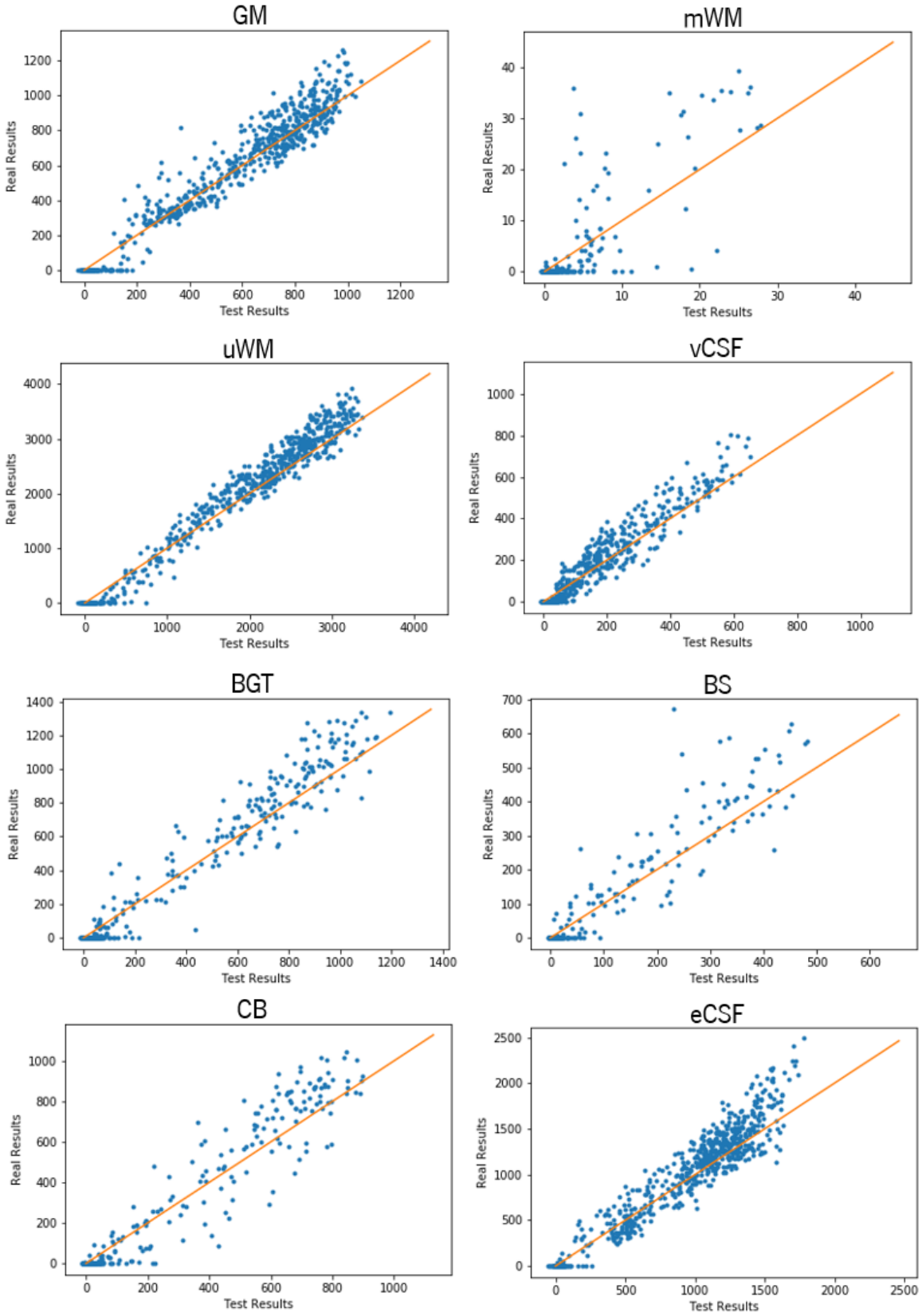


Figure 3.19 - Test (prediction) vs Real (ground truth) graphs of the 8 tissues for the 2D approach with pre-processing, using images at 30 weeks PMA. The name of the tissue is on top of its corresponding graph. Values are in mm^3.

On the whole range of volumes, GM and uWM continue to be the easiest tissues to predict. Both eCSF and vCSF improved considerably in relation to the previous approach, by looking at both graphs and metrics. The low normalized RMSE for vCSF, BS, BGT and CB has the same explanation as before (many slices have no presence of the tissue). mWM continues to be bad probably due to the same reasons as before (bad segmentations and small number of mWM samples with positive volumes in training).

3.3.2.2 40 WEEKS

The pre-processing techniques allowed the 40 weeks data to fit in memory. The 153 studies are divided in 107 for training, 23 for validation for 23 for testing. Hyperparameters are specified in table 3.6. Table 3.7 presents the correlation and normalized RMSE values for the 8 tissues. Prediction vs ground truth graphs for the test set are depicted in figure 3.20.

Table 3.6 - Parameters used in the 2D approach with pre-processing, using the 40 weeks PMA dataset

Minibatches	100000
Batch size	40
CNN	VGG-like
Filters per convolution	16
Learning rate	0,001
Optimizer	RMSProp
Filters' size	3 × 3
Max-pooling size	2 × 2

Table 3.7 - Values of correlation and normalized RMSE for the 2D approach with pre-processing, using the 40 weeks PMA test set

	CB	mWM	BGT	vCSF	uWM	BS	GM	eCSF
Correlation	0,989	0,965	0,980	0,969	0,988	0,984	0,983	0,977
Normalized RMSE (%)	4,6	5,5	8,0	5,5	7,0	3,6	8,9	4,7

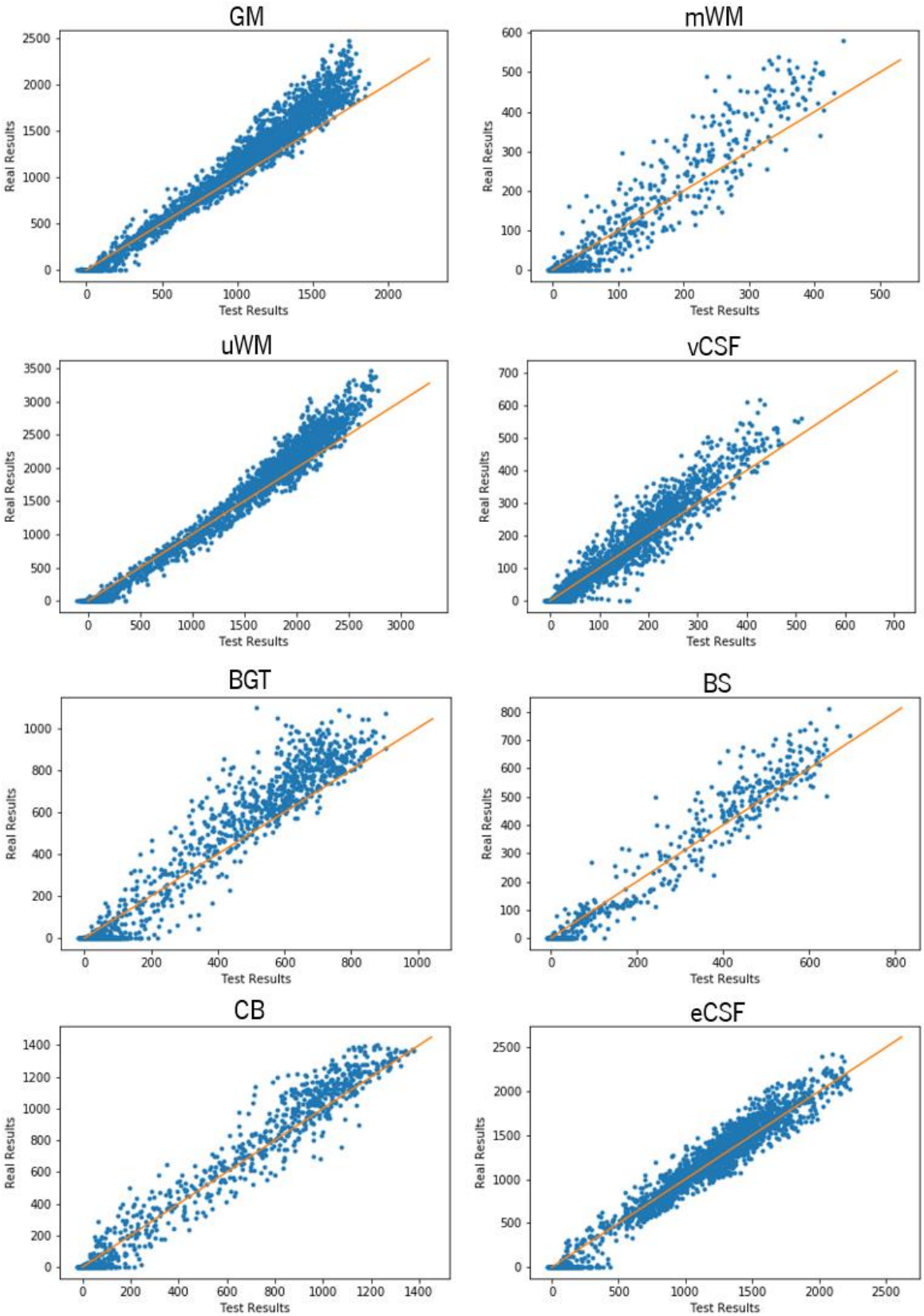


Figure 3.20 - Test (prediction) vs Real (ground truth) graphs of the 8 tissues for the 2D approach with pre-processing, using images at 40 weeks PMA. The name of the tissue is on top of its corresponding graph. Values are in mm^3.

Using the dataset at 40 weeks PMA, the graphs show that GM and uWM continue to be the best predicted tissues, now along with eCSF and vCSF. As explained before, the high number of 0 values in many slices explains the low values for RMSE, mainly in BS, CB, mWM and BGT. Compared to the same approach in the previous dataset, mWM now has a better-looking graph. This could be explained by the higher number of training samples with positive volumes for this data when compared to the 30 weeks PMA one. As it happened in the other dataset, for each tissue, the larger the tissue in the slice, the harder it is to have a more accurate prediction. We can see that clearly in the GM and uWM graphs, where the highest values tend to deviate from the identity function.

3.3.3 3D APPROACH USING CHUNKS

The following approach consists on using chunks/parts of 3D neonatal brain images/slices to predict the volumes of 8 different tissues.

3.3.3.1 30 WEEKS

The same split used for previous studies at 30 weeks PMA was used. Table 3.8 specifies the hyperparameters, while figure 3.21 depicts the prediction vs ground truth graphs for the test set. Table 3.9 shows the normalized RMSE and correlation results for each presented tissue.

Table 3.8 - Parameters used in the 3D approach using chunks, with the 30 weeks PMA dataset

Minibatches	24000
Chunk size	6 slices (approx. 11 % of the full brain)
Batch size	200
CNN	VGG-like
Filters per convolution	16
Architecture in the CNN	3D on the first block
Learning rate	0,001
Optimizer	RMSProp
Filters' size	3 × 3 × 2 and 3 × 3 × 1
Max-pooling size	2 × 2 × 2 and 2 × 2 × 1

Table 3.9 - Values of correlation and normalized RMSE for the 3D approach using chunks, with the 30 weeks PMA test set

	CB	mWM	BGT	vCSF	uWM	BS	GM	eCSF
Correlation	0,985	0,954	0,982	0,936	0,987	0,992	0,986	0,953
Normalized RMSE (%)	4,8	6,5	6,2	8,0	4,9	4,0	4,4	6,7

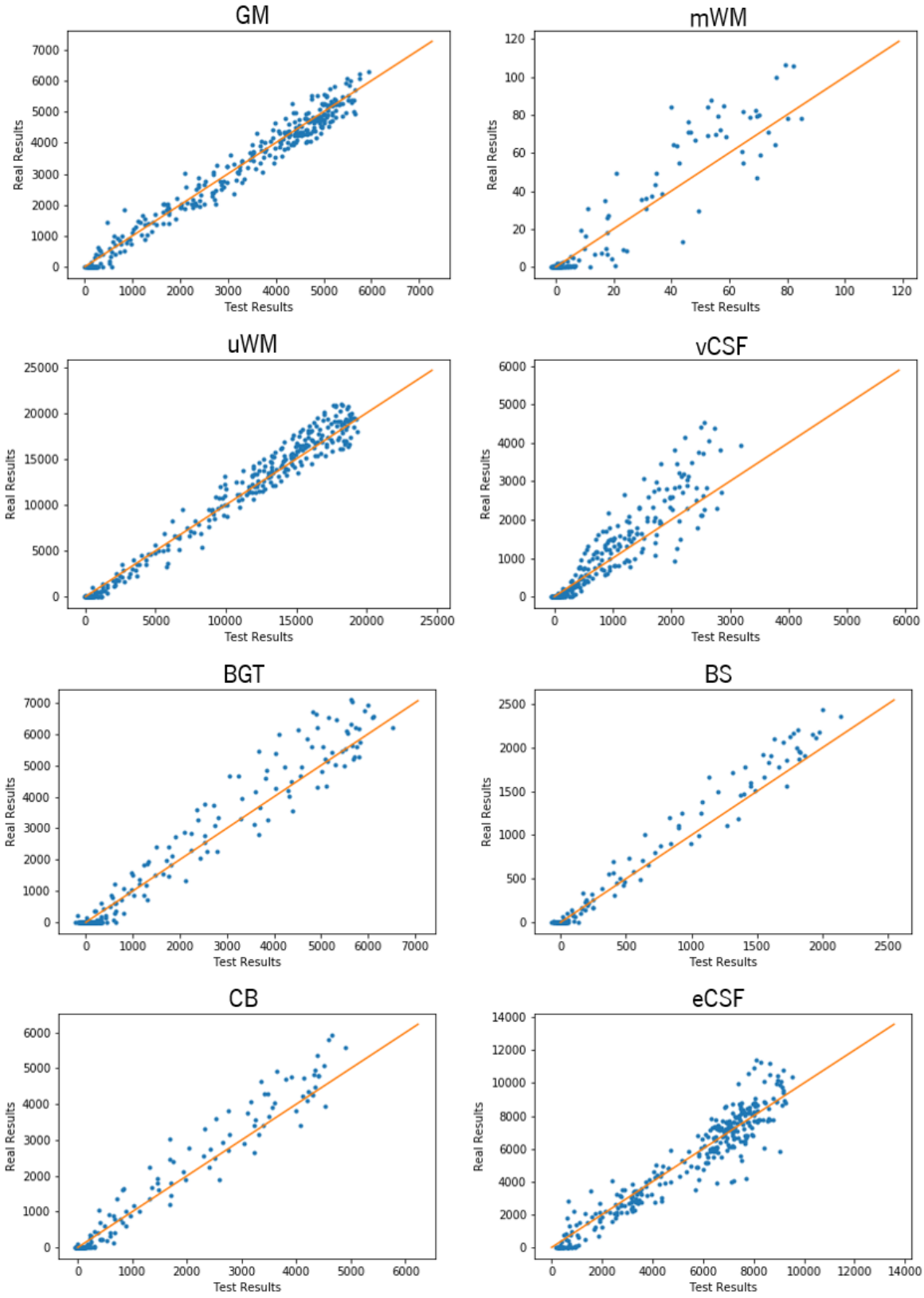


Figure 3.21 - Test (prediction) vs Real (ground truth) graphs of the 8 tissues for the 3D approach using chunks, with data at 30 weeks PMA. The name of the tissue is on top of its corresponding graph. Values are in mm^3.

Analyzing the graphs and the metrics in table 3.9, at first glance, this 3D approach using chunks improves the prediction for most tissues in general, by looking at the graphs and in terms of both metrics.

GM and uWM continue to be the easiest to predict. The low normalized RMSE of BGT, BS and CB continue to be explained by the high number of chunks with a volume of 0 mm^3 for these tissues. For these 3 tissues, even though there are a lot of these null values, the ones where the tissue is present, the predictions look good, hence their high correlation. The prediction of eCSF also looks great if we exclude some outliers. By considering the graphs, the prediction of mWM improved a lot also but continues to be unsatisfactory for data at 30 weeks PMA. The normalized RMSE may seem lower and correlation higher than expected for mWM due to high number of chunks where this tissue is absent. For each tissue in general, the larger the size of that tissue in the chunk, the harder it is to predict that specific volume. This is clearly observed in the vCSF graph and it can help explain its lower correlation and higher normalized RMSE, when compared to the others.

Other relevant tests (for bigger chunks), as well as their analysis and discussion, are presented in Appendix A.1 and A.2. The variation of correlation with the increase of the chunk size is presented in Appendix B.

3.3.3.2 40 WEEKS

The same split used for the previous study at 40 weeks PMA (section 3.3.2.2) was used. Table 3.10 specifies the hyperparameters, while figure 3.22 depicts the prediction vs ground truth graphs for the test set. Table 3.11 shows the normalized RMSE and correlation results for each presented tissue.

Table 3.10 - Parameters used in the 3D approach using chunks, with the 40 weeks PMA dataset

Minibatches	30000
Chunk size	10 slices (approx. 9 % of the full brain)
Batch size	42
CNN	VGG-like
Filters per convolution	8
Architecture in the CNN	3D on the first block
Learning rate	0,001
Optimizer	RMSProp
Filters' size	3 × 3 × 2 and 3 × 3 × 1
Max-pooling size	2 × 2 × 2 and 2 × 2 × 1

Table 3.11 - Values of correlation and normalized RMSE for the 3D approach using chunks, with the 40 weeks PMA test set

	CB	mWM	BGT	vCSF	uWM	BS	GM	eCSF
Correlation	0,984	0,978	0,982	0,940	0,986	0,990	0,987	0,961
Normalized RMSE (%)	6,0	4,5	6,2	6,9	4,8	3,3	5,0	6,3

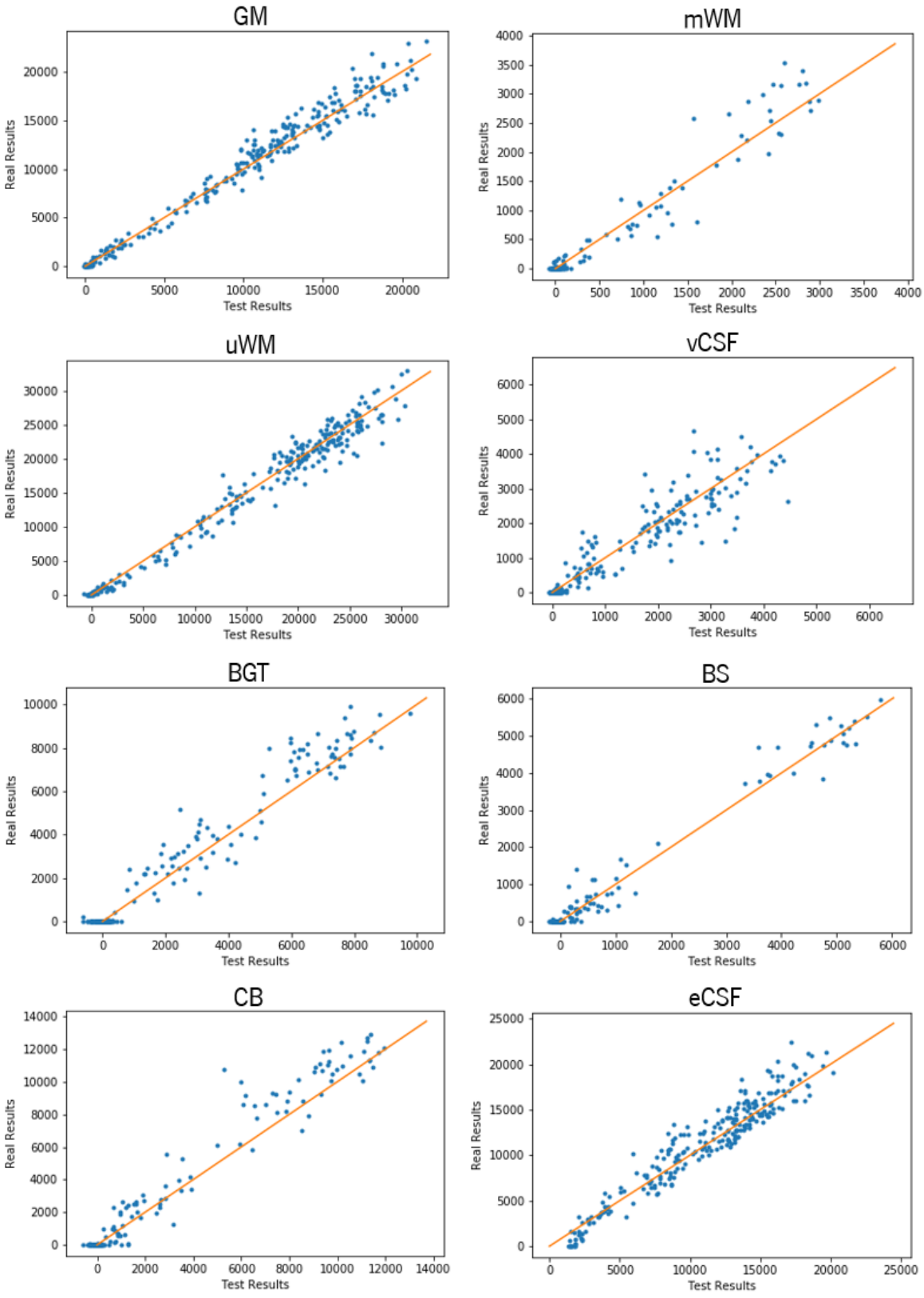


Figure 3.22 - Test (prediction) vs Real (ground truth) graphs of the 8 tissues for the 3D approach using chunks, with data at 40 weeks PMA. The name of the tissue is on top of its corresponding graph. Values are in mm^3.

By assessing the graphs, this approach for the dataset at 40 weeks PMA, when compared to the 2D approach using pre-processing techniques (subsection 3.3.2.2), generally has a slight improvement in the predictions, as it seems the predicted volumes are closer to the real ones. Again, GM and uWM are the easiest to predict. However, the overall improvement is only slight, because, for example, the graphs and metrics of some tissues like vCSF and eCSF do not follow this betterment trend.

Compared to the same approach on the 30 weeks dataset, CB, BS and BGT graphs have a similar behavior. mWM has a better prediction, but that is probably connected to the higher number of samples with positive volumes in training, as stated before. vCSF is an exception. For this approach, using both datasets, the larger is its volume, vCSF tends to have a less accurate prediction.

3.3.4 3D APPROACH USING FULL 3D BRAINS

The same split used in the approach using chunks for dataset at 30 weeks PMA (section 3.3.3.1.) was considered. For this approach, due to lack of data, the test set samples (and corresponding labels) are augmented so there are enough examples for a reliable analysis of the results. Hence, the red dots in the graphs correspond to the original test set, and the blue ones correspond to the augmented ones. In this approach, memory constraints only allowed the use with the dataset at 30 weeks PMA. Table 3.12 specifies the hyperparameters, while figure 3.23 depicts the prediction vs ground truth graphs of the 8 volumes for the test set and figure 3.24 the graphs for the other 5 descriptors. Table 3.13 and table 3.14 show the normalized RMSE and correlation results for each presented tissue and descriptor, respectively.

Table 3.12 - Parameters used in the 3D approach using full brains with the 30 weeks PMA dataset.

Minibatches	25000
Batch size	22
CNN	VGG-like
Filters per convolution	16
Learning rate	0,001
Optimizer	RMSProp
Architecture in the CNN	3D on the first block
Filters' size	3 × 3 × 3 and 3 × 3 × 1
Max-pooling size	2 × 2 × 2 and 2 × 2 × 1

Table 3.13 - Values of correlation and normalized RMSE of the 8 volumes for the 3D approach using full brains, with the 30 weeks PMA test set

	CB	mWM	BGT	vCSF	uWM	BS	GM	eCSF
Correlation	0,832	0,767	0,727	0,147	0,836	0,685	0,812	0,604
Normalized RMSE (%)	18,4	18,5	19,0	11,9	11,3	19,4	12,6	15,7

Table 3.14 - Values of correlation and normalized RMSE of the 5 other descriptors for the 3D approach using full brains, with the 30 weeks PMA test set

	IS	OS	MT	GI	MC
Correlation	0,704	0,766	0,293	0,385	0,478
Normalized RMSE (%)	19,3	12,8	29,5	20,8	34,6

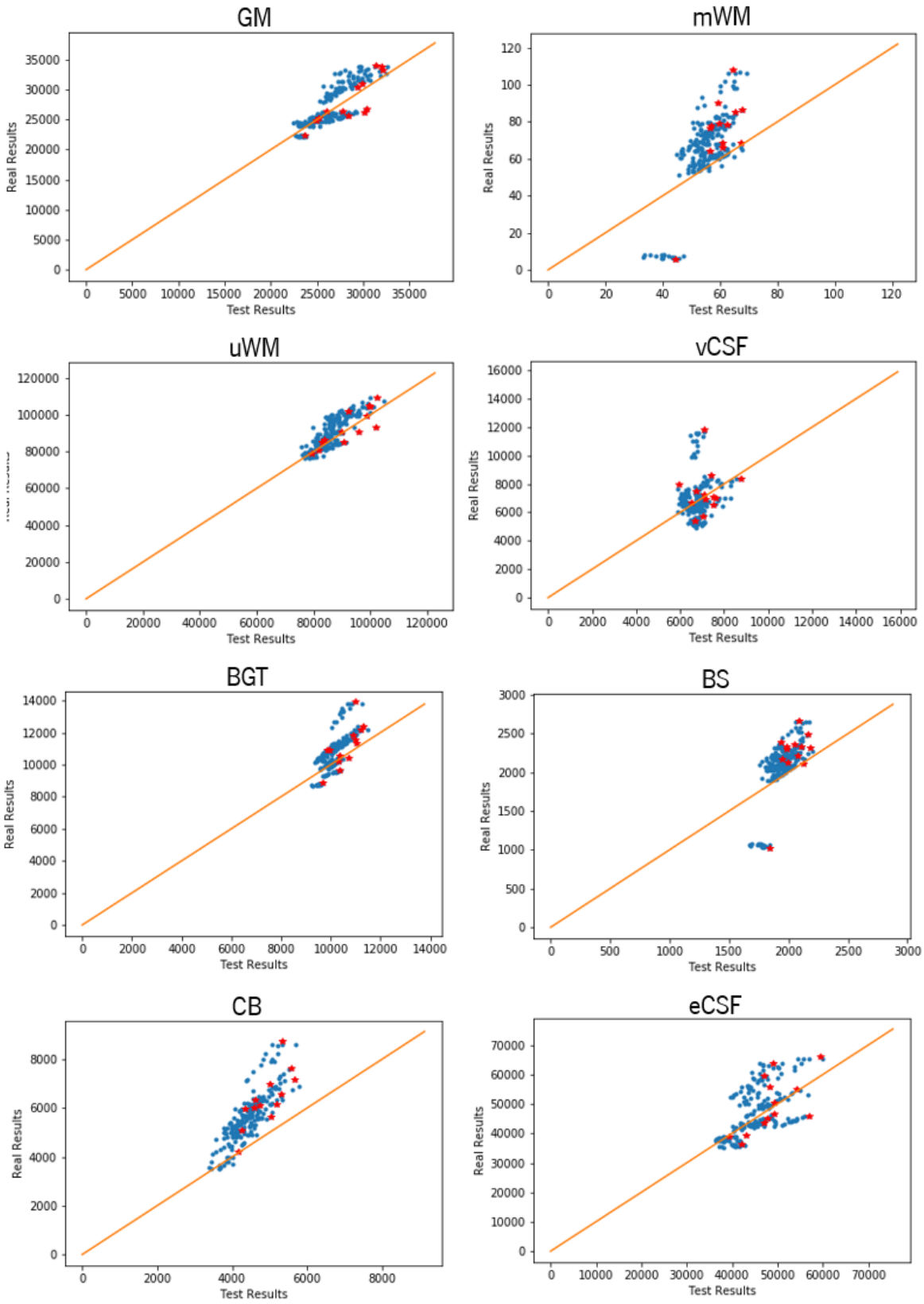


Figure 3.23 - Test (prediction) vs Real (ground truth) graphs of the 8 tissues for the 3D approach using full brains, with data at 30 weeks PMA. The name of the tissue is on top of its corresponding graph. Values are in mm^3.

Using entire full brains, the overall predictions expectedly became worse, mainly due to the fact that the same network is now dealing with larger volumes. GM and uWM continue to be the best predicted tissues by observing the graphs and because their correlation is among the highest and normalized RMSE the lowest, when compared to the other ones. CB also has a good correlation, but it is a little deviated from the optimal y=x function (that's why normalized RMSE is higher when compared to the best predicted tissues). For both BGT and BS, their correlation (and normalized RMSE) would be improved if it wasn't for the outliers situated, respectively, higher and lower than the other test samples. In the case of vCSF, the outliers have a very important role in explaining the very low correlation of this tissue when compared to the other ones. In addition, for this case, it hard to build a linear relationship between all predictions and ground truth values. As for mWM, it is possible to denote a line for the predictions, however, these predictions should not be considered trustworthy due to the fact of the bad segmentations for this tissue at 30 weeks PMA.

As previously remarked, the red dots correspond to the actual test set and the blue ones to augmented examples based in the actual test set. While sometimes these augmented examples can help in increasing correlation (for example, in the cases of CB and BS), other times can reduce it by creating more outliers (in the cases of eCSF and GM).

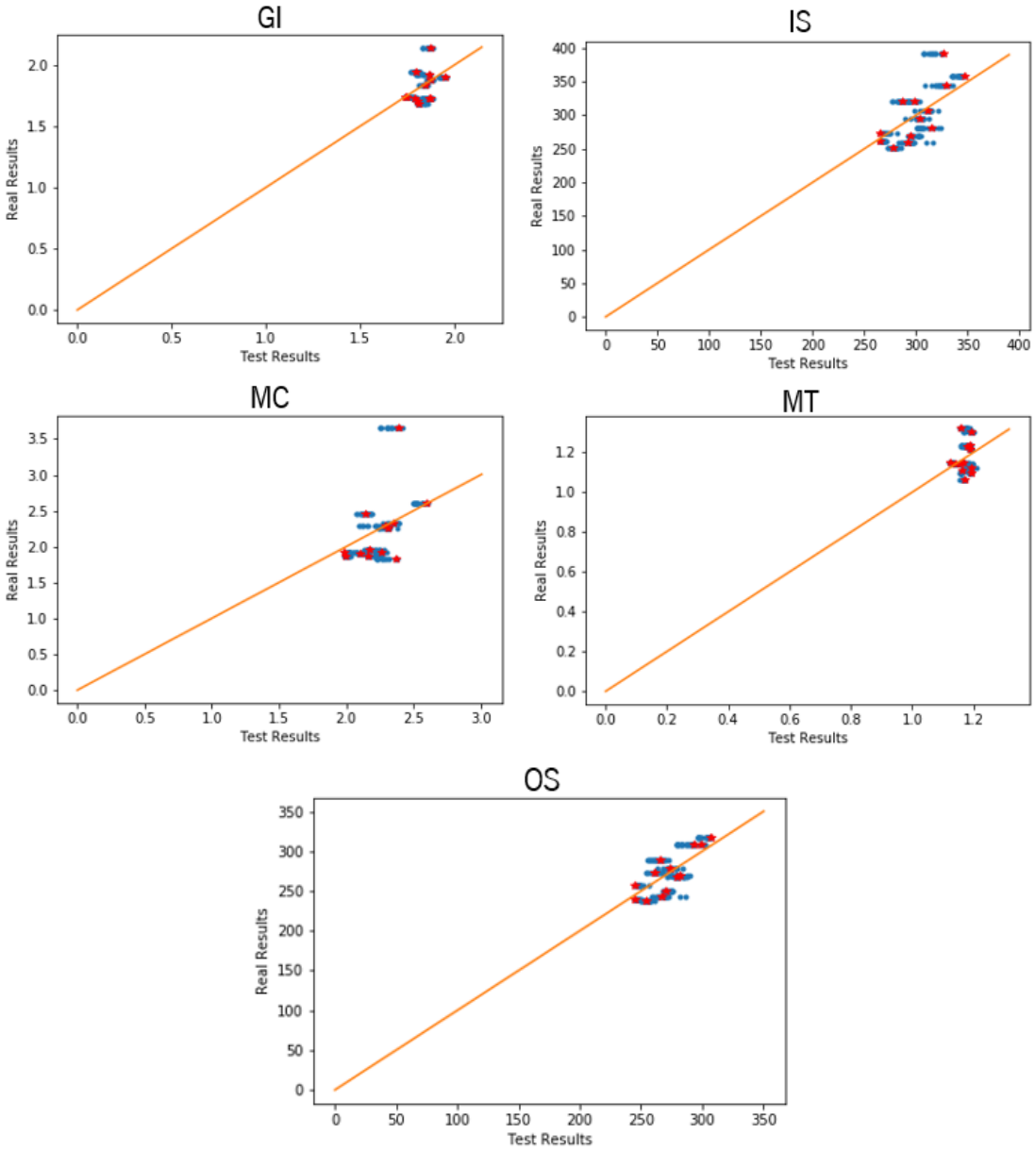


Figure 3.24 - Test (prediction) vs Real (ground truth) graphs of the 5 other descriptors for the 3D approach using full brains, with data at 30 weeks PMA. The name of the descriptor is on top of its corresponding graph.

Concerning the other descriptors, OS and IS seem to be the best among all, by looking at the graphs and considering the both correlation and normalized RMSE. Their graphs show that if in a full brain the descriptor is higher, the prediction is also higher in general. The augmented tests (blue dots) seem to worsen the results for these 2 tissues. MC also seems to have some good predictions, excepting one or two outliers. These outliers may explain the lower correlation and higher normalized RMSE for MC. For the descriptors GI and MT, the system is only capable of predicting values inside the interval of ground truth values of the test sets. These tissues are the hardest to predict for the system. This could be explained by the fact that the range of values for these 2 descriptors is so low (around [1,5;2,1] for GI and [1.0;1,3] for MT), that the network cannot distinguish between them.

3.3.5 OTHER CONSIDERATIONS

Due to memory limitations, it is necessary to use some techniques, one of them being downscaling. These downscaling degrades the images and that could be a factor that affects the training and, consequently, the predictions of the DNN, thus affecting the results.

As mentioned in the methods, more particularly in subsection 3.2.3.3, 3 different architectures in the CNNs of the 3D approaches are considered. Using 3D on the first block (first 3 convolution layers and pooling one) is the architecture that allows better overall results when in comparison to both other architectures (3D only on all convolutional layers and 3D only on all pooling layers).

Also reported in the methods, in the predictions of the entire 3D brain (approach of section 3.2.4), some data augmentation techniques are added. One of these techniques that can play an important role in the results for the approach is the change of image intensity. Since the system is trained with this augmentation, these changes in intensity can play an important role in the training part, influencing the testing. For this same approach, the small amount of data can be another issue that hampers the optimal results. In fact, only 60 full 3D brains (studies) are used in training. More diversity of studies is required to improve the performance.

In terms of the different tissues it is important to underline that GM and WM (and sometimes eCSF) are clearly the easiest tissues to predict, for all the models constructed. It may be related to their larger size and consequent higher number of samples with varied (positive) volumes, when compared to the other tissues.

The VGG-like network is the CNN used in all the results shown in the sections before. Since in theory ResNet architecture is more recent and powerful than VGG, the ResNet-like CNN was used just in the 3D approaches, mainly to try to improve the results for the full 3D neonatal brain. However, this ResNet adaptation did not lead to better predictions than the VGG-like CNN. Techniques like batch normalization and PReLUs, which are also tested with both CNNs, however, did not lead to any improvements.

The most used optimizer in the tests is RMSprop. Adam was also executed but it led to a slightly worse optimization. In terms of learning rate, n=0,001 is the one that brought a better optimization for the dataset used. A higher learning rate of n=0,01 seemed too difficult convergence, due to

fluctuations around the minimum. A lower learning rate of n=0,0001 would take a lot of time to converge. In fact, n=0,001 is the same learning rate recommended in [27], for RMSprop.

There is no single kernel which leads to better results. It differs depending on the situation. However, if what may differentiate the examples are some small and local features, small kernels should be utilized [25]. This way, in the 2D approaches, 3x3 kernels are used in the 2D convolutions. In the 3D ones, 3x3x3 (and 3x3x2 and 3x3x1) kernels are applied with the 3D convolutions. The number of kernels per convolution were assigned to 4, 8, 32 and 64. Graphs, correlation and normalized RMSE were evaluated in order to determine the appropriate number of kernels per convolution. Better looking graphs, along with a higher correlation and lower normalized RMSE are possible for 8 and 16 filters.

Having in mind a clinical application of these approaches on systems, the test times for each slice (2D approach with techniques), chunk (3D approach using chunks) and full brain (3D approach using full brains) are noted in table 3.15.

Table 3.15 – Test times of a slice, chunk and full brain, for each approach, respectively

	30 weeks	40 weeks
2D with techniques	0,016 seconds/slice	0,026 seconds/slice
3D chunks	0.21 seconds/chunk of 6 slices	0,37 seconds/chunk of 10 slices
Full 3D brains	1,15 seconds/brain	———

The quantifications predicted in the 2D approach with pre-processing techniques can be summed to have the quantification of a certain part of the brain or a certain tissue or even a full brain, depending on the physician's objective. Respectively, it is just necessary to predict the volumes from the slices where that part of the brain or tissue or full brain are present. The same can be said about the 3D approach using chunks.

4 Quantification of brain characteristics using rat MRI

Similar to the work done in the quantification of neonatal brain, the objective of this chapter is to quantify 3 volumes of the rat brain, directly from MRI scans of these rodents. Figure 4.1 portrays a scheme of this quantification for a clear understanding. The red arrow depicts the direct quantification, the objective of this chapter

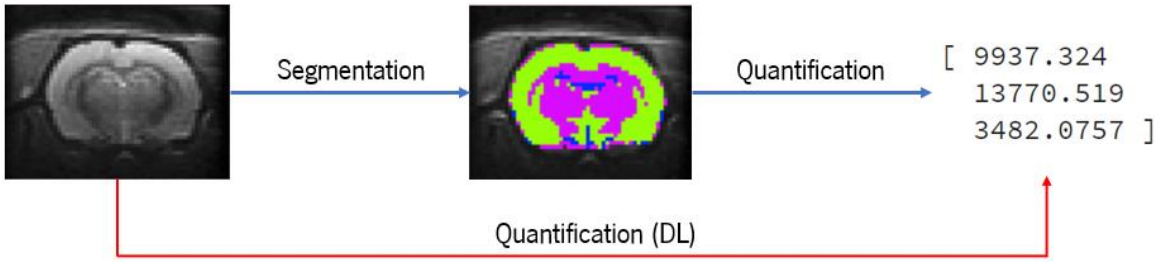


Figure 4.1 – Quantification of the rat brain from a MRI scan: blue arrows represent the usual steps required for the quantification; red arrow shows the DL quantification of the rat brain directly from a MR scan, the goal of this chapter.

4.1 MATERIALS

Similar to the neonatal brain project, all the data used in this project is described. The mouse brain images are considered the main source of data and are used for the deep learning tasks. The ground truth volumes serve as labels.

4.1.1 RAT BRAIN IMAGES

The dataset with rats MRI scans was acquired from a part of a project called SIGMA, partially financed by *Fundação para a Ciência e Técnologia* and *Agence National de Recherche* (ref: FCT-ANR/NEU-OSD/0258/2012). Wistar rats were the species of rodents used in this study. Using a 2x2 surface coil associated with software *Paravision 6*, the 139 scanning sessions were carried out on a *Bruker Biospec 11.7T* pre-clinical scanner. SE-EPI diffusion sensitive acquisitions were used, with Time to Repetition=5s, Time to Echo=20ms, in-plane resolution of 0.375x0.375 mm, slice thickness of 0.5 mm over 40 slices and a Field-of-View of 24 mm. Each acquisition acquired and averaged 10 volumes. Using WM, GM and CSF priors, *SPM Segment* was used with the objective of creating the ground truth for segmentation [64]. Additional information can be found on table 4.1. In figure 4.2, it is represented an example of data: a brain slice.

Figure 4.2 - Example of a rat brain slice.

Table 4.1 - Additional information about the rat data

	Mouse brain dataset
Number of scans	139
Reconstruction matrix	64 × 64 × 40
Reconstruction voxel size (mm^3)	3,75 × 3,75 × 4,96875

4.1.2 VOLUMES

This time, the volumes expected to be predicted are GM, WM and CSF. An example of a rat brain segmentation slice is shown I figure 4.3 just to give information about the location of these 3 tissues.

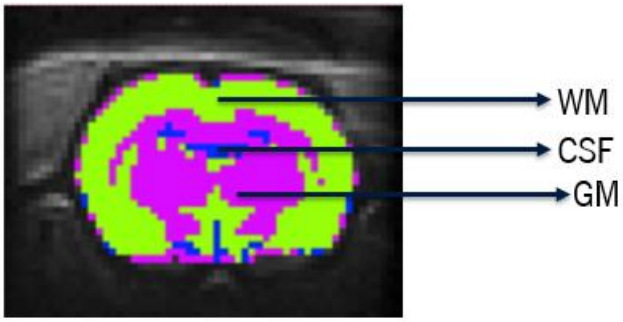


Figure 4.3 – Location of GM, WM and CSF in a brain segmentation slice of the rat.

The volumes are provided in an *Excel* file with the counted voxels of GM, WM and CSF for each mouse brain study. These voxels are quantified based on ground truth segmentations of these studies. This file is used in the 3D approach.

A similar file with the voxels per 2D slice for each study is also considered, also quantified based on ground truth segmentations. This one is used in the 2D approach.

4.2 METHODS

The objective is to predict the 3 mentioned volumes from the full 3D rat brain, i.e., the neural network is trained in a 3D procedure, so it receives full brains and tries to predict their WM, GM and CSF volumes. A 2D procedure which quantifies volumes per slice is also executed. Instead of presenting the predicted volumes for each slice in the 2D approach, these quantifications are summed for the full brain. This is due to a following comparison with a segmentation method.

4.2.1 3D APPROACH

4.2.1.1 LOADING PROCESS

A loading technique similar to the one described in subsection 3.2.1.1 is used. In this case, it is only necessary to load the brain images as there is no need for any padding. All 139 studies are divided in a train (89), validation (22) and test (28) dataset.

As previously remarked, *SimpleITK* is used to get the width, height and depth of each study. For this whole set of mouse brain images, the dimensions corresponded to 64 × 64 × 40 (Width × Height × Depth). Since all the studies have the same size, there was no need to pad any of them.

Also using *SimpleITK*, all the information related to each study is considered and transformed into *numpy* array with an ideal converter. The results of the loading are sets of concatenated slices for training, validation and test. One example of a slice is represented in figure 4.2.

4.2.1.2 PRE-PROCESSING: IMAGE NORMALIZATION

Although MRI scans may look similar from the doctors (and patients) point of view, a scan on the same subject is susceptible to some changes in intensity, depending on its acquisition. In other words, there can be some large variations on intensity within MRI scans, due to, intra and inter-scan variables. For example, using the same scanner, a later scan and/or a different positioning from the subject can lead to these intensity variations. In addition, in DL, it is required to deal with large amounts of data, where some of the studies are acquired with different parameters and scanners. This way, reducing these intensity variations in MRI scans could be an influential pre-processing step for a more accurate MRI analysis [65], [66]. On the mouse brain studies, the normalization is performed by dividing all the intensity values by the maximum intensity of the studies.

4.2.1.3 Pre-processing: 3D transformation

This process is already described in section 3.2.3.1 for the neonatal brain. In order to transform the slices to 3D so they can be fed to the CNN, a new variable (depth) is created. This can be achieved by stacking the slices according to the size of the full 3D brain. This stacking is shown in figure 4.4 in the 3 planes.

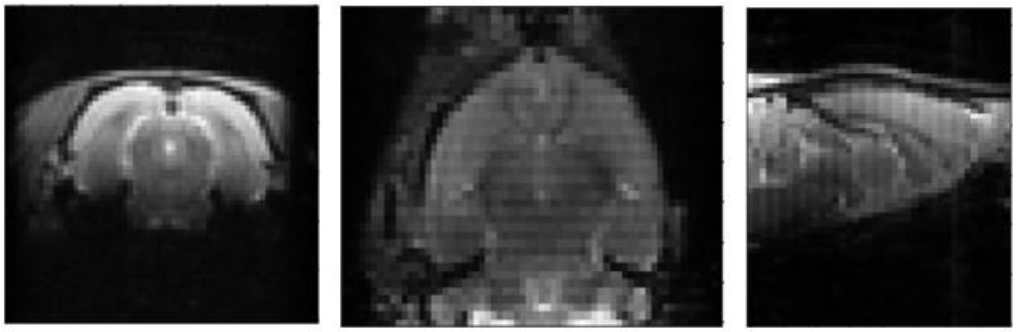


Figure 4.4 - Neonatal brain represented in 3 planes: A- axial; B- coronal; C- sagittal.

4.2.1.4 Data augmentation

With the objective of creating more samples for the training step of the system, data augmentation was performed in 4 different ways:

- Scaling, using only reductions. The factor varies between 0.8 and 1, being 1 the exact same image and 0.8 a reduction of 80% of that image.
- Intensity variation, using the equation (4) described in subsection 3.2.3.2. The variable *number* ranges from [0,97;1,03].
- Rotations, from -7 to 7 degrees.
- Flipping.

4.2.1.5 Neural-network, training and validation

The neural network used for this problem is another adaptation ResNet-34 [61], depicted in figure 4.5.

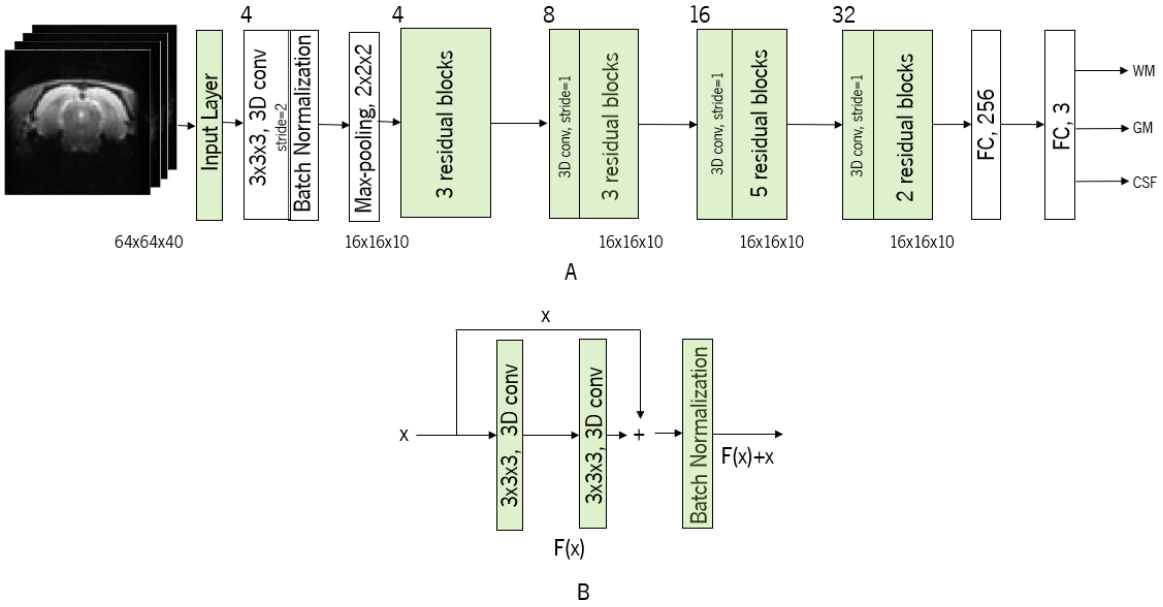


Figure 4.5 - Architecture of the adapted 3D neural network, based on Resnet-34: A- a general overview; B- exemplification of a residual block. The number of filters associated to each step is located on the top left of that step.

As shown in figure 4.5 A, the 3D images suffer a first convolution and pooling after the input layer. Then, the architecture has 4 big steps (represented in green), each of them consisting in the repetition of several residual blocks. The residual learning blocks that create the steps mentioned before are exemplified in figure 4.5 B. All the 3D convolutions use ReLUs as activation functions. Batch normalization is used after the initial convolution and, in each residual block, after the sum of the residuals. The network ends with 2 fully connected layers, being the final output the 3 descriptors, which correspond to the volumes for prediction.

The sizes of the filters (3x3x3), the learning rate (0,001) and the optimizer (RMSProp), are chosen based on the successes of this approach on the neonatal brain studies. The GPU used for a faster training was Quadro P6000.

4.2.1.6 Prediction and Model Evaluation

The same metrics used to assess the results are the same ones used in the neonatal brain approach. In other words, prediction vs ground truth graphs are plotted, along with correlation and normalized RMSE.

4.2.2 2D APPROACH

The 2D approach has the same steps as the 3D approach described in the subsection before (4.2.1): loading, image normalization, training, validation, predictions and model evaluation. There are only 3 differences.

The first one consists in the removal of the 3D stacking transformation (described in subsection 4.2.1.3). Since the first step performs a loading of data in slices and the method of training is in 2D, there is no need to turn it 3D. The second difference is that for obtaining the best results using this approach, there was no need of performing data augmentation.

The last aspect is clearly the neural network used. The 2D CNN is presented in figure 4.6 and it is another adaptation of the VGG-like [54].

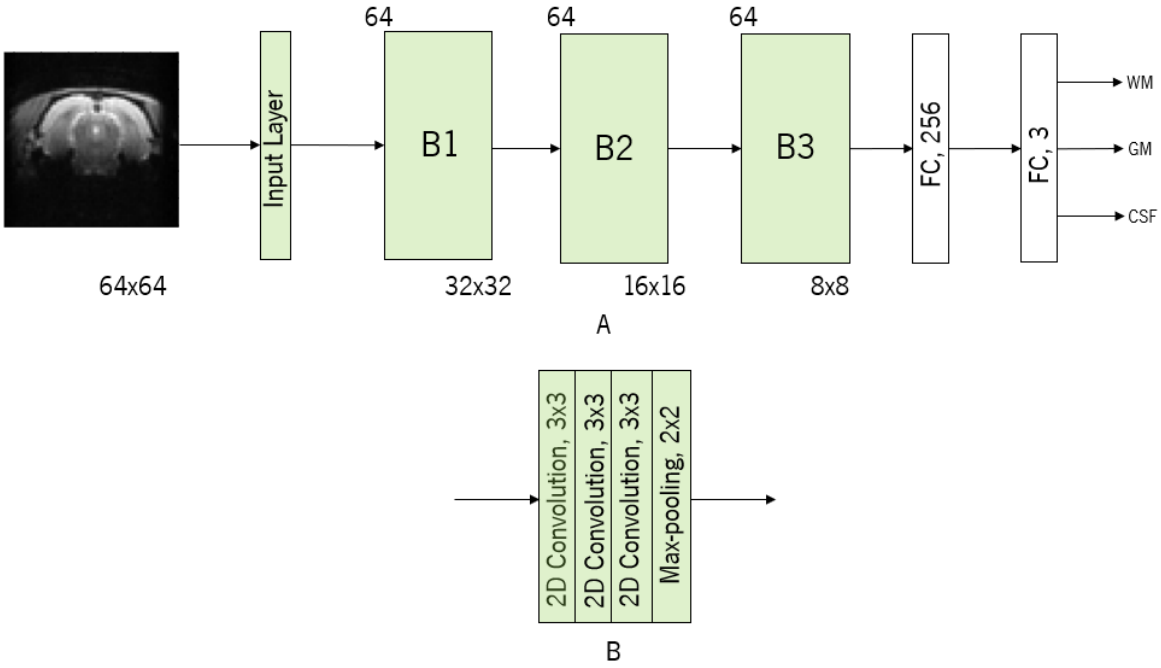


Figure 4.6 - Architecture of the adapted 2D neural network, adapted from the original VGG network: A- a general overview; B- convolutions and pooling used in each step (B1, B2 and B3). The number of filters associated to each step is located on the top left of that step.

As depicted in figure 4.6, after the input layer, 3 similar steps (B1, B2 and B3) are executed. Each step consists of 3 consecutive 2D convolutions with 3x3 filters, following by a max-pooling by a 2x2 window. Each convolution produces 64 filters, associated to ReLU as activation function. Finally, the network ends with 2 fully connected layers. The first one has 256 units, while the last one has just 3, corresponding to the volumes that must be predicted. The learning rate (0,001) and the optimizer (RMSProp) were the same ones used in the 3D approach.

4.3 RESULTS AND DISCUSSION

4.3.1 3D APPROACH

This approach consists on predicting the volumes of 3 different tissues from full 3D mouse brains. For the dataset, 89 studies were used for training, 22 for validation and 28 for testing. The hyperparameters fixed for this test are specified in table 4.2. Figure 4.7. shows the prediction vs ground truth graphs for the 3 tissues, using unseen samples (test set) by the model. Table 4.3 shows the correlation and normalized RMSE results for each graph presented.

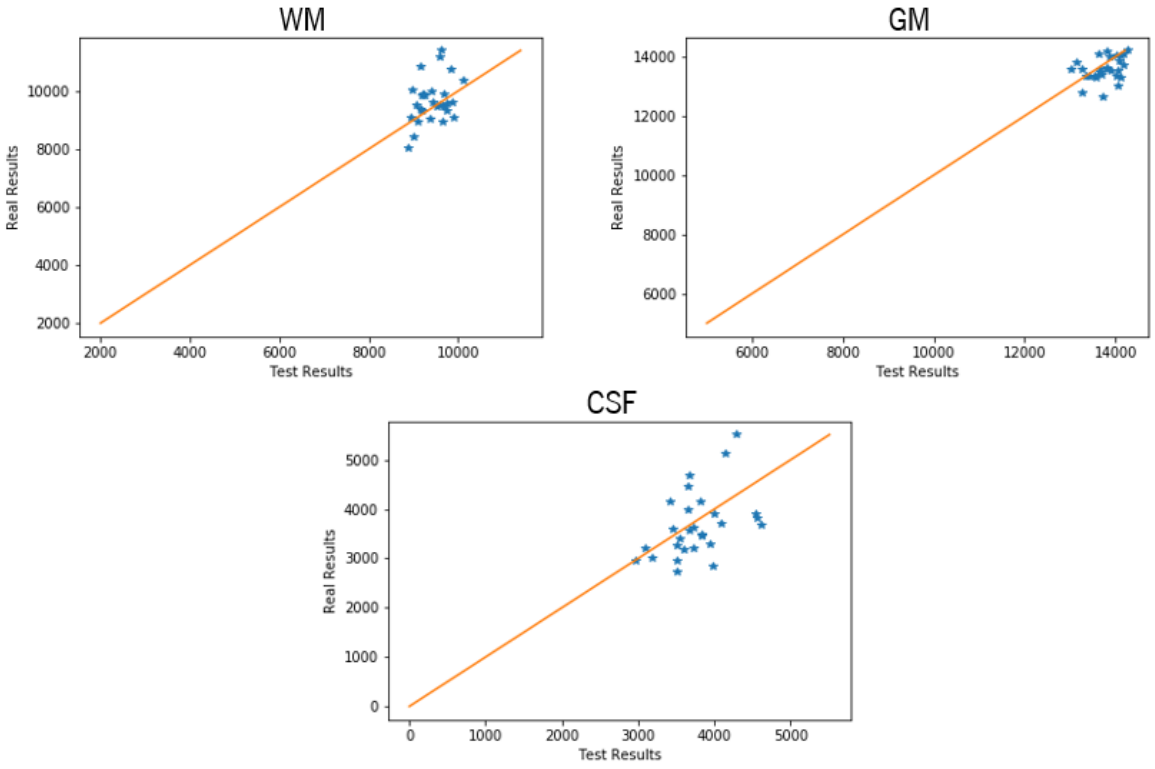


Figure 4.7 - Test (prediction) vs Real (ground truth) graphs for the 3 tissues of the full 3D mouse brain, using a 3D approach. The name of the tissue is on top of its corresponding graph. Values are in voxels.

Table 4.2 - Parameters used in the 3D approach for the rats

Minibatches	70000
Batch size	21
Filters' size	3 × 3 × 3 and 3 × 3 × 1
Max-pooling size	2 × 2 × 2 and 2 × 2 × 1

Table 4.3 - Values of correlation and normalized RMSE for the 3D approach, using the rats test set

	WM	GM	CSF
Correlation	0,361	0,339	0,432
Normalized RMSE (%)	6,7	3,0	10,4

The normalized RMSE evaluates the errors towards the identity function considering the range of volumes that are being dealt with. Considering this metric, GM is the best predicting tissue. The reason behind that fact is because GM has the largest range of volumes (range of 15305 voxels in the training set). The absolute error between the 3 tissues is very similar, although they have different range of volumes.

The correlation of the dots is low. This fact does not overwhelm the results as it may seem. The explanation is related to the number of examples in the test set and their volume values. Since the number of examples is low and those examples have very similar values, it is hard to assess the variation of both variables.

4.3.2 2D APPROACH

This following procedure has the objective of predicting the volumes of 3 different tissues of the rat brain, using 2D mouse brain slices. It is considered the same split used in the 3D approach. The hyperparameters fixed for this test are specified in table 4.4. Figure 4.8. shows the prediction vs ground truth graphs for the 3 tissues, using unseen samples (test set) by the model. Since in the next section (i.e. section 4.3.3), both these 3D and 2D quantification approaches for the rat brain are going to be compared to another developed method (a segmentation method), these graphs present the 2D quantification for the for the full 3D brain, which is the unity of comparison between the 2 methods in that section. The quantifications for each 2D slice were obtained through Deep Learning techniques and summed in the end, to quantify for the full rat brain. Table 4.5 shows the correlation and normalized RMSE results for each graph presented.

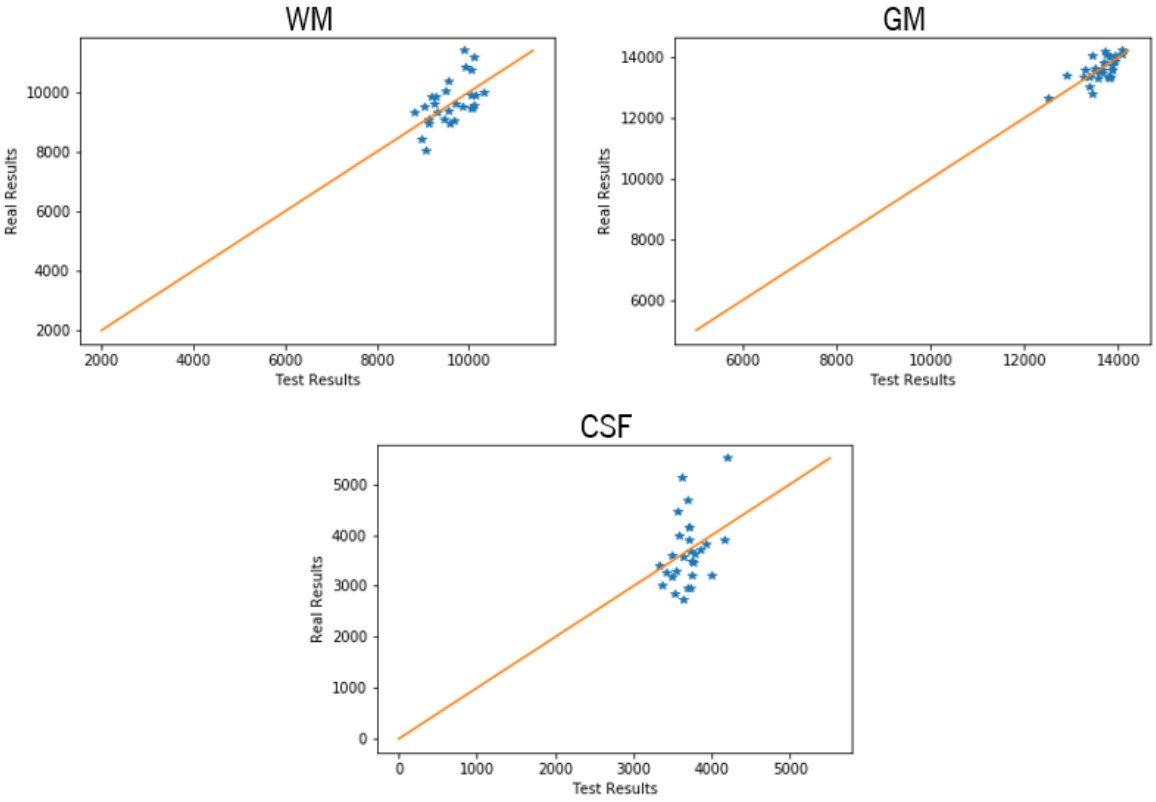


Figure 4.8 - Test (prediction) vs Real (ground truth) graphs for the 3 tissues of the full mouse brain, using a 2D approach. The name of the tissue is on top of its corresponding graph. Values are in voxels.

Table 4.4 - Parameters used in the 2D approach for the rats

Minibatches	100000
Batch size	20
Filters' size	3 × 3
Max-pooling size	2 × 2

Table 4.5 - Values of correlation and normalized RMSE for the 2D approach, using the rats test set

	WM	GM	CSF
Correlation	0,555	0,648	0,373
Normalized RMSE (%)	5,7	2,0	10,4

Using a 2D approach, allowed an improvement on the prediction of 2 tissues as it is possible to observe in the graphs and metrics. Both GM and WM got a higher value of correlation and lower normalized RMSE. In fact, using a 2D approach, more different samples (i.e. slices) are used for training, which allows a more accurate performance.

vCSF did not improve at all, it stayed relatively the same. Curiously, in both 3D and 2D approaches, the exact same outliers are easily noticed in the CSF graphs. These exceptions are studies with ground truth volumes near or higher than 5000 voxels, which means that the system is not prepared to predict these particular cases (probably because there were not similar cases like these in the training set).

However, correlations are not optimal. The reason behind this, as previously referred for the 3D approach, is the fact of having a fewer number of examples in the test set and their close volumes.

4.3.3 Comparison with a segmentation method

The following subsection has the objective of comparing the quantification method developed in the subsections before with a segmentation one, developed by Mariana Rodrigues in her dissertation "Brain Semantic Segmentation: A Deep Learning approach in Human and Rat MRI studies" (2018).

The quantification approach (blue line in figure 4.9) predicts the number of voxels for each tissue of the rat brain, using DL techniques directly from an MRI scan. The segmentation method (red lines in figure 4.9) performs a segmentation on the images using DL techniques and uses a non-DL quantification on the segmentations to count the number of voxels. For each method, the same pre-processing of data was used before the DL steps, respectively.

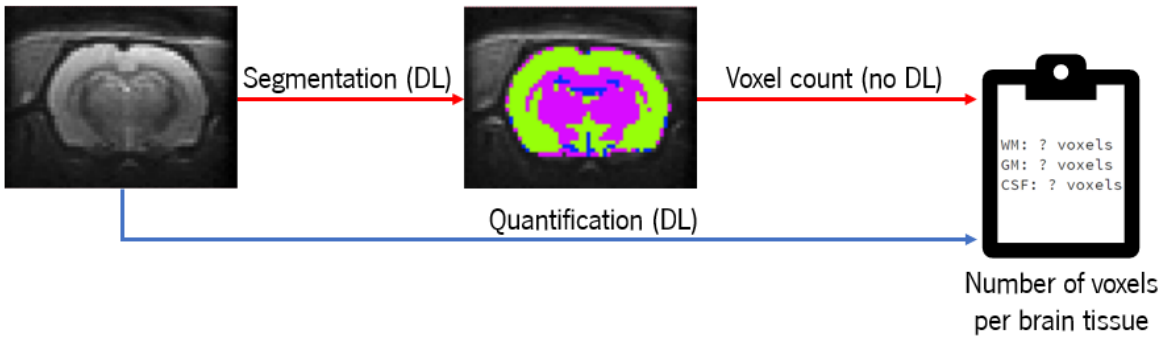


Figure 4.9 - Representation of both methods compared in this section.

The 2 methods were compared using 2D and 3D approaches. Table 4.6 shows the comparison of the methods for the 2D approach, while table 4.7 shows the comparison for the 3D approach. Since the objective is only a comparison between methods by counting the number of voxels for each tissue per brain, the metric used was the normalized RMSE, to access the error between the counted number of voxels for each method and the ground truth.

Table 4.6 – Values of RMSE of each tissue of the rat brain for the segmentation and quantification methods, using a 2D approach

	WM	GM	CSF
Segmentation	2,4%	2,0%	7,9%
Quantification	5,7%	2,0%	10,4%

Table 4.7 - Values of RMSE of each tissue of the rat brain for the segmentation and quantification methods, using a 3D approach

	WM	GM	CSF
Segmentation	4,3%	1,6%	9.1%
Quantification	6,7%	3,0%	10,4%

Considering a 2D approach, the quantification method performs similarly to the segmentation one for GM. For WM and CSF, the segmentation method performs better. Using a 3D approach, the normalized RMSE associated to the segmentation method is lower than the quantification one. In sum, even though it is a more automatic method, the quantification is associated to a higher normalized RMSE.

These results were expected due to the fact that the segmentation method is in fact a voxel classification process and deals with probabilities. This method has the objective of classifying each voxel between 4 different classes (GM, WM, CSF and background). Each voxel is classified into a certain class depending on the probabilities of belonging to that class and the others. Therefore, a voxel can be correctly segmented even if there is a certain probability of not being the correct class. In the quantification method, which is related to regression, the system directly predicts a continuous

number instead of 4 classes. Therefore, all this uncertainty will be translated to the continuous number.

It is important to notice that these results do not explicitly conclude which method is better. Even though normalized RMSE is a good metric for a comparison between segmentation and quantification in these terms, it is not the most accurate metric to access the performance of a segmentation method. For example, a segmentation can have the very similar number of voxels as the ground truth for a certain tissue but, this segmentation may be dislocated (i.e. miss some voxels where the tissue is present and segment voxels from other tissues).

5 Conclusion

This project consists on creating one or more DL systems/approaches that would help in the quantification of volumes and other descriptors of 2 neonatal brain datasets. An extension of this idea to a mouse brain dataset was executed in the last stages of the dissertation.

Since the project involved DL techniques and brain MRI, it was necessary to be immersed in these fields. To build all models, knowledge on image processing, DNN architectures, DL tuning techniques and statistical metrics was vital.

In the neonatal case, the first set of models consists in 2D approaches, which predict the volumes of 8 brain tissues per slice. The 2D approach where image processing techniques is used, displays better results, when compared to the one without a previous image processing. The second set of models consists in a 3D approach using chunks of the neonatal brain. The 8 volumes are predicted per chunk of brain of a certain size. Depending on the size of the chunk, smaller chunks led to a slight improvement in performance while larger ones led to less accurate results. A model using full 3D neonatal brains to predict 8 volumes and 5 more descriptors is also constructed. As expected due to an increase in size, the prediction of a full 3D brain is more difficult when compared to the previous approaches. VGG-like and ResNet-like are the CNNs used for these neonatal approaches. For all the models developed for the neonatal brain, GM and WM stand out as the easiest tissues to predict, while mWM is hardest to predict.

As for the dataset containing rat's studies, the objective is to quantify the volumes of 3 tissues. Two methods are performed: a 3D method based on the full brain and a 2D method based on slices. For the 3D and 2D methods, adaptations of, respectively, ResNet-34 and VGG are constructed. Due to the small number of examples in the test set, low correlations are obtained. However, the range of normalized RMSE is acceptable. These DL quantification methods with rat data are also compared with a DL segmentation approach.

In sum, the different approaches load brain MRI data and quantify the desired descriptors, with a considerable performance, proving that this quantification method can be generalized to different brains (i.e. neonatal and rat brain). In addition, it is not necessary to indicate which features are relevant for the predictions because the CNNs do that task on their own. In the neonatal case, a considerable step was taken to achieve the goal of the big project where this work is inserted, which is to predict, using DL techniques, if a preterm infant has cognitive and social impairments from neonatal MRI data. However, it is important to notice that in spite of these systems having a more

ambitious final objective and being an improvement in terms of automation and execution time when compared to classification techniques, they are based in regression, which means that it considers all uncertainties related with the small probabilities that classification tasks usually ignore.

Even though the results are positive, there are always some topics that should be further investigated and may contribute for improvements and/or new discoveries. For future work, firstly, like in every DL projects, a more focused tuning of the (hyper)parameters in the models can lead to a slight improvement of the results. These parameters range from learning rates, number of minibatches to small changes in the convolution layers of the CNN. However, more time was required for the tuning of these rat brain approaches, since they were developed in the final stages of the dissertation.

Specifically, for the neonatal approaches, especially in the model for the full 3D brain, where there are less examples to train and it is hard to create data augmentation, another deep CNN, like *Inception* [67] could lead to for more accurate predictions. The same can be said for the 3D method on the dataset with rat studies.

It is not possible to execute the 3D approach using full 3D brains for the data at 40 weeks PMA. Since this approach uses full 3D brains and the MRI scans at this stage of development are way larger than at 30 weeks PMA, a more optimal technique is required to perform this approach.

REFERENCES

[1] H. Müller, X. Gao, and S. Luo, “From medical imaging to medical informatics,” *Comput. Methods Programs Biomed.*, vol. 92, no. 3, pp. 225–226, 2008.

[2] T. Tolxdorff, T. M. Deserno, H. Handels, and H.-P. Meinzer, “Advances in Medical Image Computing,” *Methods Inf. Med.*, vol. 48, no. 4, pp. 311–313, 2009.

[3] F. J. Wippold II, “Head and neck imaging: The role of CT and MRI,” *J. Magn. Reson. Imaging*, vol. 25, no. 3, pp. 453–465, 2007.

[4] B. Larroque *et al.*, “Special care and school difficulties in 8-year-old very preterm children: The Epipage Cohort study,” *PLoS One*, vol. 6, no. 7, 2011.

[5] P. Moeskops *et al.*, “Development of cortical morphology evaluated with longitudinal MR brain images of preterm infants,” *PLoS One*, vol. 10, no. 7, pp. 1–22, 2015.

[6] H. Kidokoro, J. Neil, and T. Inder, “A New MRI Assessment Tool to Define Brain Abnormalities in Very Preterm Infants at Term,” *AJNR. Am. J. Neuroradiol.*, vol. 34, no. 11, pp. 2208–2214, 2013.

[7] J. M. Van Kampen and H. A. Robertson, “The BSSG rat model of Parkinson’s disease: progressing towards a valid, predictive model of disease,” *EPMA J.*, vol. 8, no. 3, pp. 261–271, 2017.

[8] T. Petrasek *et al.*, “A rat model of Alzheimer’s disease based on Abeta42 and pro-oxidative substances exhibits cognitive deficit and alterations in glutamatergic and cholinergic neurotransmitter systems,” *Front. Aging Neurosci.*, vol. 8, no. 83, pp. 1–12, 2016.

[9] M. I. Razzak, S. Naz, and A. Zaib, “Deep Learning for Medical Image Processing: Overview, Challenges and Future,” arXiv:1704.06825v1, 2017.

[10] R. Culkin and S. R. Das, “Machine Learning in Finance: The Case of Deep Learning for Option Pricing,” 2017.

[11] P. Moeskops *et al.*, “Assessment of quantitative cortical biomarkers in the developing brain of preterm infants,” *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 8670, 2013.

[12] P. Moeskops *et al.*, “Prediction of cognitive and motor outcome of preterm infants based on automatic quantitative descriptors from neonatal MR brain images,” *Sci. Rep.*, vol. 7, no. 2163, 2017.

[13] H. C. Glass, A. T. Costarino, S. A. Stayer, C. Brett, F. Cladis, and P. J. Davis, "Outcomes for extremely premature infants," *Anesthesia and analgesia*, vol. 120, no. 6, pp. 1337–1351, 2015.

[14] L. Liu *et al.*, "Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals," *Lancet*, vol. 388, no. 10063, pp. 3027–3035, 2016.

[15] M. de Jong, M. Verhoeven, and A. L. van Baar, "School outcome, cognitive functioning, and behaviour problems in moderate and late preterm children and adults: A review," *Semin. Fetal Neonatal Med.*, vol. 17, no. 3, pp. 163–169, 2012.

[16] B. Latal, "Prediction of Neurodevelopmental Outcome After Preterm Birth," *Pediatr. Neurol.*, vol. 40, no. 6, pp. 413–419, 2009.

[17] J. Hornak, *The Basics of MRI*, Rochester, NY Cent. Imaging Sci. Rochester Inst. Technol., 2002. Online book at: http://www.cis.rit.edu/htbooks/mri/ [Accessed: 23-Apr-2018].

[18] M. J. N. L. Benders, K. J. Kersbergen, and L. S. de Vries, "Neuroimaging of White Matter Injury, Intraventricular and Cerebellar Hemorrhage," *Clin. Perinatol.*, vol. 41, no. 1, pp. 69–82, 2014.

[19] A. Langham, "MRI for premature neonatal brain injury: a case report," *J. Med. Radiat. Sci.*, vol. 64, no. 2, pp. 152–155, 2017.

[20] P. Wintermark, "The role of brain MRI scanning in the newborn," *Paediatr. Child Health*, vol. 22, no. 4, pp. 155–159, 2012.

[21] J. P. Boardman *et al.*, "A common neonatal image phenotype predicts adverse neurodevelopmental outcome in children born preterm," *Neuroimage*, vol. 52, no. 2, pp. 409–414, 2010.

[22] R. Rathbone *et al.*, "Perinatal cortical growth and childhood neurocognitive abilities," *Neurology*, vol. 77, no. 16, pp. 1510–1517, 2011.

[23] P. Manger, "Is 21st Century Neuroscience too Focussed on the Rat/Mouse Model of Brain Function and Dysfunction?," *Front. Neuroanat.*, vol. 2, 2008.

[24] C. Hoyer, N. Gass, W. Weber-Fahr, and A. Sartorius, "Advantages and challenges of small animal magnetic resonance imaging as a translational tool," *Neuropsychobiology*, vol. 69, no. 4, pp. 187–201, 2014.

[25] N. Ketkar, *Deep Learning with Python*, Apress, Berkeley, CA, 2017.

[26] J.G. Lee *et al.*, "Deep Learning in Medical Imaging: General Overview," *Korean J. Radiol.*, vol. 18, no. 4, pp. 570-584, 2017.

[27] S. Ruder, "An overview of gradient descent optimization," arXiv:1609.04747v2, 2017.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A

Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[29] W. Rawat and W. Zenghui, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, 2017.

[30] J. L. Y. Cheong and S. P. Miller, "Imaging the neonatal brain in the 21st century: why, when and how?," *Arch. Dis. Child. - Fetal Neonatal Ed.*, vol. 103, no. 1, pp. F4–F5, 2018.

[31] A. Makropoulos, S. J. Counsell, and D. Rueckert, "A review on automatic fetal and neonatal brain MRI segmentation," *Neuroimage*, vol. 170, pp. 231-248, 2018.

[32] L. Wang *et al.*, "LINKS: Learning-based multi-source IntegratioN frameworK for Segmentation of infant brain images," *Neuroimage*, vol. 108, pp. 160-172, 2015.

[33] P. Moeskops *et al.*, "Automatic segmentation of MR brain images of preterm infants using supervised classification.," *Neuroimage*, vol. 118, pp. 628–41, 2015.

[34] K. J. Kersbergen *et al.*, "Longitudinal Regional Brain Development and Clinical Risk Factors in Extremely Preterm Infants," *J. Pediatr.*, vol. 178, p. 93–100.e6, 2016.

[35] K. Keunen *et al.*, "White matter maturation in the neonatal brain is predictive of school age cognitive capacities in children born very preterm," *Dev. Med. Child Neurol.*, vol. 59, no. 9, pp. 939–946, 2017.

[36] Y. Kono, N. Yonemoto, H. Nakanishi, S. Kusuda, and M. Fujimura, "Changes in survival and neurodevelopmental outcomes of infants born at <25 weeks' gestation: a retrospective observational study in tertiary centres in Japan," *BMJ Paediatr. Open*, vol. 2, no. 1, p. e000211, 2018.

[37] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, pp. 1097-1105, 2012.

[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NIPS*, pp. 91–99, 2015.

[40] D. Amodei *et al.*, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," vol. 48, *Proceedings of Machine Learning Research*, pp. 173-182, 2016.

[41] J. Haugeland, *Artificial Intelligence: The Very Idea*, The MIT Pr. Cambridge, Mass ., 1985.

[42] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1240–1251,

2016.

[43] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu, "A deep learning method for classifying mammographic breast density categories," *Med. Phys.*, vol. 45, no. 1, pp. 314–321, 2018.

[44] M. Pekala, N. Joshi, D. E. Freund, N. M. Bressler, D. C. DeBuc, and P. M. Burlina, "Deep Learning based Retinal OCT Segmentation," arXiv:1801.09749v1, 2018.

[45] C. F. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu, "An Exploration of 2D and 3D Deep Learning Techniques for Cardiac MR Image Segmentation," arXiv:1709.04496v2, 2017.

[46] S. Wang, M. Kim, G. Wu, and D. Shen, "Scalable High Performance Image Registration Framework by Unsupervised Deep Feature Representations Learning," *IEEE Trans. Biomed. Eng.,* vol. 63, no. 7, pp. 1505-16, 2016.

[47] H. Jang, S. M. Plis, V. D. Calhoun, and J. H. Lee, "Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks," *Neuroimage*, vol. 145, pp. 314–328, 2017.

[48] Q. Dou *et al.*, "Automatic Detection of Cerebral Microbleeds from MR Images via 3D Convolutional Neural Networks," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1182–1195, 2016.

[49] M. Ghafoorian *et al.*, "Student Beats the Teacher: Deep Neural Networks for Lateral Ventricles Segmentation in Brain MR," arXiv:1801.05040v1, 2018.

[50] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. De Vries, M. J. N. L. Benders, and I. Isgum, "Automatic Segmentation of MR Brain Images with a Convolutional Neural Network," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.

[51] Y. Xu, T. Géraud, and I. Bloch, "From Neonatal to Adult Brain MR Image Segmentation in a Few Seconds Using 3D-like Fully Convolutional Network and Transfer Learning," 2017 *IEEE International Conference on Image Processing (ICIP)*, p. 4417-4421, 2017.

[52] W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, pp. 1–10, 2016.

[53] F. Liao, X. Chen, X. Hu, and S. Song, "Estimation of the volume of the left ventricle from MRI images using deep neural networks," arXiv:1702.03833v1, 2017.

[54] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image

Recognition," arXiv:1409.1556v6, 2015.

[55] F. Dubost *et al.*, "3D Regression Neural Network for the Quantification of Enlarged Perivascular Spaces in Brain MRI," arXiv:1802.05914v1, 2018.

[56] A. E. H. Scheenstra, R. C. G. van de Ven, L. van der Weerd, A. M. J. M. van den Maagdenberg, J. Dijkstra, and J. H. C. Reiber, "Automated Segmentation of in Vivo and Ex Vivo Mouse Brain Magnetic Resonance Images," *Mol. Imaging*, vol. 8, no. 1, pp. 35–44, 2009.

[57] J. W. Totenhagen, A. Bernstein, E. S. Yoshimaru, R. P. Erickson, and T. P. Trouard, "Quantitative magnetic resonance imaging of brain atrophy in a mouse model of NiemannPick type C disease," *PLoS One*, vol. 12, no. 5, 2017.

[58] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The Design of SimpleITK," *Front. Neuroinform.*, vol. 7, 2013.

[59] Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare, "SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research," *J. Digit. Imaging*, vol. 31, no. 3, pp. 290-303, 2018.

[60] D. Giavarina, "Understanding Bland Altman analysis," *Biochem. Medica*, vol. 25, no. 2, pp. 141–151, 2015.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for Image Recognition," arXiv:1512.03385v1, 2015.

[62] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," arXiv:1502.03167v3, 2015.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," arXiv:1502.01852v1, 2015.

[64] R. Magalhães *et al.*, "The dynamics of stress: a longitudinal MRI study of rat brain structure and connectome," *Mol. Psychiatry*, pp. 1–9, 2017.

[65] X. Sun *et al.*, "Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions," *Biomed. Eng. Online*, vol. 14, no. 1, p. 73, 2015.

[66] M. Shah *et al.*, "Evaluating intensity normalization on MRIs of human brain with multiple sclerosis," *Med. Image Anal.*, vol. 15, no. 2, pp. 267–282, 2011.

[67] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," arXiv:1602.07261v2, 2016.

APPENDICES

A – 3D APPROACH USING BIGGER CHUNKS

The following results shown in this section consists on the prediction of the test set of data at 30 weeks PMA, using chunks of bigger size than the one presented in subsection 3.3.3.1. Since the chunks are larger, the number of samples for testing is smaller. With the objective of having a reliable analysis of the results, more test samples are created by augmenting images and labels of the original test set. Therefore, for both approaches, the red dots correspond to the original test set, while the blue ones correspond to the augmented examples.

A.1 – 3D APPROACH USING CHUNKS OF 9 SLICES

Table of hyperparameters (table A.1), table with metrics for each tissue (table A.2) and prediction vs ground truth graphs (figure A.1) are presented for experiments in brain with a chunk size of 9 slices.

Table A.1 - Parameters used in the 3D approach using chunks of 9 slices with the 30 weeks PMA dataset

Minibatches	30000
Chunk size	9 slices (approx. 17 % of the full brain)
Batch size	42
Filters per convolution	8
Method	3D on the first block
Filters' size	3 × 3 × 3 and 3 × 3 × 1
Max-pooling size	2 × 2 × 2 and 2 × 2 × 1

Table A.2 - Values of correlation and normalized RMSE for the 3D approach using chunks of 9 slices with the 30 weeks PMA test set

	CB	mWM	BGT	vCSF	uWM	BS	GM	eCSF
Correlation	0,980	0,933	0,953	0,935	0,989	0,967	0,985	0,939
Normalized RMSE (%)	4,3	7,7	8,1	6,5	5,1	5,9	4,4	7,5

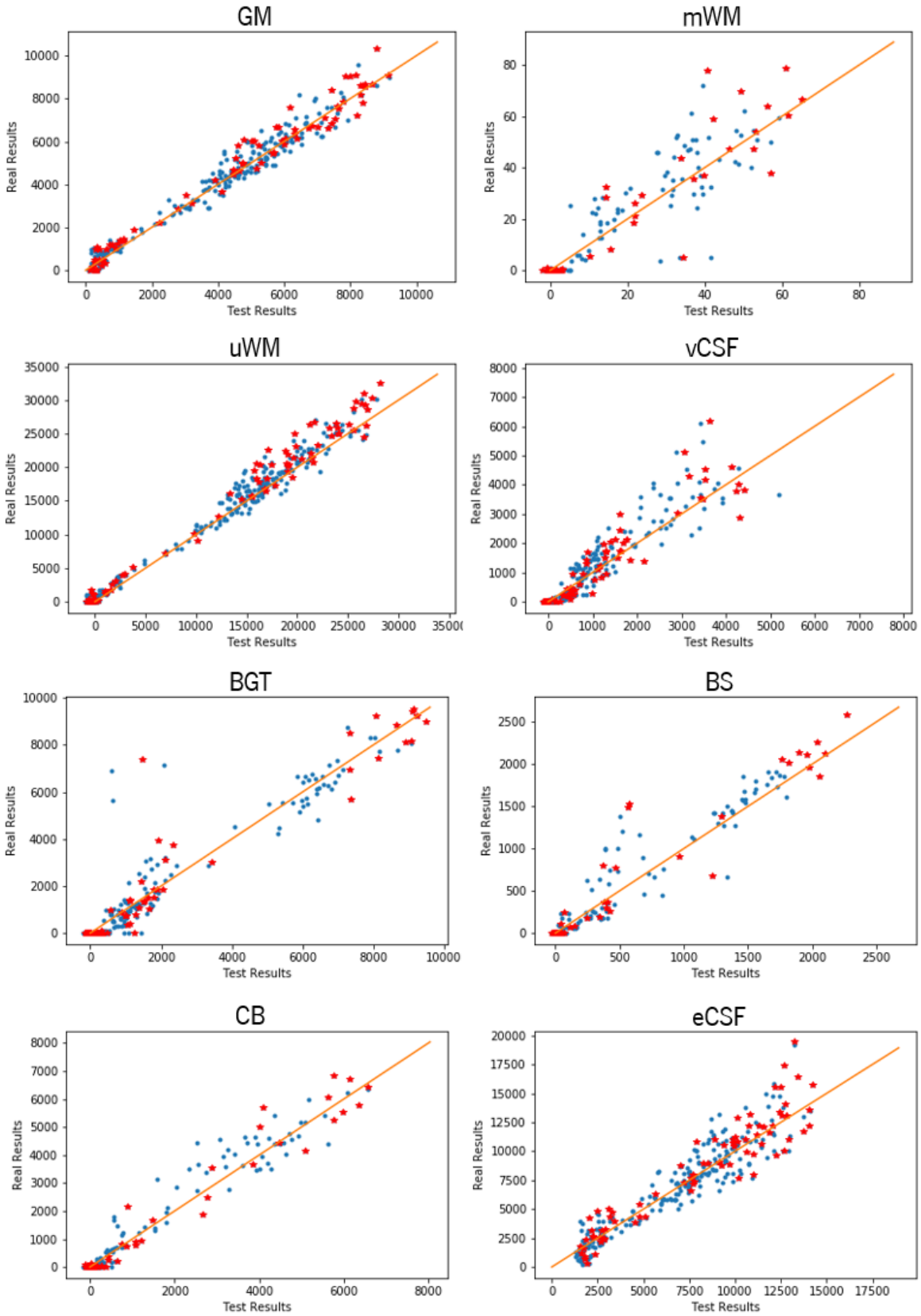


Figure A.1 - Test (prediction) vs Real (ground truth) graphs for the 8 tissues for the 3D approach using chunks of 9 slices, with images at 30 weeks PMA. The name of the tissue is on top of its corresponding graph. Values are in mm^3.

Comparing with the performance in subsection 3.3.3.1, in general, the graphs and metrics look slightly worse. That is expected because the size of the chunk is bigger, which means the range of volumes to predict is much higher. The augmentation technique in the test set proves that the system has a similar performance for both the original test set (red dots) and the augmented test set (blue dots). The larger the size of chunk, the harder the prediction, which means the appearance of some outliers in tissues that are harder to predict, such as in BGT and BS graphs.

A.2 – 3D APPROACH USING CHUNKS OF 24 SLICES

Table of hyperparameters (table A.3), table with metrics for each tissue (table A.4) and prediction vs ground truth graphs (figure A.2) are presented for experiments in brain with a chunk size of 24 slices.

Table A.3 - Parameters used in the 3D approach using chunks of 24 slices with the 30 weeks PMA dataset

Minibatches	27000
Chunk size	24 slices (approx. 45 % of the full brain)
Batch size	22
Filters per convolution	8
Method	3D on the first block
Filters' size	3 × 3 × 3 and 3 × 3 × 1
Max-pooling size	2 × 2 × 2 and 2 × 2 × 1

Table A.4 - Values of correlation and normalized RMSE for the 3D approach using chunks of 24 slices with the 30 weeks PMA test set

	CB	mWM	BGT	vCSF	uWM	BS	GM	eCSF
Correlation	0,957	0,925	0,972	0,911	0,978	0,978	0,963	0,882
Normalized RMSE (%)	8,3	11,2	7,4	7,9	8,8	6,7	6,5	10,3

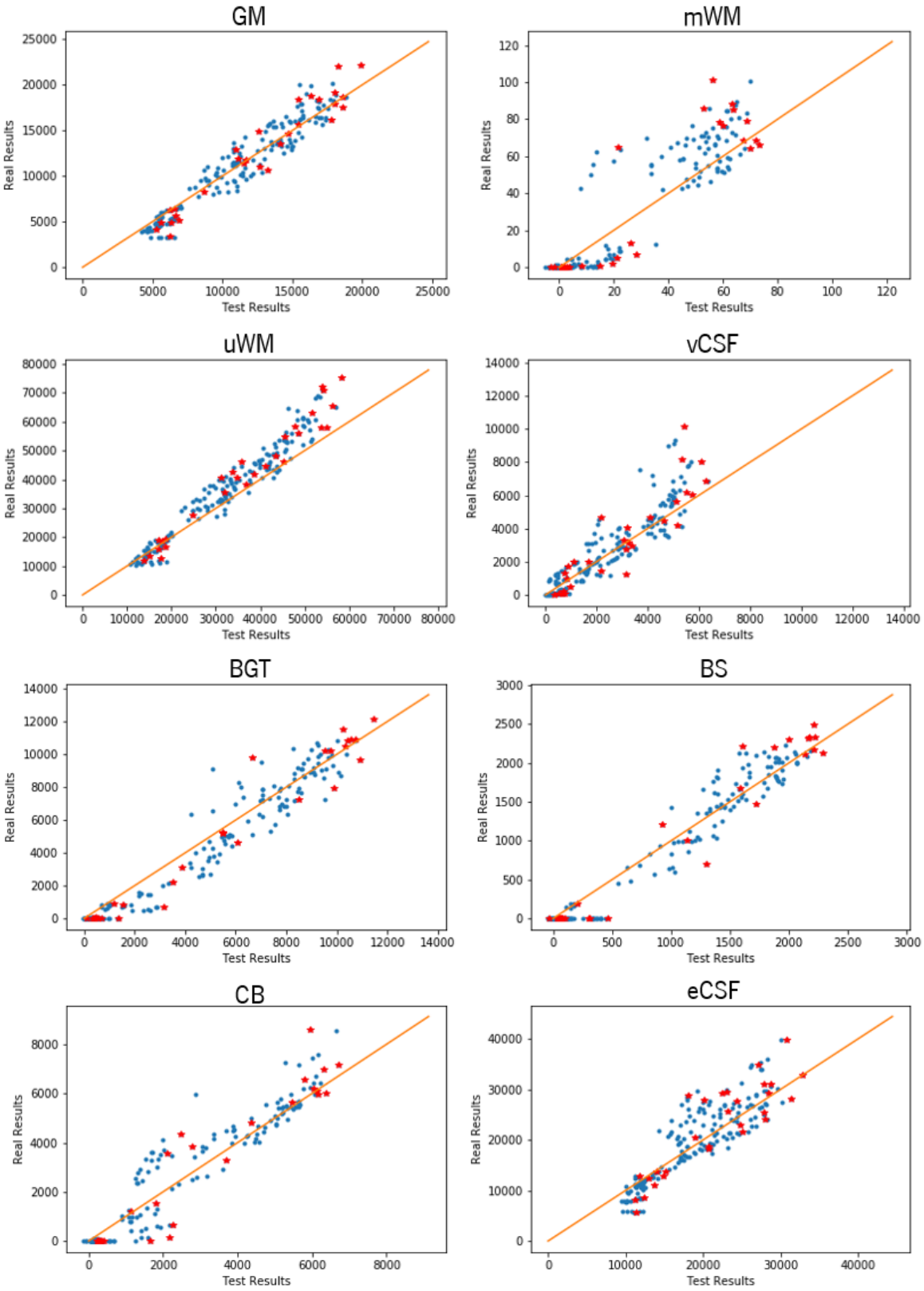


Figure A.2 - Test (prediction) vs Real (ground truth) graphs for the 8 tissues for the 3D approach using chunks of 9 slices, with images at 30 weeks PMA. The name of the tissue is on top of its corresponding graph. Values are in mm^3.

As expected, in general, the graphs are more susceptible to outliers and that is reflected in both metrics. That is due to the fact that chunks are now larger, when compared to the size of the previous experiments.

B – 3D APPROACHES USING BIGGER CHUNKS AND CORRELATION

This section only has the purpose to show the variation of correlation when increasing the size of the chunks using the data at 30 weeks PMA.

Figure B.1 tracks the variation of the mean correlation of all tissues with the size of the chunk in slices. It is possible to denote that the mean correlation of all tissues gets lower, when the size of the chunk increases. As previously remarked, this is expected because the larger chunks. The bigger the volumes that the CNN is dealing with and the less examples for training, validation and testing, make it more difficult for that network to predict with accuracy and sparser the dots in the graph will be.

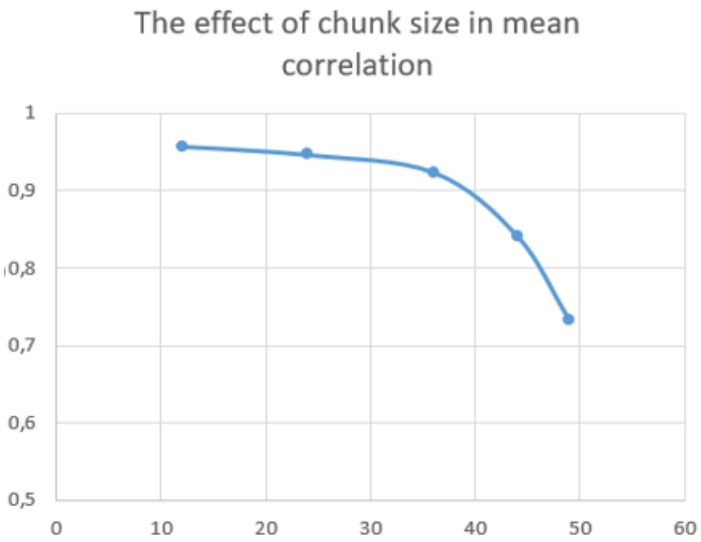


Figure B.1 - Correlation vs size of the chunk in slices: A- Variation of correlation of each tissue with chunk's size. B- Variation of the mean correlation of all tissues with chunk's size.